Feature Article:

DOI. No. 10.1109/MAES.2020.3006410

A Survey of Multimodal Sensor Fusion for Passive RF and EO Information Integration

Asad Vakil, Jenny Liu, **Oakland University, Rochester, MI 48309 USA**
Peter Zulch **, Information Directorate, Air Force Research Laboratory, Rome, NY 13441 USA**
Erik Blasch **, Air Force Office of Scientific Research, Arlington, VA 22203 USA**
Robert Ewing, **Sensors Directorate, Air Force Research Laboratory, Dayton, OH 45433 USA**
Jia Li **, Oakland University, Rochester, MI 48309 USA**

INTRODUCTION

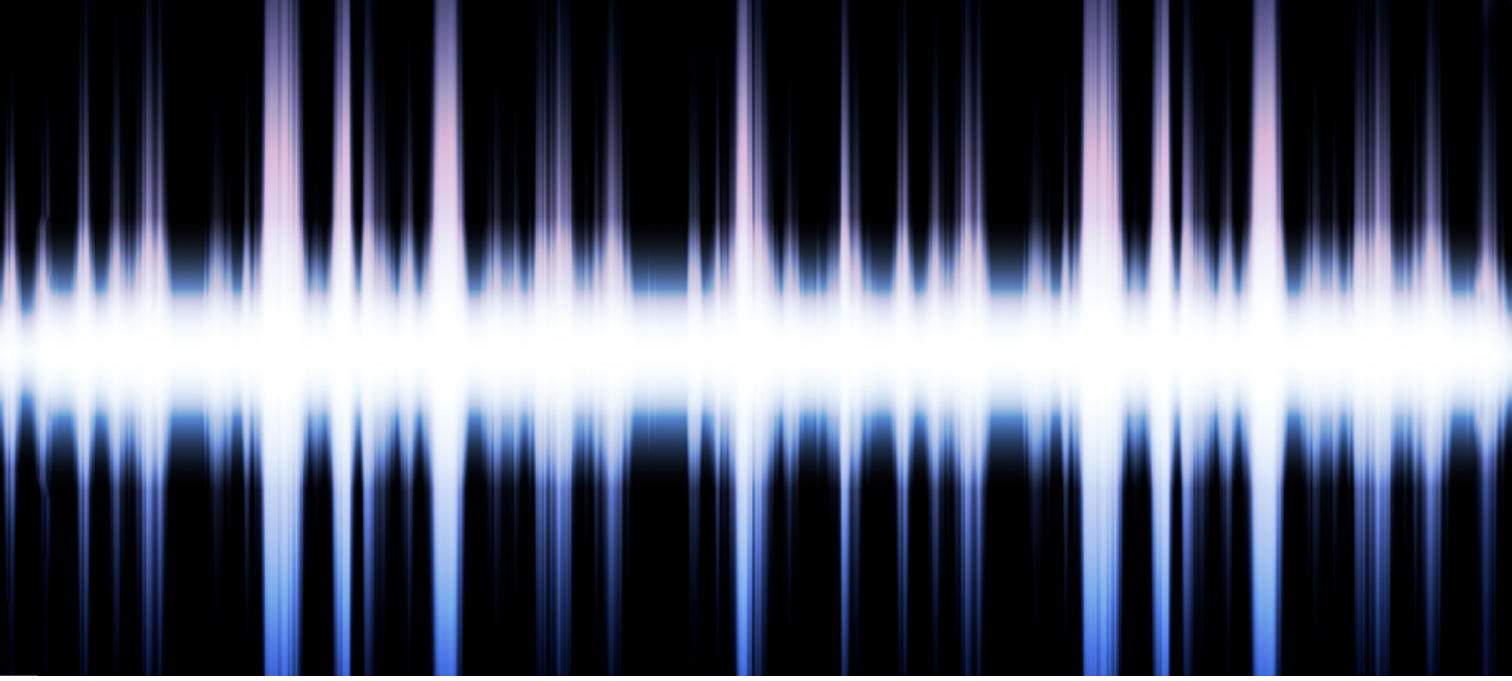
In the information age, there exist many types of sensors for collecting data from an environment, such as pressure, radar, acoustic, chemitcal, electromagnetic, thermal, proximity, and optical sensors. Each of these independent modalities has its advantages and drawbacks. Whether the modality is active or passive in nature, or vulnerable to different forms of interference, the ability to properly utilize all the available information in an efficient manner is invaluable for system performance. There are many methods of achieving sensor fusion. While there is no optimal solution to integrating information effectively, there are merits and limitations of certain approaches based on the type of modality and the nature of the dataset. The end goal of sensor fusion is to reduce uncertainty from multiple data sources in order to perform reliably and robustly.

Sensor fusion is a critical task in a wide range of applications, such as security, healthcare, weather forecasting, Internet of Things, navigation, and communication. Many technologies and services used daily are examples of sensor fusion, such as autonomous driving [1]. Among these applications, there is a prevalent benefit to taking advantage of multiple sources of information if the classification method or algorithm is capable of exploiting the relationship between the input data. How these modalities are combined can differ greatly based on the overall objective, and with the methods and sensors used. In some cases, the system of sensors might simply be redundant in nature, such as having a backup smoke detector in a room, while other systems might rely on complementary sensors such as a network of security cameras. For the purposes of this survey article, the focus will be on RF and EO modalities, and the different methods of fusion using such modalities.

The use of EO modalities such as still images, full motion video (FMV), and Infrared (IR) have a number of applications in target identification and tracking in challenging environmental conditions. Similarly, there are a number of applications for RF based modalities, such as radar or RFID aiming to achieve similar objectives over conditions that would pose a challenge for EO modalities, and provide information different than EO modalities. The heterogeneous fusion of EO/RF modalities would improve performance with the number of exploitable features it provides. While most of the traditional applications for EO/RF sensor fusion use active RF modalities, there are numerous benefits of passive RF modalities. Hence, the fusion of passive RF data with EO data is the focus of this review article.

In this article, an overview of existing multimodal EO/RF sensor fusion is presented along with a review of various sensor fusion applications, schemes, models, and approaches. The primary focus will be on the application of fusing EO and passive RF data for the purposes of detection and tracking. The aim of this article is to provide a general insight into the state-of-the-art technologies for EO/RF sensor fusion, and to discuss the features and architecture proposed for EO/RF sensor fusion for the purposes of object assessment. The rest of this article is

Authors' current addresses: Asad Vakil, Jenny Liu, and Jia Li, Department of Electrical and Computer Engineering, Oakland University, Rochester, MI 48309 USA (e-mail: avakil@oakland.edu). Peter Zulch, Information Directorate, Air Force Research Laboratory, Rome, NY 13441 USA. Erik Blasch, Air Force Office of Scientific Research, Arlington, VA 22203 USA. Robert Ewing, Sensors Directorate, Air Force Research Laboratory, Dayton, OH 45433 USA.
Manuscript received February 29, 2020, revised June 11, 2020; accepted June 18, 2020, and ready for publication June 29, 2020.
Review handled by Dietrich Fraenken.
0885-8985/21/$26.00 © 2021 IEEE



Credit: Image licensed by Ingram Publishing

organized as follows. Common definitions, conceptualizations, and related literature in EO/RF fusion are discussed. Details are provided for the experimental design and architecture of fusion of EO and RF neural network (FERNN), and the results of those experiments are discussed. Finally, we present concluding remarks and discusses future research directions.

REVIEW OF STATE-OF-ART TECHNOLOGIES

SENSOR FUSION ARCHITECTURE AND TERMINOLOGY

Sensor fusion is a common method for the analysis and utilization of information and is an essential part of many applications such as data mining and machine learning. As there exist many applications and unique datasets, different approaches have been developed for each of the corresponding applications. The nature of sensor modalities and sources of information determine the approach to process and correlate with different information sources.

Describing and classifying these sensor fusion methods is therefore important to distinguish the fundamental principles taken to implement sensor fusion. An important aspect of classifying fusion between sensors is the nature of the sensors themselves. The source of data can generally be categorized as heterogeneous versus homogeneous. Homogeneous data are typically simpler to fuse together, due to the data already being in a compatible form. If the sensor data are heterogeneous, most fusion methods require some level of preprocessing and a compatible means of feature extraction.

These considerations can vary depending on the specific application, the form of the input information, and how the system interprets the input information. For example, synthetic aperture radar (SAR) images are a type of RF modality, but its fusion with EO images does not necessarily require any complicated preprocessing beyond registration as both modalities are in the same 2-D matrix format.

While there is no common model of sensor fusion or any singular comprehensive system of classifying sensor fusion methods, the majority of existing models propose partitioning of the information fusion architecture based on how the source input information undergoes preprocessing, feature extraction, pattern processing, situation assessment, and decision making [2]. Classifications can be based on the communication scheme between sensors prior to fusion, how the information is processed between classification algorithms after preprocessing and feature extraction by the levels where fusion takes place in the fusion architecture. Level-based classifications generally categorize fusions as low-level, mid-level, and high-level sensor fusion, normally corresponding to the terms data-level, feature-level, and decision-level fusion, respectively [3].

Among the earliest widely adopted sensor fusion classification criteria is from Dasarathy [4]. Dasarathy divides data fusion into five categories, data in-data out, data in-feature out, feature in-feature out, feature in-decision out, and decision in-decision out. The Dasarathy model classifies sensor fusion techniques into five categories by their respective input and output in order to avoid ambiguity between data selection, feature extraction, feature fusion, pattern recognition, and decision analysis.

In the context of EO/RF sensor fusion, data input for EO can vary from still images to video, while data input for RF can receive signal strength (RSS) values. The EO features that could be extracted include decluttered images, corner detection, segmented regions of interest, etc. For RF features, time of arrival (TOA), time difference of arrival (TDOA), angle of arrival (AOA), and Doppler are often calculated. The decisions used as an input could be detection or motion estimation results based on individual sensor modality, while the output of fusion provides low-level object assessment.

Table 1.

Preclassification Methods		
Methodology	**Characteristics**	**References**
Data Level Fusion	End-to-end learning, in the context of EO/RF the raw processing of modalities such as image/video/laser with radar/I/Q data/SAR/RSS to achieve classification.	[5], [6], [7], [8]
Feature Level Fusion	Achieving classification by extracting meaningful features such as decluttered images obtained via spatio-temporal filtering, edge detection and TDOA/TOA/AOA to improve classification.	[9], [10], [11], [12], [13]

Aside from the aforementioned sensor fusion criteria, the types of sensor fusion can be split between preclassification and postclassification. As the name suggests, *preclassification* sensor fusion takes place before any classification occurs, while postclassification fusion takes place afterward. Preclassification entails both sensor-level and feature-level fusion. Several examples of EO/RF fusion in the preclassification category are listed in Table 1.

For sensor fusion schemes that use low- or data-level fusion, the expectation is that the fused data are more informative and synthetic than the original sets of information alone. End-to-end learning has been conducted in the past for EO modalities, as the end-to-end learning classification of images is one of the most rudimentary neural network (NN) models. NN have also been implemented for RF applications, such as obtaining radio spectrum feature vectors [7], wireless signal identification [5], and cognitive radios [8].

Feature extraction for EO/RF modalities vary with applications. The procedure of spatio-temporal alignment, data association and correlation, and grouping techniques such as clustering, or state estimation, might be implemented in order to improve the performance. The expectation is that the input of structures would help with classification, tracking, or identification; thereby transforming the raw data into meaningful features for fusion and generating a decision.

Postclassification, also referred as decision-level fusion, has several variations, which are summarized above in Table 2. While the majority of EO/RF sensor fusion applications are decision- or upstream-level fusion at most, in situations where the number of classes are high, the use of postclassification fusion is desirable. Abstract-level fusion is the simplest form of postclassification fusion, which includes methods such as majority voting and weighted majority voting [9]. An example of abstract-level sensor fusion for EO/RF modalities can be seen in [14], in which SAR and EO images are fused through multilevel decision fusion before the classification with majority voting.

Rank-level fusion generates and assigns ranks to classes normally using either class set reduction or class set reordering methods to select the smallest possible subset of decisions that contain a correct class or to generate a ranking of classes in which the correct class has the highest possible ranking [15]. Another type of postclassification information fusion is measurement-level fusion [17], which can be further divided into classification and combination approaches. Classification sensor fusion uses multiple classification methods to consolidate scores from each method, also referred to as score-level fusion

Table 2.

Postclassification Methods		
Methodology	**Characteristics**	**References**
Abstract Level Fusion	Comparison of different classifier decision outputs	[7], [14], [15]
Rank Level Fusion	Comparison of decisions from different classifiers based on the class set reduction or class set reordering	[9], [15], [16]
Measurement Level Fusion	Integration of classifiers by normalizing individual classification schemes	[15], [17]
Dynamic Classifier Selection	Multilevel fusion of different classifiers that are optimized based on their respective performance at different levels	[15], [18], [19], [20], [21], [22], [23]

Table 3.

Sensor Relationship Fusion Based Schemes		
Methodology	**Characteristics**	**References**
Competitive Fusion	Independent classification from separate sources such as video or radar input that improve reliability and error detection, normally through methods such as decision level fusion or voting.	[10], [24]
Complementary Fusion	In the context of EO/RF fusion, the spatial and spectral benefits of both modalities can be used improve the performance of a fusion system by exploiting the overlap between the sensors.	[19], [25]
Cooperative Fusion	Fusion of EO, RF, and other modalities in order to provide a complete picture of the environment that the individual features and input data alone cannot.	[16], 26], [27]

[28]. In the case of combination approach, the framework integrates and normalizes and uses the weighted sums of the individual classification schemes [29]. Finally, the dynamic classifier selection uses the results of the classifier most likely to provide an accurate result for the specific pattern input [18], a method also known as a winner take all approach, or the associative switch [30].

Sensor fusion schemes can also be distinguished by the relationship of the sensor modalities used with respect to the application, which leads to the categories of competitive, complementary, and cooperative fusion. These relationship based schemes are summarized in Table 3. *Competitive fusion*, sometimes referred to as redundant fusion, uses each individual modality to deliver independent measurements of the same property. The primary benefit of such fusion schemes is the improved reliability and accuracy, typically used in high-level fusion, such as voting. An example of competitive sensor fusion can be found in [10], where a noncooperative EO/RF sensor fusion was used to improve the detection performance of unmanned aerial system (UAS). Camera input and radar input were mapped into a NED (north-east-down) coordinate system while the combined observations were processed through the integrated system. The use of a competitive data fusion scheme improved the system's situational awareness compared to the standalone single modality system, while maintaining the same reliability and false positive identifications as the single modality system.

Complementary sensors do not directly depend on each other but can be combined to give a more complete image of the phenomenon under observation. Many examples of *complementary sensor fusion* can be found in decision making algorithms, since this type of fusion typically occurs at the raw data level. Some notable examples include deep learning, hidden Markov models, and support vector machine (SVM). Fusion algorithms that can handle end-to-end learning from heterogeneous sensor input belong to this category. An example [31] uses a fully convolutional neural network (CNN) and a traditional extended Kalman filter to combine LiDAR, camera, and radar data for road detection of autonomous vehicle. The architecture for its fusion framework uses the complementary inputs and is designed to tolerate the individual sensor's failure. The redundant sensor fusion scheme enabled an extremely robust and reliable system of detection.

Cooperative sensor fusion uses the information provided by two or more independent sensors to derive information that would not otherwise be available from the sensors if they operated independently. Cooperative fusion is typically used in applications such as triangulation, from which multiple RF receivers can locate what the individual receivers cannot. Due to the nature of cooperative fusion and the fact that the fusion scheme exploits the unique data that different modalities and information sources can provide, many cooperative sensor fusion methods are used for the heterogeneous data. An example [26] realizes sensor fusion between EO/RF sensors, radar, and IR via an interacting multiple model (IMM) algorithm. The radar is used as a noncooperative sensor to improve robustness of the fusion model for the purposes of implementing a sense-and-avoid system for UAS. With the use of sensor inputs cooperating with each other for the purposes of air traffic detection and a noncooperative radar input used as a means of independently verifying the state of traffic, the IMM tracking could produce accurate position and velocity vectors which provide a more reliable trajectory prediction than the IMM tracking with only radar input.

Other communication classification systems for sensor fusion include decentralized, also known as distributed, centralized, and hierarchical. Traditionally, most sensor fusion architectures are *centralized*, pooling the available information in order to output a decision or classification, with the centralized architecture providing measurements to a common unit to achieve sensor fusion. In comparison, a *decentralized* architecture implies there is no communication between the sensor nodes, with each node using its own processing abilities to fuse local information with the

information from its peers. Decentralized sensor fusion is accomplished autonomously, occurring at no single point. In contrast, distributed nodes interchange data at a given communication rate. The benefits of a decentralized sensor fusion scheme can be seen in [32] in which the use of sparse approximation (Joint Sparse Representation) is capable of achieving the same estimation result as a centralized algorithm while significantly reducing the communication cost.

Hierarchical methods are a hybrid of the other techniques [33], instead of performing sensor fusion at different levels within the architecture's hierarchy, some using feedback while others are simply feedforward. The nature of the system still centralizes its output to some degree, but with the added benefit of reducing the calculations needed by designating separate nodes to achieve a form of decision or feature-in decision-out fusion. There are many variations on this sort of sensor fusion architecture, Kahler *et al.* [34], for example, used a two-step probabilistic cue integration for the purposes of achieving object tracking in three dimensions while Song *et al.* [35] used a hierarchical architecture to reduce the computational complexity of a decentralized data fusion algorithm by being placed between the clusters and fusion center.

Another notable sensor fusion categorization discriminates between upstream and downstream data fusion. *Upstream data fusion* is the processing, exploitation, and fusion of sensor data as closely to the raw sensor data feed as possible. In contrast to downstream (post-decision) fusion, upstream data fusion processes the input information and minimizes the data loss that can result from conventional data reduction methods. Upstream fusion improves the performance by accessing the data at an appropriate point in the processing chain near the data source, strategically chosen to acquire the desired data within the earliest point in the fusion architecture [36]. Upstream fusion was used by Garagic *et al.* [12] to integrate FMV and passive RF data, using multimodal emitter tracking and localization architecture. The algorithm combines deep learning and feature manifold representations to achieve upstream sensor fusion.

METHODOLOGIES IN EO AND PASSIVE RF SENSOR FUSION

When dealing with modalities that involve radio frequency (RF) and electro-optical (EO) sensors, the focus for fusion has traditionally been on active RF sensors. Doppler radar and imaging radar (e.g., side-looking airborne radar), as well as other similar active RF sensors, are well suited for tracking a moving target when used in concert with a form of EO modality. However, the combined exploitation of the two sensor modalities can still be improved [37]. That being said, RF modalities excel in providing range, angular, and spectral resolution of information from RF modalities and the benefits of combining RF data with higher spatial resolution of EO-based sensors is extremely desirable for detection and tracking.

There are a number of RF-based modalities that are used in applications such as tracking, proximity, localization, and detection. While many EO modalities are intuitively easier for humans to understand and to implement for similar applications, unlike RF modalities; RF approaches to such problems are less susceptible to problems such as ocular interference. RF-based sensors are not limited by factors like visual interference from natural phenomenon such as fog, clouds, snow, or any other form of weather that would otherwise interfere in the collection of EO data. In addition, RF-based sensors can provide repetitive coverage over a wide geographical area, and in doing so, can determine the precise distance and velocity of a target. As mentioned earlier in this section, many RF modalities in detection and tracking applications utilize active RF sensing.

One example of active RF/EO fusion was given by Seo [38], where SAR data are used in combination with multispectral (MS) images using random forest regression. This approach fuses together the SAR image containing the surface roughness characteristics while retaining the spectral characteristics of the MS images, before being fed into a modified random forest regression algorithm, outperforming the comparison algorithms with the same KOMPSAT-5 and Landsat-8 OLI datasets. The use of SAR imaging to facilitate fusion between RF and EO sensor data is a commonly used approach for automatic ground target recognition. While Orynbaikyzy *et al.* [39] used a more traditional algorithm, Kim *et al.* [40] utilized a double weighted NN fusion scheme that uses sum-based linear fusion to generate features and a NN- based fusion at the decision level.

Another example of active EO/RF fusion is Bui *et al.* [41] in which SAR and multisource satellite imagery were fused together at the data, feature, and decision levels. As mentioned earlier [41], 2-D radar, EO, and IR sensor data were used with an extended Kalman Filter and State Vector Fusion to track a target in 3-D Cartesian coordinates in a Monte Carlo simulation. In Zhang *et al.* [13], the use of Rao–Blackwellized particle filtering was implemented in order to fuse the asymmetrical fields of view for the radar and EO modalities. The application of adaptive waveform design and control incorporated dynamic agility selection in order to improve the performance of the system's ability to track an unknown number of targets using the two modalities. Other notable algorithms include the use of sparse representation in order to combine medium wavelength IR (MWIR) cameras and RF Doppler sensors for vehicle tracking [42], which uses a joint sparse approximation approach for multimodality images.

There are a number of advantages for the implementation of passive RF modalities such as passive radar or RFID. Passive RF modalities are difficult to detect, require lower power and have lower costs than the ones associated

Table 4.

Methods of Achieving Detection and Tracking via EO/Passive RF Sensor Fusion		
Input Data	**Method**	**References**
Passive Radar and EO/IR sensor input	Unmanned Aircraft Vehicle sense and avoid application using SVM classifier	[6]
FMV and Passive RF	Sheaf-based heterogeneous sensor fusion using passive RF collected via Doppler Radar and FMV for target detection and tracking.	[38]
FMV and Passive RF	Joint Manifold Learning based heterogeneous data fusion approach to form a joint sensor data manifold for vehicle detection and tracking.	[41]
FMV and Passive RF	Deep learning approach using feature manifold representations for multiobject tracking and detection.	[42]
FMV and Passive RF	Autoencoder based Dynamic Deep Directional-unit network to achieve unsupervised upstream sensor fusion for the detection and tracking of vehicles	[20]

with the construction and usage of active radar, and are harder to implement countermeasures against, such as jamming and spoofing which can corrupt the collection of RF-based modalities and transmitted imagery. Combining the EO/RF modalities improves the overall reliability and has been implemented in a few applications for target detection, estimation, and tracking, as summarized in Table 4.

For the research by Barott *et al.* [6], a SVM is used as a final method of classification. Similar to previously mentioned papers, Fasano [10] and Kemkemian and Nouvel [26], which use different metrics for accuracy, they share the end goal is for the sense-and-avoid of unmanned aircraft. The purpose of the fusion is to use two complementary instruments, passive radar, and an EO/IR system to not only the detection of aircraft, but also the identification of the model and relative threat to the unmanned aircraft. The architecture for fusion first preprocesses the thermal and visible images, isolating the propulsion and aircraft before extracting the characteristics. These features are then correlated with the relative distance and orientation of the radar return, and subsequently to create a multispectral aircraft signature, which is used as an input for the SVM classifier.

The use of an autoencoder-based dynamic deep directional-unit network [20] was capable of learning compact, abstract feature representations from the high-dimensional spatiotemporal data of full motion video, and I/Q data. The architecture exploits the access to elements of interest within regions of interest using temporal tracking and supervised classification before being fed into a decentralized supervised discrimination layer that applies Bayesian program learning in order to implement upstream multimodal data fusion. Among the network's achievements, a notable benefit of the approach is that the network is capable of reconstructing missing modalities given the observed signatures.

Other research into achieving EO/RF fusion for vehicle tracking and detection using FMV and P-RF include joint manifold learning [43],sheaf-based approach with its data [21], SVM classifier [6]. In [43] and [21], the use of simulation data is used for the primary method of training and testing, while in Barott *et al.* [6] real data collected from Daytona Beach International Airport are used. In [43], the use of a joint manifold learning fusion approach is used for the mixed simulation data. The use of a digital imaging and remote sensing image generation (DIRSIG) dataset provides video measurements and three distributed RF sensors. The intrinsic low-dimensional data, the 2-D images of the vehicles, are extracted by manifold learning algorithms from high-dimensional data by implementing a linear transformation of the vehicle positions. The RF data are similarly handled by manifold learning, and then the implementation of linear regression is used for tracking. These results were compared with a number of methods, such as maximally collapsing metric learning or neighborhood preserving embedding, calculating position errors with respect to the ground truth after implementing noise.

Finally, in Robinson *et al.* [21], the use of simulated multisensor data is used to locate a moving emitter. The method of fusion implemented is Sheaf Theory, a tool for systematically tracking locally defined data attached to the open sets of a topological set. For the purposes of implementing sensor fusion, the data samples and model of data are used as the inputs of a sheaf-based fusion architecture. The model of the data is used to construct the sheaf while the data samples are converted into samples for partial assignment. The outputs of the two are then used to search over the global sections using the optimizer, before using the results to report values over the stalks. During testing, the observed stalks for each

Table 5.

Methods of Evaluating EO/RF Fusion Algorithms		
Applications	**Method**	**References**
Multilevel Mapping	Multilevel map classification evaluation using producer's and user's accuracies.	[7]
Indoor Tracking	Location estimation error compared to ground truth	[44], [45]
Air Traffic Detection	Cramer–Rao lower bound, performance indices, range estimation error in terms of mean and standard deviation.	[6], [26], [32]
Target Detection and Tracking	Comparison of positional errors between the ground truth and the mapped manifold learning results.	[43]
Automated Activity Recognition	Computational cost, accuracy, and comparison of memory fingerprints for evaluating fusion efficiency	[46]

sensor measure the real-time offset and complex I/Q samples for each RF sensor while the EO data are collected, keeping the xy-location of each detected pixel for the video input. For the modeled stalks, which provide a comparison for each pair of sensors, the time offset relative to the video detection and the time aligned I/Q samples for each group of RF sensors are tracked. The third vertex tracks the true location of the ground truth, the emitter, and transmitted signal.

EVALUATION METRICS OF EO/RF SENSOR FUSION

Sensor fusion algorithms are generally evaluated in terms of accuracy and robustness. During experimentation, certain trials will likely be completed in order to compare the robustness of the model. Methods such as adding visual occlusion or noise will normally be implemented, with other methods of distortion to ensure that the model can use that training in unfamiliar situations. But sensor fusion is the process of combining measurements from multiple nodes, each of which have a certain level of uncertainty. When these sources of information are fused together, it is not always clear how these uncertainties will interact and influence the overall performance of the sensor fusion algorithm. It is important to collect information in order to gain insight into the performance of an implemented fusion architecture. This section discusses the evaluation methods for EO/RF fusion and their respective applications, summarized below in Table 5.

For the vast majority of the many sensor fusion applications that are NN based, F-1 Score is traditionally used to score the performance. The measurements of precision and recall, the measurements that compare the true positive rate and the sensitivity are commonly accepted for most classification problems. However, not all EO/RF sensor fusion applications are focused on classification. Some are made for tracking and estimation purposes, and therefore need to be compared to a ground truth. For machine learning however, there are a few more factors that need to be addressed when testing a NN-based sensor fusion architecture.

One of the major characteristics and concerns for machine learning in general is the nontransparency of the models. Deep learning itself suffers from several major limitations, requiring vast amounts of training data, having poor ability to represent uncertainty, being easily fooled by adversarial examples, and being difficult to optimize. Because of the black-box like nature of such networks, it can be difficult to perform reasoning with them and these aforementioned qualities mean that they are prone to generalizing poorly or overfitting the data. In order to improve these issues with uncertainty, the implementation of Aleatory Variability and Epistemic Uncertainty are commonly used approaches. Aleatoric variability is uncertainty inherent in observation noise while Epistemic Uncertainty is ignorance about the correct model generated by the data, the parameters, the convergence, etc. In Tagasovska and Lopez-Paz [47] specifically, the use of Simultaneous Quantile Regression and Orthonormal Certificates are used as a loss function to estimate Aleatoric Variability and Epistemic Uncertainty, respectively. There is an important distinction between a system malfunction, a failure that is recognized, and a normal operation where false-positives can occur, and the use of uncertainty is an important measure to help interpret the results of a NN-based fusion method.

In [48], concerns about the stability of a fusion system that is trained end-to-end is not encouraged by its readers, due to the potential incompatibility with assumptions about the stable hierarchical architectures of components. In Yang *et al.* [49], the creation of an explainable neural network (xNN+) in order to improve the understanding and provide sufficient model interoperability is explored. The estimation of multiple parameters via a modified mini-batch gradient descent method derived from the backpropagation for calculating derivatives and the use of the Cayley transform is implemented to preserve the

projection orthogonality. The approach improves interpretability of the model while maintaining the prediction accuracy. While there has been research into creating a more transparent models [50], the vast majority of machine learning methods are all-training based and nontransparent, which makes them being able to use metrics in order to properly interpret the results all the more important.

For the multilevel map classification via SVM [7], Kulin *et al.* compared the producer's accuracy and user's accuracy, the complements of omission and commission error, respectively. *Producer's accuracy* is the map accuracy from the map maker's point of view. This metric for mapping measures how often real features on the ground are accurately shown on the classified map and probability that a certain land cover on the ground is classified correctly. Conversely, the *user's accuracy* is accuracy from the point of view of the map user. The metric can be summarized as reliability, calculating the total number of correct classifications for a specific class and then dividing it by the row total.

For the evaluation of their joint manifold learning framework, Shen *et al.* [43] compared their results for vehicle detection and estimation to the ground truth. In order to better test the reliability of their algorithm, the implementation of white noise and position shifts are applied in order to test if the framework can apply its learned intrinsic mapping from one scene to a similar one. In addition, comparisons were made with other traditional sensor fusion algorithms in multiple scenarios, relating the estimated trajectory with the ground truth and plotting position errors over the respective frame index.

For the passive multispectral radar/EO/IR sensor fusion architecture for UAS in [6], the use of a Cramer–Rao lower bound (CRLB) on localization accuracy is used. In similar applications, such as [26] or [32], which seek to avoid collision between unmanned aircraft, a simulation system for the EO and radar input is used to implement tracking, trace range, and azimuth error for the RF modality in terms of the mean and standard deviation. Performance indices were used to measure the performance of the UAS system, emphasizing the characteristics of the response deemed to be relevant to optimize the control system and providing feedback.

Besides metrics that directly relate to the evaluation of accuracy measurements, another aspect of performance to consider is computational cost of the process and delay caused by the fusion architecture. As previously mentioned, the design of decentralized or hierarchal fusion architectures are less taxing in terms of computational cost, but in exchange introduce more delays in communication. Certain models actually exploit the nature of decentralized architectures, improving the individual node estimators by capturing the correlation between the sensor observations of matching parameter values for different manifolds [51].

Table 6.

Simulations in DIRSIG Data		
Simulation	**Number of Vehicles**	**TX Waveform**
1	1	Tone
2	1	Tone
3	1	2G
4	2	2G, None
5	2	2G, 3G
6	3	2G,3G, None
7	3	2G, 3G, 4G
8	3	2G, 3G, 4G
9	2	2G, 3G
10	1	2G
11	3	2G, 3G, 4G
12	3	2G, 3G, 4G
13	3	2G, 3G, 4G

Aside from the metrics of computational cost and efficiency, the size of training data is also relevant. The main problem that all fusion methods face is the constraints in the execution of data acquisition, processing, and the implementation of the algorithm. In Martín *et al.* [46], the accuracy, computational costs, and memory fingerprints for the traditional classifiers Naïve Bayes, Decision Table, and Decision Tree were calculated for different sensor data and optimization method. These metrics provide insights into how the individual modalities interact with the fusion, particularly with the memory fingerprints. While there is no universal benchmark for fusion evaluation, these are some measures that provide for better understanding the fusion architecture and assessing the impact of different modalities.

FUSION OF EO AND PASSSIVE RF NN

In 2017, Michigan Tech Research Institute created a DIRSIG simulation of Medium Wave IR FMV input and emulated three corresponding P-RF sensors (see Table 6). The DIRSIG dataset contains 13 simulations that cover a variety of visual obscuration scenarios, while receiving RF signals at three different locations. These DIRSIG simulations are oriented around tracking one or more moving targets, automotive vehicles, and provide various

Figure 1.
Comparison of DIRSIG Simulation 9 Frame 295 versus Frame 310.

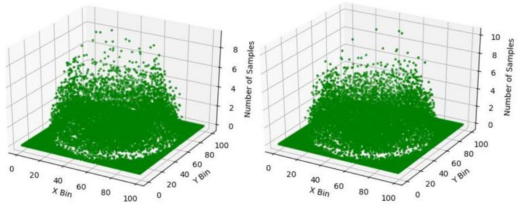


Figure 2.
Comparison of DIRSIG Simulation 9's RF histograms for frames 295 and 310.

opportunities, visual obscuration, to test the uncertainty reduction of the chosen sensor fusion method. For the purposes of implementing EO/RF fusion, the FMV input was treated as an EO input. As the vehicles in the simulations are small enough when compared to the background of the simulation, the thermal properties of the vehicle will not change the relative size of the target the network needed to detect.

For the EO and RF sensor fusion research presented in this article, the raw P-RF data are preprocessed to obtain I/Q histograms over time. The histograms are then aligned in time with the simulated EO data for the purposes of detection and estimating the number of vehicles. When compared to the ground truth of the simulation, neither of the standalone modalities was able to achieve accuracy above 90%. But the fusion of EO/RF neural network (FERNN) was able to achieve an accuracy of 95%.

Histograms has been widely used in image processing and image retrieval [52], [53]. In [22], a histogram of RSS values was used for a Wi-Fi-based indoor positioning system called KAILOS (KAist Indoor Locating System). The system uses crowd-sourced fingerprints via signal fluctuation matrix (SFM) and an extended Viterbi algorithm to achieve accurate indoor positioning. SFM is essentially a universal histogram of the RSS values irrespective of locations and access points that calculates the probability of observing an online RSS of an access point at a location as a log-odd probability.

While there exist many methods and algorithms of sensor fusion, the difficulties in establishing a correspondence between the EO and RF inputs indicate that a deep learning approach would be more suited for the research [23]. While methods such as Dempster–Shafer theory provide the opportunity to reason with uncertainty [54], the sheer size of the samples and RF data in raw format make finding a traditional correlation between EO and RF data very difficult.

The DIRSIG dataset contains 13 different simulations. Table 6 shows the number of vehicles and the waveforms transmitted by vehicles in each scenario. For the purposes of the experiment, six simulation scenarios were selected from the DIRSIG data based on the number of vehicles and the transmitted waveforms in the simulation. These simulations were selected in order to balance and ensure the robustness of the model for the training and validation data of the NN. From the six chosen scenarios, simulations 2 and 10 were used for the training and testing of the FERNN's accuracy in detecting one vehicle. For the scenario of two vehicles, simulations 4 and 9 were selected for testing and training, and for the scenario of three vehicles, simulations 11 and 12 were chosen. Simulation 2 contains a single vehicle which transmits a tone, while simulation 10 has a single vehicle which transmits a 2G waveform. Simulations 4 and 9 share a common 2G waveform while simulations 11 and 12 share the 2G, 3G, and 4G waveforms. The purpose of choosing these scenarios with different waveforms was to ensure that a level of robust change was added into the training in order to prevent the NN from ignoring the EO input.

In the DIRSIG dataset, there are similar scenarios where fusion of EO and P-RF can provide better accuracy over single modality. The EO sensor in the simulations is limited by visibility in detecting the number of vehicles when compared to the ground truth. Even when the targets reach the scope of the EO sensor, there are several instances of optical obscuration that are caused by the simulated foliage. As seen in Figure 1, frame 295 and frame 310 are generated by EO sensor corresponding to a scene with two moving vehicles. But only one vehicle is visible to the sensor in frame 310 due to obscuration caused by foliage. However, the RF histograms remain largely unchanged as can be seen in Figure 2, because the optical obscuration does not affect the histogram generated for the same circumstances 15 frames later. In order to overcome this difficulty of EO sensor and to better gauge the accuracy of the network with respect to the ground truth, improving the accuracy and robustness of the system via EO and passive RF fusion is necessary.

The *fusion of EO/RF neural network* (FERNN) proposed seeks to accurately estimate the number of moving targets in the scene through EO and passive RF fusion. Four states were generated in order to best describe the detection of the given scenario. The state of the simulation is denoted by a one-hot vector $\boldsymbol{s} = [\, s_1 \quad s_2 \quad s_3 \quad s_4 \,]$, whose elements are all false values (0) with the exception of one true value (1) for the element corresponding to a

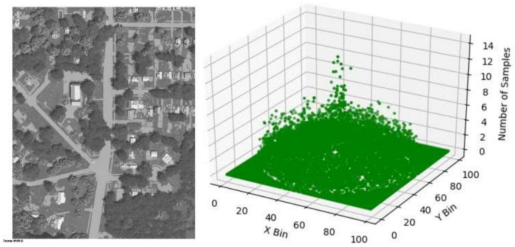


Figure 3.
The extracted EO frame and the corresponding P-RF histogram (Simulation 9, NADIR RF reciever).

specific state. The first state, i.e., $\boldsymbol{s} = [\,1 \quad 0 \quad 0 \quad 0\,]$, is defined as having no cars in the simulation, which is included for the purposes of determining the accuracy of the EO modality as the ground truth versus what the EO modality can detect. Within the DIRSIG simulations, the ground truth is that there is always at least one vehicle in all the simulations. The other three states are for the presence of a single car, two cars, and three cars, respectively, within the simulation.

For data from EO sensors, each simulated video input was first processed into image frames, resized, and converted into grayscale for simulation video that was not originally grayscale. For the purposes of reducing training time and conserving processing power, the image frames were resized to match the P-RF histograms.

Prior to being fed into FERNN, the raw P-RF data are processed to generate histograms of the I/Q data. The histogram depicts the estimation of the probability distribution of the P-RF data. The histograms are then fed into the NN in conjunction with the corresponding EO frames. Figure 3 shows an extracted EO frame and the corresponding histogram of P-RF sensor that have been aligned in time. In order to achieve heterogeneous feature-level fusion, a deep neural network is trained over the pairs of 2-D matrices from the two modalities. The fusion NN itself is a sequential model that compares the predicted states with the labels and then modifies its weights accordingly, as shown in Figure 6 and described in part C. For the purpose of decision-level fusion, the NNs are trained for standalone modalities, i.e., an EO NN and a P-RF NN.

EO NEURAL NETWORK

A separate CNN is trained for the detection and estimation of the number of vehicles based on EO data. This image-based NN can achieve 91% accuracy upon using the modified labels that are exclusively just for an image classification. For these modified labels, if a frame is generated for a period in which the vehicle is temporarily obscured by foliage, the frame is labeled as having no vehicle detected, which is inaccurate when compared to the ground truth of the simulation. Likewise, simulation frames in which the vehicle has not entered the scope of the EO sensor are also labeled as not detecting a vehicle.

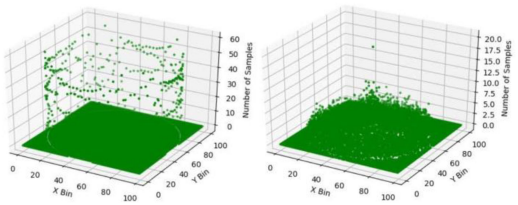


Figure 4.
Comparison of I/Q histograms collected by SIGINT in simulation 1 (left) and simulation 9 (right).

In order to accomplish classification, both FERNN and the standalone EO network begin preprocessing, resizing, and labeling the frames. After that the networks categorize the images by the number of vehicles detected. Unlike the ground truth for the overall scenario simulation, the standalone EO NN will output that no vehicle is detected when vehicle is not visible to the EO sensor. For the purposes of comparison testing, the standalone EO NN retains the original training it received to classify an image by the number of vehicles it detects. When being tested against the ground truth for each simulation, in terms of the number of vehicle(s) traveling in the area, the accuracy of the EO network decreased to an accuracy of only 72%. This result is expected, as the simulation has a number of optically obscured examples inside of the simulation set.

RF FEATURE EXTRACTION AND NN

The most basic signal that can be collected for the RF sensing is known as in-phase and quadrature components. These I/Q components are the basis of complex RF signal modulation and demodulation, and the backbone of modern communication systems. During previous experiments, our group had successfully trained a CNN to detect the human occupancy of an enclosed indoor space using the raw I/Q data of passive RF signals. For the DIRSIG dataset, however; the I/Q data in its raw format were ineffective for the purposes of vehicle detection. The I/Q data were processed to generate a 2-D histogram, which is an estimation of the probability density function of the P-RF data. In the DIRSIG simulation, there are three SIGINT sensors to generate the P-RF data. These sensors are placed orthogonally, one in north, and one in west, and one in the nadir. The generated 2-D histograms are then fed into the fusion networks in order to facilitate the homogenous fusion between the three P-RF sources.

As seen in Figures 4 and 5, the 2-D histograms show visually different patterns due to the different waveforms transmitted by vehicles. To illustrate the differences in histograms, the histogram value was plotted in the z-dimension to provide a clear visual difference. The histogram of P-RF in Simulation 1 (see Figure 4, left, single

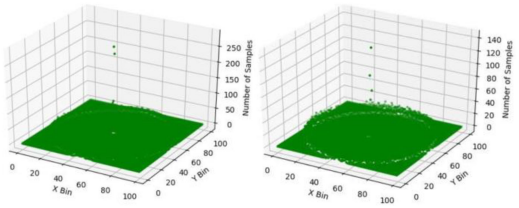


Figure 5.
Comparison ofI/Q histograms collected by SIGINT in simulation 4 (left) and simulation 11 (right).

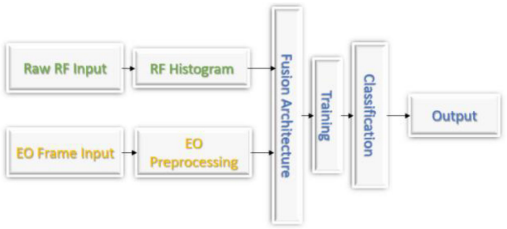


Figure 6.
Fusion of EO and RF neural network (FERNN) architecture.

vehicle, tone signal) is noticeably different from Simulation 9 (see Figure 4, right, two vehicles, 2G and 3G signals). Some simulations, such as Simulation 4 (see Figure 5, left, two vehicles, 2G on one of the vehicles in question) and Simulation 1 1(see Figure 5, bottom right, three vehicles, 2G, 3G, and 4G signals) share one common RF signal types, which is harder for the detection and discrimination of multiple types of vehicles. The histograms of Simulations 4 and 9 are still noticeably different despite both simulations have the same number of moving vehicles.

For the standalone RF NN, in earlier iterations, the classification accuracy would rise to 100% accuracy within four epochs for differentiation. These earlier versions of the software, however, relied on the unique histograms formed by the different transmission waveforms. In order to balance training and testing data for the standalone RF and EO networks, the simulations used for testing were limited to the ones whose transmission waveforms are as different as possible. Simulations 1 and 2 for example are unique compared to other simulations in which only one car is detected because the waveform is a pure tone. Simulations 11, 12, and 13 all have a variety of signals, 2G, 3G, and 4G, while some simulations such as 4 or 10 are limited to a single waveform (2G). When trained under these constraints, the accuracy of the standalone P-RF NN is 83%.

FEATURE LEVEL FUSION

In order to achieve sensor fusion from heterogeneous modalities, both the resized and preprocessed grayscale EO frames and P-RF histograms are fed into a sequential NN. The data from both the RF and EO modalities are stacked into a sequence of arrays that acts as the training data for the NN. After being standardized and normalized, the sequential model begins the training for feature-level fusion.

FERNN takes the input arrays containing the values from the preprocessed RF and EO data and then flattens them, using ReLu and SoftMax before compiling the model with the Adam Optimizer. Unlike classical stochastic gradient descent (SGD), which maintains a single learning rate for all weight updates, Adam Optimizer utilizes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients. This approach combines the advantages of two other existing extensions of the SGD, adaptive gradient algorithm, and root mean square propagation. The loss function of the model is sparse categorical cross-entropy. SoftMax was applied to implement classification of different states. Compared to the results of the standalone RF and EO NNs, the feature-level fusion network can achieve 95% accuracy, with regards to the ground truth of the simulation.

DECISION-LEVEL FUSION AND COMPARISON RESEARCH

Traditional learning methods and probabilistic classifiers were implemented to compare FERNN with methods of decision-level fusion. Logistic regression (LR), Naïve Bayes (NB), random forest (RF), Gaussian naïve Bayes (GNB), and support-vector machine (SVM) were implemented for decision-level fusion experiments. In addition to these methods, FERNN was modified to implement downstream fusion as well.

Naïve Bayes is a probabilistic classifier that applies Bayes' Theorem under the assumption that the data are independent of each other. GNB instead works under the assumption that the continuous values for each class have a Gaussian distribution. Random forest is an ensemble learning method for classification that focuses on the generation of decision trees. These decision trees are used to avoid overfitting and generate a prediction. Logistic regression is a statistical model that is used to implement regression analysis to model the probability of a class or event. SVM is a supervised learning model that analyzes data for the regression analysis and classification. However, unlike logistic regression, naïve Bayes, and GNB, it is a nonprobabilistic linear classifier.

In order to better use the available data, the approach for decision-level fusion was ensemble learning methods, soft and hard voting, in addition to a NN approach and late (or downstream) fusion via SVM. Soft voting uses the individual classifiers calculations for the probability of the outcomes and averages out the resulting outputs. Hard voting uses majority vote in order to choose a model from the

Table 7.

RF Accuracy Comparison	
Situation	**Accuracy**
Comparison of simulations 9 and 10, differentiating between detecting one vehicle and two vehicles	95%
Comparison of simulations 2, 9, and 10, differentiating between detecting one vehicle, two vehicles, and three vehicles	92%
Combined data from simulations 2,4,9,10,12, and 13, differentiating between detecting one, two vehicles, and three vehicles	83%

ensemble to make the final prediction with the available data. Besides soft and hard voting, the data from the feature-level fusion experiments, the standalone EO and standalone RF NNs, were used to implement decision-level fusion. The predictions for each of the individual models were fed into an SVM model that used the concatenated values. Besides SVM, a NN model used the same prediction values to classify the dataset.

RESULTS

The trained standalone EO NN alone could reach 91% accuracy in determining the current number of detected vehicles based on the available image provided. However, when tested against the ground truth, with the knowledge of a vehicle moving outside of the camera's angle or obscured by local foliage, the overall accuracy of the NN would decrease to only 72%. Similarly, the standalone P-RF NN, could only reach an accuracy of 83% when taking into account all the histograms for the selected simulations, where scenario 1 describes a ground truth of one vehicle, scenario 2 describes a ground truth of two vehicles, and scenario 3 describes a ground truth of three vehicles.

As seen in Table 7, the accuracy of the standalone P-RF NN was significantly higher when trained with a small number of simulations. When comparing simulations 9 and 10 to differentiate between one vehicle and two vehicles, the NN could perform at 95% accuracy. However, when the number of categories increases to 3 and the NN was trained with three different simulations, the performance decreased to 92% accuracy. In this situation, histograms formed by the P-RF feature extraction are still unique enough to ensure an acceptable accuracy. In order to ensure that the training data were robust enough for the purposes of vehicle detection, six simulations were combined in the training process, each of which containing visually different histograms. The addition of all these training and testing data reduced the accuracy of the NN to 83%.

As shown in Table 8, the overall accuracy of different sensors on their own is unsatisfactory, with EO only managing to score a 72% accuracy against the ground truth and RF only reaching 83% accuracy when trained and tested with large number of different scenarios. With FERNN, the accuracy can reach a much higher level of 95%, when the EO frames, the data of P-RF sensors located at nadir, north, and west are all fed into the NN.

Based on the results of the feature-level fusion, a significant increase in accuracy was dependent on the number of feature sources for training available. As seen in Table 9, the F1 score for using only the EO and Nadir SIGINT is 80%. Similarly, the F1 scores for the EO and North and EO and West are at 78% and 82%, respectively. Based on the results presented in Tables 10–12, the accuracy of the NN only reaches satisfactory values when all four sources of SIGINT (see Table 13) and the corresponding frames are fed into the training

$$F_1\text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}. \quad (3)$$

Table 8.

Accuracy Comparison Between Standalone Modalities and Feature Level Fusion	
Situation	**Accuracy**
Standalone EO	72%
Standalone RF	83%
Feature Level Fusion Architecture	95%

Table 9.

EO and Nadir sigint Fusion			
Scenario	**Precision**	**Recall**	**F1-Score**
1	0.70	0.72	0.71
2	0.87	0.79	0.83
3	0.83	0.88	0.85
Accuracy		0.80	
Macro AVG	0.80	0.80	0.80
Weighted AVG	0.80	0.80	0.80

Table 10.

EO and North sigint Fusion			
Scenario	**Precision**	**Recall**	**F1-Score**
1	0.87	0.65	0.74
2	0.67	0.92	0.78
3	0.82	0.86	0.84
Accuracy		0.78	
Macro AVG	0.79	0.81	0.79
Weighted AVG	0.80	0.78	0.78

Table 11.

EO and West sigint Fusion			
Scenario	**Precision**	**Recall**	**F1-Score**
1	0.79	0.75	0.77
2	0.92	0.82	0.86
3	0.75	0.92	0.83
Accuracy		0.82	
Macro AVG	0.82	0.83	0.82
Weighted AVG	0.83	0.82	0.82

Table 12.

EO and RF sigint Fusion			
Scenario	**Precision**	**Recall**	**F1-Score**
1	0.96	0.96	0.96
2	0.99	0.94	0.96
3	0.91	0.96	0.93
Accuracy		0.95	
Macro AVG	0.95	0.95	0.95
Weighted AVG	0.95	0.95	0.95

In order to accurately evaluate the performance of FERNN, the F1 score, precision, and recall were calculated based on (1)–(3) for the statistical analysis. The F1 score is the harmonic mean of the precision and recall, the measurements of positive predictive value and sensitivity for machine learning. Precision is the measurement of type-I error, false positives, while recall is the measurement of type-II error, false negatives. Figure 7 shows the F1 scores of all the NN trained for vehicle detection and scenario categorization. The fusion of EO and all three P-RF sensors yield the best result.

Besides feature-level fusion, more traditional methods were explored to analyze the effectiveness of using the RF features and EO input. Logistic regression, random forest, Naïve Bayes, and GNB were all applied to the same datasets as the EO-RF fusion NNs. As these ensemble methods are more traditional in nature and not based on NNs, the evaluation of classification accuracy is conducted by k-fold cross validation.

K-fold cross validation is a procedure meant to estimate the skill of a machine learning model on unseen data. The limited samples are used to estimate how the model is expected to perform in general when used to make predictions on data that is not used during training. The process first shuffles the dataset randomly, splitting up the dataset into k groups. For each of these unique groups, one is designated as a test dataset, while the others are treated as part of the training dataset. The model is fit on the training data and then tested on that dataset, saving the evaluation score and discarding the model. After repeating the process, the skill of the model is summarized using the sample of the model evaluation scores.

As can be seen in Table 13, Logistic Regression and GNB consistently performed better than Naïve Bayes and Random Forest in terms of accuracy. The accuracy for these classifiers improved when there were fewer inputs from the RF histograms. Considering the nature of the RF features and the EO inputs, it is possible that because the data received as a 2-D array that the classifiers that assume the data to be independent, random forest and Naïve Bayes, performed comparatively poorer than logistic regression and GNB.

Once the feature-level fusion comparison experiments were completed, the architecture for the feature-level fusion was adopted to implement decision-level fusion. As seen in Figure 8, at least one source of EO data were classified with the RF features, and once that classification

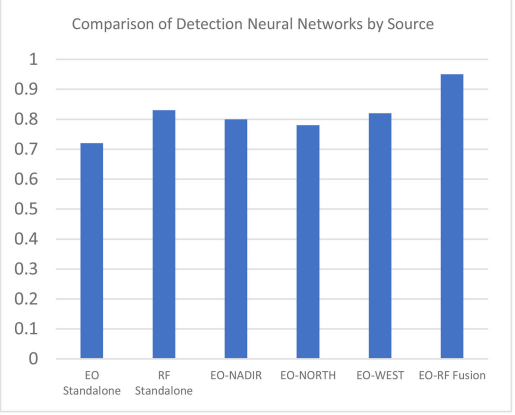


Figure 7.
Comparison of F1 scores for different NNs implemented for vehicle detection.

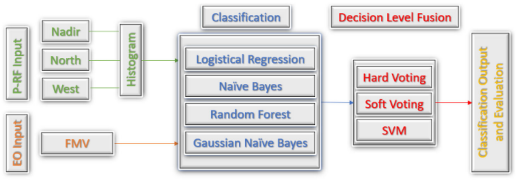


Figure 8.
Overview of comparison research architecture.

Table 13.

Classification Accuracy of Traditional Probabilistic Classifiers and Learning Methods	
Method and Input Data	**Accuracy**
Logistic Regression (EO and Nadir)	0.81(± 0.13)
Logistic Regression (EO, Nadir, and North)	0.75 (± 0.17)
Logistic Regression (EO, Nadir, North, and West)	0.73 (± 0.17)
Naïve Bayes (EO and Nadir)	0.64 (± 0.04)
Naïve Bayes (EO, Nadir, and North)	0.64 (± 0.08)
Naïve Bayes (EO, Nadir, North, and West)	0.63 (± 0.07)
Random Forest (EO, Nadir, North, and West)	0.64 (± 0.21)
Random Forest (EO, Nadir, and North)	0.69 (± 0.16)
Random Forest (EO, Nadir, North, and West)	0.67 (± 0.17)
Gaussian Naïve Bayes (EO and Nadir)	0.73 (± 0.14)
Gaussian Naïve Bayes (EO, Nadir, and North)	0.71 (± 0.15)
Gaussian Naïve Bayes (EO, Nadir, North, and West)	0.70 (± 0.14)

is completed the results are used for decision-level fusion. For the purposes of comparison research, hard and soft voting was chosen, using the classification output of logistical regression, Naïve Bayes, random forest, and GNB to implement decision-level fusion.

As seen in Table 14, the results for hard and soft voting showed little difference in terms of total accuracy. Given the input methods being weighed against each other, and the accuracies they had individually, the result of the decision-level fusion is dependent on the accuracies of the methods implemented. Soft voting performed marginally better than hard voting, as for the given inputs only one of the methods performed above 0.80 in accuracy.

Besides implementing voting for decision-level fusion, SVM learning was used in order to test the effectiveness of decision-level fusion. For the SVM decision-level fusion, the prediction values of the independently trained standalone EO and standalone RF NN were fed as a concatenated array of values. As seen above in Table 15, the late decision fusion implemented via SVM with the standalone EO and RF NN classification weights could achieve an accuracy of 88%.

Out of the results of using feature- and decision-level fusion with the RF histograms and EO frames, the highest accuracy of all the methods tested was FERNN. Out of the decision-level fusion methods, late-fusion NN (LFNN) was the closest in terms of accuracy, with an F1 score of 90.7%. From the results, it can be concluded that for this dataset of the RF and EO features, a NN benefits more

Table 14.

Decision-Level Fusion Comparison	
Method and Input Data	**Accuracy**
Hard Voting (LR, RF, NB, GNB) (EO and Nadir)	0.73 (+/- 0.14)
Hard Voting (LR, RF, NB, GNB) (EO, Nadir, and North)	0.71 (+/- 0.15)
Hard Voting (LR, RF, NB, GNB) (EO, Nadir, North, and West)	0.70 (+/- 0.12)
Soft Voting (LR, RF, NB, GNB) (EO and Nadir)	0.74 (±0.13)
Soft Voting (LR, RF, NB, GNB) (EO, Nadir, and North)	0.72 (± 0.11)
Soft Voting (LR, RF, NB, GNB) (EO, Nadir, North, and West)	0.71 (± 0.13)

Table 15.

SVM Fusion for EO and RF Data			
Scenario	**Precision**	**Recall**	**F1-Score**
1	0.76	0.84	0.81
2	0.93	0.96	0.94
3	1.00	0.73	0.85
Accuracy			0.88
Macro AVG	0.90	0.88	0.88
Weighted AVG	0.90	0.88	0.88

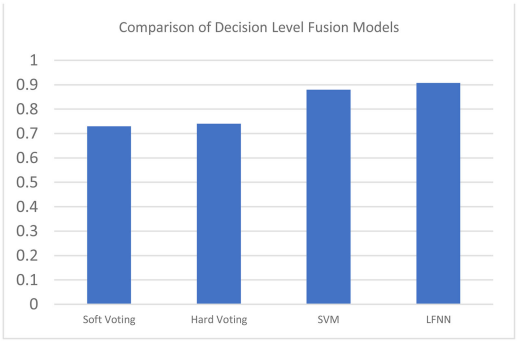


Figure 9.
Comparison of accuracy for different NNs implemented for decision level fusion.

from feature-level fusion over decision-level fusion. Likewise, the results indicate the preference of upstream fusion to downstream fusion.

Within ensemble learning fusion, soft voting, which uses the individual outputted probabilities over simple majority voting; still performed as well if not better. Based on the approach and results, the association of the RF features with the corresponding EO frame produced more accurate results compared to the direct estimations of the output class in question. From the ensemble learning experiments and the standalone EO and RF network results, it can also be concluded that the association of changes in the RF histogram features was relied on more than the EO, which was less accurate on its own compared to the ground truth.

When compared to the decision-level fusion models that were the SVM and LFNN, the results showed that the ensemble decision fusion schemes significantly underperformed. The LFNN can achieve 90.7% accuracy with the same training set that the SVM decision-level fusion model was able to achieve 88% accuracy with. In comparison, the soft and hard voting decision-level fusion with traditional classifiers however, both failed to achieve even 80% accuracy. Compared to all the decision-level fusion methods tested, FERNN, the proposed feature-level fusion performed with the highest accuracy, achieving a 95% F1 score versus the LFNN's 90.7% F1 score.

CONCLUSION

Multimodal sensor fusion, especially in the context of EO and passive RF fusion, is an active research field that is growing with many different innovative applications and approaches. The sheer volume and variety of methods makes it often difficult to pick and choose for a particular situation or dataset, a problem that is made worse by the complex sensor sources. It goes without saying that there is no singular or general solution for determining the optimal approach for information fusion, as the answer is always dependent on the situation, the dataset, and modalities used. Just explaining how some of these sensor fusion schemes are classified or described is a difficult task as there currently is no singular all-encompassing organization or classification of methods for this field.

Besides surveying the state-of-the-art literature based on the contributions to multimodal sensor fusion, the focus was on EO and passive RF fusion. While this research was primarily focused on the application of deep learning in information fusion, there exist many suitable classification schemes for exploiting the advantages of EO and passive RF modalities. For the purposes of EO and passive RF fusion, the use of raw data (i.e., upstream fusion) and features provides more robust and reliable results when compared to decision or postclassification fusion schemes.

Related literatures have also found similar results when using passive RF and any form of EO modality, such as previously mentioned [12] and [43]. For the purposes of using passive RF in applications that do not use forms of RF, such as Doppler or SAR imaging, the value that lower level fusion provides is greater than the metal-learning of higher level (i.e., situation) or decision-level (i.e., product) fusion for the purposes of fusing these two modalities. While it is a fundamental issue for any multimodal sensor fusion application, appropriate synchronization of different modalities is still a subject of interest. Determining when and how much data need to be processed from different modalities in order to optimize correlation and best help extract relevant features is an issue that has not been explored exhaustively. The use of spiking deep belief networks, such as [55] for example, could have potential unsupervised use in probabilistically reconstructing the passive RF data to perform classification.

While there are many approaches to machine learning and classification, some other possible directions using passive RF in sensor fusion could be addressed by including machine vision. For EO and passive RF, a few machine vision approaches have been researched, such as [56] and [57]. But the question of how to properly integrate context in the fusion process in order to improve a classification algorithm's ability to find relevant features and better discriminate between different classes is important [58]. Using the machine vision approach for passive RF creates a need to formalize the concept of context and also to explore how the changing context could influence the fusion process, as well as determining what model would be best suited to handle such a change.

While feature- and raw-level fusion have shown promising results, the question of what value correlation at the decision level could have for classification has not been explored thoroughly for passive RF/EO fusion. While it may be difficult to apply for the passive RF modality there could be an intrinsic value in using a dynamic classifier selection approach, similar to [22] which uses a dynamic settings hidden Markov model (HMM) classification algorithm for object detection with passive RFID tags. Even if decision-

level fusion is not necessarily the best approach for classifying passive RF input, the metalearning training could have value in a multilevel sensor fusion, such as in [59].

This article has covered existing passive RF/EO sensor fusion works, identified relevant issues that deserve further investigation, and proposed a feature level fusion network for integrating information from passive RF histograms and EO sensors and compared its performance to traditional methods of classifying linear input. From the results of FERNN, it can be concluded that the application of P-RF histograms as a feature can significantly improve the accuracy of the NN, particularly when fused at the feature level. The performance of the proposed EO and P-RF fusion network is superior to the performance of the NN of single sensor modality for both feature- and decision-level fusion in regard to vehicle detection and scenario categorization.

ACKNOWLEDGMENT

This work was supported by Grant FA9550-18-1-0287.

REFERENCES

[1] D. Feng, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[2] E. Blasch, E. Bosse, and D. A. Lambert, *High-Level Information Fusion Management and Systems Design*. Norwood, MA, USA: Artech House, 2012.

[3] Y. Zheng, E. Blasch, and Z. Liu, *Multispectral Image Fusion and Colorization*. Bellingham, WA, USA: SPIE Press, 2018.

[4] B. V. Dasarathy, "Sensor fusion potential exploitation-innovative architectures and illustrative applications," *Proc. IEEE*, vol. 85, no. 1, pp. 24–38, Jan. 1997.

[5] M. Kulin, T. Kazaz, I. Moerman, and E. D. Poorter, "End-to-end learning from spectrum data: A deep learning approach for wireless signal identification," *Spectrum Monit. Appl.*, vol. 6, pp. 18484–18501, 2018.

[6] W. C. Barott, E. Coyle, T. Dabrowski, C. Hockley, and R. S. Stansbury, "Passive multispectral sensor architecture for radar-EOIR sensor fusion for low SWAP UAS sense and avoid," in *Proc. IEEE/ION Position, Location Navigat. Symp.*, Monterey, CA, USA, 2014, pp. 1188–1196.

[7] M. Kulin, T. Kazaz, I. Moerman, and E. D Poorter, "End-to-end learning from spectrum data a deep learning approach for wireless signal identification in spectrum monitoring applications," *Special Section Real-Time Edge Analytics Big Data Internet Things*, vol. 6, pp. 18484–18501, 2017.

[8] C. Xu, M. Zheng, W. Liang, H. Yu, and Y. Liang, "End-to-end throughput maximization for underlay multi-hop cognitive radio networks with RF energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3561–3572, Jun. 2017.

[9] M. Przybyła-Kasperek and A. Wakulicz-Deja, "Comparison of fusion methods from the abstract level and the rank level in a dispersed decision-making system," *Int. J. General Syst.*, vol. 46, no. 4, pp. 386–413, 2017.

[10] G. Fasano *et al.*, "Multi-sensor-based fully autonomous non-cooperative collision avoidance system for unmanned air vehicles," *J. Aerosp. Comput., Inf.*, vol. 5, no. 10, pp. 338–360, 2008.

[11] A. Mikhalev and R. Ormondroyd, "Fusion of sensor data for source localization using the hough transform," in *Proc. 9th Int. Conf. Inf. Fusion*, Florence, Italy, 2006.

[12] D. Garagic *et al.*, "Upstream fusion of multiple sensing modalities using machine learning and topological analysis: An initial exploration," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, 2018.

[13] J. J. Zhang, A. Papandreou-Suppappola, and M. Rangaswamy, "Multi-target tracking using multi-modal sensing with waveform configuration," in *Proc. EEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, 2010, pp. 3890–3893.

[14] B. Waske and S. van der Linden, "Classifying multilevel imagery from SAR and optical sensors by decision fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1457–1466, May 2008.

[15] E. Marasco, A. Abaza, and B. Cukic, "Why rank-level fusion? And what is the impact of image quality?" *Int. J. Big Data Intell.*, vol. 2, p. 106, 2015.

[16] J. Machaj, P. Brida, and R. Piche, "Rank based fingerprinting algorithm for indoor positioning," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat.*, 2011, pp. 1–6.

[17] B. Habtemariam, R. Tharmarasa, M. McDonald, and T. Kirubarajan, "Measurement level AIS/radar fusion. signal processing," *Signal Process.*, vol. 106, pp. 348–357, 2015.

[18] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric system," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, 2005.

[19] G. Gennarelli, M. G. Amin, F. Soldovieri, and R. Solimene, "Passive multiarray image fusion for RF tomography by opportunistic sources," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 641–645, Mar. 2015.

[20] D. Garagic *et al.*, "Unsupervised upstream fusion of multiple sensing modalities using dynamic deep directional-unit networks for event behavior characterization," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, 2019.

[21] M. Robinson, J. Henrich, C. Capraro, and P. Zulch, "Dynamic sensor fusion using local topology," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, 2018.

[22] D. Han, S. Jung, and S. Lee, "A sensor fusion method for Wi-Fi-based indoor positioning," *ICT Express*, vol. 2, no. 2, pp. 71–74, 2016.

[23] E. Blasch, S. Ravela, and A. Aved, *Handbook of Dynamic Data Driven Applications Systems*. New York, NY, USA: Springer, 2018.

[24] Y. Zhou, Y. Chen, R. Gao, J. Feng, P. Zhao, and L., Wang, "SAR Target recognition via joint sparse representation of monogenic components with 2D canonical correlation analysis," *IEEE Access*, vol. 7, pp. 25815–25826, 2019.

[25] T. Oskiper, H.-P. Chiu, Z. S. S. Zhu, and R. Kumar, "Multi-modal sensor fusion algorithm for ubiquitous infrastructure-free localization in vision-impaired environments," in *Proc. IEEE/RSJ Int. Conf. Int. Robots Syst.*, Taipei, 2010.

[26] S. Kemkemian and M. Nouvel, "Sense-and-avoid system based on radar and cooperative sensors," in *Encyclopedia of Aerospace Engineering*. Hoboken, NJ, USA: Wiley, 2015, pp. 1–14.

[27] T. Deyle, H. Nguyen, M. Reynolds, and C. C. Kemp, "RF vision: RFID receive signal strength indicator (RSSI) images for sensor fusion and mobile manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, St. Louis, MO, USA, 2009, pp. 5553–5560.

[28] Y. Zheng and E. Blasch, "An exploration of the impacts of three factors in multimodal biometric score fusion: Score modality, recognition method, and fusion process," *J. Adv. Inf. Fusion*, vol. 9, no. 2, pp. 106–123, 2015.

[29] V. N. Balasubramanian, S. Chakraborty, and S. Panchanathan, "Conformal predictions for information fusion," *Ann. Math. Artif. Int.*, vol. 74, no. 1/2, p. 45–65, 2014.

[30] L. Xu, A. Krzyzak, and C. Y. Suen, "Associative switch for combining multiple classifiers," in *Proc. IJCNN-91-Seattle Int. Joint Conf. Neural Netw.*, Seattle, WA, USA, 1991, pp. 43–48.

[31] B. S. Jahromi, T. Tulabandhula, and S. Cetin, "Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles," *Sensors*, vol. 19, no. 20, 2019, Art. no. 4357.

[32] G. Fasano, D. Accardo, A. Tirri, A. Moccia, and E. Lellis, "Radar/electro-optical data fusion for non-cooperative UAS sense and avoid," *Aerosp. Sci. Technol.*, vol. 46, pp. 436–450, 2015.

[33] M. Liggins, II, D. Hall, and J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2017.

[34] O. Kahler, J. Denzler, and J. Triesch, "Hierarchical sensor data fusion by probabilistic cue integration for robust 3D object tracking," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, Lake Tahoe, NV, USA, 2004, pp. 216–220.

[35] I. Song, "Robust hierarchical data fusion scheme for large-scale sensor network," *J. Sensor Sci. Technol.*, vol. 26, no. 1, p. 1–6, 2017.

[36] A. J. Newman and G. E. Mitzel, "Upstream data fusion: History, technical overview, and applications to critical challenges," *Johns Hopkins Apl Tech. Dig.*, vol. 31, no. 3, pp. 215–233, 2013.

[37] D. L. Hall and J. Llinas, "An Introduction to multisensory data fusion," *Proc. IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[38] D. K. Seo, "Fusion of SAR and multispectral images using random forest regression for change detection," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 10, 2018, Art. no. 401.

[39] A. Orynbaikyzy, U. Gessner, and C. Conrad, "Crop type classification using a combination of optical and radar remote sensing data: A review," *Int. J. Remote Sens.*, vol. 40, no. 17, pp. 1–43, 2019.

[40] S. Kim, W.-J. Song, and S.-H. Kim, "Double weight-based SAR and infrared sensor fusion for automatic ground target recognition with deep learning," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 72.

[41] B. Bui, N. T. D. C. Pham, B. Q. Nguyen, and S. T. Le, "Tracking a 3D target with fusion of 2D radar and bearing-only sensor," in *Proc. IEEE Int. Conf. Ind. Technol.*, Lyon, 2018, pp. 1532–1537.

[42] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Inf. Fusion*, vol. 40, pp. 57–75, 2018.

[43] D. Shen, E. Blasch, P. Zulch, M. Distasio, and J. L. R. Niu, "A joint manifold leaning-based framework for heterogeneous upstream data fusion," *J. Algorithms Comput. Technol.*, vol. 12, no. 4, pp. 311–332, 2018.

[44] A. Fink and H. Beikirch, "Hybrid indoor tracking with Bayesian sensor fusion of RF localization and inertial navigation," in *Proc. 6th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Syst.*, Prague, 2011, pp. 823–827.

[45] B.-S. Choi, J.-W. Lee, J.-J. Lee, and K.-T. Park, "A hierarchical algorithm for indoor mobile robot localization using RFID sensor fusion," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2226–2235, Jun. 2011.

[46] H. Martín, A. M. Bernardos, J. Iglesias, and J. R. Casar, "Activity logging using lightweight classification techniques in mobile devices," *Personal Ubiquitous Comput.*, vol. 17, no. 4, pp. 675–695, 2012.

[47] N. Tagasovska and D. Lopez-Paz, "Single-model uncertainties for deep learning," 2019, *arXiv:1811.00908*.

[48] R. Saley, R. Queiroz, and K. Czarnecki, "An analysis of ISO 26262: Using machine learning safely in automotive software," SAE Tech. Paper 2018-01-1075, 2018.

[49] Z. Yang, A. Zhang, and A. Sudjianto "Enhancing explainability of neural networks through architecture constraints," 2019, *arXiv:1901.03838*.

[50] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, and K. Sycara, "Transparency and explanation in deep reinforcement learning neural networks," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, New York, NY, USA, 2018.

[51] M. A. Davenport, C. Hegde, M. F. Duarte, and R. G. Baraniuk, "Joint manifolds for data fusion," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2580–2590, Oct. 2010.

[52] Q. Feng, Q. Hao, Y. Chen, Y. Yi., Y. Wei, and J. Dai, "Hybrid histogram descriptor: A fusion feature representation for image retrieval," *Sensors*, vol. 18, no. 6, 2018, Art. no. 1943.

[53] N. Habeeb, H. Aljebori, H. Saad, and P. Picton, "Multi-sensor fusion based on DWT, fuzzy histogram equalization for video sequence," *Int. Arab J. Inf. Technol.*, vol. 15, no. 5, pp. 825–830, 2018.

[54] E. Blasch, R. Cruise, U. Majumder, and T. Rovito, "Methods of AI for multi-modal sensing and action for complex situations," *AI Mag.*, vol. 40, no. 4, pp. 50–65, 2019.

[55] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, and M. Pfeiffer, "Real-time classification and sensor fusion with a spiking deep belief network," *Frontiers Neurosci.*, vol. 7, 2013, Art. no. 178.

[56] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat, "Vision and RFID data fusion for tracking people in crowds by a mobile robot," *Comput. Vision Image Understanding*, vol. 114, no. 6, pp. 641–651, 2010.

[57] A. Isasi, S. Rodriguez-Vaamonde, J. López-De-Armentia, and A. Villodas, "Location, tracking and identification with RFID and vision data fusion," in *Proc. Eur. Workshop Smart Objects, Syst., Technol. Appl.*, Ciudad, Spain, 2010.

[58] L. Snidaro, J. G. Herrero, J. Llinas, and E. Blasch., *Context-Enhanced Information Fusion: Boosting Real-World Performance with Domain Knowledge*, New York, NY, USA: Springer, 2016.

[59] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Multi-level sensor fusion with deep learning," *IEEE Sensors Lett.*, vol. 3, no. 1, Jan. 2019, Art. no. 7100304.