Turun yliopisto
University of Turku

KONGSBERG

Deep Convolutional Neural Network-based Fusion of RGB and IR Images in Marine Environment

Fahimeh Farahnakian, Jussi Poikonen, Markus Laurinen, and Jukka Heikkonen

IEEE Intelligent Transportation Systems Conference (ITSC2019)
28 October, 2019, Auckland, New Zealand

Introduction (1/2)

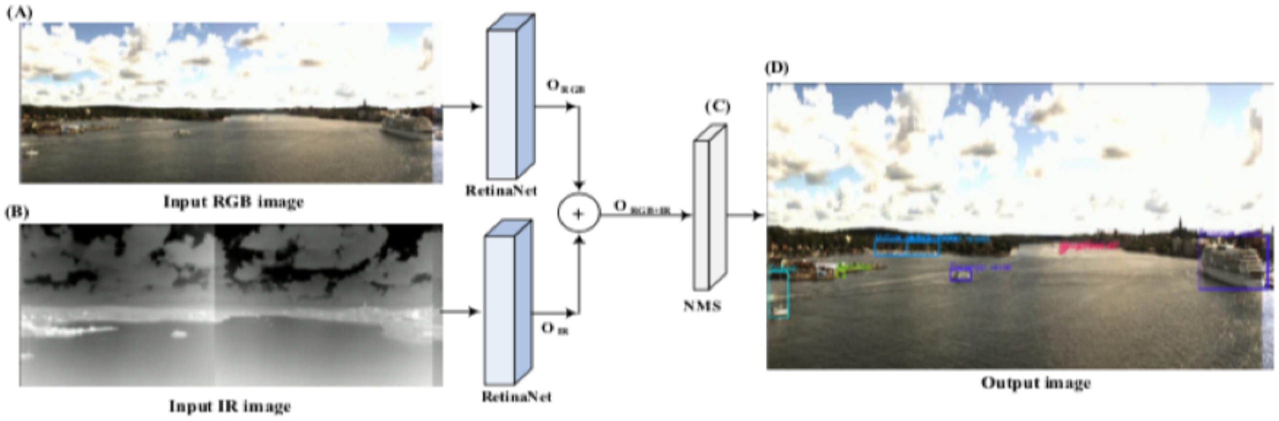
- Safety and security are critical issues in maritime environment.
- Automatic and reliable object detection is one of the efficient way for improving these issues in intelligent systems.
- Reliable and robust object detection is one of the most challenging task for autonomous vehicles due to the uncertain and dynamic environment and rapid movement of objects.
- Maritime object detection is quite different from its street counterpart
- Most of false positive because of
  - light reflection from water
  - remnants left by a boat on the surface of the water
  - seaweed that floats at the surface

Introduction (2/2)

- One of the main technologies that improve the understanding of the surrounded environment and therefore the robustness of object detection is **sensor fusion**.
- As each individual sensor has some weakness, combining the data from different sensors can optimise the situational awareness under all conditions.
- For example, visible cameras provide high resolution images for the object classification task. Although, infrared cameras can increase nigh-time navigation safety and detect warm objects at night time with high accuracy.

We believe that multi-sensor data fusion can develop a reliable perception capability for object detection in autonomous vehicles.

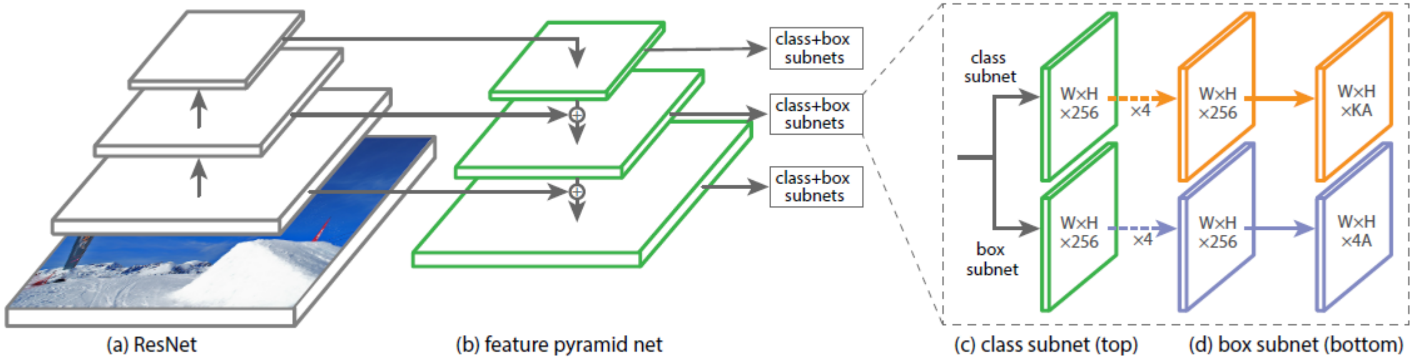
Proposed late fusion framework



1. The framework first precepts and obtains a detailed description of the environment surrounding the vessel by using RGB and IR cameras.
2. Then it employs RetinaNet for each input camera image separately in order to extract the interest targets proposals.
3. After that, it concatenates the obtained target proposals from two cameras and generates a final set of possible proposals.
4. Finally, the non-maximum suppression procedure is applied on the set of proposals to remove redundant proposals (duplicated detections on the same target).

RetinaNet

- RetinaNet is a simple dense detector which contains a backbone network and two sub-networks.
- First, the backbone network computes a convolutional feature map over an entire input image.
- Then, the first sub-network performs convolutional object classification on the backbone's output and the second sub-network applies convolutional bounding box regression.



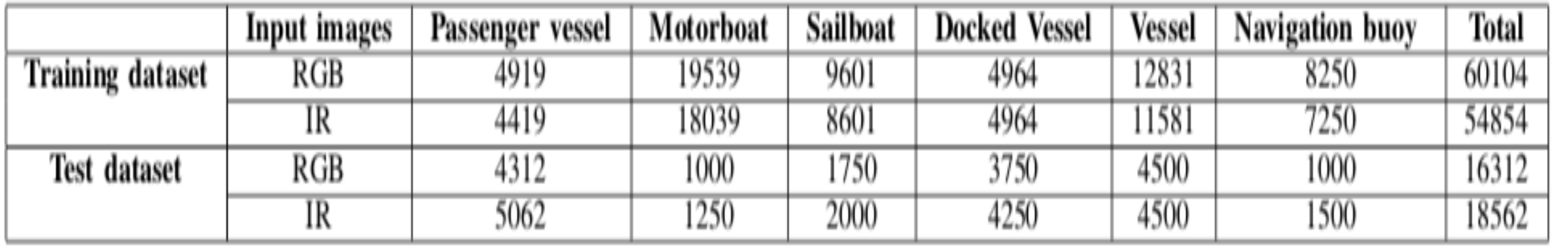
Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dolla ́r. Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision (ICCV), pages 2999–3007, 2017.

Data (1/2)

- Data was collected using a sensor system onboard a vessel in the Finnish archipelago.
- This sensor system includes RGB (visible spectrum) and IR (thermal) camera arrays, providing output which can be synchronized and stitched to form panoramic images.
- The individual RGB cameras have full HD resolution, while the thermal cameras have VGA resolution. Both camera types have horizontal field of view approximately 35 degrees.
- The data shows maritime scenarios with various objects such as ships and other vessels. For the experiments, we selected 5000 and 1750 images for training and testing the network, respectively. Size of each image is 3240 × 944 pixels. The original training images are augmented via a number of random transformations for training RetinaNet.

Data (2/2)

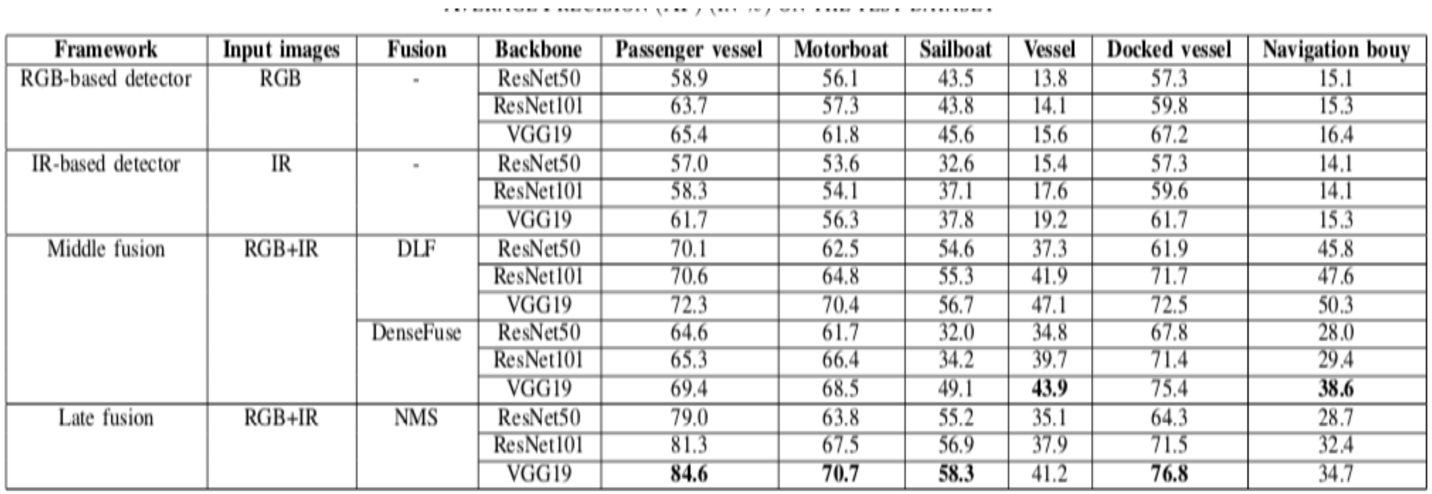
- We manually annotated the boundind boxes and the labels for the six object classes (five types of vessels, navigation buoy) on our dataset.
- Note that any far away vessels that could not be visually recognized as ”passenger vessel”, ”motorboat”, ”sailboat” or ”docked vessel”, were placed under the general label ”vessel”.



	Input images	Passenger vessel	Motorboat	Sailboat	Docked Vessel	Vessel	Navigation buoy	Total
Training dataset	RGB	4919	19539	9601	4964	12831	8250	60104
	IR	4419	18039	8601	4964	11581	7250	54854
Test dataset	RGB	4312	1000	1750	3750	4500	1000	16312
	IR	5062	1250	2000	4250	4500	1500	18562

Experimental Results

- We compare **two uni-modal** frameworks with **our multi-modal** framework.
- The **uni-modal** framework utilizes only the **visible** or **infrared** images to detect the interest objects around of vessel.
- Our multi-modal framework combines the information from **two input infrared and visible** images using the proposed image fusion methods.
- RetinaNet is trained based on three different backbone networks in our experiments: **ResNet50**, **ResNet101** and **VGG19** for our experiments.



Framework	Input images	Fusion	Backbone	Passenger vessel	Motorboat	Sailboat	Vessel	Docked vessel	Navigation bouy
RGB-based detector	RGB	-	ResNet50	58.9	56.1	43.5	13.8	57.3	15.1
			ResNet101	63.7	57.3	43.8	14.1	59.8	15.3
			VGG19	65.4	61.8	45.6	15.6	67.2	16.4
IR-based detector	IR	-	ResNet50	57.0	53.6	32.6	15.4	57.3	14.1
			ResNet101	58.3	54.1	37.1	17.6	59.6	14.1
			VGG19	61.7	56.3	37.8	19.2	61.7	15.3
Middle fusion	RGB+IR	DLF	ResNet50	70.1	62.5	54.6	37.3	61.9	45.8
			ResNet101	70.6	64.8	55.3	41.9	71.7	47.6
			VGG19	72.3	70.4	56.7	47.1	72.5	50.3
		DenseFuse	ResNet50	64.6	61.7	32.0	34.8	67.8	28.0
			ResNet101	65.3	66.4	34.2	39.7	71.4	29.4
			VGG19	69.4	68.5	49.1	**43.9**	75.4	**38.6**
Late fusion	RGB+IR	NMS	ResNet50	79.0	63.8	55.2	35.1	64.3	28.7
			ResNet101	81.3	67.5	56.9	37.9	71.5	32.4
			VGG19	**84.6**	**70.7**	**58.3**	41.2	**76.8**	34.7

Experimental Results

- We evaluated the performance of two proposed deep network based image fusion approaches (DLF[1] and DenseFuse[2]) in the middle fusion framework on our test dataset using six common quality metrics:

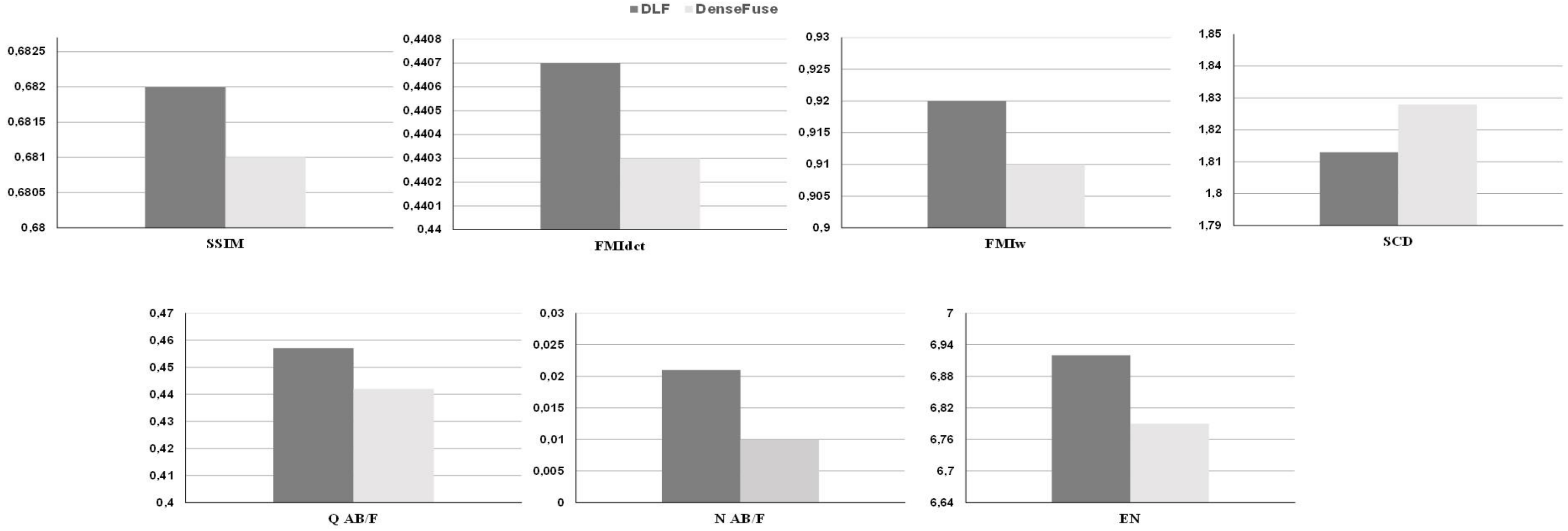
1. Structural SIMilarity (SSIM) compares the contrast, structure and luminance between image sources
2. Feature Mutual Information (FMI) is a non-reference performance metric that calculates the mutual information between the fused image and the source image. It measures the amount of information conducted from source images to fused image. Here, wavelet (FMI_w) and discrete cosine (FMI_{dct}) features are used.
3. Entropy (EN) measures the amount of information presented in the image. Lower entropy indicates better fusion.
4. Quality ($Q^{AB/F}$) index represents the visual information associated with the edge information. It measures the amount of edge preservation from source images to the fused image.
5. Noise ($N^{AB/F}$) index is the fusion artifacts measure proposed by which measures the noise or artifacts added in fused image
6. Sum of the Correlations of Differences (SCD) index calculates the sum of the correlation of differences based on the complementary information transferred from the source images.

[1] . H. Li, X.J Wu, and J. Kittler. Infrared and visible image fusion using a deep learning framework. CoRR, 2018.

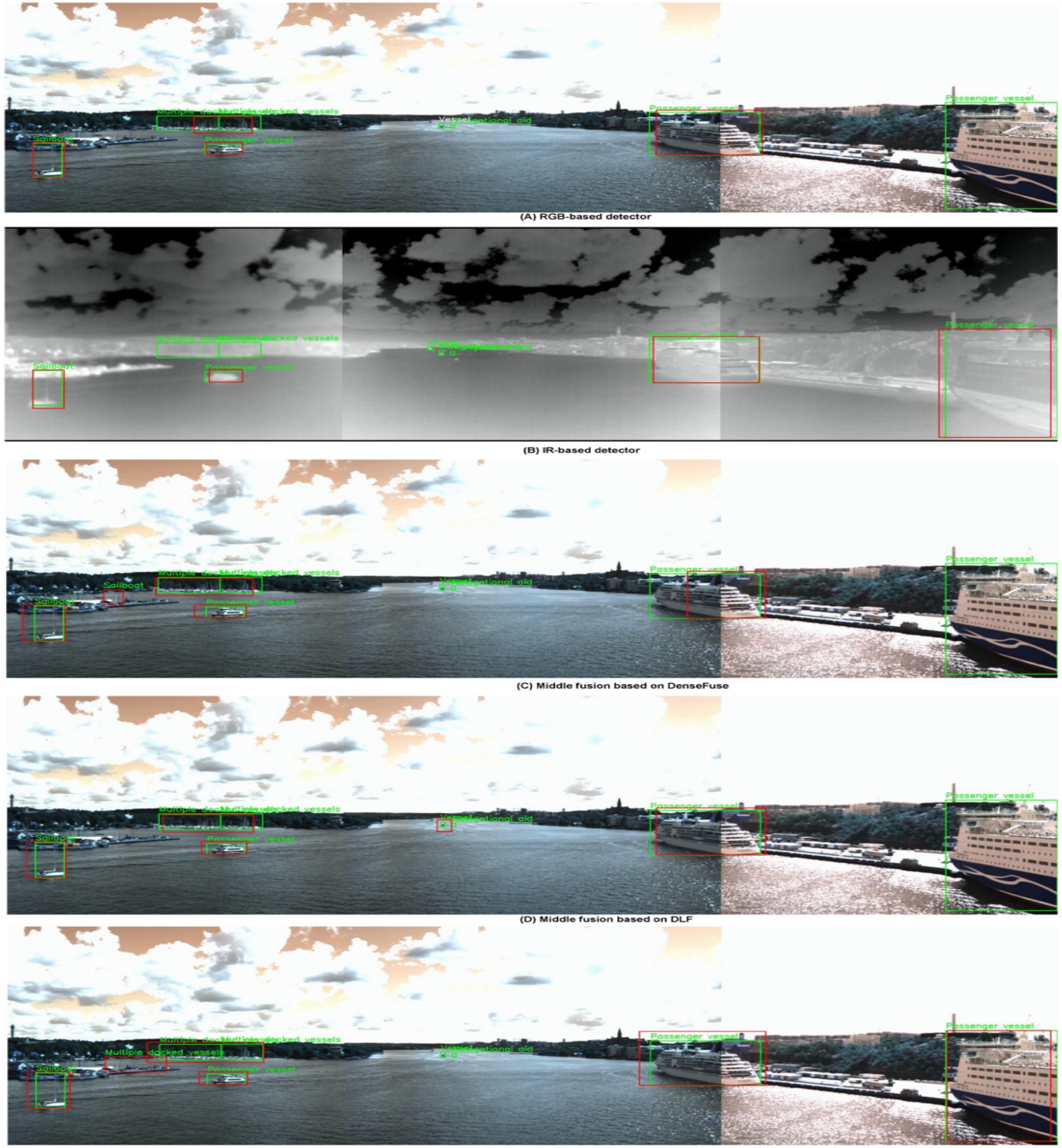
[2]. H. Li and X.Jun Wu. Densefuse: A fusion approach to infrared and visible images. CoRR, 2018.

Experimental Results

- DLF fusion methods with the highest values of SSIM, FMI_w and FMI_{dct} performs better than DenseFuse.
- Moreover, the SCD value of DLF is less than DenseFuse method. The reason is DLF can extract more structural information and features.
- However, DenseFuse obtains more natural and contain less artificial noise since it has the lowest values of $N^{AB/F}$, $Q^{AB/F}$ and En.
- In general, DLF performs better than DenseFuse for target detection in the proposed middle fusion framework.



Qualitative Results



(A) RGB-based detector

(B) IR-based detector

(C) Middle fusion based on DenseFuse

(D) Middle fusion based on DLF

(E) Late Fusion

Conclusion

- An **late fusion** framework is proposed in order to detect the interest objects in marine environment.
- To demonstrate the effectiveness of the proposed framework, we compared it with two uni-modal frameworks applied on only visible or infrared images.
- And two middle fusion frameworks
- We also evaluate the effects of more powerful backbone networks on the performance of RetinaNet in our framework.
- The experimental results on real marine data show that our multi-modal framework can achieve higher detection accuracy comparison with two another uni-modal frameworks.
- Our framework is effectively able to detect and classify objects into one of vessel type or navigation buoy in the real marine dataset, as long as their apparent image size is more than 10 × 10 pixels.

Future Works

1. The effects of object size and distance on the performance of our framework will be studied.
2. As it is very challenging to accurately detect small objects, an improved network structure of RetinaNet will be investigated in the future for this purpose. Further, we will extend our fusion framework by using data from lidar and radar besides RGB and IR cameras to improve the detection results.
3. In addition, more effective fusion schemes based on DL could be further developed to pursue better fusion performance.

Turun yliopisto
University of Turku

KONGSBERG

Deep Convolutional Neural Network-based Fusion of RGB and IR Images in Marine Environment

Fahimeh Farahnakian, Jussi Poikonen, Markus Laurinen, and Jukka Heikkonen

Thank you!