Chapter 1

Sensor Data Fusion and Autonomous Unmanned Vehicles for the Protection of Critical Infrastructures

Konstantinos Ioannidis[1], **Georgios Orfanidis**[1], **Marios Krestenitis**[1], **Stefanos Vrochidis**[1] **and Ioannis Kompatsiaris**[1]

Abstract Current technology in imaging sensors offers a big variety of information that can be derived from an observed scene. Captured images from modalities exhibit diverse characteristics such as type of degradation; salient features etc. and can be particularly beneficial in surveillance systems. Such representative sensory systems include infrared and thermal imaging cameras, which can operate beyond the visual spectrum providing functionality under any environmental conditions. Multi-sensor information is jointly combined to provide an enhanced representation, particularly utile in automated surveillance systems such as monitoring robotics. In this chapter, a surveillance framework based on a fusion model is presented in order to enhance the capacities of unmanned vehicles for monitoring critical infrastructures. The fusion scheme multiplexes the acquired representations from different modalities by applying an image decomposition algorithm and combining the resulted sub-signals via metric optimization. Subsequently, the fused representations are fed into an identification module in order to recognize the detected instances and improve eventually the surveillance of the required area. The proposed framework adopts recent advancements in object detection for optimal identification by deploying a deep learning model properly trained with fused data. Initial results indicate that the overall scheme can accurately identify the objects of interest by processing the enhanced representations of the fusion scheme. Considering that the overall processing time and the resource requirements are kept in low levels, the framework can be integrated in an automated surveillance system comprised by unmanned vehicles.

1.1 Introduction

Surveillance applications usually employ multimodal imaging sensors to enhance the exploitation performance and expand the functionalities of vision systems under any environmental conditions, target variations and viewpoint obscurations. Different sensors are used to ensure a wider spectral coverage and reliable imaging even in adverse acquisition conditions. The main drawback of the additional robustness is a considerable increment of the image data. Multi-sensor image fusion deals with this data overload by combining images of the same scene acquired using different sensors in a single, composite fused image [1]. The latter provides comprehensive information about the scene such that no multiple post-processing modules are required.

One typical example is a night vision application of context enhancement where Mid-wavelength (MWIR) or Long-wavelength (LWIR) surveillance cameras enhance objects for detection. The emitted energy of an object is received from the infrared sensor and is converted into a temperature value based on the sensor's calibration equation and object's emissivity. However, these sensors are not sensitive to the low temperature background while sensors capturing visual spectrum can provide a relatively clear perspective of the environment. On the contrary, visual representations display increased sensitivity in illumination variations. Eventually, multi-sensor image fusion approaches aim at multiplexing all the information across the electromagnetic spectrum for applications such as monitoring critical infrastructures.

Considering the recent advancements and their increased utilization, unmanned vehicles can be enhanced with additional functionalities and operational capabilities if the overall system could process such fused representations. The effectiveness of visual related tasks in robotics could be significantly increased leading to a system with improved operability. In this chapter, we propose a complete framework (Fig. 1.1) that combines two individual modules in order to accomplish optimal surveillance of critical infrastructures by utilizing unmanned vehicles and their sensors. Towards this objective, a multi-sensor image fusion algorithm is initially inserted to multiplex the acquired information and retain the most significant features from images under different spectrum. The combined

[1] Information Technologies Institute-Centre for Research and Technology Hellas, Thessaloniki, Greece
{kioannid, g.orfanidis, mikrestenitis, stefanos, ikom}@iti.gr

representations are processed further from an object identification module in order to categorize the detected instances and increase the situational awareness of the operational personnel. The identification module relies on a state-of-the-art deep learning architecture, which exhibits sufficient accuracy for such applications. Our initial evaluation tests indicate that the framework is able to provide additional functionalities to the relevant systems. Moreover, the produced images of the multi-sensor fusion scheme are privily beneficial in visual navigation and object identification improving the corresponding accuracy.

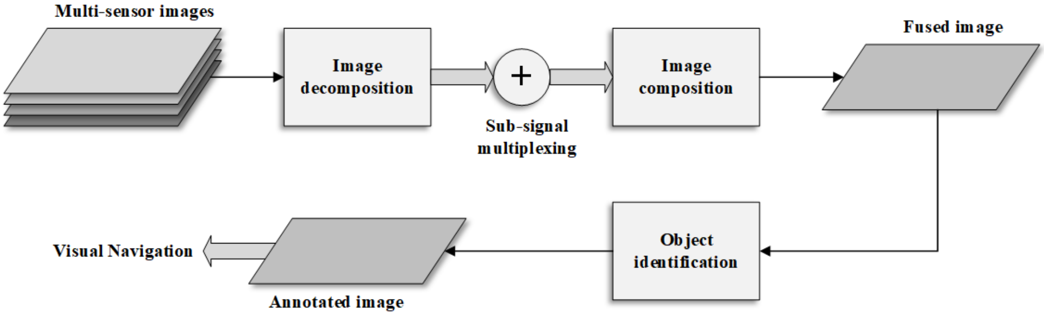

Fig. 1.1. Enhanced identification framework.

The rest of the chapter is organized as follows. In Section 1.2, we present in details the multi-sensor image fusion scheme that was deployed for our framework while in Section 1.3, the main pillars of the object identification are provided. The identification results of the complete architecture are presented and commented in Section 1.4. The chapter is completed with Section 1.5 where conclusions are outlined.

1.2 Multi-sensor image fusion

Fusion of infrared and visible images has found a large utilization in both civilian and non-civilian applications. For the latter category, monitoring critical infrastructures and terrain surveillance comprise the most prominent applications due to the importance of the systems' objectives. This is attributed to significant variations between the infrared and visible light sensors. Visible images can provide more clear information regarding the background of the captured scene while, the details of the under-detection objects could not be perceived sufficiently as the visual sensor is strictly affected by lighting conditions [2]. As infrared images are responsive to thermal variations, the corresponding information can be collected without illumination constraints. Since the visible and infrared representations have complementary information, image fusion could provide an enhanced scene representation.

Based on the adopted transform strategy, existing approaches can be categorized into four main classes [3]: (i) multi-scale decomposition methods, (ii) sparse representations, (iii) methods that performs fusion directly to image pixels either in spatial or transform domains and (iv) combinations of the aforementioned techniques. Another key factor that significantly affects the fusion results comprises the fusion strategy based on which eventually the pixel values are combined to produce the fused information. Multiple variations and numerous combinations have been proposed to provide enhanced scene representations properly exploited by surveillance systems.

The first approach relies on decomposing the input images based on a type of a multi-scale transform in which features are represented in a joint space-frequency domain. Subsequently, the resulted representations are combined based on a fusion rule in which the activity level of coefficients and the correlations among adjacent pixels are considered. Finally, the inverse transformation is applied to the fused representation in order to derive the required enhanced representations. The most widely used decomposition methods involve the application of the Laplacian pyramids [4], the Discrete Wavelet Transform [5] and stationary wavelet decomposition [6]. Wavelet based transforms suffers from some fundamental shortcomings such as shift variance and aliasing leading to pure decomposition results and so, fused representations. Variations of these methods have been proposed to overcome these disadvantages including dual-tree complex wavelet [7], contourlet [8] and shearlet [9]. On the contrary, sparse representation based models display increased fusion accuracy nonetheless; their implementation requires increased computational costs and resources liming their application in real-time surveillance systems. The sparse representation can describe the images by a sparse linear combination of atoms selected from an over-complete dictionary and representing the weighted coefficients. Initially used in [10] for fusing images, sparse representation and its variations were widely used as fusion models such as [11] and [12].

In addition, numerous image fusion methods have been proposed and do not rely on either multi-scale representation or sparse representations. A fraction of these methods computes directly the weighted average of the input pixels or combines their values in other transform domains. For the first case, the required weights are determine based on the activity level of different pixels [13]. Support vector machines [14] and artificial neural networks [15] were also deployed to select the pixels with the highest activity in which the wavelet coefficients were used as features. Besides the above approaches, algorithms that initially convert the input images into other domains where fusion is applied have also been successfully evaluated. Such domains are the principal component analysis [16] and the Intensity-Hue-Saturation transform [17].

1.2.1 Image Decomposition

The aforementioned approaches require a significant amount of computational resources imposing constraints for their application in demanding systems where real-time processing is a prerequisite. Despite their accuracy in combining the multi-sensor information, their application in robotic systems where resources are limited is considered impossible in many cases. Towards surpassing such constrains, we describe a multisensor image fusion scheme which comprises the initial processing phase of the proposed identification framework. The method comprises a modified version of the two-dimensional version of the Empirical Mode Decomposition algorithm [18], called Fast and Adaptive Bidimensional Empirical Mode Decomposition (FABEMD) [19]. The input images are decomposed into several sub-signals which are combined based on a weighted summation scheme. The corresponding coefficients are calculated so that the entropy of each sub-signal pair is maximized. After multiplexing each sub-component, the fused image is produced by summing all the sub-signals. Fig. 1.2. depicts a flowchart of the framework's submodule for fusing images acquired from visual and thermal sensors.

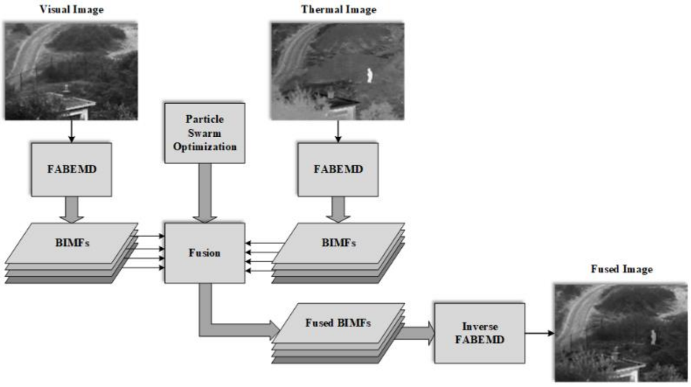

Fig. 1.2. Flowchart of the image fusion module.

The FABEMD comprises a shifting process that decomposes the original signal into multiple sub-components with specific features, called Bidimensional Intrinsic Mode Functions (BIMFs) and a residue signal. As the process is progressed, lower frequency components are included in the subsequent BIMFs meaning that the higher frequency components of the initial signal are contained in the firstly extracted BIMFs. According to [18], the sub-components must display the following features in order to be denoted as BIMFs: **(i)** the number of local extrema (maxima and minima) and the number of zero crossing must be equal or differ by at most one, **(ii)** there should be only one oscillation mode, **(iii)** the mean value of the upper and the lower envelopes should zero or close to zero at any point and **(iv)** the BIMFs must be orthogonal among each other and as a set.

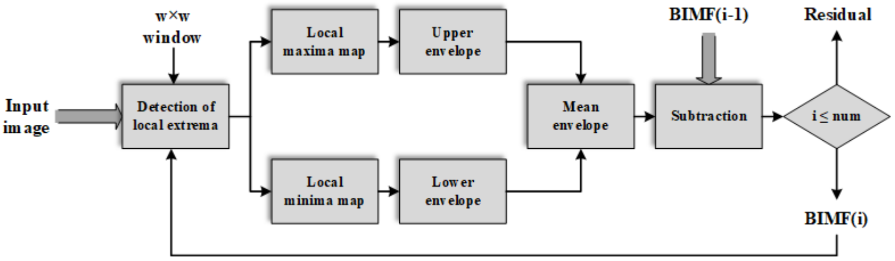

Fig. 1.3. FABEMD process.

The decomposition starts with the assumption that the input image corresponds to one BIMF (for Fig. 1.3., *i=1*). The size *w* and the number of the required BIMFs to extracted are predefined. Thus, the entire process will be terminated when the fixed number of BIMFs is attained. Initially, both local minima and local maxima are calculated by searching the minimum and maximum pixel value, respectively, within a window of size $w \times w$. The two extrema maps include pixel values equal to the local extrema and zeroes for all other pixels. These arrays are processed further in order to construct the upper (P_j) and the lower (Q_j) envelope of the processed signal. Both instances, P_j and Q_j, are computed with the application of appropriate order statistics and smoothing filters. For both cases, the size of the filters is determined based on the maxima and the minima maps. For each non-zero element, the Euclidean distance to the nearest non-zero element is calculated producing the so-called adjacent maxima ($d_{adj\text{-}max}$) and minima ($d_{adj\text{-}min}$) distance array, respectively. The filter size is defined based on the four below options:

$$\begin{aligned} w_{en} &= d_1 = min\left\{min\{d_{adj-max}\}, min\{d_{adj-min}\}\right\} \\ w_{en} &= d_2 = max\left\{min\{d_{adj-max}\}, min\{d_{adj-min}\}\right\} \\ w_{en} &= d_3 = min\left\{max\{d_{adj-max}\}, max\{d_{adj-min}\}\right\} \\ w_{en} &= d_1 = max\left\{max\{d_{adj-max}\}, max\{d_{adj-min}\}\right\} \end{aligned} \qquad (1)$$

With the determination of the window size for envelope formation, order statistics filters MAX and MIN are applied to the corresponding BIMF F_{Tj} to obtain the upper and lower envelops. Essentially, the application of the order statistics filters to the distance maps replaces the zero values with the nearest non-zero element within the distance determined by Eq. 1. The process produces abrupt transitions between adjacent pixel values failing to smoothly osculate the original signal. Thus, both maps are smoothed with the application of a smoothing filter of size w_{en}. The resulted smoothed envelopes will be processed further calculating the corresponding mean envelope which is subtracted from the BIMF(i-1) to form BIMF(i). If the BIMF index is smaller or equal to the predefined total number of the BIMFs, the subtraction result is subjected to the same shifting process. Otherwise, it is considered as the residual signal and the process is terminated. The process is considered as entirely reversible since the summation of all the extracted sub-components results to the source image. This feature is eventually utilized to produce the required fused representation of the source images. Fig. 1.4 presents the extracted BIMFs for an example source image captured under the infrared spectrum.

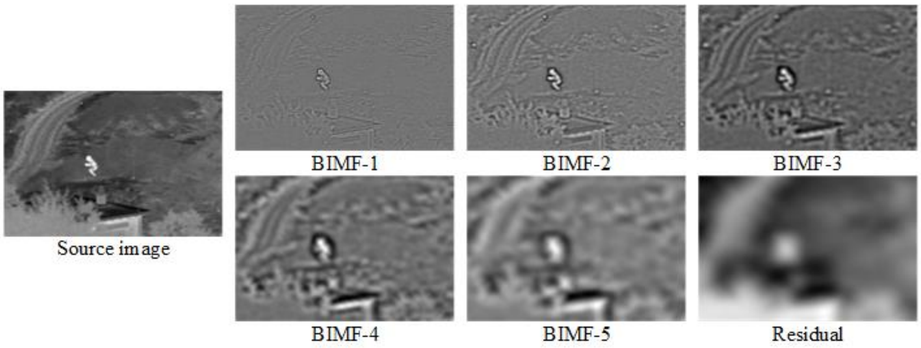

Fig. 1.4. Example of BIMFs.

1.2.2 Fusion scheme

The visible and the infrared representations are decomposed into sub-signals with the application of FABEMD producing equal number of BIMFs for both instances. The decomposed sub-signals with the same indexes are sequentially fused based on the following fusion rule:

$$F(i) = a * BIMF_{vis}[i] + b * BIMF_{the}[i] \tag{2}$$

where $BIMF_{vis}$ and $BIMF_{the}$ corresponds to the *i* BIMF of the visual and the thermal images, respectively. Variables *a* and *b* represents the weighting coefficients for the multiplexing to be defined. The main objective is to compute the values of *a* and *b* to maximize the efficiency of the rule. The problem is considered as an optimization and thus, a metric function should be defined. The presented fusion scheme aims at maximizing the entropy metric since it represents the amount of information produced by a stochastic source of data and is given by:

$$H(F) = -\sum_{i=1}^{n} P(F_i) log[P(F_i)] \tag{3}$$

where *F* denotes the discrete random variable (fused BIMFs) and *P(X)* the probability mass function. To identify the values of *a* and *b* which maximize the entropy, a potential solution will be to identify all combinations meaning a brute force solution. However, this requires extensive usage of resources which are limited in surveillance systems like robots. The proposed fusion scheme adopts a computational method which improves the search of the candidate solution (maximized entropy). The method, called Particle Swarm Optimization (PSO) [20], comprises a metaheuristic as it makes no assumption about the problem being optimized and can search large spaces of candidate solutions. In this work, the particles are related to the weights with the fitness function of entropy and so, candidate solutions are doublets of the form [a,b]. Fig. 1.5. includes the fused BIMFs of the Fig. 1.4 example. The final fused image is resulted by simply adding all the resulted fused BIMFs as inverting the FAEMD process.

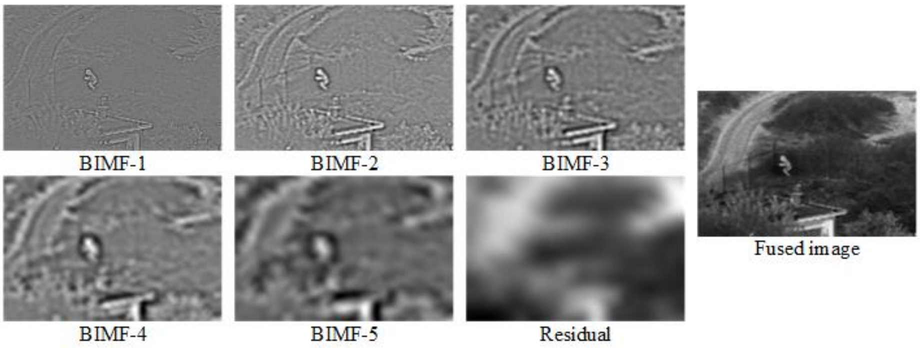

Fig. 1.5. Fused BIMFs and their final fused image.

1.3 Object identification for surveillance objectives

Overall, intelligent surveillance systems involve two sequential processes of visual data, object detection and the subsequent recognition of the detected instances. Traditional object detection methods initially extract some features (such as Histogram of oriented gradients-HOG etc.) from the captured scene and subsequently classify them based on the training data. Due to the low-level capacity of the features, such models fail to identify accurately the detected instances making their use limited in surveillance operations. Such limitations constrain the identification results and decrease the accuracy of the following classification process. Misclassified instances and false positives are the most common problematic results of these models which eventually affect the results of the monitoring system.

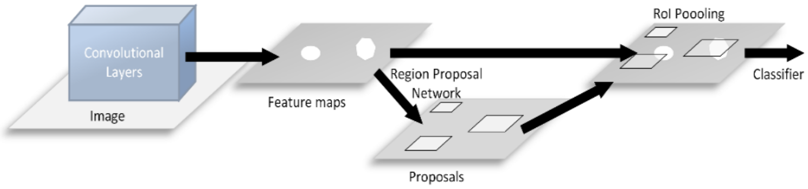

Fig. 1.6. Higher level representation of Faster R-CNN.

Recent advances in the research field overcome such type of drawbacks due to the model's capabilities to extract high-level internal features. Deep learning architectures and more specific, Convolutional Neural Networks (CNNs) can perform sufficiently in terms of accurate identification. These characteristics render the models as suitable candidates for object identification in surveillance systems. Nonetheless, the basic CNNs require the classification of a large number of image regions significantly decreasing the processing rate and increasing the processing cost. Therefore, alterations and enhanced architectures were proposed aiming at minimizing the required scanned image regions. Fast R-CNN [21] and Faster R-CNN [22] are CNN alternatives which perform significantly sufficient and in near real-time operations.

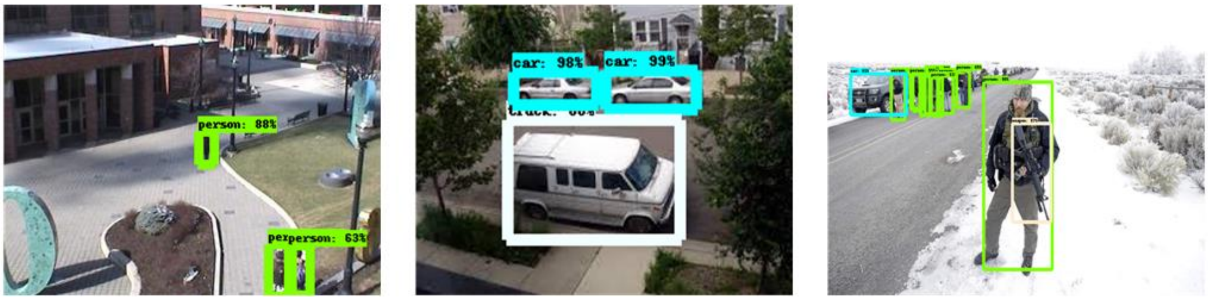

Fig. 1.7. Identification results processing images from visual spectrum.

Considering the distinctiveness of such surveillance systems, our framework adopts a Faster R-CNN model in order to identify the required instances from visual data. The resulted processing time as well as the identification accuracy of the model render the approach as one of the most proper selection for such system objectives. The model was tweaked and properly trained with public available datasets as well as with manually annotated images (visual and fused) to cover a wider range of objects. In general, the deployed identification model consists of two discrete modules. The first module constitutes a deep fully convolutional network that extracts potential regions of interests while the second unit is the Fast R-CNN detector. The Region Proposal Network (RPN) is comprised by a fully convolutional network where two selections were investigated; the first with five shareable (ZF model) and the second with thirteen convolutional layers (VGG-16). Aiming at a decreased processing time framework, both networks (RPN and Fast R-CNN) share a common set of convolutional layers by default in order to mitigate the high computational cost of such architectures. Indicative identification results of the model's application in images of visual spectrum are provided in Fig. 1.7.

1.4 Experimental results

In general, the identification model results increased confidence levels of the detected instances when the training dataset is comprised by a vast number of images. Towards this objective, the Faster R-CNN model was trained with various datasets most of which are publicly available and include representations captured under the visible spectrum. The selection of training the model with images from different spectra is manifold. Considering that the Faster R-CNN comprises a convolutional neural network, the first layers of the network can export low and mid-level image features which are better represented in mid-infrared images. On the other hand, higher-level features are extracted more efficiently in visible spectrum and fused images thus, their classification in training and testing processes is more accurate.

For the conducted experiments, approximately 20K images were utilized for training purposes while 130 images comprised the testing dataset. In addition, the model was trained to identify three separate instances from the fused representations which are the most common in surveillance objectives: humans, weapons and cars. Thus, three separate classes were identified. Two different sets of experiments were conducted. Initially, the model was tested with visible spectrum images while in the second case, it was evaluated with visual images in low-light conditions and with their corresponding fused representations. The overall performance was measured in terms of average precision and confidence level. Results are provided and commented below.

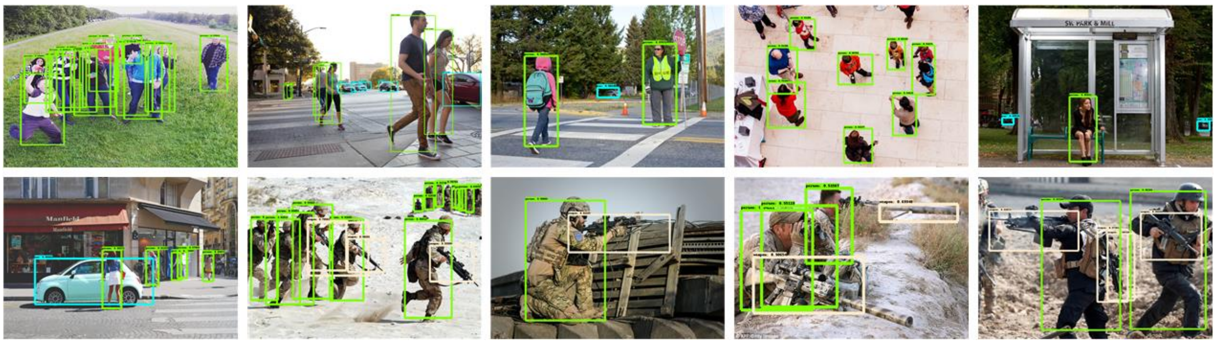

Fig. 1.8. Identification results using visible spectrum images (From top left corner to bottom right corner: Image 1-10).

Fig. 1.8. presents some indicative results with testing visual images. The average precision for the class "human" was equal to 0.735 while for the "weapons" class was equal to 0.383. "Car" instances were identified with an accuracy of 0.719. For representation purposes, the resulted identification confidence scores are provided in Table 1. **1**. per image of Fig. 1.8. In general, the model managed to identify "humans" and "car" instances accurately achieving state-of-the-art identification scores. On the contrary, low identification scores for the class "weapons" are due to the lack of proper training dataset. The detection system achieved a frame rate equal to 5 fps, one the highest in similar schemes rendering it appropriate for near-real time application such as the surveillance of critical infrastructure.

Table 1. 1. Resulted confidence levels of visual example images.

Objects	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6	Image 7	Image 8	Image 9	Image 10
Human	0.7552	0.9365	0.9785	0.8856	0.9583	0.8019	0.7679	0.9906	0.5335	0.9777
Weapon	-	-	-	-	-	-	0.8776	0.8225	0.8030	0.7888
Car	-	0.9687	0.9025	-	0.7522	0.9660	-	-	-	-

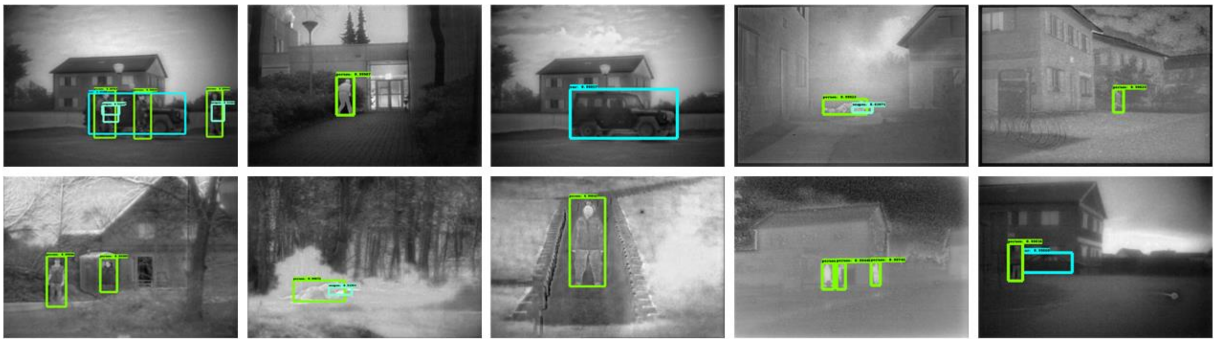

Fig. 1.9. Identification results using fused images (From top left corner to bottom right corner: Image 1-10).

Nonetheless, the exclusive use of images captured under the visible spectrum inserts many limitations in low-lighting conditions leading to inaccurate identification results. This was verified from the second set of the conducted experiments where both fused and their corresponding visual images were fed to the identification model. A small set of images that resulted from processing fused images and their identification bounding boxes are depicted in Fig. 1.9. The confidence levels are given in Table 1. **2** as well as the identification results using only as inputs the visual corresponding images. The fused mode can accurately identify human figures where the visual mode fails due to the scenery itself (fog, high occlusion etc.). Similar results are valid for the "car" class also. Finally, weapon detection results prove the superiority of the fused model over the visual since no proper identification was achieved.

Table 1. 2. Resulted confidence levels of fused example images (UD: undetectable, FD: faulty detection).

	Visual detection			Fused detection		
	Human	Weapon	Car	Human	Weapon	Car
Image 1	UD	-	0.8932	0.9982	0.5257	0.9948
Image 2	UD	-	-	0.9990	-	-
Image 3	-	-	0.9714	-	-	0.9986
Image 4	UD	UD	-	0.9992	0.6287	-
Image 5	UD	-	-	0.9982	-	-
Image 6	UD	-	0.6716 (FD)	0.9960	-	-
Image 7	UD	UD	-	0.9967	0.5195	-
Image 8	UD	-	-	0.9984	-	-
Image 9	UD	-	-	0.9950	-	-
Image 10	UD	-	UD	0.9982	-	0.9986

In general, the identification model fails to detect the required objects of interests in cases where the visual sensors fail to capture the visible light of the scene. Low levels of visible light, fog, dusty environments are the most common, non-ideal conditions of image acquisition leading to insufficient confidence levels of detection. Thus, both navigational and surveillance objectives of the unmanned vehicles are restricted affecting the ultimate goal of protecting critical infrastructures. Enhancing the inputs of the identification model with mid-infrared visualizations can transcend such constraints integrating additional capabilities in the surveillance system, as it is resulted by the conducted experiments.

1.5 Conclusions

In this chapter, a novel framework comprised by two discrete modules was presented. The initial fusion module operates as a complementary pre-processing stage to overcome limitations inserted by the exclusive usage of images under the visible spectrum. The application of the FABEMD algorithm produces multiple sub-signals of the visible and the mid-infrared representations each of which includes specific features. Every visual sub-signal is multiplexed with its corresponding mid-infrared by a weighted summation scheme. Optimum values of weights are calculated using the PSO algorithm to decrease the computational cost of an extensive search approach. The final fused representations are fed into the identification module where a Faster R-CNN was deployed. The model was properly trained in order to be able to identify the most common objects for surveillance tasks, e.g. humans, cars and weapons. The identification model performs significantly more efficient due to the inclusion of the mid-infrared information into the visual representations. In addition, the computational cost is kept at low levels rendering the framework proper for real and/or near real-time systems such as unmanned vehicles. Current work can be extended by multiplexing additional information from sources that can capture wavelengths other than visible and/or thermal.

Acknowledgements

This work was supported by ROBORDER and BEAWARE projects funded by the European Commission under grant agreements No 740593 and No 700475, respectively.

References

1. Stathaki, T. (2008). *Image fusion: Algorithms and Applications*. Academic Press, Inc.
2. Meng, F., Guo, B., Song, M., & Zhang, X. (2016). Image fusion with saliency map and interest points. *Neurocomputing, 177,* pp. 1-8.
3. Li, S., Kang, X., Fang, L., Hu, J., & Yin, H. (2017). Pixel-level image fusion: A survey of the state of the art. *Information Fusion, 33,* pp. 100-112.
4. Mertens, T., Kautz, J., & Van Reeth, F. (2007). Exposure fusion. *15th Pacific Conf. on Computer Graphics and Applications,* pp. 382-390.
5. Ben Hamza, A., He, Y., Krim, H., Willsky, A. (2005). A multiscale approach to pixel-level image fusion. *Integr. Comput.-Aided Eng. 12, 2,* pp. 135-146.
6. Li, S., Kwok, J.T., & Wang, Y. (2002). Using the discrete wavelet frame transform to merge Landsat TM and SPOT panchromatic images. *Information fusion, 3, 1,* pp. 17-23.
7. Lewis, J.J., O' Callaghan, R.J., Nikolov, S.G., Bull, D.R, & Canagarajah, N. (2007). Pixel- and region-based image fusion with complex wavelets. *Information fusion, 8, 2,* pp. 119-130.
8. Li, T., & Wang, Y. (2011). Biological image fusion using a NSCT based variable-weight method. *Information fusion, 12, 2,* pp. 85-92.
9. Wang, L., Li, B., & Tian, L.-F. (2014). Multi-modal medical image fusion using the inter-scale and intra-scale dependencies between image shift-invariant shearlet coefficients. *Information fusion, 19,* pp. 20-28.
10. Yang, B., & Li, S. (2010). Multifocus image fusion and restoration with sparse representation. *IEEE Trans. On Instrumentation and Measurements, 59, 4,* pp. 884-892.
11. Li, S., Yin, H, & Fang, L. (2012). Group-sparse representation with dictionary learning for medical image denoising and fusion. *IEEE Trans. On Biomedical Engineering, 59, 12,* pp. 3450-3459.
12. Nejati, M., Samavi, S., & Shirani, S. (2015). Multi-focus image fusion using dictionary-based sparse representation. *Information fusion, 25,* pp. 72-84.
13. Gangapure, V.N., Banerjee, S., & Chowdhury, A.S. (2015). Steerable local frequency based multispectral multifocus image fusion. *Information fusion, 23,* pp. 99-115.
14. Li, S., Kwok, JT., Tsang, I.W., & Wang, Y. (2004). Fusing images with different focuses using support vector machines. *IEEE Trans. On Neural Networks, 15, 6,* pp. 1555-1561.
15. Li, S., Kwok., J.T., & Wang, Y. (2002). Multifocus image fusion using artificial neural networks. *Pattern Recognition Letters, 23, 8,* pp. 985-997.

16. Shahdoosti, H. R., & Ghassemian, H. (2016). Combining the spectral PCA and spatial PCA fusion methods by an optimal filter. *Information fusion, 27,* pp. 150-160.
17. Tu, T.-M., Huang, P.S., Hung, C.-L., Chang, C.-P. (2004). A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. *IEEE Trans. On Geoscience and Remote sensing letters. 1, 4,* pp. 309-312.
18. Huang, N.E., Shen, Z., Long, S., Wu, M.C., Shih, H., Zheng, Q., Yen, N.-C., Tung, C.C, & Liu, H.H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. Of the Royal Society of London A: mathematical, physical and engineering sciences, 454, 1971,* pp. 903-995.
19. Bhuiyan, S., Adhami, R., & Khan, J. (2008). Fast and adaptive bidimensional empirical mode decomposition using order-statistics filter based envelope estimation. *EURASIP Journal on Advances in Signal Processing*, 1.
20. Kennedy, J., & Eberhart, R. (1995). Particle Swarm Optimization, *Int. Conf. on Neural Networks, 4,* pp. 1942-1948.
21. Girshick, R. (2015). Fast R-CNN. *In Proc. Of the IEEE Int. Conf. on Computer Vision,* pp. 1440-1448.
22. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, *39, 6,* pp. 1137-1149.