Benchmarking operational SAR ship detection

Harm Greidanus*, Peter Clayton1, Marte Indregard2, Gordon Staples3, Naoki Suzuki4, Paris Vachon5, Chris Wackerman6, Tove Tennvassas7, Jordi Mallorquí8, Naouma Kourti*, Robert Ringrose1, Harm Melief9

*) European Commission Joint Research Centre (JRC)
TP 267, I-21020 Ispra (VA), Italy
Tel +39-0332-789739, Fax +39-0332-789658, Email harm.greidanus@jrc.it


Abstract – The performances of eight satellite SAR ship detection systems (most of them operational) are compared, by running a benchmark test on RADARSAT images of various modes. No single system is best under all circumstances.

I. INTRODUCTION

One of the “classical” uses of radar is to find ships; to this purpose, radars are placed on ships and on surveillance aircraft, and indeed also satellite SAR is used to this end. Research and development into vessel detection with satellite SAR has been spurred, first in a limited way by SEASAT, then with more strength by ERS-1, JERS-1 and RADARSAT-1. Several commercial providers are now offering ship detection as a value-added service. The user community potentially encompasses authorities involved in control of maritime traffic, pollution, fisheries, smuggling, as well as the defense sector. Operational use of satellite SAR by these communities is, however, still quite limited. One of the reasons has been the time lag between data acquisition and availability to end user; however, since very recently, faster and cheaper computing and communications have made it possible to get detected vessel positions to the end user within 30 minutes after satellite overpass. Another reason may be the immaturity of automatic detection algorithms. Although operational, the algorithms used are not perfect and even not always fully automatic.

This paper is about comparing the performance of (near)-operational satellite SAR ship detection systems and establishing the state-of-the-art. For the end-users, it is important to know what quality they may expect from current satellite ship detection services. For the providers, it is important to know how they can improve their products. The relevant questions are therefore:
• What is the performance of present-day ship detection systems?
• Is there much variation between the various providers?
• Are some approaches more successful than others, and does this depend on specific circumstances?
• Can we expect scope for improvement, and how may this be realized?

II. METHOD

These questions are addressed in a research project named DECLIMS (“Detection and Classification of Marine Traffic from Space”), which in addition to the companies listed on top of this paper that work on SAR, also includes partners active in the optical domain.

Under DECLIMS, a benchmark test was carried out, which encompassed all partners running their respective vessel detection systems on a standard set of satellite SAR imagery.

The ship detection process consists of many steps, notably geo-location, background estimation, application of CFAR thresholds, clustering of pixels into targets, removal of false alarms, and estimation of ship parameters. For each step in the analysis, different approaches are being used by different providers, e.g. related to the localness of the background estimation, the noise and clutter models used, assumptions on the shape of the targets, etc. A cursory overview of the approaches used under DECLIMS is given in Table 1. It should be noted that the conciseness of this table prevents it from doing justice to the details and options of, and the full differences between the systems.

The test set consists of 17 RADARSAT images, in various modes (S5, S7, W2, W3, SNA, SNB). Most of the images have ground truth associated in the form of ship positions supplied by VMS, the Vessel Monitoring System which applies to fishing ships [1]. Also a few ship positions from coastal radar are available.

Each vessel detection system provides a list of targets for each of the images it analyzes. For each target the location is specified, as pixel position in the image and in geographical coordinates (cf. Fig. 1). Some systems also give target size, and most give a reliability figure, which was standardized for the present exercise into 4 levels: false alarm, probable false alarm, probable target, certain target.

The detection results are compared by assuming that detections from different systems closer than 200 m (based on pixel coordinates) correspond to the same target. Each target defined in this way, i.e. found by at least one of the systems, is also visually inspected in the original image and assigned a visual reliability. Also, each such target is correlated with available VMS positions to assign a ground truth. The correlation between VMS position and detected target position takes into account distance between these two positions, the time difference between image acquisition and VMS position (which are not exactly simultaneous with image acquisition,
with time differences between 0 and 60 minutes), and an assumed maximum vessel speed of 10 knots. Sometimes two VMS positions are available, before and after image acquisition; in this case the position is linearly interpolated. The association between detected target and ground truth is not always unique, particularly when vessels are clustered and VMS is not timely. Also here a 4-level reliability is assigned, based on the time difference – highest reliability for timely VMS signals (< 15 min difference).

The association between detected target and ground truth is quantified by their assigned reliabilities, the detection statistics (i.e., the visual interpretations and the ground truth) and with respect to VMS ground truth (P_D/V, P_F/V). P_D/T counts the fraction of VMS positions detected by the system in question; P_D/V counts the fraction of visual targets detected, where it should be remembered that the set of visual targets is the collection of all targets found by one or more of the systems, after removal of some that visual inspection deemed to be false alarms. Note that these numbers only have a limited significance in the absolute sense. P_D/V and P_F/V are not ‘hard’ detection and false alarm rates because we do not know the quality of visual interpretation (in fact we can be sure that it is not perfect). P_D/T may appear high in case the image contains many more vessels than which report a VMS position. The most meaningful statistic in the absolute sense is P_D/T – this objectively counts the number of known ships that are detected. However, even that one can be contaminated in case a different vessel is detected instead of the VMS vessel close to its position or (less likely) in case the VMS position is wrong.

Considering the uncertainties in the references for the detection statistics (i.e., the visual interpretations and the ground truths) and the uncertainties in the system detections, all quantified by their assigned reliabilities, the detection statistics are compiled in a ‘fuzzy’ way, by assigning penalties in case system-provided reliability and reference reliability of the target differ by two levels or more. E.g., if a system does not detect a target which has visual classification ‘certain’, its score is lowered by 1; if a system detects such a target but classifies it as ‘probable false alarm’, its score is lowered by 0.75; if a system detects this target as ‘probable target’, no penalty is assigned. (The choice of 0.75 for the magnitude of the lower penalty is somewhat arbitrary.)

III. RESULTS–OVERALL

In this paper, the results are anonymized, because of commercial implications in combination with too small a test sample which may lead to unfairness at the level of individual systems. The images in the test set differ in many respects: image mode, sea state, region, size of vessels, density of vessels, processor; this makes it impossible to attribute a certain detection performance to a certain cause. However, it is the variations between the different systems which are of relevance here, and the test sample is deemed to be large enough to point to general trends and conclusions in that respect.

First we look at detection rates of the systems w.r.t. visual interpretation. There are ‘easy’ images which score high detection rates (>97 %) from all systems. Most common are images with detection rates in the range 85 – 95 %. Then there are some more ‘difficult’ images on which most systems score 75 – 90 % or only even 70 – 80 %. On these images, some systems even score considerably lower, 60 % or down to 40 %; however, it is always different systems which have the low scores. So, although maybe some systems score over the whole somewhat higher than others, none of the tested systems is always (or even very often) better than all the others – w.r.t. visual interpretation. In addition to and not counted with the above are images which have a non-standard lookup table (logarithmically scaled); these lead to very low scores with some systems, which apparently haven’t taken this scaling properly into account. The quoted rates are based on a maximum of 963 targets summed over all of the images – maximum because some systems only analyzed a subset of the images.

In comparing these detection rates, it should be remembered that each detector must make a trade-off between detection and false alarm rates. One expects a higher detection rate to go hand in hand with more false alarms. This kind of correlation is in fact found only for 3 images. In 6 images the different systems have approximately the same number of false alarms (while detection rates may vary), and in 5 images the correlation between detection and false alarm rate is actually inverted. Therefore, the above conclusions are vindicated; some images are more ‘easy’ for some systems, whereas other images are easier for others.

Upon inspection of individual detections, it can be seen that, unsurprisingly, the strongest targets are found by all systems; however, for the weaker targets, often some of them are found by some systems and others by other systems, i.e.,

![Image](Fig. 1. RADARSAT W2 image with detections as found by different systems, each system with its own symbol (yellow triangle, red circle, green diamond, blue cross, blue star; black square is ground truth position; W Coast Canada).)
there is no series of ever weaker targets which are found by progressively fewer systems.

The above discussion was w.r.t. visual interpretation. Now turning to a comparison with known VMS (and coastal radar) positions, 8 of the images have detection rates in the 80 – 100 % bracket, and 5 in the range 40 – 75 %. Looking at variation between the systems, there are 9 images where variation is within 10 percentage points; 2 images with variations of 15 percentage points; and 2 images where the systems differ by as much as 30 percentage points in the detection rates. Concerning the significance of these numbers, there are in total within all images 99 VMS positions of high reliability, 31 of medium and 31 of lower reliability (totaling 137.8 positions after applying penalty weighting). The four images that show the large variations in $P_{DT}$ contain in total 28.1 weighted VMS positions.

From the above we can conclude that also w.r.t. an objective measure (the VMS reports), detection performance is not only imperfect (not a surprise), but often significantly varying between different detectors. Like in the case w.r.t. visual, there is not one system which is best or worst for all of the images.

IV. RESULTS – DETAIL

Trying to establish causes of underperformance, we find specific causes, applicable to a certain system, and general causes, applicable to all systems.

In several cases it is easy to recognize the specific performance as a consequence of the algorithmic approach (cf. Table 1). E.g., the use of a template leads to some missed targets which have a shape not conforming to the template; the use of a very local background estimation leads to increased false alarms caused by bright spots in local depressions of the backscatter; etc. However, these drawbacks in many cases are offset with specific advantages associated with these algorithms – if not, the present study indicates areas of improvement for these systems. So, here we find at least a partial explanation to the conclusions of the previous paragraph that some systems perform better on some images, while other systems on others.

The benchmark study also very clearly shows some problems common to all systems. Of course, there are fundamental limitations on the detection rate – false alarm rate dependency which are inherent to the radar imaging and can never be overcome. But in addition, the common problems are the following (more or less in order of severity):

Special clutter. The thresholding is designed to deal with clutter, and e.g. a thresholding based on the K distribution is quite general as it can handle a slightly rough sea surface as well as one with resolved swell. However, some ocean surface features and textured clutter still pose difficulties. Some systems have limited local texture analysis built in and are able to discard, or at least to indicate as unreliable, some alarms related to sea surface features. Nevertheless, most systems give false alarms on some types of ocean surface features that a human operator can recognize as such. It would seem that there is still room for improvement in this respect.

On the other hand, there are also types of surface clutter that even baffle the human analyst. Fig. 2 shows an example, tentatively interpreted as caused by atmospheric convection.

Land masking. In order not to mistake land for a target, the land is masked out. There are two issues with this. First, the geo-positioning of the satellite image has limited accuracy, in particular in near real time (most relevant for the application). This can lead to image shifts in azimuth direction, and a consequent displacement of the land mask and associated false alarms near the coast. Secondly, available land masks are of limited accuracy; there are problems with small islands, intertidal areas and coastal man-made constructions, also leading to false alarms (Fig. 3 left). Some systems cope with this by excluding a safety zone near the coast, but this then clearly leads to potential missed detections.

Automation. Ideally, a detector should be able to run fully automatically. In two respects, this is still often not possible. First, most systems have the experience that CFAR thresholds need to be adjusted depending on region and situation for optimal results. This adjustment needs human intervention. Secondly, some systems still feel safer when an operator inspects ‘difficult’ targets after automatic detection, to weed out false alarms.

Image artifacts and interference – leading to false alarms. The most common image artifacts are side lobes and ambiguities, caused by very strong targets. In most cases side lobes are not a big problem as they are agglomerated with the central peak anyway based on a distance criterion. Only in fine beam images the side lobes can extend so far as to cause unrelated alarms. Fewer systems take into account ambiguities, which can produce a ghost image of a strong target some specific distance away in (mainly) azimuth. As side lobe and ambiguity behavior are determinate, correction for these effects is in principle possible. However, to correct for ambiguities caused by targets on land, or even outside of the image, is somewhat more problematic.

Compound targets. The signature of a ship in a radar image may be composed of several peaks due to separated scattering centers, it may be elongated due to azimuth
smearing or extended due to a disturbance of the water (a short wake – long wakes were not present in the test set). In this way, it may mimic the signature of two vessels in close proximity. The different systems deal differently with these situations; they may declare different targets, or a single one (Fig. 3, right). Which corresponds better to the situation is difficult to say, in many cases also after visual inspection.

Fig. 3. Left: Image chip from RADARSAT F of 4 x 4 km showing displacement of the coast line due to geo-positioning error, as well as coastal structures not present in the land mask leading to false alarms (Tarifa, Spain). Right: Image chip from RADARSAT W3, 500 x 500 m around a target. Target positions as found by 5 systems are indicated by the various symbols. One system (diamonds) declares 3 targets, the other systems just a single one.

V. CONCLUSIONS

The main conclusions of the SAR ship detection benchmark, extending over 8 different ship detection systems, can be summarized as follows. Apart from the fundamental limitations on detection vs. false alarm rates, all or most systems share the common problems of:

1. False alarms from certain types of ocean surface features;
2. Imperfect land masking due to inaccuracies in coast line data bases and in near real time image geolocation;
3. For optimal results, detection thresholds should be adapted to the region / situation;
4. False alarms from image artifacts, interference and ambiguities.

Different systems show different performance, but not in a consistent way – for some images / situations some systems perform better, while for other images other systems give better results. This can be attributed to the different algorithmic approaches used in the systems.

The present study shows room for improvements in the existing ship detectors, and indeed indicates exactly, through detailed inspection of the targets missed or spuriously declared by the systems, where problems occur for each of the detection systems. Such a detailed analysis is beyond the scope of this paper but can be undertaken by the system developers for themselves.

REFERENCES

[1] G. Lemoine et al., “Near real time vessel detection using spaceborne SAR imagery in support of fisheries monitoring and control operations”, these proceedings

ACKNOWLEDGMENT – The national Fishery Management Centres that provided the VMS positions are gratefully acknowledged.

<table>
<thead>
<tr>
<th>System</th>
<th>Company¹</th>
<th>Background estimation²</th>
<th>Noise model³</th>
<th>Target model⁴</th>
<th>Target info⁵</th>
<th>Domain⁶</th>
<th>Status⁷</th>
<th># Im⁸</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUMO</td>
<td>JRC⁹⁰</td>
<td>Tiles</td>
<td>θ = κ (µ+σ)</td>
<td>Template</td>
<td>R</td>
<td>Ampl</td>
<td>Op</td>
<td>17</td>
</tr>
<tr>
<td>MaST</td>
<td>QinetIQ⁰⁰</td>
<td>Stencil</td>
<td>θ = µ + κ σ</td>
<td>Box</td>
<td>R L W</td>
<td>Power</td>
<td>Op</td>
<td>10</td>
</tr>
<tr>
<td>OMW</td>
<td>CCRS⁸⁰ / SATLANTIC⁹⁰</td>
<td>Tiles</td>
<td>K distribution</td>
<td>Pixel</td>
<td>R L W H</td>
<td>Power</td>
<td>Op</td>
<td>7</td>
</tr>
<tr>
<td>MDA/RSI</td>
<td>MDA/RSI⁰⁰⁰</td>
<td>Tiles excl. high values</td>
<td>K distribution</td>
<td>Pixel</td>
<td>Power</td>
<td>Op</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>SUMO/K</td>
<td>JRC⁹⁰ / TNO⁰⁰</td>
<td>Tiles excl. high values</td>
<td>K distribution</td>
<td>Pixel</td>
<td>R L W H</td>
<td>Ampl</td>
<td>Dev</td>
<td>7</td>
</tr>
<tr>
<td>FF/VSPT</td>
<td>KSAT⁹⁰ / KSPT⁰⁰</td>
<td>Stencil</td>
<td>θ = µ + κ σ</td>
<td>Box</td>
<td>L H</td>
<td>Power</td>
<td>Op</td>
<td>17</td>
</tr>
<tr>
<td>GD</td>
<td>GD⁰⁰</td>
<td>Stencil</td>
<td>θ = µ + κ σ</td>
<td>Box</td>
<td>R L</td>
<td>Power</td>
<td>Op</td>
<td>17</td>
</tr>
<tr>
<td>Visual</td>
<td>MHI⁰⁰</td>
<td>Tiles</td>
<td>K distribution</td>
<td>Geometrical feature</td>
<td>L W</td>
<td>Power</td>
<td>Res</td>
<td>6</td>
</tr>
<tr>
<td>Split-look</td>
<td>MHI³⁰</td>
<td>Small tiles</td>
<td>Sub-apertures uncorrelated</td>
<td>Sub-apertures correlated</td>
<td>SLC</td>
<td>Res</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Wavelets</td>
<td>UPC⁰⁰</td>
<td>Small tiles</td>
<td>Wavelet sub-bands spatially uncorrelated</td>
<td>… correlated</td>
<td>R L W H</td>
<td>Power</td>
<td>Res</td>
<td>-</td>
</tr>
</tbody>
</table>

Table #9T1. Satellite SAR ship detection systems considered in DECLIMS. (1): developer, user (i.e., one who runs the system). (2): Background is estimated on large Tiles, on Small tiles, or on a hollow moving Stencil. (3): Detection threshold θ, local mean µ , local st. dev. σ , constant factor κ. (4): Template: fitting a prescribed shape for the target; Box: the average within a small box (e.g. 2x2 pixels) is compared to the background; Pixel: single pixel values are compared to the background. (5): Reliability, Length, Width, Heading. (6): Detection works on Amplitude, Power or Single Look Complex. (7): Operational, under Development, Research. (8): Number of images analyzed in this benchmark.