Evaluating Aerial Vessel Detector In Multiple Maritime Surveillance Scenarios

Gonçalo Cruz

Portuguese Air Force,
2715-021 Sintra, Portugal

Alexandre Bernardino

Institute for Systems and Robotics,
Department of Electrical and Computer Engineering,
Instituto Superior Tcnico,
1049-001 Lisboa, Portugal

***Abstract*—In this paper we present an autonomous detection approach for airborne surveillance in maritime scenarios. This approach is robust to sun glare, waves and scale variation. Additionally, we introduce a new metric to evaluate detection and tracking results that is more adequate for these scenarios. The proposed detection method is evaluated using videos from different monitoring missions and its results are compared with a state-of-the-art neural network. This comparison is done using a traditional and the proposed evaluation metric.**

1. Introduction

Despite the environmental [1], economical [2] and social [3] importance of maritime environments, monitor such vast areas is still a challenge. Maritime monitoring has implied the use of multiple assets, namely coastal stations, manned vessels, manned aircraft and satellites. All these platforms have relied primarily on radar technology, creating numerous approaches for automatic detection [4]. The use of active sensors has also placed power, space and weight requirements on the platforms, which limited the use of smaller and cheaper vehicles. The high cost of these technologies prevented a stronger control of the seas that could help with security [5], safety [6] and ecological [7] problems.

Lately, small aircraft (in particular unmanned) became easily accessible and particularly suited to carry visible spectrum cameras. Theses sensors are also ubiquitous but automatic detection in aerial images is still an open problem, therefore algorithms adequate to this problem are needed.

Detection in aerial images over land had some developments, with some approaches assuming the background is approximately static and objects are moving [8] [9]. Maritime scenarios preclude some of these approaches, has they contain challenging situations like glare, parts of the background that are also moving and objects can be quite diverse (from life rafts to oil tankers).

Even with the enumerated difficulties, several attempts to solve this problem have been made. Having several options raises questions like ”Which technique is better?” or ”What is the performance of a given method in different scenarios?”. Computer Vision researchers have faced this issue numerous times, defining metrics that help understanding the performance. The adoption of a given metric is usually related to the characteristics of the problem at hand. In the present case, we would like to evaluate detections of objects with significant size and aspect ratio variations and therefore there is a need to weight the localization error by the size of the bounding box.

This paper presents a detector that tries to overcome the difficulties of maritime surveillance scenarios as well as an adequate way to measure its performance. The contributions are:

- the use of a Convolutional Neural Network (CNN) with a method that exploits the time coherence present in video sequences;
- introduction of a metric to compare detection results with the ground truths, weighting the errors by the size of the object of interest;
- testing of the detector in different scenarios, using labeled videos with a considerable duration.

This paper is organized as follows. In Section 2, we review the main approaches for detection, in particular the ones that are applied to maritime scenarios. We also report the existing methodologies used to evaluate detection results. Section 3 contains the description of the detector. Section 4 focuses on the metric used to evaluate detection results. In Section 5, we describe the scenarios that were used for evaluation, present and discuss the results. Finally, in Section 6, we present some concluding remarks.

2. Related work

Objection Detection is one of the fundamental tasks in Computer Vision, which led to many approaches being suggested to this problem. Detection in aerial images has followed the trends of general computer vision but having some specificities. Some of the most relevant are the limited computational power if we consider that processing runs on the aircraft and also processing time that should small enough allow a timely action (*e.g.* start tracking). One important driver of evolution and improvement of the algorithms performance was the creation of evaluation methodology that allow an exact comparison of performance [10] [11]. Also like in other areas, an adequate evaluation

methodology is also important to compare different detection schemes. In the next two subsections we will discuss the method suggested for detection using aerial images and the existing evaluation methodologies, respectively.

2.1. Detection in Maritime Scenarios

Following some of the developments made in general Computer Vision, some works like [12], have detected people on land using Histogram of Oriented Gradientss (HOGs). Others like [13], used cascaded classifiers to detect people on foot and land vehicles and complement visible images with thermal images. Even with these techniques performing well in some cases, in the case of maritime detection, the objects' appearance variability is much higher. Others approaches, like [14] depend on the movement of the targets which is not well suited for maritime environments as some targets may be still and undesired events like wave crests and sun glare may have a significant movement. It is therefore difficult to characterize possible targets with respect to size, shape, colors, textures or movement in the image space.

Even with the aforementioned peculiarities, several specialized maritime detectors have been proposed. In [15], a set of features is designed to distinguish nautical objects from the ocean. However, the authors need to use several other layers to discard clutter. Similar approach is followed in [16] to detect marine mammals, using color features on a first stage and shape features secondly. In [17], the authors detect castaways by exploring the information contained in video sequences, more specifically, by using a Hidden Markov Model (HMM).

Just like in other applications, neural networks have been applied to the recognition of aerial images, even before modern CNNs were available [18] [19]. Like in other areas, the results with older network configurations were limited. More recently, taking advantage of more advanced network configurations, CNNs have been used for maritime detection using airborne images [20] [21] but with limited results when compared with other areas. This is caused by factors like the high variability of objects' appearance and phenomena like sun glare and waves. Additionally, most detectors were designed to be used in isolated pictures and only very recently some approaches based on CNNs like Kang *et al.* [22], took advantage of processing sequences instead of isolated images.

The method suggested in this work makes use of a existing network configuration and improves its performance by using a Multiple Hypothesis Tracker (MHT) to capture the dependency of detection across time.

2.2. Performance Metrics

Two main types of approaches have been considered to evaluate detection performance. The first type are Pixel-based methods, in which the detection output (groups of pixels of a given class) is compared with a labeled segmented image. The evaluation is then done as a binary classification problem and can be evaluated with missed detection rates, false positives or Receiver Operating Characteristic (ROC) [23]. This kind of method is not very attractive for maritime surveillance specially because creating accurate ground truth segmented images is a tremendous amount of work and detection stage usually does not need such detailed description of the object.

This drawback leads us to the second type of evaluation: Region-based methods. For these methods, ground truth typically consists of a rectangular region containing the object of interest and the output of the algorithm is usually rectangular as well (usually designated as bounding boxes). In this case, detections hardly are exactly equal to the ground truth, therefore there is no direct assignment (as with pixels) and a matching strategy must be followed. Some of the most used options consider the distance between both rectangles [24] or the overlap between both regions [25]. With this matching, evaluation is still considered as a binary classification problem, defining true and false positives, correct and missed detections. Subsequently, many performance figures can be defined (*e.g.* Precision, Recall or F-score) but are still dependent on the threshold considered for the matching. For instance, [10] considers that there should be an overlap between both Bounding Boxes (BBs) of 50%.

To avoid defining arbitrary thresholds, other metrics have been introduced that focus on evaluating localization [26]. A recent tracking evaluation approach [27] overcomes some of the mentioned limitations. In particular, the mentioned work assumes the creation of two curves: *Precision* and *Success*. The first one, plots the ratio of BBs as a function of distance between detections and Ground Truths (GTs). The second curve, shows the ratio of BBs as a function of overlap. The main advantage of this method is the use of continuous values for distance and overlap and not an arbitrary discrete value. In this work, we will follow this idea, using a distance between BBs that is weighted by the size and shape of the GT and plot the ratio of detections as a function of that distance.

3. Detection Approach

Our approach consists on using a Convolutional Neural Network to generate detection proposals and then associate these proposals using a Multiple Hypothesis Tracker (MHT). The network was based on DetectNet [28], a network which is itself based on GoogleNet but instead of classifying, it produces bounding boxes. In its unchanged version, depicted in Figure 1, the network's last layer receives a coverage map and produces bounding boxes. Each cell of this map represents a area of the original image and has a value between 0 and 1, where higher values mean that the cell is more likely to contain an object of interest. On its unaltered version, DetectNet processes each image without any memory of the past and does not take advantage of the previous detections.

To overcome this issue, we have slightly changed DetectNet as represented in Figure 2. This modifications consist on creating tracks with the successive detections and predicting

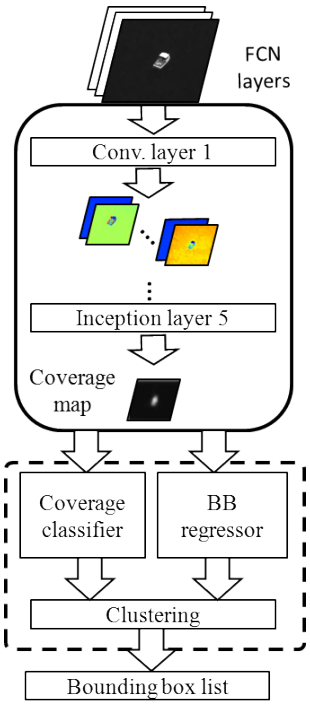


Figure 1. Simplified diagram of DetectNet. The solid line box encompasses the convolutional part of the network, up to the moment where the coverage map is produced. The final stage of the network (represented by the dashed line box) receives the coverage map, creates BBs and scores them.

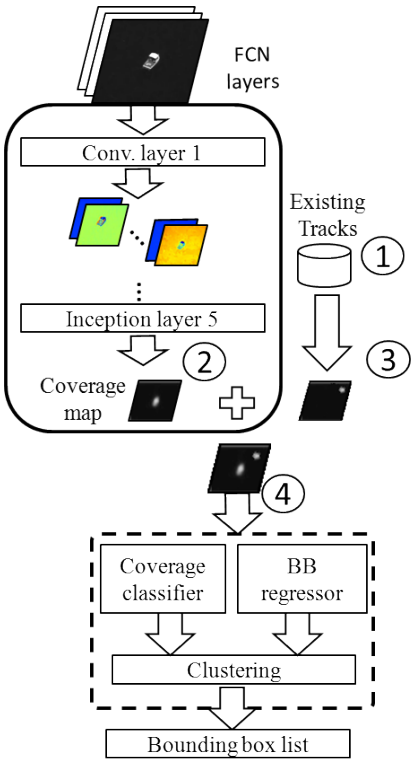


Figure 2. Simplified diagram of the network based on DetectNet. As in the previous figure, the solid line box encompasses the convolutional part of the network. Parallel to this convolutional part, there is the generation of a map, given the predicted object position (this only takes place, if there are any tracks). Finally, the last part of the network (represented by the dashed line box) receives the merge of the two maps and creates scored BBs.

the position of the objects in the future frame. The prediction model is

$$z(t+1) = z(t) + \Delta z(t) \tag{1}$$

where $z(t) = (x(t), y(t))$ is the position of a given object and $\Delta z(t) = z(t) - z(t-1)$. This location is then converted into the corresponding cell of a map (with the same size as the coverage map) and the cell's value is defined as 1.0. This process is depicted in the right-hand side of Figure 2. The map that was composed with the predicted position is then mixed with the network's coverage map, produced by the pipeline represented in the left-hand size of Figure 1. The resulting map is then fed into the last layer, represented as the dashed box in Figure 1.

In our problem, the MHT is used mainly to increase robustness by creating associations between detections and not to create very long tracks discriminating different objects. Yet it also improves the results by creating short duration tracks that ultimately increase the persistence of an object in the coverage map. The tracker is implemented by building a graph with the detections $\bar{\mathbf{D}}^t$. As depicted in Figure 3, at each time instant, a level of the graph is built. The level is composed of nodes that correspond to the detections from that time instant. At a given level, a node will have as parents all the nodes from the top level, unless its distance is above a certain threshold. Additionally, each node also contains the detection's score.

The tracking is then done by searching paths from a given detection at time instant t, all the way to older levels. This process is represented in Figure 3 with (a) showing the paths obtained from $\bar{D}^t_{k_0}$ and (b) showing the paths from $\bar{D}^t_1$ up to level 3. Because of the combinatorial nature of finding all paths, the search is done only up to a given time horizon, limiting the considered number of levels. With a limited search depth, the graph will create tracks $T_j^{t,\cdots,t-R}$ that are composed by detections $\left\{\bar{D}^t_{l_0}, \bar{D}^{t-1}_{l_1}, \cdots, \bar{D}^{t-R}_{l_R}\right\}$, where $\bar{D}^{t-R}_{l_R}$ is the *lth* detection in level R. The selection of a depth R should allow a time interval bigger than phenomena like glare and waves, excluding most false detections.

The probability of each track $T_j^{t,\cdots,t-R}$ being generated by correct detections $\left\{\bar{D}^t_{l_1}, \bar{D}^{t-1}_{l_2}, \cdots, \bar{D}^{t-R}_{l_R}\right\}$ is evaluated using the score of each detection and a weighting function that penalizes parent-child pairs that are farther from each other. This probability can be written as

$$P(T_j^{t,\cdots,t-R}) = P(\bar{D}^t_{l_r} \approx G^t_j) \prod_{r=1}^{R} P(\bar{D}^{t-r}_{l_r} \approx G^{t-r}_j) C(\bar{D}^{t-r}_{l_r}, \bar{D}^t_{l_r}) \tag{2}$$

with $P(\bar{D}^t_{l_r} \approx G^t_j)$ representing the probability of a detection matching a given ground truth label G^t_j and $C(\bar{D}^{t-1}_{l_1}, \bar{D}^t_{l_0})$ representing the cost of associating two con-

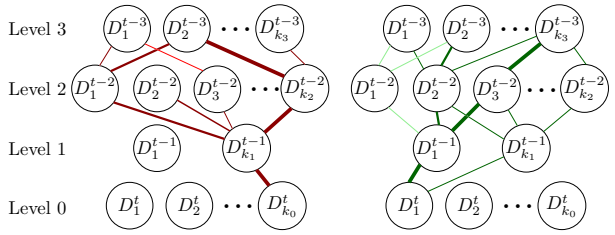


Figure 3. Simple example of a graph similar to those used in MHT.

secutive detections $\bar{D}_{l_1}^{t-1}$ and $\bar{D}_{l_0}^{t}$. To compute C, we have used a Gaussian function that depends on the distance between the center point of the bounding boxes, *i.e.*,

$$C(\bar{D}_{l_1}^{t-1}, \bar{D}_{l_0}^{t}) = e^{(-d_x^2+d_y^2)/2\sigma_x\sigma_y} \tag{3}$$

with d_x and d_y being calculated as

$$d_x = (x_{l_1}^{t-1} + \frac{w_{l_1}^{t-1}}{2}) - (x_{l_0}^{t} + \frac{w_{l_0}^{t}}{2}) \quad \text{and} \tag{4}$$

$$d_y = (y_{l_1}^{t-1} + \frac{y_{l_1}^{t-1}}{2}) - (y_{l_0}^{t} + \frac{h_{l_0}^{t}}{2}) \quad . \tag{5}$$

The σ_x and σ_y should be a compromise between being able to accommodate for camera motion and small enough to distinguish boats that are close to each other.

By calculating Equation (2), for each detection at a time instant t, we get possible tracks with different scores, represented in Figure 3 by thinner and thicker edges. Using MHT, if we have a false detection that is wrongly assigned a high score by the CNN but spatially is far from detection in previous frames, then the track with origin in this detection will have a lower probability. Conversely, if there is a correct detection with a low score but that is at approximately the same location as previous detections, then the detection might still have a significant probability.

The usage of the MHT allowed us to relax the tunning of the threshold present in DetectNet's last layer, setting it to a low value. With a low threshold, many detections are created and fed into the MHT that associates them, given the score of each detection and also the distance between detections in consecutive frames. This leads to many possible combinations but only the ones with higher combined score are considered.

In Figure 4, we supply an example where the contribution of the tracks to the detection process is visible. In Figure 4(a) is shown the result of the tracker, with a green line connecting the centers of previous detections. Figure 4 (b) contains the coverage map produced by the part of the network contained in the solid line box in Figure 2 (the same as would be produced by the unchanged DetectNet). While the bigger blob of this map has a high value, the smaller blob has values closer to the background. If no additional care was taken, it would be very difficult to distinguish this situation from a case where a distractor would be visible. With the incorporation of the tracks' information into the map in Figure 4(c), then the knowledge of the previous time instants is propagated into the neural network. The cells corresponding to the position predicted using Equation (1), are assigned a value one and then merged with the coverage map produced by the neural network. The final result is the map represented in Figure 4(d) that incorporates information of the convolutional part of the network and also from the tracks and where the smaller blob has a higher value.

4. Performance Metrics

Detection and tracking are very relevant tasks in Computer Vision and have benefited a lot from the creation of benchmarks and establishment of solid performance metrics. Nevertheless, different metrics highlight different characteristics and there is still no one-fits-all metric.

One of the most used for detection was introduced by Dollar *et al.* [10]. This method characterizes a given detection scheme by checking if detections can be matched to ground truths. According to this matching, Precision-Recall or Missed Rate-False Positive Per Image curves are created. However, the matching is based on a hard threshold. If there is enough overlap between detection and ground truth, the matching is correct. If the overlap is not enough, then the detection is considered incorrect. One disadvantage of this method is that detections have the same quality as long as both satisfy the minimum overlap with GT. Another drawback is that if a detection has no overlap, even though it might have exactly the same size and aspect ratio, then is considered as bad as another that might be far from the ground truth.

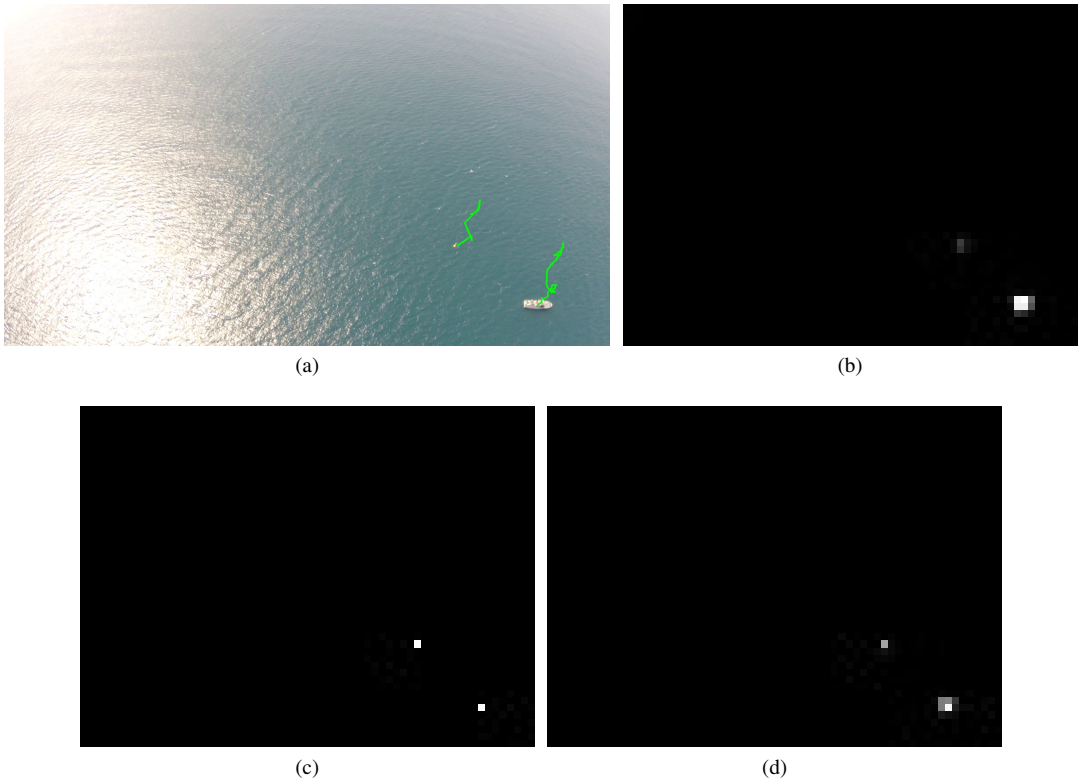
For the tracking task, one common approach is Object Tracking Benchmark (OTB) [27]. As stated in Subsection 2.2, *Precision* is verified for different distance thresholds and *Success* with different overlap values. Even though this method has softer metrics, *Sucess* penalizes equally two cases that do not overlap (even if in one case BBs are close and in other are very far apart) and *Precision* penalizes equally two cases given the distance between BBs (though in one case, the size of bounding boxes might be very similar and in the other it might be very different).

4.1. Proposed Metric

To have a metric in which detections are not classified binarily but differences in localization, areas and aspect ratios of detections and ground truths are considered, we introduced a distance based on Mahalanobis distance. This distance, that we designate as *Ground Truth Weighted Bounding Box Distance*, is defined as

$$D_{\text{GT weighted}} = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad . \tag{6}$$

In this equation, x and μ represent the data of detections and Ground Truths and are defined as $x = \begin{bmatrix} x_C & y_C & w_C & h_C \end{bmatrix}^T$ and $\mu = \begin{bmatrix} x_G & y_G & w_G & h_G \end{bmatrix}^T$. The central coordinate of detections and ground truths are



(a) (b) (c) (d)

Figure 4. Representation of an example where the detection is improved by using information provided by the tracks. In (a) are illustrated part of the tracks that exist at location 1 of Figure 2; the tracks are represented as green lines connecting the centers of previous detections. In (b) is shown the coverage map that was produced by the neural network at location 2 of Figure 2. The coverage map has higher values in areas where the more likely to have boats present. The map in (b) contains two blobs but the one that corresponds to the life raft, because of its size and the lighting conditions, has smaller values. The map in (c) is generated based on the tracks and is binary, with ones only in the prediction location for each object; this map is obtained in location 3 of Figure 2. Finally in (d) there is the merge of the previous maps, which is created at location 4 of Figure 2.

represented by (x_D, y_D) and (x_G, y_G), respectively. Likewise, (w_D, h_D) and (w_G, h_G) represent the width and height of detections and ground truths. The subtraction of these two variables is therefore used to compare the error in position and size of the detection with respect to the GT. The absolute difference of BBs is multiplied by a matrix Σ defined as

$$\Sigma = \begin{bmatrix} w_G^2 & 0 & 0 & 0 \\ 0 & h_G^2 & 0 & 0 \\ 0 & 0 & w_G^2 & 0 \\ 0 & 0 & 0 & h_G^2 \end{bmatrix}, \tag{7}$$

i.e., only the size of GT is used to weight the difference between detections and Ground Truths.

4.2. Example

In Figure 5, we present the *Ground Truth Weighted Bounding Box Distance* computed in four different cases. These are just toy examples to display properties that are useful in the context of maritime surveillance using airborne images . The black boxes, represent the ground truth and the colored represent the detections. In the red case, the detection box is equal to the ground truth and is shifted 15 units in both axes. In the blue case, the shift is equal but the size of both boxes is the double, which make the cost smaller. The green case, is similar to the first but the shift of 21 units ($\sqrt{(15)^2 + (15)^2}$) was done only in the smaller axis, making the cost bigger than the first case. In the bottom

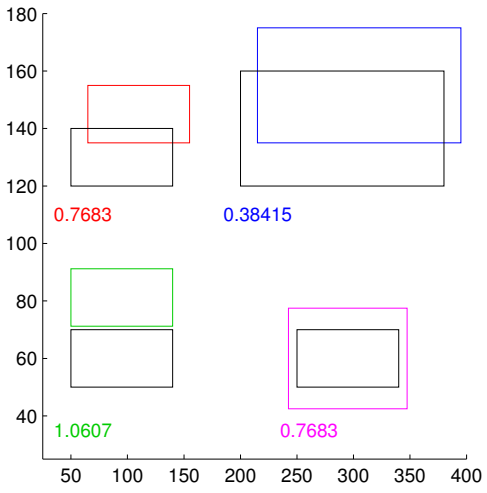


Figure 5. Example of the *Ground Truth Weighted Bounding Box Distance* for 4 different cases.

right example, both boxes are perfectly centered and only differ in size (15 units in both axes). With a size difference similar to the position difference of the first case, the cost is also similar.

5. Evaluation

The evaluation in this work was done with two goals in mind. The first is to provide information about the adequateness of the detection mechanism presented in Section 3. The second goal is to display the need of having a more adequate evaluation metric which is presented in Section 4. In the following subsection, we describe the conditions in which the videos were captured, their main characteristics and the settings that were used for the algorithm. In the latter subsection, we present the evaluation results and discuss the advantage of the proposed metric and the performance of the detector in scenarios very close to real maritime surveillance missions.

5.1. Experimental Setup

The testing of the detector was done using four different video sequences. These sequences were captured near the portuguese coastline during spring and summer, using a small size Unmanned Aerial Vehicle (with a weight of 25 kg). The aircraft flew at a constant speed of $20m/s$ and its altitude varied from 200 to 1500 feet above the ocean surface.

To show the applicability of the detector to the real world, in each video sequence a different scenario was created. In the first (sequence WIDE), a wide area monitoring mission is simulated with the aircraft flying at $1500ft$. with a wide angle camera. In the second scenario, a vessel is followed closely, at an altitude of $200ft$ (seq. NEAR). In the third scenario, we simulate potentially illegal activities, with the vessel deploying a small skiff (seq. SUSP). In the last sequence, we simulate a search and rescue mission, with a vessel deploying a life raft (seq. SAR). [1].

In Table 1, more details about the videos are presented. In particular, we listed the number of frames of each video, number of frames containing only one boat, number of frames containing two or more boats and the average distance between them. Even though the problem considered in this work is detection and we do not care about pointing out the identity of each boat in the sequence, the data about multiple boats is relevant to access the algorithm's robustness when boats are close to each other. The average dimensions of the boat are also presented to provide a hint on the sensitivity to the apparent size of the object.

To achieve a compromise between the duration of distractors and the computational complexity of calculating all the combinations of paths in the graph, the time horizon that was used in Equation 2 was $R = 6$. Based on the distances between boats, their sizes and also the apparent motion caused by the camera motion, for Equation 3 we selected σ_x and $\sigma_y = 60$.

1. The four sequences are available at http://vislab.isr.ist.utl.pt/seagull-dataset/ and are designated, respectively as `bigShipHighAlt_clip2.mp4`, `lanchaArgos_clip3.mp4` , `2015-04-22-16-05-15_jai_eo_.mp4` and `GP030175_part01.mp4`

TABLE 1. MAIN DATA OF THE VIDEO SEQUENCES THAT WERE CONSIDERED FOR THE TESTING OF THE DETECTOR. DISTANCES ARE MEASURED IN PIXEL.

	WIDE	NEAR	SUSP	SAR
# resolution	1920 ×1080	1920 ×1080	1024 ×768	1920 ×1080
# frames	2250	1400	16368	4850
# frames with only one boat	1798	1237	12129	3816
# frames with 2+ boats	123	0	3746	794
average and minimum distance between boats	1747 1717	- -	187 0	211 0
average width and height of boats	18 19	111 50	20 17	35 35

5.2. Results discussion

The evaluation of our detection approach was conducted both with traditional detection metric proposed in [10] and with our proposed metric. The evaluation method used by Dollar *et al.* requires the BBs produced by the algorithm and the GT to be bigger than 50% to consider a detection as correct. On the present scenario, objects have a smaller apparent size and therefore is harder to meet this requirement. We have tried smaller thresholds of 10 and 50%, respectively. As presented in Figure 7 and 8, the differences in results between using an overlap of 10% and 30% makes difficult an interpretation about the performance of the algorithm.

One might argue that in the case of object detection, the overlap is not of paramount importance, although, this could lead to a misleading situation. If evaluation is done using a very small threshold, then a given algorithm producing very adequate BBs would lead to the same evaluation results as another algorithm that produced BBs of worst quality (as long as the overlap threshold was satisfied).

When inspecting Figure 9, it easier to understand what are the tests in which the algorithm performed better. For instance, looking at the results obtained for sequence WIDE detectnet only with detectnet neural network, it is possible to verify that approximately 65% of detections achieve a *Ground Truth Weighted Bounding Box Distance* of 0.1 or better (in this case less distance is better). For the detectnet with MHT applied to the same sequence (WIDE detect w/ MHT) then approximately 70% of detections achieve a *Ground Truth Weighted Bounding Box Distance* of 0.1.

Generally, the proposed method produces good results as visible in Figure 6. The exception to this behavior is when vessels are in areas affected by the sun glare. All sequences contain glare but it is specially severe in WIDE, SUSP and SAR, which causes a low recall in Figure 9. Additionally, sequence SUSP poses another challenge as it contains several boats and when they become adjacent, only one is detected. Nevertheless, an important feature is that the

Figure 6. Example of detections for sequence WIDE, NEAR, SUSP and SAR, respectively.

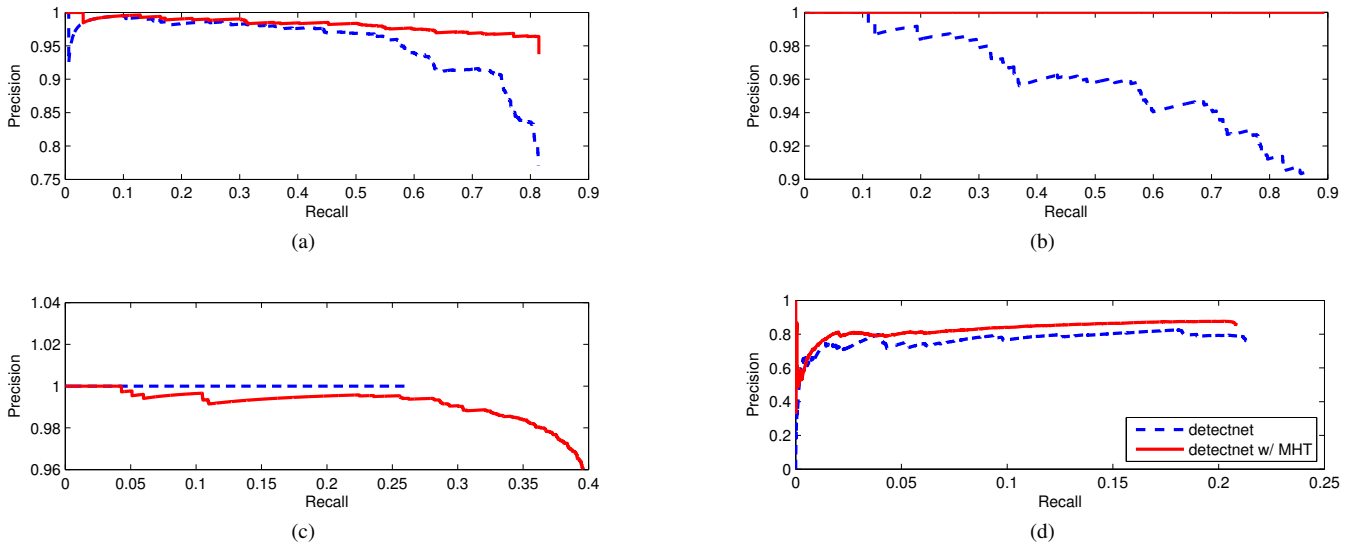


Figure 7. Results of evaluation using a traditional detection metric [10], with a overlap threshold of 10%. Results were obtained for sequence (a) WIDE, (b) NEAR, (c) SUSP and (d) SAR respectively, using the standalone neural network (blue dashed line) and using the network with MHT (red solid line).

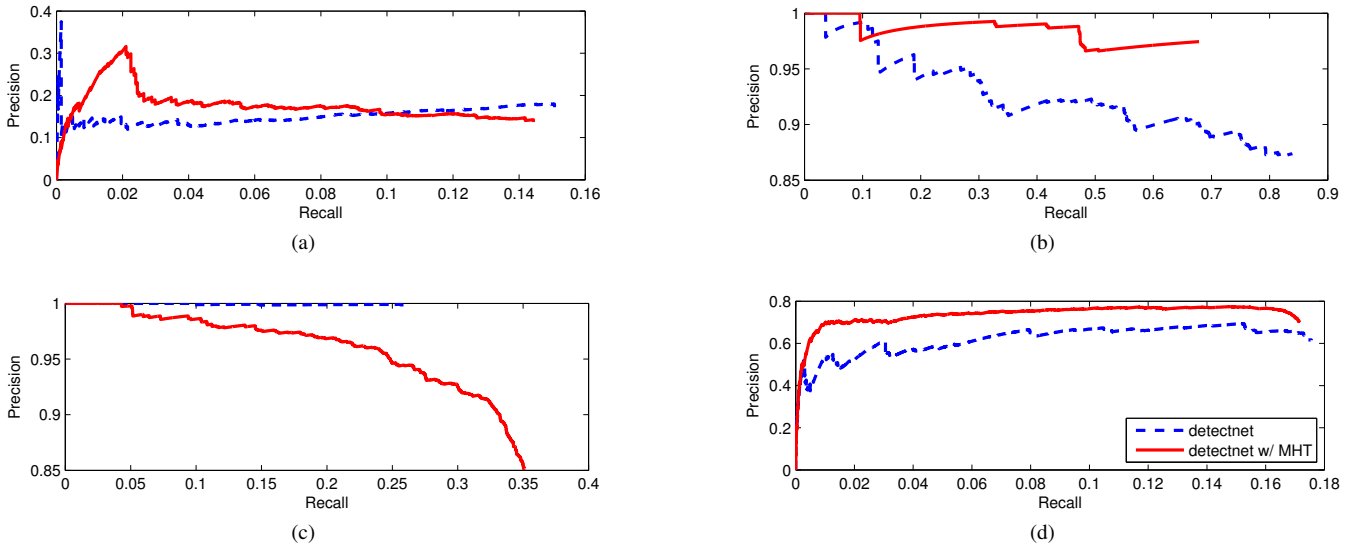


Figure 8. Results of evaluation using a traditional detection metric [10], with a overlap threshold of 30%. Results were obtained for sequence (a) WIDE, (b) NEAR, (c) SUSP and (d) SAR respectively, using the standalone neural network (blue dashed line) and using the network with MHT (red solid line).

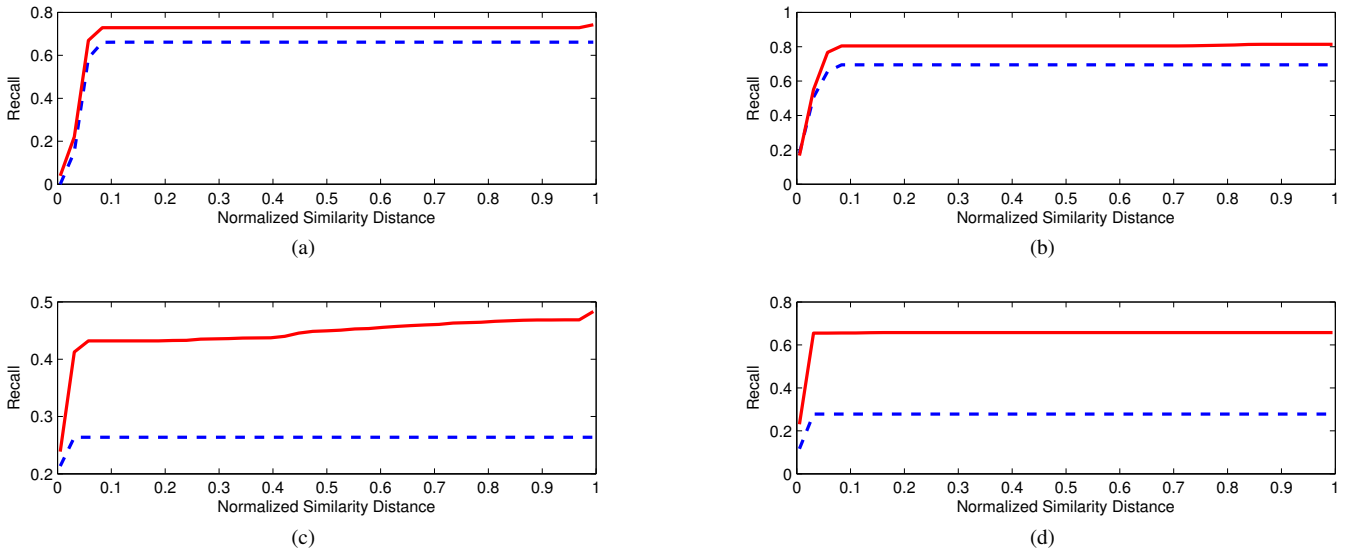


Figure 9. Results of *Ground Truth Weighted Bounding Box Distance* for sequence (a) WIDE, (b) NEAR, (c) SUSP and (d) SAR respectively, using the standalone neural network (blue dashed line) and using the network with MHT (red solid line). The horizontal axis represents the *Ground Truth Weighted Bounding Box Distance* and the vertical axis represents the recall obtained for a given operating point.

MHT improves results in every situation, which validates the proposed detection method.

6. Conclusions

This paper presents a detector for aerial images captured in maritime surveillance missions. This detector is based on a CNN architecture that created BBs with different scales but it is still not very robust due to scale and perspective variations as well as due to distractors, like glare and wave crests. To overcome its limitations, we have used a MHT that tracks all the possible detections for a limited time horizon, computing the probabilities of combining detections (based on the difference of position) in successive frames. The combinations that have low probabilities are then discarded, limiting the number of combinations and making the computation feasible.

We have also evaluated this detector in four very different conditions, where it showed an interesting performance without changing any setting. The worst result was obtained in a sequence were two boats are close to each other and consequently the association probabilities computed by the

MHT are not very discriminative.

In this work, we have also presented an evaluation metric that is not based on a binary decision but based on the difference of position and size of the BBs produced by the algorithm and the ground truths. We evaluated the results with the proposed metric and with one of the most used metrics for detection. On the already existing metric, the difficulty on selecting adequate threshold for the binary decision was apparent. Using the proposed metrics, this difficulty did not exist and better understanding of the behavior of the algorithm is possible.

References

[1] H. O. Pörtner, D. M. Karl, P. W. Boyd, W. W. L. Cheung, S. E. Lluch-Cota, Y. Nojiri, D. N. Schmidt, and P. O. Zavialov, "Ocean systems," *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 411–484, 2014.

[2] I. M. Association *et al.*, "International shipping facts and figures–information resources on trade, safety, security, and the environment," *London: International Maritime Association*, 2011.

[3] D. Hinrichsen, "The coastal population explosion," in *Trends and Future Challenges for US National Ocean and Coastal Policy: Proceedings of a Workshop*, vol. 22. NOAA, January 22, 1999, Washington, DC, 1999, pp. 27–29.

[4] R. O. Nanette Arnesen, "Literature review on vessel detection."

[5] "Piracy and armed robbery against ships," ICC International Maritime Bureaus INTERNATIONAL MARITIME BUREAU, Tech. Rep., 2016.

[6] I. O. f. M. IOM, "Mediterranean update migrant deaths rise to 3,329 in 2015," October 2015. [Online]. Available: https://www.iom.int/news/mediterranean-update-migrant-deaths-rise-3329-2015

[7] H. K. White, P.-Y. Hsing, W. Cho, T. M. Shank, E. E. Cordes, A. M. Quattrini, R. K. Nelson, R. Camilli, A. W. Demopoulos, C. R. German *et al.*, "Impact of the deepwater horizon oil spill on a deep-water coral community in the gulf of mexico," *Proceedings of the National Academy of Sciences*, vol. 109, no. 50, pp. 20 303–20 308, 2012.

[8] A. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 666–673.

[9] P. Liang, H. Ling, E. Blasch, G. Seetharaman, D. Shen, and G. Chen, "Vehicle detection in wide area aerial surveillance using temporal context," in *Information Fusion (FUSION), 2013 16th International Conference on*. IEEE, 2013, pp. 181–188.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 743–761, 2012.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.

[12] O. Oreifej, R. Mehran, and M. Shah, "Human identity recognition in aerial images," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 709–716.

[13] A. Gaszczak, T. P. Breckon, and J. Han, "Real-time people and vehicle detection from uav imagery," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2011, pp. 78 780B–78 780B.

[14] T. Pollard and M. Antone, "Detecting and tracking all moving objects in wide-area aerial video," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 15–22.

[15] M. Dawkins, Z. Sun, A. Basharat, A. Perera, and A. Hoogs, "Tracking nautical objects in real-time via layered saliency detection," in *SPIE Defense+ Security*. International Society for Optics and Photonics, 2014, pp. 908 903–908 903.

[16] F. Maire, L. Mejias, A. Hodgson, and G. Duclos, "Detection of dugongs from unmanned aerial vehicles," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 2750–2756.

[17] P. Westall, P. O'Shea, J. J. Ford, and S. Hrabar, "Improved maritime target tracker using colour fusion," in *High Performance Computing & Simulation, 2009. HPCS'09. International Conference on*. IEEE, 2009, pp. 230–236.

[18] M. V. Shirvaikar and M. M. Trivedi, "A neural network filter to detect small targets in high clutter backgrounds," *Neural Networks, IEEE Transactions on*, vol. 6, no. 1, pp. 252–257, 1995.

[19] C. Clark, "The detection of ship trail clouds by artificial neural network," *International Journal of Remote Sensing*, vol. 20, no. 4, pp. 711–726, 1999.

[20] F. Bousetouane and B. Morris, "Fast cnn surveillance pipeline for fine-grained vessel classification and detection in maritime scenarios," in *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*. IEEE, 2016, pp. 242–248.

[21] F. Maire, L. Mejias, and A. Hodgson, "A convolutional neural network for automatic analysis of aerial imagery," in *Digital lmage Computing: Techniques and Applications (DlCTA), 2014 International Conference on*. IEEE, 2014, pp. 1–8.

[22] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[23] H. L. Van Trees, *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[24] F. Bashir and F. Porikli, "Performance evaluation of object detection and tracking systems," in *Proceedings 9th IEEE International Workshop on PETS*, 2006, pp. 7–14.

[25] V. Manohar, P. Soundararajan, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo, "Performance evaluation of object detection and tracking in video," *Computer Vision–ACCV 2006*, pp. 151–161, 2006.

[26] A. Baumann, M. Boltz, J. Ebling, M. Koenig, H. Loos, M. Merkel, W. Niem, J. Warzelhan, and J. Yu, "A review and comparison of measures for automatic video surveillance systems," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–30, 2008.

[27] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[28] J. Barker and S. Prasanna. (2016, August) Deep learning for object detection with digits. NVIDIA. [Online]. Available: https://devblogs.nvidia.com/parallelforall/detectnet-deep-neural-network-object-detection-digits/