2018 21st International Conference on Intelligent Transportation Systems (ITSC)
Maui, Hawaii, USA, November 4-7, 2018

An Efficient Multi-sensor Fusion Approach for Object Detection in Maritime Environments

Mohammad-Hashem Haghbayan[1], Fahimeh Farahnakian[1], Jonne Poikonen[1], Markus Laurinen[2], Paavo Nevalainen[1], Juha Plosila[1] and Jukka Heikkonen[1]

Abstract— **Robust real-time object detection and tracking are challenging problems in autonomous transportation systems due to operation of algorithms in inherently uncertain and dynamic environments and rapid movement of objects. Therefore, tracking and detection algorithms must cooperate with each other to achieve smooth tracking of detected objects that later can be used by the navigation system. In this paper, we first present an efficient multi-sensor fusion approach based on the probabilistic data association method in order to achieve accurate object detection and tracking results. The proposed approach fuses the detection results obtained independently from four main sensors: radar, LiDAR, RGB camera and infrared camera. It generates object region proposals based on the fused detection result. Then, a Convolutional Neural Network (CNN) approach is used to identify the object categories within these regions. The CNN is trained on a real dataset from different ferry driving scenarios. The experimental results of tracking and classification on real datasets show that the proposed approach provides reliable object detection and classification results in maritime environments.**

autonomous vessel, object detection, multi-sensor fusion, region proposals, convolutional neural networks, maritime environment.

I. Introduction

Designing reliable autonomous navigation systems have drawn a lot of industrial and academic interest in recent years. Today's advanced technologies provide great progress in autonomous vehicles field including surrounding environment perception, path planning and vehicle control in real-time. Multi-sensor fusion is one of the key technologies in this field. It can achieve a detailed environment description and accurate detection of interest objects based on the information from different sensors.

Robust object detection is a critical step of autonomous navigation systems [1]. The sub-sequence actions such as object classification and tracking would be impossible in these systems without efficient object detection. However, designing an automatic object detection method is one of the most challenging tasks for most applications. In maritime environment, object detection is a challenging problem due to the dynamic nature of the sea caused by waves, weather conditions and boat wakes [2]. Moreover, variety of objects and their appearance, camera motion and direction and geographical locations are other factors which make the detection process challenging. On the other hand in maritime environment different sensors must be employed to get enough information from the surrounding. These multiple sensors demand intelligent sensor fusion techniques to enhance the information from the surrounding for the navigator unit, specially while the vessel is in different environmental conditions. We believe to enhance the safe navigation in maritime environments, there must be a tight relationship between the robust object detection and sensor fusion units .

To address this problem, we proposed a multi-sensor fusion approach to achieve the complementary properties of objects by considering multiple sensors. A single sensor can not provide sufficient information for designing reliable vessel/vehicles under all possible conditions. For example, radar can accurately measure the distance and velocity of objects in various weather conditions. However, it has insufficient resolution for extracting object features in order to perform the object classification task. RGB and Infrared (IR) cameras can provide better resolution in a suitable range of distance. Our fusion approach is applied at decision level to fuse the object detection results obtained from four essential sensors: radar, LiDAR, RGB camera and IR camera. The object detection result determines the interest object localization surrounding the autonomous vessel/vehicles based on each sensor independently.

The proposed multi-sensor fusion approach uses the Probabilistic Data Association (PDA) [3] and generates object region proposals based on the fused object detection result. We also apply a Convolutional Neural Network (CNN) on the top of region proposals for classifying the interest objects within the regions. Inspired by the success of applying CNN in many number of challenging classification problems [4]–[6], we employed CNN for this purpose. The performance of CNNs depends strongly on the network topology. Therefore, we investigate the effect of both number of layers and neurons on the CNN performance. The obtained results on a real training data set show that CNN can achieve better detection accuracy in our problem when it has two convolutional layers, two max-pooling layers, one fully connected layer, and a softmax layer.

To the best of our knowledge, currently there are no existing works on using real data from four sensors to

*This work is part of the Advanced Autonomous Waterborne Applications Initiative (AAWA) and the New 3D Analytics Methods for autonomous Ships and Machines projects funded by the Tekes (Finnish Funding Agency for Technology and Innovation).

[1]Mohammad-Hashem Haghbayan, Fahimeh Farahnakian, Jonne Poikonen, Paavo Nevalainen, Juha Plosila, and Jukka Heikkonen are with Department of Future Technologies, University of Turku, Turku, Finland {mohhag, fahfar, jukapo, ptneva, juplos, jukhei}@utu.fi

[2]Markus Laurinen Researcheris with Rolls-Royce Oy Ab, Rauma Finland {markus.laurinen}@rolls-royce.com

detect and classify the objects in a maritime environment. In particular, the contributions of this paper are three-fold: (1) an efficient object detection method; (2) an accurate object classification; and (3) evaluation on a real data. For efficient object detection, we develop a PDA-based fusion approach to generate meaningful object region proposals. This approach fuses the object region proposals that are obtained from LiDAR, radar, RGB camera and IR camera. The proposed PDA-based approach can reduce the false detection rate by data fusion of different sensors based on the uncertainty of the measurement origin in different environmental conditions. For accurate object classification, the region proposals are fed into a CNN that is trained on real datasets.

The proposed object detection and classification solutions were evaluated using real data that is collected in the Finnish archipelago by a ferry equipped with four sensor as a part of Advanced Autonomous Waterborne Applications Initiative (AAWA) project [7]. This project tested sensor arrays in a range of operating and climatic conditions in Finland and has created a simulated autonomous ship control system which allows the behaviour of the complete communication system to be explored after surrounding object detections. The experimental results show that our multi-sensor fusion approach can achieve better detection results than other more classical methods tested.

The remainder of the paper is organized as follows. Section II discusses some of the most important related works. The proposed framework is introduced in Section III. Section IV presents object detection techniques for each sensor data individually. Section V describes how the proposed multi-sensor fusion approach fuses the obtained detection results from all sensors and generates candidate regions. Section VI presents the details on applying the proposed CNN for classifying the objects within the regions extracted from the fusion approach. Experimental setup and results are presented in Section VII and Section VIII . Finally, the conclusions is presented in Section IX.

II. Related Work

Automatic object detection is a critical issue for designing reliable autonomous vessel/vehicles. Multi-sensor fusion is an efficient approach to obtain accurate object detection by combining available information originating from various sensors. Different multi-sensor fusion methods have been studied for autonomous applications in [8], [9]. The most common multi-sensor fusion approaches are based on probabilistic techniques [10]. Generally, the multi-sensor fusion methods can be divided into three main groups based on the level of data abstraction used for fusion. (1) Measurement fusion methods first convert the data from each sensor to a common form and then the actual fusion of data is performed in the common representation. (2) Feature level fusion methods extract the relevant feature of each sensor individually and then the obtained features are combined into a single vector as an input of a fusion module. Therefore, the measurement and feature level fusion methods fuse raw sensor data or concatenate feature descriptors. However, they can not handle in incomplete measurements if one sensor modality becomes useless due to malfunctions, breakdown or severe weather conditions [11]. (3) Decision level fusion methods independently perform object detection from each sensor and the outputs of each sensor are fused at the decision level for final classification. Therefore, they can prevent the autonomous system from becoming non-functional when information conflicts are introduced to more than one sensor. In addition, the reliability and plausibility of each sensor can be considered. For this reason, we use a decision level multi-sensor fusion approach for fusing the object detection results obtained from four sensors. In [11], a decision level multi-sensor fusion method is presented to fuse the classification outputs of independent classifiers, such as 3D point clouds of LiDAR (Light Detection And Ranging) and image data using CNN. In another work [12], the authors propose a decision level fusion approach to reduce the number of misdetections that can lead to false tracks. In comparison with these works, the advantage of our fusion approach is that the description of the objects can be enhanced by adding knowledge from four sensor sources. For example, LiDAR data can give a good estimation of the distance to the object and its visible size. Therefore, we use other main sensors such as radar, IR camera and RGB camera to find more information about objects surrounding autonomous vessel.

Classification of the object is another main task in autonomous vessel/vehicles to identify objects of interest surrounding the vehicles. Primarily, object classification is conducted using a trained classification model on an offline dataset. Over the years, CNN is one of the most common models for object classification and detection specifically for 2D data, like video and images. CNN can automatically extract salient features from images and classify them. In [11], first authors use two independent CNNs for classifying 3D point clouds and image data in an autonomous vehicle system. Then they fuse the classification outputs from unary classifiers. Their method based on CNN achieve better performance than the previous methods. In another work, Girshick et al. [13] propose Region-based Convolutional Neural Networks (R-CNN), which led to substantial gains in object detection accuracy. The R-CNN approach first identify region proposals (i.e regions of interest that are likely to contain objects) and then classify these regions into object categories or background using CNN. One disadvantage of R-CNN is that it computes the CNN independently on each region proposal, leading to time-consuming and energy-inefficient computation. In order to reduce running time of R-CNN, Faster R-CNN [14] ignores the time spent on region proposals. In addition, R-CNN only plays as a classifier and it cannot predict object bounds. Prabhakar et al. [15] propose a system based on Faster R-CNN for detection and classification of on-road objects. The outputs of the system are the rectangular bounding boxes (BB) and class information of objects which are useful parameters for the motion planning of self-driving vehicles. Their deep learning network is found to be robust to variation in object's view, lighting

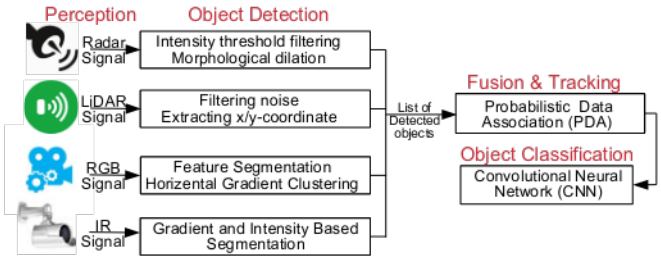

Fig. 1. Proposed decision level fusion framework

and climatic conditions. AlexNet [16], VGGNet [17]and GoogLeNet [18] are other popular deep CNNs for image classification and detection.

III. Proposed Decision Level Fusion Framework

Our framework can recognize and classify objects appearing around an autonomous vessel based on the three main following modules. Fig. 1 illustrates our whole framework (details will be found in later sections).

Object detection: the environment is continuously observed by four sensors (radar, LiDAR, RGB camera, and IR camera) to obtain a detailed description of the environment. After preprocessing of raw sensor data, this module identifies the locations of objects. The output of this module is a list of detected objects via four available sensors.

Fusion and tracking: this module contains a PDA-based mutli-sensor fusion approach to fuse the object detection results obtained from four sensors. Before fusing the data, the location of detected objects is mapped onto radar coordinates. This come from the fact that bird view of the objects provides more suitable environment for applying probabilistic models of movement[1]. Then, the fused detection result is mapped onto the input image from the RGB camera in order to generate region proposals. The region proposals provide some regions to localize objects. Once an object has been detected, it can also be tracked based on our approach in order to derive more valuable object properties such as its location and velocity.

Object classification: once objects have been detected by the proposed multi-sensor fusion approach, the type of each object within regions is determined by a CNN. The generated regions are used as input to the CNN architecture to extract the deep features. We focus on three main objects in maritime environment: boat, seamark and lands.

IV. Object Detection

This section presents the object detection method for each individual sensor.

Radar-based object detection: The marine radar data frames are mapped first from polar to 2D cartesian coordinates. Then an intensity threshold filtering is applied to remove noise and extract the objects from radar data. After that the extracted objects are mapped to a binary image. The intensity threshold is determined through empirical experiments. A morphological dilation technique is applied on the created binary image to cluster the detected objects into more coherent groups. The refined 2D-plane coordinates determined from the radar data is used as the basis for mapping all other sensor data in a format applicable for sensor fusion.

LiDAR-based object detection: To remove noise from the point cloud LiDAR data, a low-pass/median filter is employed. Filtering the LiDAR data can effectively mine the information from the data for object detection. Later, the height component of the LiDAR data is discarded in this module and the x/y-coordinates of the LiDAR point cloud features are treated similarly to radar data.

IR camera-based object detection: Feature segmentation is firstly applied on grayscale IR camera images. The IR-camera segmentation is based on both gradient and intensity-based feature extraction. Image areas with significant and uniform horizontal gradients, which are not typical for the water surface are extracted with grayscale convolution and threshold operations. Moreover, high-intensity "hot" features are extracted with a threshold operation. The results of the gradient and intensity evaluation are combined into a single binary (1b) feature image. Mathematical morphology operations are then applied, to remove noise and to cluster remaining features into more robust object blobs. After IR camera images have been segmented, they are stitched into a single binary feature image and a Connected-Component-Labeling (CCL) operation is applied to extract a bounding box for each binary object. The bounding boxes are then given to a standard Kalman filter to remove temporal noise, such as blinking or very short-lived features.

RGB camera-based object detection: a similar approach for IR data segmentation is applied for RGB camera. The approach is based on the extracting local (horizontal) gradients clusters differing from the typical water surface. Large high-intensity features (discarding image saturation) are also extracted with a threshold operation and combined logically with the gradient data. As the intensity gradients approach cannot efficiently detect and track some small objects such as seamark that hardly is distinguishable from the water, a red/green feature segmentation approach is employed. In addition, the image-based evaluation and processing tasks are applied on RGB camera data in order to take into account environmental issues such as day or night conditions and sun glare induced sensor saturation. Finally, the object detection is performed by extracting the binary features from RGB cameras. The detected objects are stitched together and given to a Kalman filter for temporal filtering.

Fig. 2 shows an example of the output of object detection module for each individual sensor in our framework.

V. Fusion and Tracking

Each sensor, based on its ability, can detect limited objects with some levels of deficiency. Combining data captured

[1]This comes form the fact that the movement of the objects in camera coordinate, i.e., front view, in comparison against the radar coordinate, i.e., bird view, does not follow a linear pattern, because of non-linearity of the transformation of xyz coordinates on camera coordinates, and is not suitable for Kalman filtering and other probabilistic movement prediction methods, and to do that the camera calibration parameters must be considered

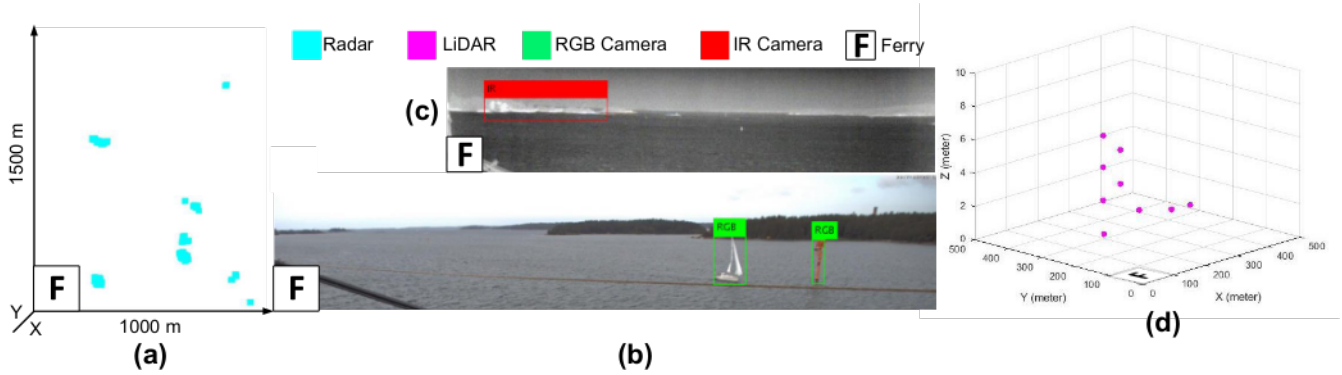

Fig. 2. An example of the output of object detection module in the proposed framework. (a) location intensities of the detected objects by radar; (b) two detected objects (boat and seamark) by RGB camera; (c) one detected object (land) by IR camera and (d) 3D point cloud data for two detected objects (boat and seamark). In the bottom left corner of all figures, 'F' box indicates the location of the ferry.

from multiple sensors corrects such deficiency of individual sensors to improve the performance in calculating the location and orientation of objects. In addition, noise and weaknesses vary under different environmental conditions for each individual sensor. For example, RGB camera or IR camera can be used for day and night-time imaging. However, radars can detect object better than cameras in different weather conditions (e.g. heavy rain or snow) as they use S-or X-bands. The data from LiDAR can give a good estimation of the distance to the object and its visible size. To achieve accurate object detection by using measured data from four sensors: RGB camera, IR camera, radar and LiDAR, we applied an extended version of probabilistic data association (PDA) technique for multi-sensor systems. In this technique, the probability that each measured sensor data is attributable to an object is calculated in real-time based on the current environmental condition and distance estimation of the target object from the ferry. The probabilistic data is used to track the objects over time that is called PDA filtering (PDAF).

Before performing the proposed fusion approach, all sensor data is projected to radar domain where size and position of the detected objects relative to the maritime can be recognized in a 2D map. Since the LiDAR data is in 3D point cloud format, it can be easily projected to radar domain by eliminating the third dimension. RGB and IR cameras data are mapped to radar domain with the inverse perspective mapping (IPM) technique [19], [20]. To do that the transformation matrix for each camera is calculated based on the location of the camera with respect to the surface of the water, that is the coordinate of radar in 2D bird view domain. Fig. 3(a) shows an example of mapping the detected objects by all sensors on radar domain. This is an example of *input* in our fusion approach that described in Algorithm 1.

The input of the Algorithm 1 is the list of location of detected objects by four sensors in radar domain and the output is the list of fused detection result. A target of interest, while using the standard PDA method, has the *Markov-property* [21] that means its state, x_k, i.e., the object position, velocity, and so on at time step k is only dependant on the

Algorithm 1 Sensor fusion process.

Input: $input$: Extracted geographical location of objects in radar domain, env: Environmental condition;
Outputs: FD : List of fused detection result;
Constant: G_d: Gating values for each distance zone d;
R: List of measurement co-variance matrix for different zones and environmental conditions;

Body:
1: $FD \leftarrow \emptyset$;
2: **for** each distance zone d **do**
3: **for** $Objs$:Objs $\in$ $input$ and $Objs \in$ the distance zone d **do**
4: $FD \leftarrow$ PDAF($Objs$, G_d,R$_{(env,Z_d)}$)$\cup FD$;

state in time step $k-1$. In *Markov-property* based systems, the state can be modeled as follows:

$$x_k = f(x_{k-1}) \tag{1}$$

where f is the iterative map function that evolves the state of the object in discrete time.

In PDA, a *measurement validation region* for each detected object must be defined [3]. For this purpose, the history of the position and behaviour of the moving object should model by a normal distribution by considering *Markov-property*. The *measurement validation region* is elliptical region V defined by Mahalanobis distance [22] as follows:

$$V(k,\gamma) = \{z : [z - \hat{z}_{k|k-1}]^T S_k^{-1} [z - \hat{z}_{k|k-1}] \leq \gamma\} \tag{2}$$

$$S_k = H_k P_{k|k-1} H_k^T + R_k \tag{3}$$

where $\hat{z}_{k|k-1}$ and $p_{k|k-1}$ represent the value of mean and co-variance for normal distribution, respectively. P_k and R_k indicates the state co-variance matrix and measurement co-variance matrix in standard Kalman filter [3]. P_k can model the noise in object localization and tracking. R_k represents the real noise by each sensor in a multi-sensor environment.

In Equation 2, parameter γ determines the size of the valid region for measurement and called *gate threshold*. Algorithm 1 performs PDA based on the distance from the detected objects to the autonomous vessel (ferry in our

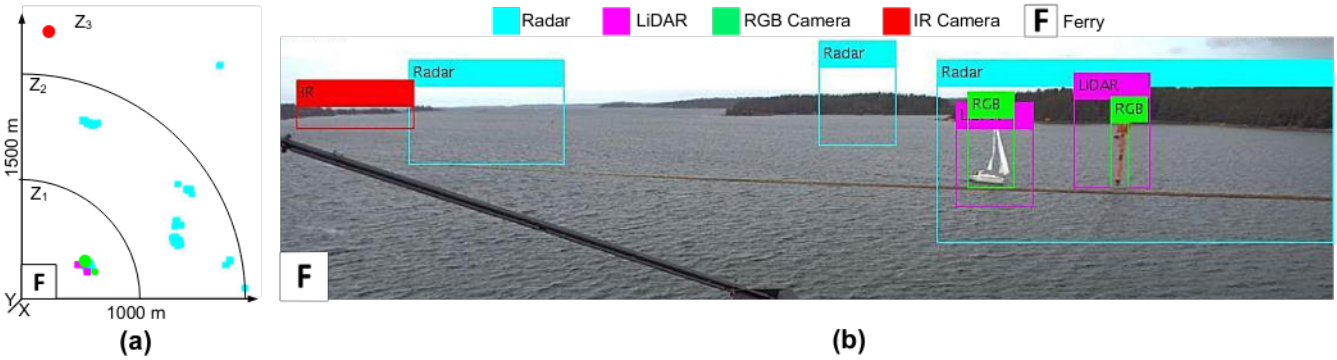

Fig. 3. Mapping the detected objects by all sensors in Fig. 2 on (a) radar coordinates and (b) RGB camera image. 'F' box in the bottom left corner of each figure indicates the location of the ferry.

case) as the amount of clutter and false alarms in farther distance is higher. It splits the tracking region into the divided zones based on radial distance from the ferry with different *gate threshold* value for each zone. The farther zone from the ferry, the largest *gate threshold*. Algorithm 1 uses the different sensor's error co-variance matrix for each distance zone and environmental condition (Line 2-4). The environmental condition (e.g. night/day) is an input of the algorithm. It means we consider different error co-variance matrices based on the current environmental condition. The environmental condition is obtained based on preprocessing of images from RGB camera.

In each zone, a PDAF function is applied to calculate the location of each object (Line 4). An example is shown in Fig. 3(a) for three zones. If in time step k, the i^{th} validated measurement Θ_i inside the *measurement validation region* is the target originated measurement with the probability of $P(\Theta_i|z_k)$. The object location in time step k , $\hat{x}_{k|k}$ is calculated as follows:

$$\hat{x}_{k|k} = \sum_i \hat{x}^i_{k|k} \times P(\Theta_i|z_k) \tag{4}$$

where $\hat{x}^i_{k|k}$ is the envisioned state if the i^{th} measurement is correct and can be calculated according to the normal Kalman filter equation as follows:

$$\hat{x}^i_{k|k} = \hat{x}_{k|k-1} + K_k v^i_k \tag{5}$$

$$K_k = P_{k|k-1} H^T_k S^{-1}_k \quad , \quad v^i_k = z^i_k - \hat{z}_{k|k-1} \tag{6}$$

where K_k is the filter gain and v_k is the innovation or the measurement residual on time step k. The prediction of $\hat{x}_{k|k-1}$ which is the conditional mean of the state at time step k as well as covariance matrix P_k is done the same as traditional Kalman filter as follows:

$$\hat{x}_{k|k-1} = A_{k-1} x_{k-1|k-1} \tag{7}$$

$$P_{k|k-1} = A_{k-1|k-1} P_{k-1} A^T_{k-1} + Q_{k-1} \tag{8}$$

where Q_{k-1} is the covariance matrix of noise on measurement on time step $k-1$ and is different from R. Equation 4 is generally takes *all* the probable measurements into account to find the position of an object with an appropriate probability. Fig. 4(a) shows an example of the output of Algorithm 1 that fused the detection results of four sensors (see Fig. 2) and generates the location of objects.

Algorithm 2 is performed to extract the object region proposals. First of all, the corresponding bounding boxes for each sensor data and final fused object detection results (the output of Algorithm 1) are mapped on RGB camera image via the perspective mapping (PM) technique, i.e, opposite to IPM. It is worth mentioning that radar data gives blobs in the area where there exists some object. After IPM to transfer radar data to RGB image, the bounding boxes of radar data is extracted from the transfered blobs on RGB image. Fig. 3(b) is an example that shows how the detected objects in Fig. 2 are mapped on RGB camera image. After that the nearest bounding box from the fused object detection results with the highest priority will be selected as a region proposal. The priority is defined based on the type of the sensors. Since the experimental results show that the RGB camera image is the most accurate candidate for defining a bounding box spatially, it has the highest priority. After that, the generated bounding boxes by IR camera, LiDAR and radar have higher priorities, respectively. For example in Fig. 3(b), there are three bounding boxes around the boat. However, the bounding box of RGB is a tightest bounding box in comparison with other two LiDAR and radar bounding boxes. Finally the generated ROIs will be passed to the classification module. Fig. 4(b) shows the final extracted region proposals based on the fused detection result without labels.

VI. Object Classification

The extracted object region proposals via the proposed fusion approach are classified by using Convolutional Neural Network (CNN). An example of the output of classification module is shown in Fig. 4(b). We applied a CNN on top of region proposals in order to extract a feature vector from respective region proposals. The proposed CNN consists of

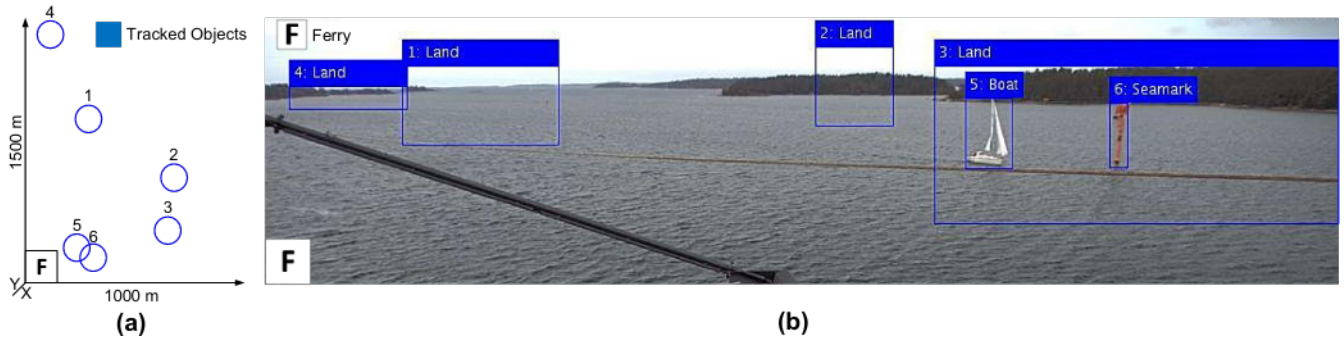

Fig. 4. The result of multi-sensor fusion for the example shown in Fig. 3 on radar coordinates and the corresponding regions of interest and classification results on RGB camera image. 'F' box in the bottom left corner of each Fig. indicates the location of the ferry.

Algorithm 2 Cropping process.

Inputs: FD: List of fused detection result from Algorithm 1, BBs : List of BBs of detected objects ;
Outputs: RPs: List of region proposals;

Body:
1: $RPs \leftarrow \emptyset$;
2: **for** each $Obj \in FD$ **do**
3: $RPs \leftarrow RPs \cup$ from BBs select nearest highest priority BB from Obj;
4: remove Obj from FD;

two convolutional layers, two max-pooling layers, one fully-connected layer, and a softmax layer. Each convolutional layer is followed by a max-pooling layer.

All convolutional and fully connected layers utilize the rectified linear unit (ReLU) as an activation function. ReLU can make training faster than other activation functions, such as tanh and sigmoid [6]. It maps negative values to zero and maintains positive values as

$$y = max(0, x), \tag{9}$$

where y and x are the neuron output and input, respectively. Each neuron in the second, forth, and sixth convolutional layers is modeled by max-pooling to reduce the amount of parameters and computation in the network.

In order to reduce overfitting in the fully-connected layers, we employed a popular regularization method called "dropout" that proved to be very effective. Dropout [23] provides a way to approximately combine many different neural networks efficiently. The key idea is to randomly drop out hidden and visible units (along with their connections) from the network during training.

VII. Experimental Setup

To evaluate our proposed framework, we collected a real dataset from a ferry operating in the Finnish archipelago [7]. In this dataset, RGB camera images are 1920×1080 captured via 5MP image sensor with 92° lens angle. The frame rate of the captured video is two frames in seconds. The deployment locations included the open sea. The IR camera image resolution is 512×640 working between -50°C to 70°C temperature. The radar range is upto 1.7KM with angular sampling interval of 0.4°. It is worth mentioning that, since the navigator must discard the nearby object such as the mast of the boat and passengers, all the nearby sensor data up to 18m are filtered.

The proposed CNN model is trained based on the images of three interest objects. These images are extracted from real videos that are recorded by the RGB camera. These videos represent various weather conditions from 4^{th} October 2016 to 25^{th} July 2017 for each sensor. Examples of images are shown in Fig. 5. Each image includes a object of three classes: boat, seamark and land. These images are generated by creating minimal bounding boxes around an object that is detected by the RGB camera-based object detection method.

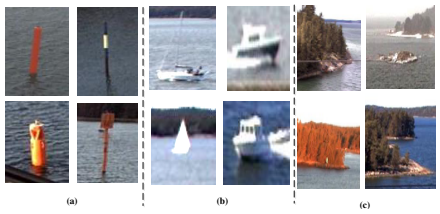

Fig. 5. Example of Images for training the proposed CNN of three interest objects (a) seamark, (b) boat and (c) lands

The following pre-processing steps were performed on the images before using them for CNN training:

1) **Resizing:** as the smallest cropped image size was 32×32 pixels, we change the size of all images into 32×32 that later are fed to our CNN tool.
2) **Feature Normalization:** the numeric features must be normalized for removing the effect of original feature value scales. The pixel values are in the range of 0 to 255 for each of the red, green, and blue channels. The pixel values were normalized into the range 0 to 1.
3) **Class encoding:**the non-numerical class types are converted into the numeric categorizes. We used one hot encoding to convert four categorical classes into four binary classes, with only one active.
4) **Data augmentation:** apart from regularization, another efficient way in order to avoid overfitting is data augmentation. We create more images from the original images via a number of random transformations [24]. Random transformations were applied on

the original training images include rotation, cropping, swirl, vertical flip and horizontal flip. The number of images for seamark, boat and land classes after data augmentation are 4572, 3759, and 4757 respectively.

We tune different hyperparameter in order to evaluate the performance of CNN model. The performance of the model is changed depending on the value of hyperparameters. In order to tune the hyperparameters for all models in this paper, we utilize 10-fold cross-validation approach subjected to the dataset of 17,328 images. After that, the most fitted value of hyperparameters are selected, the final model is trained with all 17,328 images. The performance of CNNs highly depends on the network topology. For this reason, we tried to find the network topology that is optimal to our object detection problem. The layers' structure of proposed CNNs is described in Table I. We got 85.81%, 92.87%, 92.78%, 86.74% and 85.33% test accuracy for $Model1$, $Model2$, $Model3$, $Model4$ and $Model5$ respectively. Therefore, the CNN with two convolutional layers is the optimal model, i.e., $Model2$.

In order to reduce overfitting, we use dropout in the fully-connected layers of $Model2$. The value of the dropout ranges from 0.0 to 0.9. We see that as dropout is 0.5, the model can get better accuracy. Moreover, the value of batch size and epochs are 25 and 10, respectively. In addition, the best optimizer for our neural network model in order to learn properly and tune the internal parameter is the Adam [25] based on our experiments. Moreover, we tune the learning rate parameter that is used in the Adam with the grid search. Learning rate controls the speed of wight updating at the end of each batch. We tried a suite small standard learning rate from 0.001 to 0.3 in steps of 0.1. The best performance of the model is achieved when the learning rate is 0.001.

The parameters of proposed PDA-based fusion approach are are shown inTable II. Two environmental conditions 'day' and 'night' is considered for setting different error estimate for sensors.

VIII. Experimental Results

Our framework is evaluated on a real test dataset which is collected by the ferry. The proposed framework first performed object detection based on each sensor data. Then, it used a PDA-based fusion and tracking approach at decision level in order to generate region proposals. The generated regions are mapped onto the RGB images that obtained from RGB camera. For each region, we wrap the image to fixed pixel size 32x32 that is required to make it compatible with the trained CNN. With each warped region, we extract features from the CNN with two convolutional and one fully connected layers. In addition, the RGB images of test data set was manually tagged in order to provide a ground truth reference. The number of three interest objects of Land, Boat, and Seamark are 851, 103, and 266 respectively.

Table III and Table IV show that the detection and classification results obtained by the radar, LiDAR, cameras and our fusion approach. The first row of each sensor shows that how many of objects are detected or classified in each class. For clarity sake, the number of detections and classification are also represented by percentages at the second row. Three objects of interest were taken into account: seamark, land and boat. Table III shows that the correct and false detection of three objects. The correct detection determines how many of each object is detected correctly. The false detection represents how many of all objects are not detected. The results show that the detection rate of three objects is improved by our fusion approach. The false detection rates (12.6%) is due mainly to the noisy radar target detection and the reflection in raw LiDAR data which creates ghost objects. However, the fusion approach allows to obtain a highly correct classification rate for all objects. It is worth mentioning that sensors in maritime environment behave differently compared to when they are employed in vehicles. For example the surface of the water does not reflect the laser beams for LiDAR data that results in low amount of reflected data. Another issue is the distance of objects of interest that is farther compared to objects of interest in vehicles. Such facts reveals high diversity among the contribution of sensors to detect an object.

Table IV summarizes the results collected after testing our trained CNN with on-line data. Correct classifications represent well classified objects of three classes when the region proposals obtained from each sensor and our fusion approach. False classifications show the number of percentage of object that are miss-classified for each class. When the CNN is applied on the regions obtained by our approach, it can correctly classify 89.4%, 100% and 91.1% of seamark, boat and land, respectively. Moreover, we can achieve a high classification accuracy from CNN on region proposals obtained by each sensor. Therefor, the classification rate of all objects by the proposed CNN are nearly perfect (86-100%).

IX. Conclusion

This paper presents an efficient multi-sensor fusion approach based on the probabilistic association method. In order to achieve reliable object detection, this approach fuses the data from four sensors and generates object region proposals. Moreover, a deep convolutional neural network is proposed for classifying the objects within the regions. We evaluated the performance of our proposed approach by conducting experiments with real data obtained by testing sensor arrays in a range of operating and climatic conditions in Finland. The obtained results show that our approach has a clear potential. As a future work, we plan to extend this approach to apply in time-series data on real marine environment in collaboration with Rolls-Royce. Meanwhile, it would be interesting to investigate uncertainty of CNN output in order to track the objects of interest across consecutive sensor data will facilitate detection and recognition. Fusing the classification results is another possible future works.

References

[1] Daniel Socek, Dubravko Culibrk, Oge Marques, Hari Kalva, and Borko Furht. A hybrid color-based foreground object detection method for automated marine surveillance, 2005.

TABLE I

PROPOSED CNNS FOR OBJECT CLASSIFICATION

Name	Conv1 32(3×3)	Pool1 (2×2)	Conv2 32(3×3)	Pool2 (2×2)	Conv3 64(3×3)	Pool3 (2×2)	Conv4 64(3×3)	Pool4 (2×2)	Conv5 128(3×3)	Pool5 (2×2)	Conv6 256(3×3)	Pool6 (2×2)	Conv7 512(3×3)	Pool7 (2×2)	FC1 128	FC2 3
Model1	✓	✓	✓	✓	✓	✓	✓	✓							✓	✓
Model2	✓	✓			✓	✓										✓
Model3	✓	✓			✓	✓			✓	✓						✓
Model4	✓	✓			✓	✓			✓	✓	✓	✓				✓
Model5	✓	✓			✓	✓			✓	✓	✓	✓	✓	✓		✓

TABLE II

PDA PARAMETERS IN THE PROPOSED FUSION APPROACH

Probability of gating (PG)	0.85
Probability of detection (PD)	0.75
Δt (time interval sampling(Second))	0.5
Maximum number of target objects	10
Gate thresholds for three Zones(m^2)	20, 30, 50

TABLE III

DETECTION RESULTS OF FOUR SENSORS AND OUR MULTI-SENSOR FUSION APPROACH

Sensor	Correct			False
	Seamark	Boat	Land	All
Radar	97	0	737	386
	36.4%	0.0%	86.6%	31.6%
LiDAR	17	0	69	1134
	6.3%	0.0%	8.1%	92.9%
IR camera	180	37	263	740
	67.6%	35.9%	30.9%	60.6%
RGB camera	180	53	101	886
	67.6%	51.4%	11.8%	72.6%
Our approach	**209**	**67**	**790**	**154**
	78.5%	**65.0%**	**92.8%**	**12.6%**

TABLE IV

CNN CLASSIFICATION RESULTS ON DETECTED OBJECTS OF FOUR SENSORS AND OUR FUSION APPROACH

Sensor	Correct			False			Total accuracy
	Seamark	Boat	Land	Seamark	Boat	Land	All
Radar	81	n/a	737	16	n/a	0	818
	83.5%	n/a	100%	16.50%	n/a	0.0%	98.0%
LiDAR	17	n/a	69	0	n/a	0	66
	100%	n/a	100%	0.0%	0.0%	0.0%	100%
IR camera	140	37	238	40	0	25	415
	77.7%	100%	90.4%	22.2%	0.0%	9.50%	86.4%
RGB camera	158	53	95	22	0	6	306
	87.7%	100%	95.65%	12.3%	0.0%	4.35%	91.6%
Our approach	**187**	**67**	**776**	**22**	**0**	**75**	**1030**
	89.4%	**100%**	**91.1%**	**10.5%**	**0.0%**	**9.6%**	**96.6%**

[2] R. T'Jampens, F. Hernandez, F. Vandecasteele, and S. Verstockt. Automatic detection, tracking and counting of birds in marine video content. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, Dec 2016.

[3] Y. Bar-Shalom and X.R. Li. *Multitarget-multisensor Tracking: Principles and Techniques*. Yaakov Bar-Shalom, 1995.

[4] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[5] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[7] S. Jokioinen, J. Poikonen, M. Hyvönen, A. Kolu, T. Jokela, J. Tissari, A. Paasio, H. Ringbom, F. Collin, M. Viljanen, R. Jalonen, R. Tuominen, M. Wahlström, J. Saarni, S. Nordberg-Davies, and H. Makkonen. Remote and autonomous ships - the next steps. *white paper*.

[8] T. Herpel, C. Lauer, R. German, and J. Salzberger. Trade-off between coverage and robustness of automotive environment sensor systems, Dec 2008.

[9] A. Mukhtar, L. Xia, and T. B. Tang. Vehicle detection techniques for collision avoidance systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2318–2338, Oct 2015.

[10] R Omar Chavez-Garcia and Olivier Aycard. Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking. *IEEE Transactions on Intelligent Transportation Systems*, PP(99):1–10, 2015.

[11] Sang-Il Oh and Hang-Bong Kang. Object detection and classification by decision-level fusion for intelligent vehicle systems. *Sensors (Basel)*, 17(1):s17010207–s17010207, 2017.

[12] Raphaël Labayrade, Dominique Gruyer, Cyril Royere, Mathias Perrollaz, and Didier Aubert. Obstacle Detection Based on Fusion Between Stereovision and 2D Laser Scanner. In Sascha Kolski, editor, *Mobile Robots: Perception & Navigation*. Pro Literatur Verlag, 2007.

[13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, June 2017.

[15] G. Prabhakar, B. Kailath, S. Natarajan, and R. Kumar. Obstacle detection and classification using deep learning for tracking in high-speed autonomous driving. In *2017 IEEE Region 10 Symposium (TENSYMP)*, pages 1–6, July 2017.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[18] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[19] Trevor Taylor. *Mapping of indoor environments by robots using low-cost vision sensors*. PhD thesis, Queensland University of Technology, 2009.

[20] Massimo Bertozz, Alberto Broggi, and Alessandra Fascioli. Stereo inverse perspective mapping: Theory and applications. 16:585–590, 06 1998.

[21] Yaneer Bar-Yam. *Dynamics of Complex Systems*. Perseus Books, Cambridge, MA, USA, 1997.

[22] P.C. Mahalanobis. On the generalised distance in statistics. 1936.

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[24] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6, Nov 2016.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.