remote sensing

MDPI

Article

Fusing Multimodal Video Data for Detecting Moving Objects/Targets in Challenging Indoor and Outdoor Scenes

Zacharias Kandylakis *, Konstantinos Vasili and Konstantinos Karantzalos

Remote Sensing Laboratory, National Technical University of Athens, 15780 Zographos, Greece; k.vasili913@gmail.com (K.V.); karank@central.ntua.gr (K.K.)

* Correspondence: zach.kandylakis@gmail.com; Tel.: +30-210-7721673

Received: 31 December 2018; Accepted: 16 February 2019; Published: 21 February 2019

Abstract: Single sensor systems and standard optical—usually RGB CCTV video cameras—fail to provide adequate observations, or the amount of spectral information required to build rich, expressive, discriminative features for object detection and tracking tasks in challenging outdoor and indoor scenes under various environmental/illumination conditions. Towards this direction, we have designed a multisensor system based on thermal, shortwave infrared, and hyperspectral video sensors and propose a processing pipeline able to perform in real-time object detection tasks despite the huge amount of the concurrently acquired video streams. In particular, in order to avoid the computationally intensive coregistration of the hyperspectral data with other imaging modalities, the initially detected targets are projected through a local coordinate system on the hypercube image plane. Regarding the object detection, a detector-agnostic procedure has been developed, integrating both unsupervised (background subtraction) and supervised (deep learning convolutional neural networks) techniques for validation purposes. The detected and verified targets are extracted through the fusion and data association steps based on temporal spectral signatures of both target and background. The quite promising experimental results in challenging indoor and outdoor scenes indicated the robust and efficient performance of the developed methodology under different conditions like fog, smoke, and illumination changes.

Keywords: hyperspectral; SWIR; thermal; video; multisensor; detection; tracking; moving object

1. Introduction

Numerous monitoring and surveillance imaging systems for outdoor and indoor environments have been developed during the past decades based mainly on standard RGB optical, usually CCTV, cameras. However, RGB sensors provide relatively limited spectral information especially for challenging scenes under problematic conditions of constantly changing irradiance, illumination, as well as the presence of smoke or fog.

During the last decade, research and development in optics, photonics and nanotechnology permitted the production of new innovative video sensors which can cover a wide range of the ultraviolet, visible as well as near, shortwave and longwave infrared spectrum. Multispectral and hyperspectral video sensors have been developed, based mainly on (a) filter-wheels, (b) micropatterned coatings on individual pixels, (c) optical filters monolithically integrated on top of CMOS image sensors. In particular, hyperspectral video technology has been employed, for the detection and tracking of moving objects in engineering, security and environmental monitoring applications. Indeed, several detection algorithms have been proposed for various applications with moderate to sufficient effectiveness [1,2]. In particular, hyperspectral video systems have been employed for developing object tracking solutions through hierarchical decomposition for chemical gas plume tracking [3]. Multiple object tracking based on background estimation in hyperspectral video sequences as well as multispectral change detection through joint dictionary data have also been addressed [4,5]. Certain processing pipelines have also been proposed to address the changing environmental illumination conditions [2]. Scene recognition and video summarization have been also proposed based on machine learning techniques and RGB data [6,7].

These detection capabilities are gradually starting to be integrated with other video modalities like, e.g., standard optical (RGB), thermal, and other sensors towards the effective automation of the recognition modules. For security applications, the integration of multisensor information has been recently proposed towards the efficient fusion of the heterogeneous information towards robust large-scale video surveillance system [8]. In particular, multiple target detection, tracking, and security event recognition is an important application of computer vision with significant attention on human motion/activity recognition and abnormal event detection [9]. Most algorithms are based on learning robust background models from standard optical RGB cameras [10–14] and more recently from other infrared sensors and deep learning architectures [15]. However, estimating a foreground/background model is very sensitive to illumination changes. Moreover, extracting the foreground objects as well as recognizing its semantic class/label is not always trivial. In particular, in challenging outdoor (like in Figure 1) and indoor scenes with a rapidly changing background such algorithms fail to model the background efficiently, resulting to several false positives or negatives.

Recent advances in machine learning have provided robust and efficient tools for object detection (i.e., point out a bounding box around the object of interest in order to locate it within the image plane) based on deep neural network architectures. Since, the number of object occurrences in the scene is not a priori known, a standard convolutional network, followed by a fully connected layer, is not adequate. In order to limit the search space (the number of image regions to search for objects) the recent R-CNN [16] method proposed to extract just 2000 regions from the image which were generated by a selective search algorithm. Based on R-CNN, the Fast R-CNN method was proposed which instead of feeding the region proposals to the CNN, feeds the input image to the CNN, to generate a convolutional feature map. Thus, it is not necessary to feed 2000 region proposals to the convolutional neural network. Instead, the convolution operation is done only once per image [17]. Improving upon this, a more recent approach utilizes a separate network to predict the region proposals. These are reshaped using a RoI pooling layer which is then used to classify the image within the proposed region and predict the offset values for the bounding boxes (FR-CNN [17]). In the same direction, the Mask R-CNN deep framework [18] decomposes the instance segmentation problem into the subtasks of bounding box object detection and mask prediction. These subtasks are handled by dedicated networks that are trained jointly. Briefly, Mask R-CNN is based on FR-CNN [17] bounding box detection model with an additional mask branch that is a small fully convolutional network (FCN) [19]. At inference, the mask branch is applied to each detected object in order to predict an instance-level foreground segmentation mask.

All of the previous state-of-the-art object detection algorithms use regions to localize the object within the image. The network does not look at the complete image but instead, looks at parts of the image which have high probabilities of containing an object/target. In contrast, the YOLO ('You Only Look Once') method [20] employs a single convolutional network which predicts the bounding boxes and the class probabilities for these boxes. YOLO has demonstrated state-of-the-art performance in a number of benchmark datasets, while advancements have been already proposed [21] towards improving both accuracy and computational performance for real-time applications.

Towards a similar direction and additionally aiming at exploiting multisensor imaging systems for challenging indoor and outdoor scenes, in this paper, we propose a fusion strategy along with an object/target detection and verification processing pipeline for monitoring and surveillance tasks. In particular, we build upon recent developments [4,22] on classification and multiple object tracking from a single hyperspectral sensor and have moreover integrated another thermal and shortwave

(SWIR) video sensors similar to [23]. However, apart from the dynamic background modeling and subtraction scheme [23], here, we have integrated state-of-the-art deep architectures for supervised target detection making the developed framework detector-agnostic. Therefore, the main novelty of the paper lies in the proposed processing framework which can be deployed in single board computers locally near the sensors (at the edge). It is able to exploit all imaging modalities without having to process all acquired hypercubes, but only selected parts, i.e., candidate regions, allowing for real-time performance. Moreover, the developed system has been validated in both indoor and outdoor datasets in challenging scenes under different conditions like the present of fog, smoke, and rapidly changing illumination.

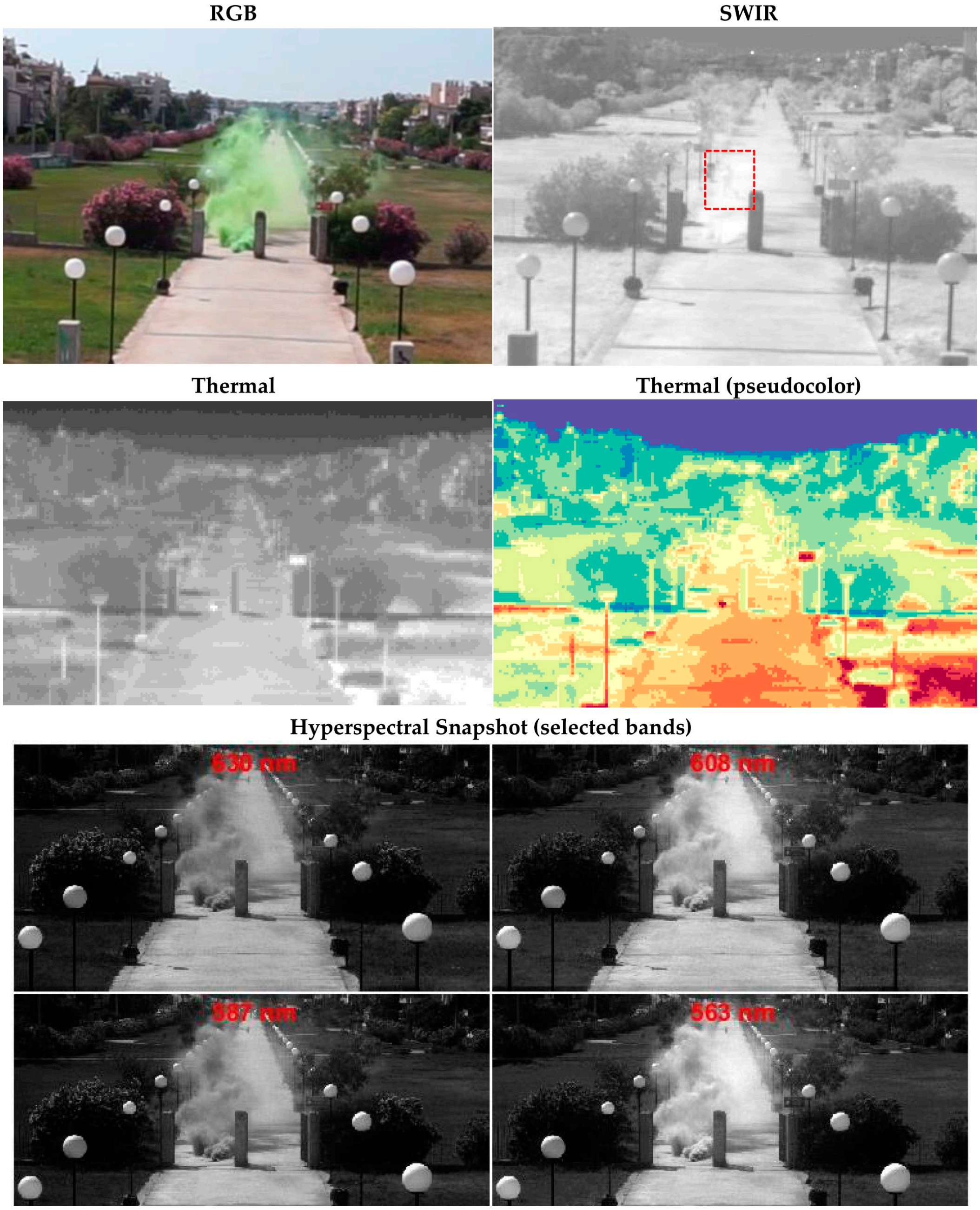
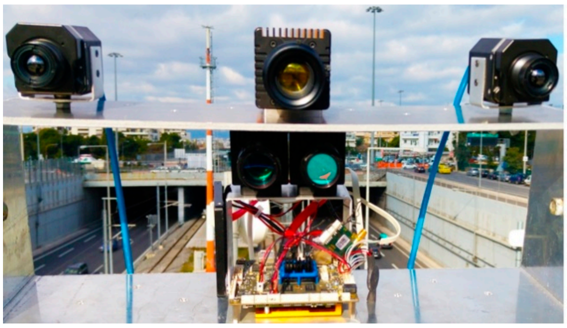


Figure 1. In challenging indoor or outdoor environments with dynamically changing conditions like different smoke, fog, humidity, etc., levels, the standard moving object detection and tracking algorithms fail to detect moving targets based on just a single imaging (usually RGB CCTV) source.

2. Methodology

2.1. The Multisensor Video System

The developed multisensor system consists of thermal, hyperspectral, and shortwave infrared (SWIR) sensors (Figure 2). In particular, the thermal sensor is the longwave infrared FLIR TAU2, with increased sensitivity (<30 mK) and the capability of capturing data in the range of 8 µm to 13 µm, at a spatial resolution of 620 × 480, and at an acquisition rate of 9 Hz. The SWIR sensor is the Xenics Bobcat 640, which covers the spectral range of 900 to 1700 nm, at a resolution of 640 × 512, and a recording rate of 100 Hz. The sensor is an InGaAs Focal Plane Array (FPA) ROIC with Capacitor Trans-impedance Amplifier topology with a pixel pitch of 20 µm, compatible with C-mount lenses and a quantum efficiency peak value at +/− 80%. The hyperspectral acquisition system consists of two imec snapshot mosaic CMOS sensors able to concurrently acquire 41 different spectral bands, in the range of 450 nm to 1000 nm with a spatial resolution of approximately 500 × 270 pixels per spectral band and a frame rate of 24 fps. The first hyperspectral sensor is a 4 × 4 snapshot mosaic with 16 bands in the 470 to 650 nm (visible range) and the second one a 5 × 5 snapshot mosaic with 25 bands in the 600 to 1000 nm range (visible–NIR range). The two sensors are integrated through a single board computer with an embedded processor and the low voltage differential signaling (LVDS) interfaces. The controller board with the field-programmable gate array (FPGA) is capable to store and transform the acquired sensor data into spatially reconstructed data cube/hypercubes. All the sensors (through USB3 or GiGE interfaces) are then connected to a mini-ATX local processing unit which handles the rest of the pipeline, as well as the dissemination of the relevant output to the cloud and the Common Operation Picture at the monitoring center.



Specs / Sensor	Thermal	Hyperspectral (2x)	SWIR
Spectral Range	8-13µm	400 – 950nm	900 – 1700nm
Spectral Resolution	1 band	41 bands	1 band
Spatial Resolution	620 × 480	500 × 270	640 × 512
Frame Rates	9 Hz	24 fps	max 100 Hz
Lens	19mm	35mm	25mm
Interface	USB3	GiGE	GiGE

Figure 2. The multisensor system on the left with the thermal (top left and right), the SWIR (middle top) and the two hyperspectral snapshot sensors (middle bottom) on top of the single board computer with the embedded processor. The sensor specifications are presented in the table on the right.

The sensors can be mounted on a fixed platform or tripod. In order to acquire oblique views of the region of interest (ROI), they should be positioned relatively higher to it. During the data acquisition process, although the system is fixed, the sensors and the video sequence was affected by the changing wind and sudden abrupt bursts, resulting into not perfectly stabilized video sequence, i.e., video frames that are not absolutely aligned/registered. Regarding the temporal synchronization, due to the different acquisition frame rates and the fact that the sensors cannot be absolutely synchronized or triggered concurrently by any hardware of software means, the acquired data per sensor (Figure 3) are associated with the central processing unit (CPU) time/clock. In Figure 3, the indicative acquired data from all sensors in an outdoor challenging scene are presented. In particular, the RGB image from a standard CCTV camera (top left) is presented along with the corresponding frames from the SWIR sensor (top right) and the thermal sensor in grayscale (left, 2nd row) and pseudocolor (right, 2nd row) colormaps. Moreover, the acquired hyperspectral data from the two snapshot mosaic (4 × 4 and 5 × 5) sensors after the hypercube reconstruction are presented in Figure 3, as well. The hypercube reconstruction process involves the formation of the 16 and 25 spectral bands, respectively, from the CMOSIS CMV2000 2.2 MP sensor.

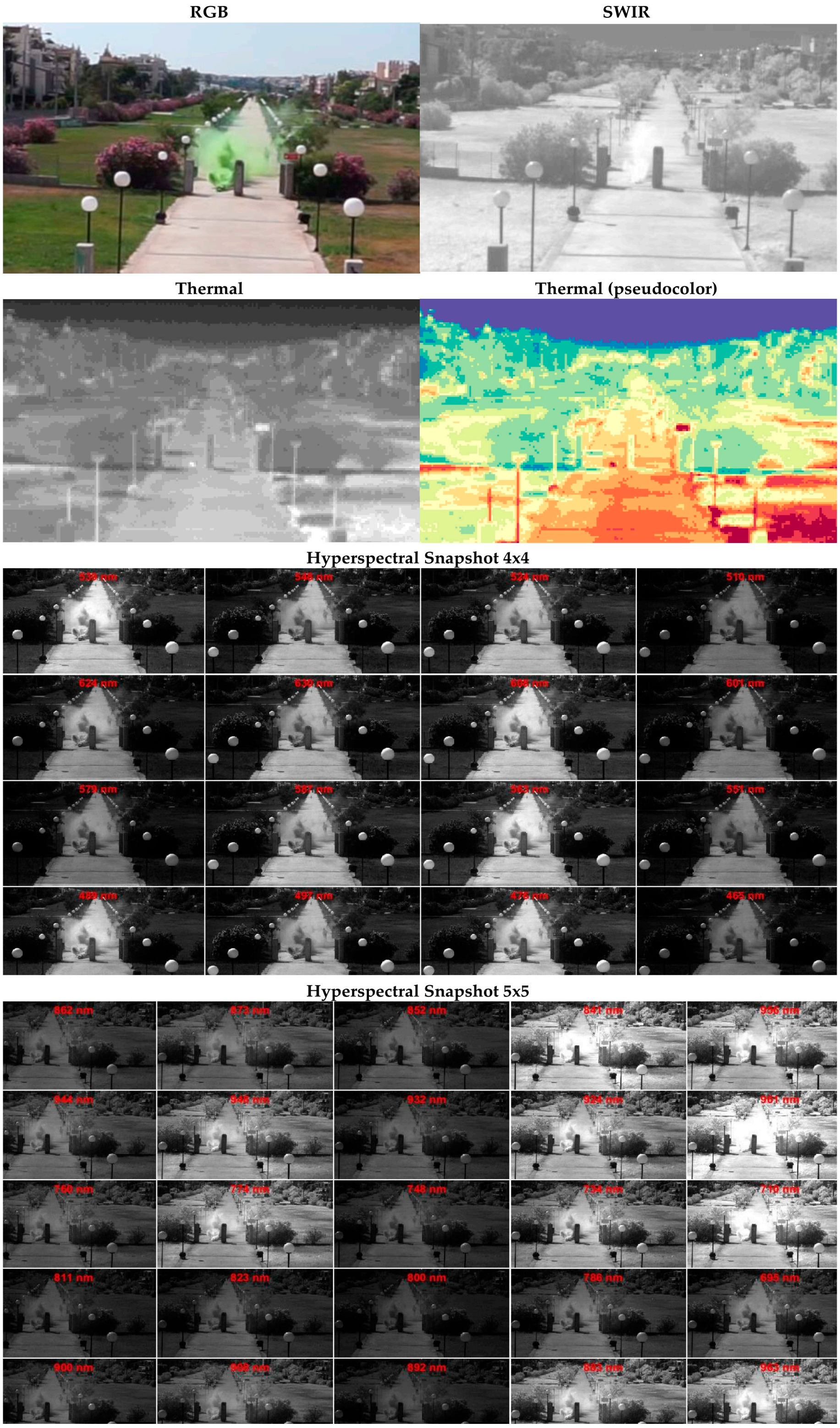


Figure 3. The concurrently acquired multimodal imaging data from the developed multisensor system including the SWIR, thermal, and the two hyperspectral (4 × 4 and 5 × 5 snapshot mosaic) sensors. The corresponding frame from a standard RGB optical CCTV sensor is presented (top left), as well.

The spectral sensors have different image sizes, lenses, and, therefore, field of views (FOVs). The main three sensors of the multisensor system and the corresponding FOVs on the Region of Interest (ROI) are presented schematically in Figure 4 (left). In particular, due to the sensors and lens configuration, the thermal sensor has a wider field of view, followed by the SWIR sensor which observes a relatively smaller area. The hyperspectral sensor has the relatively smaller FOV, while all are covering the ROI. The ROI plane is associated with an arbitrarily defined, local coordinate system (LCS).

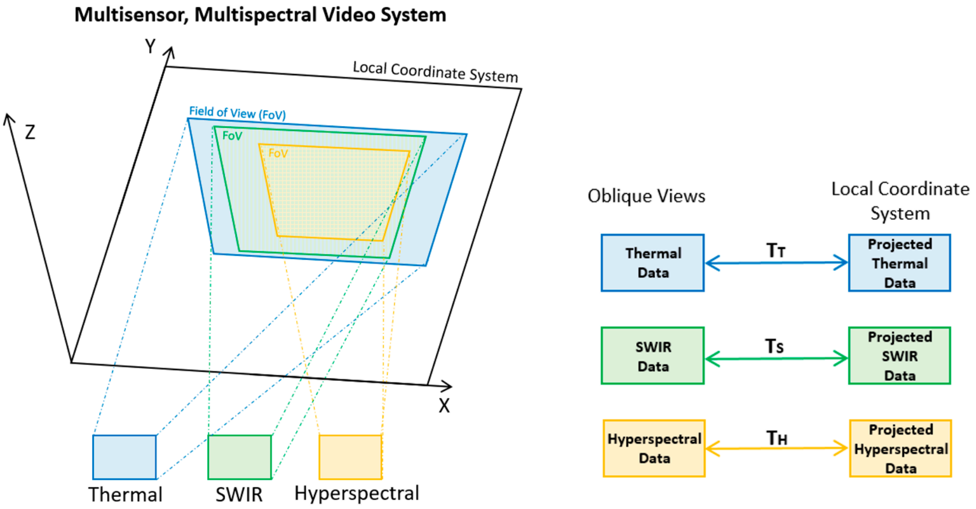


Figure 4. The multisensor system observes the region of interest (ROI), while each sensor has different field of views (FOVs). The ROI plane is associated with an arbitrarily defined, local coordinate system (LCS). In order to establish correspondences among the different FOVs, the perspective and inverse perspective transformations are estimated which can relate image coordinates among the different oblique views of all imaging sources.

2.2. Multimodal Data Processing

The developed system integrates a number of software modules like dimensionality reduction algorithms [24], registration [25,26], background subtraction, moving objection detection, and calculation of optical flows, velocity/density/direction [4].

In order to monitor efficiently and in real-time a desired ROI, the projection of all acquired frames in the same coordinate system has been addressed through the establishment of geometric correspondences among the FOVs through the Local Coordinate System (LCS). In particular, the first step before the main processing pipeline, which is schematically presented in Figure 4 (right), is the calculation of the 3 × 3 transformation matrices for the projection of all three image planes on an arbitrarily defined LCS. In addition, all the inverse transformation matrices were computed for the inverse perspective transformations from the LCS to the image planes. A significant advantage of this approach is that, the actual projection of the entire image (hypercube) is not required, omitting a computationally expensive step. Instead, only the coordinates of the possible moving objects are converted between the reference systems.

In particular, the main processing pipeline and proposed data fusion procedure is summarized in Figure 5. In order to keep the computation complexity as low as possible while allowing near-real time performance, the possible moving targets are detected on the SWIR and/or the thermal sensor (covering both day and night acquisitions). On these modalities moving object detection tasks are executed towards the detection of the possible moving objects/targets (PMT). Three detection methods have been considered and have been integrated in the processing pipeline as independent software modules, forming a detector-agnostic implementation. The first one was based on background subtraction (BS) techniques, the second one on a fast, region-based, deep convolution neural network

(FR-CNN), and the third one was based on a similar deep architecture, which only searches the image domain once (YOLO), allowing real-time object detection applications.

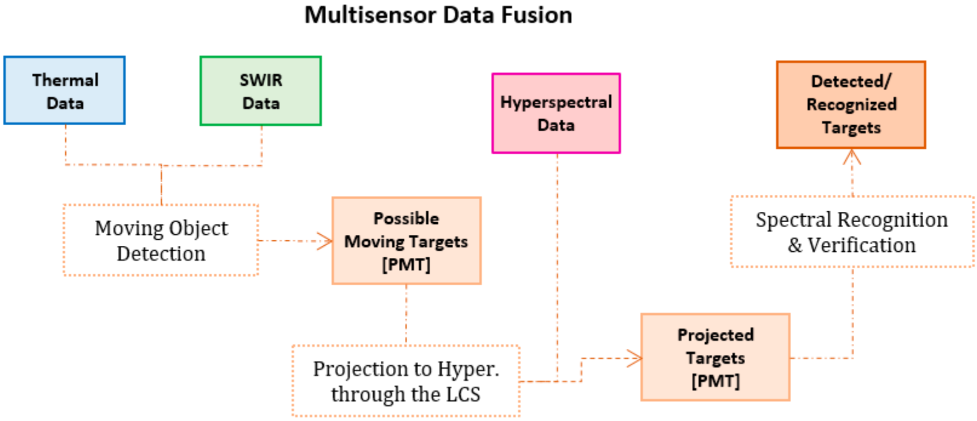

Multisensor Data Fusion
Thermal Data
SWIR Data
Hyperspectral Data
Detected/ Recognized Targets
Moving Object Detection
Possible Moving Targets [PMT]
Spectral Recognition & Verification
Projected Targets [PMT]
Projection to Hyper. through the LCS

Figure 5. The developed processing pipeline for efficient multisensor data fusion.

More specifically, the implemented BS method consisted of two processing steps; background estimation and background subtraction. The background estimation was based on a recursive technique which maintained a single background model which was updated with each new video frame. In general, these techniques are computationally efficient, have minimal memory requirements and can be deployed for real-time applications. Among the most popular background modeling techniques [11–15,27] are (a) the running Gaussian average which models camera noise as a Gaussian distribution, (b) the Gaussian mixture model (GMM) which models each channel of a pixel as a mixture of K-Gaussians, (c) the GMM with an adaptive number of Gaussians by employing a fixed number of Gaussians to model each pixel, as well as (d) the approximated median filtering which uses a recursive filter to estimate the median instead of Gaussian distributions. During all our experiments the background estimation was modeled dynamically for each pixel as a mixture of Gaussians with the number of components being unknown.

The second object detection algorithm that was integrated, was the FR-CNN method [17] which is composed of two modules: The first module is a deep fully convolutional network that is able to propose regions (region proposal network, RPN). The second one is the region-based CNN detector that considers these proposed regions. The RPN module indicates to the CNN module where to search for objects. The RPN takes an image (of any size) as input and returns a set of bounding boxes that indicate the regions were possible objects are. The predicted region proposals are then reshaped using a RoI pooling layer which is then used to classify the image within the proposed region and predict the offset values for the bounding boxes. This is a supervised procedure that requires a training stage and the tuning of several hyperparameters.

With the same supervised manner, the third framework that was integrated for object detection was the YOLO method [20] which divides the input image into an $S \times S$ grid. Each grid cell predicts only one object. The updated version YOLOv2 which was implemented here contains 24 convolutional layers followed by two fully connected layers. Moreover, each convolution block contains BatchNorm normalization and Leaky Relu activation, apart from the last convolution block. In relation to the standard YOLO implementation, the integrated here YOLOv2 has demonstrated [17] an improvement in detection accuracy by also considering a batch normalization step, a high-resolution classifier, anchor boxes, and a multiscale training procedure.

Based on the real-time object detectors (BS, FR-CNN, and YOLO), the possible detected targets (Figure 5) extracted from the thermal and SWIR sensors are then projected into the LCS (Figure 4). In particular, the bounding box or polyline coordinates are projected to the LCS through the use of the transformation matrix TS. The resulting coordinates are then projected to the hyperspectral image plane using the inverse transformation matrix TH-1. These projected targets (bounding boxes) can then directly fuse the spectral information from the hyperspectral bands, avoiding the computationally intensive hypercube coregistration with any other imaging modality. The final detected targets (Figure 6) are extracted after their spectral verification and recognition, based on data association that exploits their temporal spectral signatures (Figure 7).

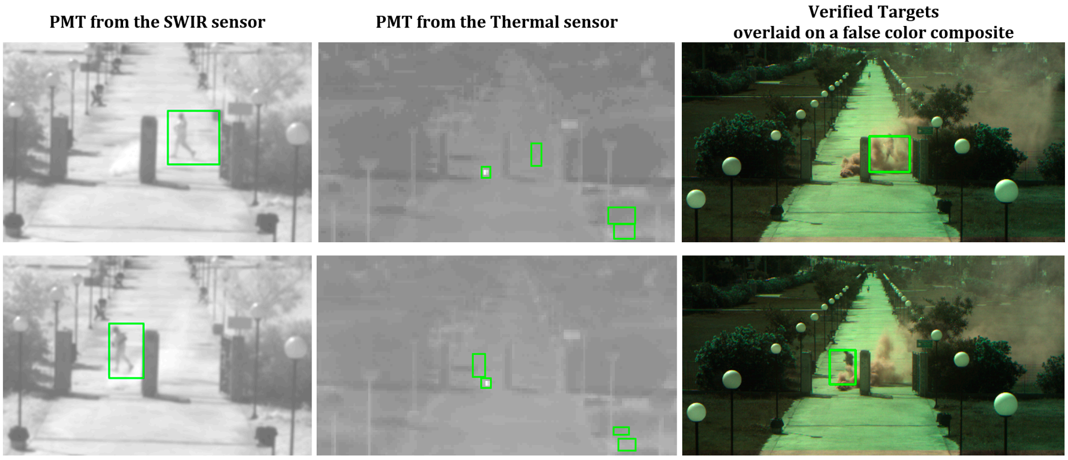

PMT from the SWIR sensor
PMT from the Thermal sensor
Verified Targets
overlaid on a false color composite

Figure 6. The possible moving targets (PMT) from the SWIR (left) and thermal (middle) sensors are projected through the LCS onto the hyperspectral data cube and the associated/verified targets are detected (right). Indicative results for frame #118 (top) and frame #123 (down) from the outdoor scene dataset are presented.

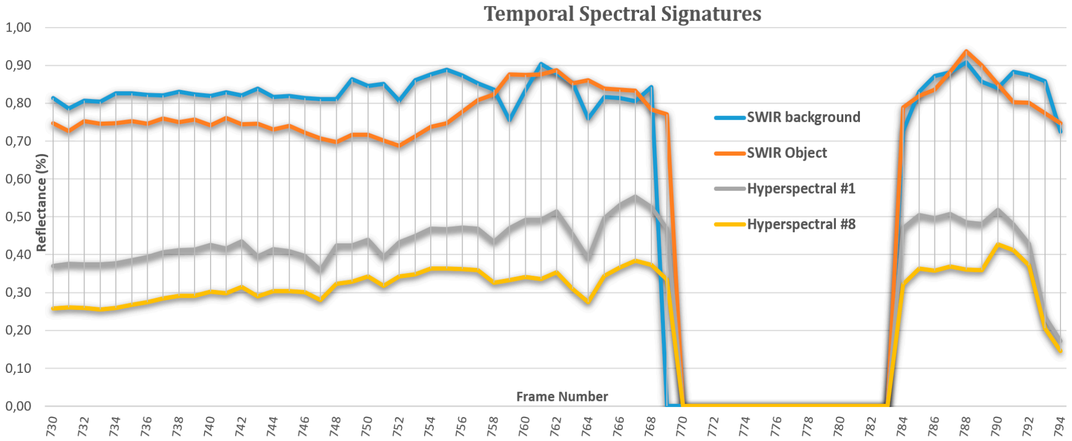


Figure 7. The temporal spectral signatures of both the target and the background are calculated and employed during the target recognition and verification step.

It should be mentioned that if the detector does not successfully detect a possible moving target (False Negative), then the spectral verification step is also unable to detect it, since the search space is limited to the considered bounding boxes. In particular, the background is considered as the rest of the area that surrounds each possible moving object inside its bounding box.

These statistics are captured and calculated at every frame and feed the fusion and data association process. The implemented data association module associates the n number of objects by selecting the optimal assignment of the n2 possible combinations. The selection is based on an equally weighted calculation of the spatial (image domain) and spectral (reflectance) distance, quantifying the likelihood that a detection (PMT) originates from a verified target from the previous frame. The smaller the distance the more likely it is to have originated from the verified targets.

The attraction of this approach is its simplicity, both conceptually and computationally, looking forward to real-time performance. However, it cannot robustly address overlapping objects and is not exploiting trajectory prediction frameworks like e.g., Kalman filtering. Moreover, since the goal here was to address monitoring tasks in challenging indoor and outdoor environments in order to preserve the spectral characteristics of every detected objects in all considered changing background/foreground conditions, apart from the background, nonrigid objects were also spectrally modeled (e.g., green/red/white smoke from smoke bombs and white smoke from fog machines) and were dynamically ignored from the spectral statistics/ data association once detected on the hyperspectral plane.

3. Experimental Results and Validation

3.1. Indoor/Outdoor Datasets, Implementation Details, and Evaluation Metrics

Several experiments have been performed in order to develop and validate the performance of the different hardware and software modules. Three benchmark datasets in challenging environments were employed for the quantitative and qualitative evaluation i.e., one outdoor (outD1) and two indoor ones (inD1 and inD2). In particular, the first one (outD1) was collected from the rooftop of a building overlooking a park walkway, situated above the Vrilissia Tunnel of Attica Tollways, in Northern Attica, Greece. The distance between camera and object was ~70 m. The conditions were sunny, but gusts of wind were disruptive to the data acquisition leading to not perfectly stabilized video sequence that was addressed by the coregistration software modules in near-real-time. For the generation of smoke, a smoke bomb was used, which generated a dense green cloud of smoke that lasted a couple of minutes. This resulted in a low number of acquired frames, in which smoke was present in the scene. The two indoor datasets (inD1, inD2) were collected at the School of Rural and Surveying Engineering, in the Zografou Campus of the National Technical University of Athens, Greece. The specific lecture room has an array of windows on two walls. The distance between the camera and ROI was approximately 15 m. The day was partially cloudy, with illumination changes from moving clouds, scattered throughout both datasets. Smoke was generated by the use of a fog/smoke machine. Dataset inD1 was collected in such a manner that the initial frames were clear, and the rest possessed a relatively small amount of smoke. The Dataset inD2 was created with the most challenging conditions: with the room almost full of smoke, with relatively lower intensity, contrast, and illumination changes derived from the partially opened windows.

Regarding the implementation details of the integrated supervised deep learning FR-CNN and YOLO methods which require a training procedure, Table 1 briefly overviews the procedure. Although, according to the literature, both deep learning methods required a huge amount of training, and we performed experiments with a limited amount of training data in order to evaluate their robustness for classes (like pedestrians, people, etc.) that have been already pretrained, albeit not in challenging, like the considered here, environments. Indeed, for the outdoor dataset (outD1) only 100 frames were used for training. This can be considered as a significantly limited number as moreover only one moving object appeared in all frames. The training process lasted seven hours for the FR-CNN and 6 h for the YOLO methods. For the indoor datasets, in order to examine the robustness of the training procedure we build one single training model by collecting a small number of the initial frames from each dataset. In particular, 156 frames were taken from inD1 and 100 frames from inD2. The training process lasted five hours for the FR-CNN and eight hours for the YOLO.

Table 1. Details regarding the training procedure of the FR-CNN and 'You Only Look Once' (YOLO) methods.

	outD1		**inD1 + inD2**	
	FR-CNN	**YOLO**	**FR-CNN**	**YOLO**
# of frames	100		256 (156 + 100)	
Duration	7 h	6 h	5 h	8 h
Steps	39,810	14,000	12,667	16,300
Loss	*from* 1.9 *to* 0.05	*from* 104 *to* 0.5	*from* 1.5 *to* 0.05	*from* 105 *to* 0.8

3.2. Quantitative Evaluation

The quantitative evaluation was performed on two stages: validation and testing. Validation was performed on a smaller number of annotated frames to provide an initial assessment of the overall achieved accuracy. Then, the validation procedure took place on relatively large number of frames where the ground truth and true positives (TPs), false positives (FPs), and false negatives (FNs) were counted after an intensive laborious manual procedure.

The quantitative evaluation results for the outdoor dataset outD1 are presented in Table 2. More specifically, the BS method resulted into the higher accuracy rates with precision at 0.90, recall at 0.97 and multiple object detection accuracy (MODA) at 0.88. The supervised methods (FR-CNN, YOLO) did not manage to perform adequately, mainly due to the limited training data (less than 100 frames with just one object) as well as the relatively small-size objects/targets. For those small objects it is difficult to retain strong discriminative features during the deep (decreasing in size) convolutional layers.

Table 2. Quantitative results from the outdoor dataset outD1.

	outD1 Dataset (Outdoor #1)—*Quantitative Evaluation*					
	Validation (69 Frames)			**Testing (156 Frames)**		
	BS	**FR-CNN**	**YOLO**	**BS**	**FR-CNN**	**YOLO**
Precision	-	0.46	0.16	0.90	0.46	0.16
Recall	-	0.44	0.16	0.97	0.44	0.16
MODA	-	0.44	0.19	0.88	0.44	0.19

In Tables 3 and 4, the results from the two indoor datasets are presented. In particular, in Table 3, the quantitative results for the inD1 dataset indicated that the BS method failed to successfully detect the moving objects in the scene. This was mainly due to the important illumination changes which were dynamically affecting every video frame. In particular, the windows around the room allowed the changing amounts of daylight of a cloudy day to affect constantly scene illumination resulting into challenging detection conditions for background modeling methods. Therefore, the BS resulted into numerous false positives and relative low accuracy rates. On the other hand, the supervised FR-CNN and YOLO methods resulted into relative high detection rates. On the validation phase, both algorithms perform in a similar manner resulting into very few false negatives (with a recall > 0.96 on both cases). Regarding false positives, the FR-CNN resulted into marginally more precise results (by approximately 5%). However, in the testing dataset, the precision scores reversed, with the YOLO detector to provide a high precision rate of 0.99. The FR-CNN resulted into more False Positives achieving a precision of 0.77. Regarding, false negatives, which is crucial for our proposed solution, the FR-CNN provided very robust results, achieving a recall rate of 0.97, which was 14% better than YOLO's 0.83. Therefore, while the reported overall accuracy favors YOLO by 8%, given our overall design and the crucial role of false negatives in monitor and surveillance tasks the FR-CNN can be considered as well.

Table 3. Quantitative results from the indoor dataset inD1.

	inD1 Dataset (Indoor #1)—*Quantitative Evaluation*					
	Validation (100 Frames)			**Testing (1390 Frames)**		
	BS	**FR-CNN**	**YOLO**	**BS**	**FR-CNN**	**YOLO**
Precision	-	0.91	0.86	0.24	0.77	0.99
Recall	-	0.96	0.99	0.83	0.97	0.83
MODA	-	0.89	0.93	0.23	0.75	0.83

Table 4. Quantitative results from the indoor dataset inD2.

	inD2 Dataset (Indoor #2)—*Quantitative Evaluation*					
	Validation (69 Frames)			**Testing (804 Frames)**		
	BS	**FR-CNN**	**YOLO**	**BS**	**FR-CNN**	**YOLO**
Precision	-	0.99	0.54	0.19	0.86	0.99
Recall	-	0.99	0.71	0.74	0.90	0.74
MODA	-	0.99	0.62	0.18	0.78	0.74

For the second indoor dataset, which had a significant presence of smoke, the calculated quantitative results are presented in Table 4. For the validation phase, the FR-CNN managed to fit better into the same, relatively small, training set. As in the previous case the YOLO did not manage to score high accuracy rates during the validation phase. However, in the large testing dataset, the YOLO method resulted overall into higher accuracy rates. The BS method failed again to provide any valuable results. The Recall and Precision detection rates were slightly lower compared to inD1 dataset, which can be attributed to the presence of smoke. Indeed, addressing heavy smoke in this indoor scene was a challenge for the detectors, resulting into a number of false negatives, as objects were less likely to stand out from the background. Moreover, sudden random spikes and illumination changes resulted in false positives; however in a lower rate that in inD1.

Moreover, in order to assess the overall performance of each detector in all datasets, in Table 5, the Overall Accuracy results are presented, for all indoor (left) and all (both indoor and outdoor) datasets (right). The performance of the considered detectors in terms of overall accuracy was reported at 0.43, 0.66, and 0.59 for the BS, FR-CNN, and YOLO methods, respectively. Of course, these results are largely constrained by the relatively lower performance of the deep architectures in the outdoor (outD1) dataset, which can be attributed to the small number of available frames (training set) for the demanding learning process of the deep convolutional networks. Despite the relatively similar performance in the overall precision rates of FR-CNN and YOLO, the first one delivered significant higher accuracy recall rates and thus performed in overall better than the other two. Still, however, the detection of small objects is a challenge which is mainly associated with the convolutional neural network processes.

Table 5. Overall accuracy results.

	Indoor Datasets			***All Datasets***		
	BS	**FR-CNN**	**YOLO**	**BS**	**FR-CNN**	**YOLO**
Precision	0.21	0.81	0.99	0.44	0.70	0.71
Recall	0.79	0.93	0.79	0.85	0.77	0.58
Overall Accuracy	0.20	0.76	0.79	0.43	0.66	0.59

3.3. Qualitative Evaluation

Apart from the quantitative evaluation, a qualitative one was also performed in order to better understand and report on the performance on the developed methodology and each integrated detector. In Figure 8, the indicative detection results based on the BS detector on the outD1 dataset are presented. For example, at frame #33, the outline of the detected object managed to describe adequately the ground truth. The finally detected target is in a similar way accurate, while the moving person is visible both on the SWIR and the hyperspectral images had being moving behind the low-density smoke. For frames #47 and #90, the outlines of the detected objects had a relatively irregular shape which didn’t perfectly match the actual ground truth. This can be attributed to the higher velocity of the detected object in this particular batch of frames as well as the movement behind relatively thick smoke clouds which challenged the developed detection methodology. As a result, the corresponding final bounding boxes included more background pixels than those in frame #33. It is also important to note that the object is not visible on any of the hyperspectral bands as the accumulated smoke is too dense.

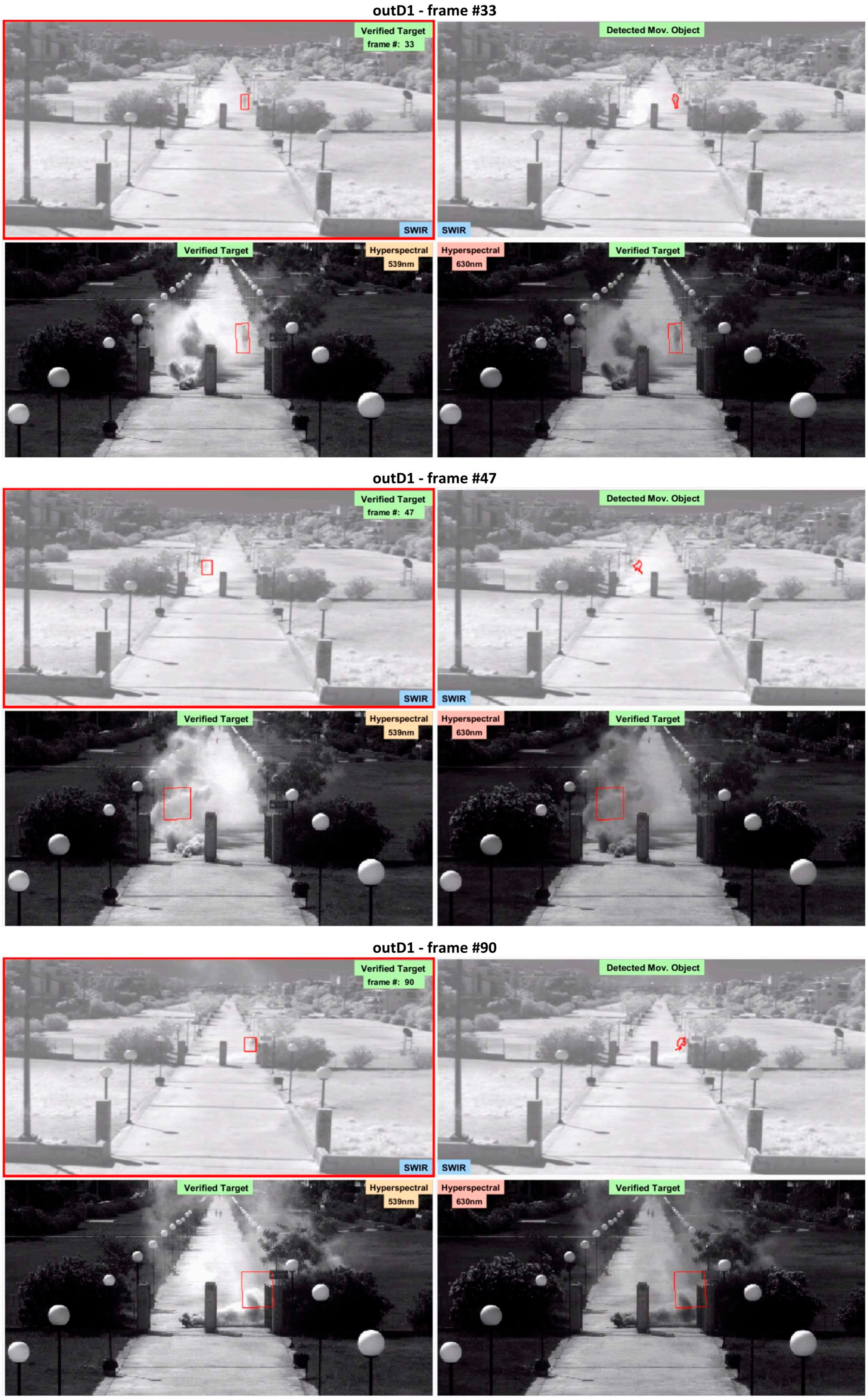


Figure 8. Detection results from the developed method (based on the BS detector) on the outD1 dataset. Frames #33, #47, and #90 are presented along with the intermediate 'detected moving objects' and finally 'verified targets' overlaid in the SWIR and hyperspectral (539 nm and 630 nm, respectively) bands.

In Figure 9, the indicative successful detection results on the outdoor outD1 dataset are presented from the integrated BS, FR-CNN, and YOLO detectors on the SWIR footage. In particular, at frame #33 the resulting detection from the BS method is presented with a red bounding box.

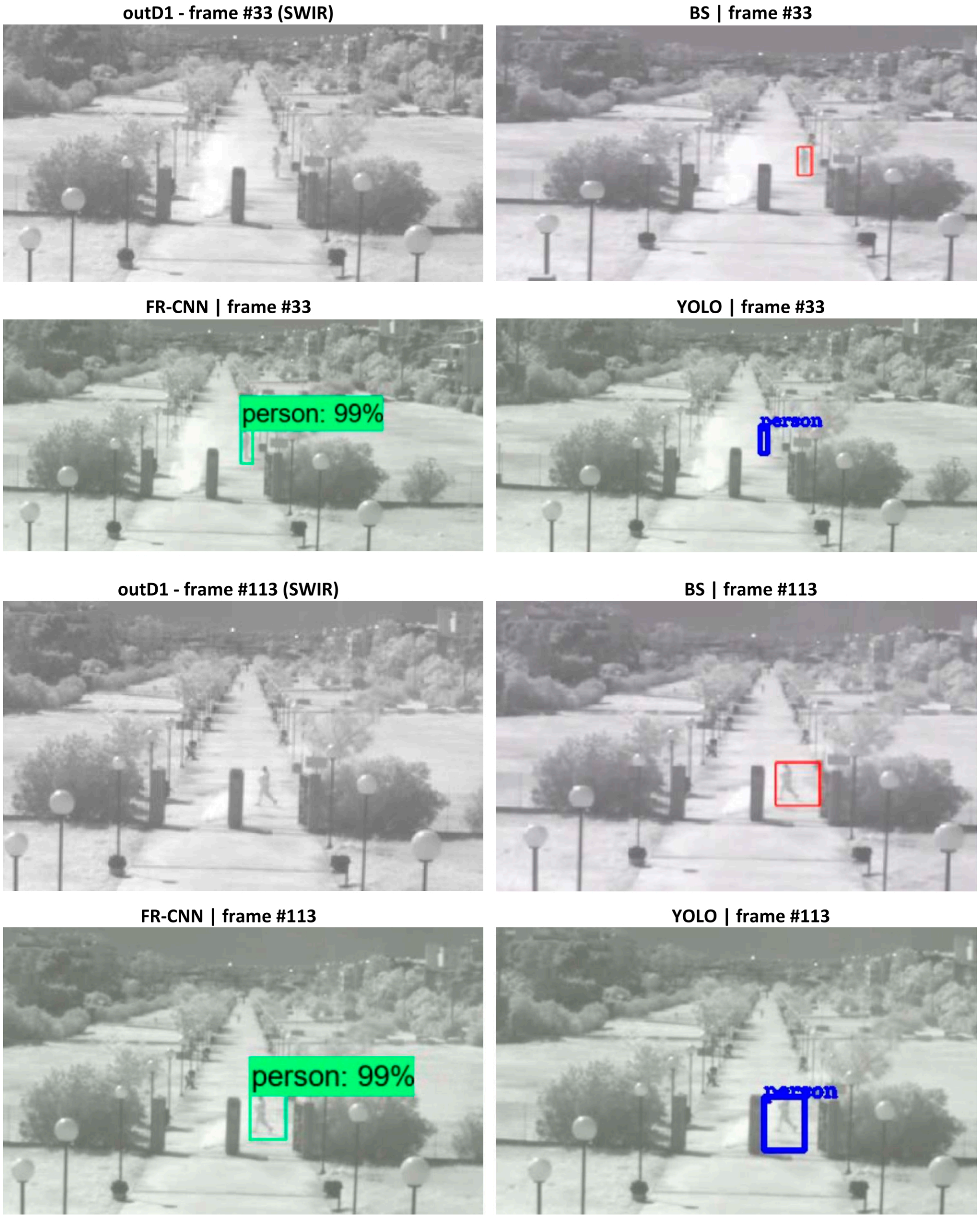


Figure 9. Indicative successful real-time detection results on the outdoor outD1 dataset from the integrated BS, FR-CNN, and YOLO detectors on the SWIR footage.

The FR-CNN result is presented with a green bounding box and the YOLO one with a blue bounding box. In all three cases, the detection result was adequate with the corresponding bounding boxes tightly describing the target. At frame #113, which is presented using the same layout as before, the resulting bounding boxes successfully included the desired target; however, they were relatively larger and contained more background than those of frame #33, affecting the object and background statistics and multitemporal modeling.

Moreover, in order to evaluate the overall performance of the developed approach, experimental results based on the BS detector are presented in Figure 10 with indicative frames (#033, #079, #087, and #108) from the outD1 dataset. In particular, the final detected moving targets are overlaid on the respective acquired SWIR image that was employed in the detection process, as well as on three hyperspectral bands centered approximately 476, 539, and 630 nm, respectively. The bounding boxes are outlined in red, and zoomed in views are also provided. It can be observed that the projection of the quadrilateral bounding box on the hyperspectral image plane, distorts it slightly into a more general polygon shape.

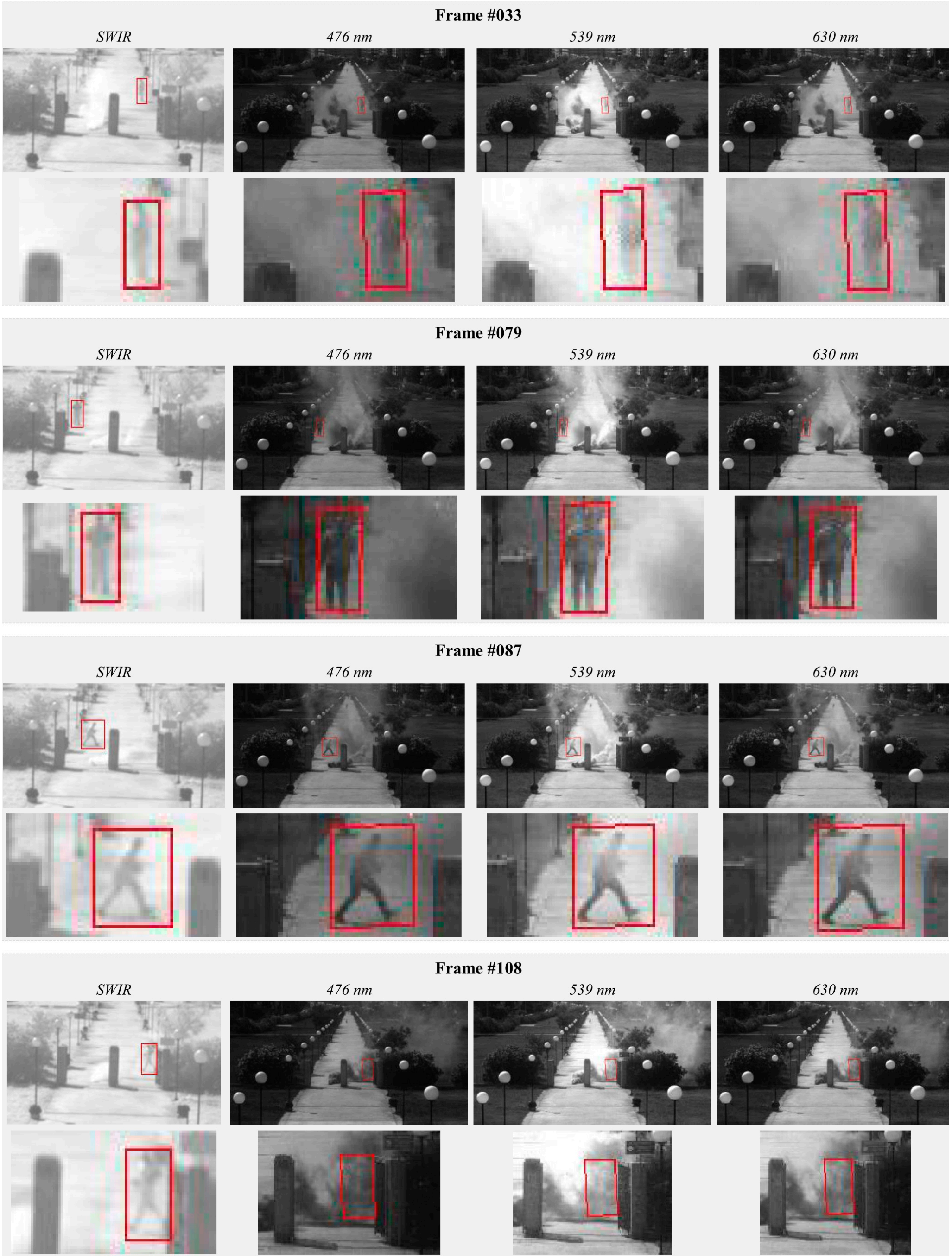


Figure 10. Experimental results after the application of the developed framework in the outD1 dataset (indicative frames #033, #079, #087, and #108). For each frame the SWIR and three hyperspectral bands (476, 539, and 630 nm) are presented. The detected targets are annotated with a red color onto the SWIR. Their projections are also shown onto the hyperspectral images. Zoom-in views are, also, provided.

In particular, for frame #033 the moving target can be hardly discriminated from the background in the SWIR imagery. However, the developed detection and data association procedure managed to correctly detect the target and correctly project its bounding box onto the hyperspectral image plane.

At frame #079, the detection worked also adequately with the bounding box containing 100% of the moving target. At frame #087, even with the target moving relatively fast (running), the developed detection procedure achieved correct localization of the SWIR image and correct projection onto the hypercube. At frame #108 which represents a rather challenging scene with the presence of dense smoke cloud, the algorithm managed to detect and successfully track the moving object. The projection on the hyperspectral cube indicated relatively high reflectance values that were not matching the actual target or background but the green dense smoke. The data association term indicated high confidence levels in the initial SWIR detection and, therefore, the final verified target correctly detected the object boundaries.

Regarding the performance on the indoor datasets in Figures 11 and 12, indicative detection results overlaid onto the SWIR data based on the integrated BS, FR-CNN, and YOLO detectors are presented. In particular, as has been already reported from the quantitative results (Section 3.2), the BS method failed to deliver acceptable results in the challenging indoor environments. This can be attributed to changes in illumination which caused numerous false positive cases. False negatives occurred also from the BS method. In the inD1 dataset, at frame #31 (Figure 10, top), the reflection of the person on the table was wrongly detected as a target by the BS method. The FR-CNN presented the best results in this frame if we consider that the resulting bounding box from the YOLO method did not describe optimally the person's silhouette. At frame #69 (Figure 10, middle), there are two moving targets in the ground truth and the BS method detected only one and in multiple instances.

The FR-CNN delivered the best results with the YOLO not adequately describing the corresponding silhouettes. At frame #130, the BS method resulted into a false negative while again both the supervised methods managed to detect the object, with the FR-CNN detecting more precisely the target's silhouette.

Regarding the performance on the challenging indoor dataset inD2 with the presence of dense smoke and rapidly changing illumination, in Figure 11 indicative detection results are presented. Similar with the previous indoor case, the BS method failed to detect adequately the moving objects, due to the fact that smoke was present, lowering the intensity and contrast of brightness values. At frame #34, all considered methods managed to detect the target with the BS delivering two different instances/segments. At frame #68, the FR-CNN resulted in a false positive which, however, was associated with a relatively lower confidence level (65% against >85%). At frame #127, the ground truth indicates two moving objects which are partially occluded since the silhouettes of the two moving people are overlapping. The YOLO and BS methods failed to detect both targets and delivered instead a relatively large bounding box containing both objects. The FR-CNN output highlighted both existing targets accurately but delivered in addition a false positive. At frame #170, the ground truth indicated two targets in this highly challenging scene and both BS and FR-CNN methods achieved their detection. The single bounding box delivered by the YOLO detector, was located in the image region between the two objects, partially including both of them. It should be noted that the aforementioned indicative results were in favor of the BS method since the goal was to compare the detection performance in terms of precision also in silhouette detection. In most cases the BS method failed to accurately detect targets, delivering numerous false positives and negatives.

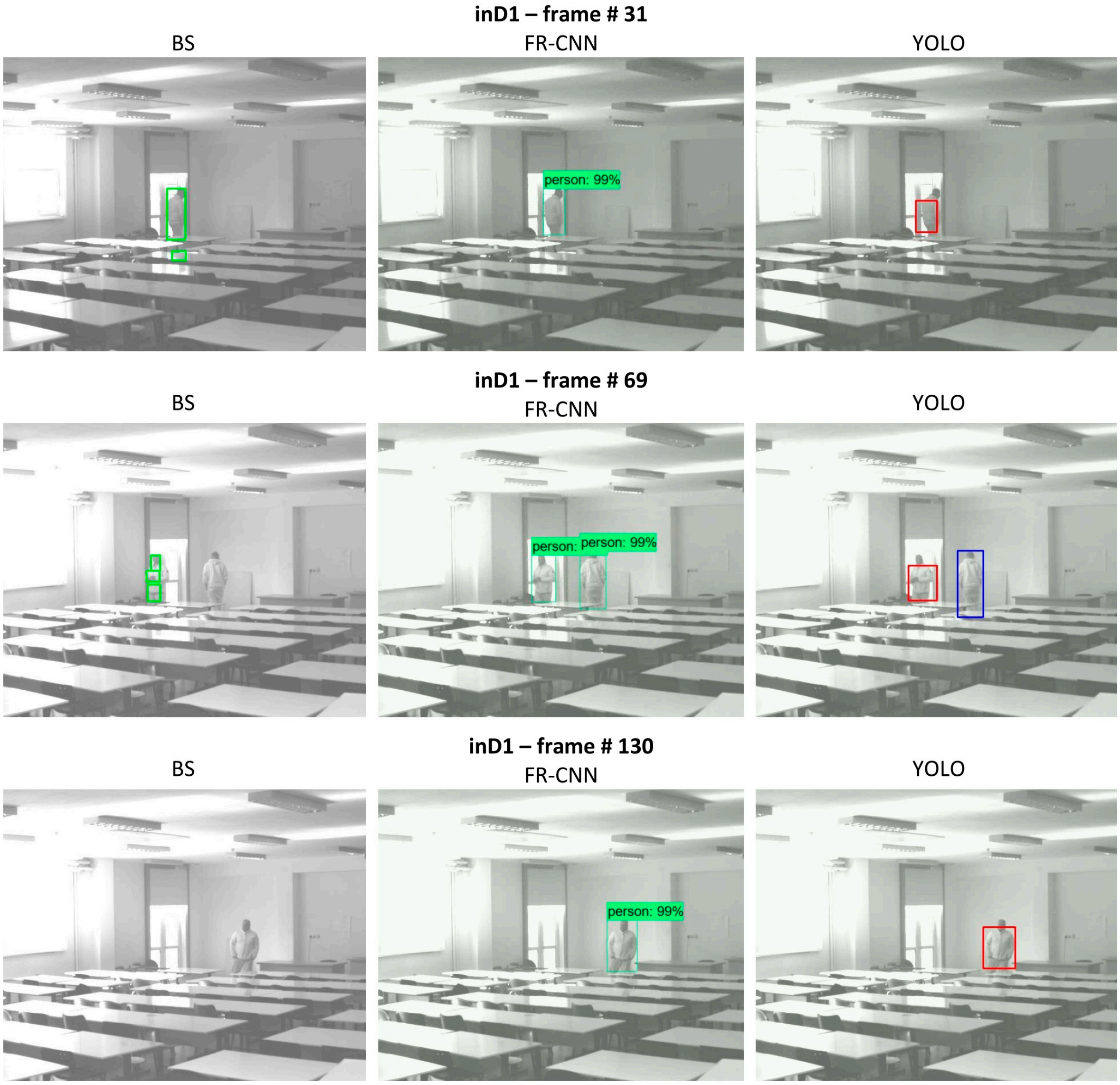


Figure 11. Indicative detection examples overlaid onto the SWIR data for the first indoor dataset (inD1) based on the integrated BS, FR-CNN, and YOLO detectors.

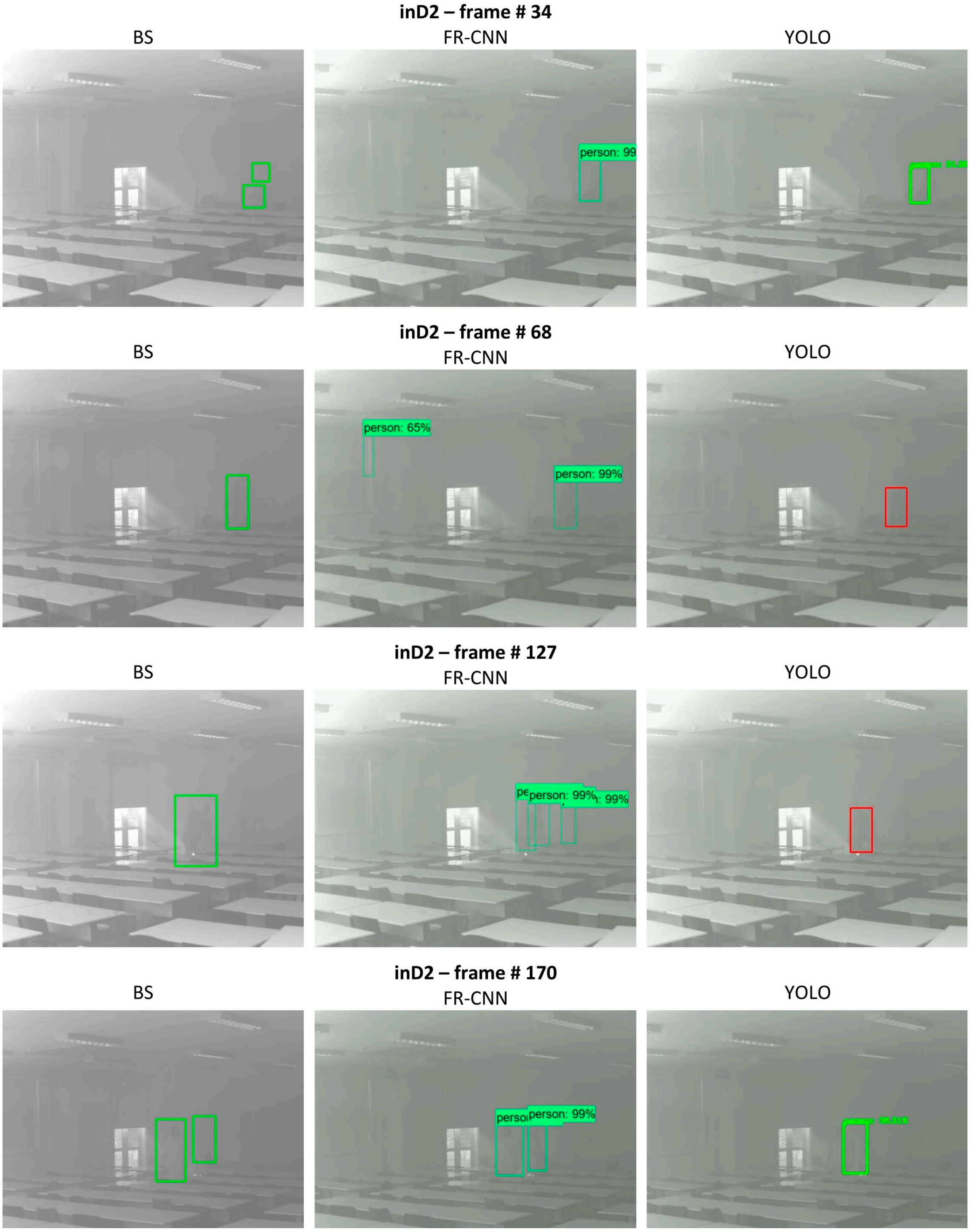


Figure 12. Indicative detection examples overlaid onto the SWIR data for the second indoor dataset (inD2) with the presence of significant smoke based on the integrated BS, FR-CNN, and YOLO detectors.

4. Conclusions

In this paper, we propose the use of a multisensor monitoring system based on different spectral imaging sensors that can be fruitfully fused towards addressing surveillance tasks under a range of critical environmental and illumination conditions, like smoke, fog, day and night acquisitions, etc. These conditions have been proved challenging for conventional RGB cameras or single sensor setups. In particular, we have integrated and developed the required hardware and software modules in order to perform near real-time video analysis for detecting and tracking moving objects/targets. Different unsupervised and supervised object detectors have been integrated in a detector-agnostic manner. The performed experimental results and validation procedure demonstrated the capabilities of the proposed system to monitor critical infrastructure in challenging conditions. Regarding the integrated detectors, the BS method although having performed successfully in an outdoor scene (with relatively smaller number of video frames and limited number of training sets) resulted into numerous false positives and false negatives, in the indoor environments with significant challenges in varying illumination conditions with or without the presence of smoke. The deep learning architectures managed even with a relatively small number of training set to tune their hyperparameters. In particular, FR-CNN and YOLO achieved high accuracy rates under the most difficult and challenging scenes. Apart from the normally required huge training set, another shortcoming that was observed regarding the FR-CNN and YOLO was the limited performance on small (few pixels) objects attributed to the several, decreasing in size, convolutional layers of their deep architectures. Among the future work is the development and validation of a multiple object tracking of different types (e.g., vehicles, bicycles, and pedestrians) in challenging scenes and environments.

Author Contributions: Z.K. designed the methodology, implemented the software, performed experiments and validation, and wrote, edited, and reviewed the manuscript. K.V. implemented part of the methodology and performed the experiments and the validation procedure. K.K. conceptualized the approach, designed the methodology, implemented part of the software, and wrote, edited, and reviewed the manuscript.

Funding: Part of this research was funded by the EU ZONeSEC (https://www.zonesec.eu/) and eVACUATE (http://www.evacuate.eu/) projects under grant agreements No 607292 and No 313161.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Manolakis, D.; Truslow, E.; Pieper, M.; Cooley, T.; Brueggeman, M. Detection Algorithms in Hyperspectral Imaging Systems: An Overview of Practical Algorithms. *IEEE Signal Process. Mag.* **2014**, *31*, 24–33. [CrossRef]
2. Pieper, M.; Manolakis, D.; Cooley, T.; Brueggeman, M.; Weisner, A.; Jacobson, J. New insights and practical considerations in hyperspectral change detection. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4161–4164.
3. Tochon, G.; Chanussot, J.; Dalla Mura, M.; Bertozzi, A.L. Object Tracking by Hierarchical Decomposition of Hyperspectral Video Sequences: Application to Chemical Gas Plume Tracking. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4567–4585. [CrossRef]
4. Kandylakis, Z.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Multiple Object Tracking with Background Estimation in Hyperspectral Video Sequences. In Proceedings of the IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Tokyo, Japan, 2–5 June 2015.
5. Lu, X.; Liu, W.; Mou, L. Semi-Supervised Multitask Learning for Scene Recognition. *IEEE Trans. Cybern.* **2015**, *45*, 1967–1976. [PubMed]
6. Li, X.; Zhao, B.; Lu, X. A General Framework for Edited Video and Raw Video Summarization. *IEEE Trans. Image Process.* **2017**, *26*, 3652–3664. [CrossRef] [PubMed]
7. Lu, X.; Yuan, Y.; Zheng, X. Joint Dictionary Learning for Multispectral Change Detection. *IEEE Trans. Cybern.* **2017**, *47*, 884–897. [CrossRef] [PubMed]
8. Fan, C.T.; Wang, Y.K.; Huang, C.R. Heterogeneous Information Fusion and Visualization for a Large-Scale Intelligent Video Surveillance System. *IEEE Trans. Syst. Man Cybern. Syst.* **2017**, *47*, 593–604. [CrossRef]
9. Liao, W.; Yang, C.; Ying Yang, M.; Rosenhahn, B. Security Event Recognition for Visual Surveillance. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *4*, 19–26. [CrossRef]
10. Gao, S.; Ye, Q.; Xing, J.; Kuijper, A.; Han, Z.; Jiao, J.; Ji, X. Beyond Group: Multiple Person Tracking via Minimal Topology-Energy-Variation. *IEEE Trans. Image Process.* **2017**, *26*, 5575–5589. [CrossRef] [PubMed]
11. Lee, D.S. Effective Gaussian mixture learning for video background subtraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 827–832. [PubMed]
12. Heikkilä, M.; Pietikäinen, M. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 657–662. [CrossRef] [PubMed]

13. Dong, P.; Wang, S.; Xia, Y.; Liang, D.; Feng, D.D. Foreground Detection with Simultaneous Dictionary Learning and Historical Pixel Maintenance. *IEEE Trans. Image Process.* **2016**, *25*, 5035–5049. [CrossRef] [PubMed]
14. Bilal, M.; Khan, A.; Karim Khan, M.U.; Kyung, C. A Low-Complexity Pedestrian Detection Framework for Smart Video Surveillance Systems. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 2260–2273. [CrossRef]
15. Zeng, D.; Zhu, M. Multiscale Fully Convolutional Network for Foreground Object Detection in Infrared Videos. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 617–621. [CrossRef]
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 27–29 October 2017.
19. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Copenhagen, Denmark, 11 August 2017.
22. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2015; pp. 4959–4962.
23. Kandylakis, Z.; Karantzalos, K.; Doulamis, A.; Karagiannidis, L. Multimodal Data Fusion for Effective Surveillance of Critical Infrastructures. In Proceedings of the ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Hamburg, Germany, 28–29 November 2017; pp. 87–93.
24. Karantzalos, K. Intrinsic dimensionality estimation and dimensionality reduction through scale space filtering. In Proceedings of the 16th International Conference on Digital Signal Processing, Santorini-Hellas, Santorini, Greece, 5–7 July 2009; pp. 1–6.
25. Karantzalos, K.; Sotiras, A.; Paragios, N. Efficient and Automated Multi-Modal Satellite Data Registration through MRFs and Linear Programming. In Proceedings of the Computer Vision and Pattern Recognition Workshops, Zurich, Switzerland, 6–7 September 2014; pp. 1–8.
26. Vakalopoulou, M.; Karantzalos, K. Automatic Descriptor-based Co-registration of Frame Hyperspectral Data. *Remote Sens. Environ.* **2014**, *6*, 3409–3426. [CrossRef]
27. Makantasis, K.; Nikitakis, A.; Doulamis, A.D.; Doulamis, N.D.; Papaefstathiou, I. Data-driven background subtraction algorithm for in-camera acceleration in thermal imagery. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2090–2104. [CrossRef]