Enhancing long-range Automatic Target Recognition using spatial context

Iain Rodger
Thales
Glasgow, UK
Iain.Rodger2@uk.thalesgroup.com

Rachael Abbott
Queen's University Belfast
Belfast, UK
rabbott02@qub.ac.uk

Barry Connor
Thales
Glasgow, UK
Barry.Connor@uk.thalesgroup.com

Neil Robertson
Queen's University Belfast
Belfast, UK
N.Robertson@qub.ac.uk

***Abstract*—This paper presents a high-performing automatic target recognition system which can be used for long-range surveillance scenarios. The main novelty of our system is that it uses contextual information from RGB images to help classify targets in long range real world LWIR images. This contextual framework provides additional information of an object's surrounding environment, leading to a significant increase in long-range target recognition accuracy. This work will be of interest to the defence community as a high-performing automatic recognition system is a highly sought-after capability.**

I. Introduction

Motivation: It is becoming increasingly familiar to undertake surveillance operations using multiple sensors [1], [2], where each additional sensor band observes a different component of the Electro-magnetic (EM) spectrum. Utilising additional spectral bands provides increased knowledge of surrounding environments. In doing so we can leverage the captured information to enhance overall situational awareness, which is our primary goal. However, in the defence surveillance setting, multiple sensors must be effectively managed by a human operator. Thus, incorporating more sensors in a surveillance platform increases the load on the human user. This could potentially lead to significant negative repercussions in dangerous defence domains, where overburdened operators are more likely to miss mission-critical events [3].

Bearing our stated goal in mind, where we aim to increase overall situational awareness of a target scene, the issue of increasing processing load with additional sensor modalities must be addressed. Towards this, we develop and extend an existing Automatic Target Detection and Recognition (ATDR) method, where it could remove some of the scene processing and comprehension tasks from a human operator. In other words, the system would be capable of intelligent and automatic signal processing. Such a system could optimise the information presented to a user but with enhanced scene perception, ultimately providing an assisted decision framework via ATDR methods. There are many examples to be found in the prior art emphasising system development for the capabilities described [1], [4], [5]. Our paper improves upon prior work and advances the field. We achieve this by enhancing a convolutional neural network [3], using context.

Related work: Context is understood to be any extra information that potentially improves overall understanding of a scene [6]. Humans possess this underlying comprehension of real-world events due to experience, which assists the human vision system on different tasks such as salient region detection [7] or person detection [8]. Machines, on the other hand, do not possess this prior knowledge and would benefit if given this further understanding to give machines a human-like capability. Whether this additional context comes from new sensor information altogether, such as inputting GPS location data as geographic context, or is produced from manipulating original pixel level data into a different representation, the end goal is always to achieve a higher level scene understanding.

Several attempts to add a semantic context into computational models for a visual task are observed in the literature. Oliva et al. [9] propose a method that takes the original low-level pixel data and creates a new, global representation of a scene that can be interpreted at a glance without the need for segmentation or region processing. This representation is named the spatial envelope and may also be thought of as like scene gist, which is also an abstract scene representation proposed by [10]. The model uses the spatial and spectral information to show accurate information about object shape or identity is not an absolute requirement to categorise a scene overall. This idea is used in a follow-up piece of work by Torralba, where a similar scheme for incorporating

978-1-5386-1663-5/17/$31.00 ©2017 IEEE

contextual information in object representations is used for object detection purposes. Low-level pixel data along with object-centric data, such as size and location, are modelled using statistics for selecting task driven regions (the focus of attention) in an image, as well as automatically inferring image scales [11].

Context can also offer improvements to tracking tasks. Recent work by [12] creates an adaptive tracking scheme via learned models for background context. It adapts the object descriptors as and when the scene background undergoes significant changes from, for example, dynamic illumination. The method uses particle filters and certain contrasting colour based components. The result is a tracker capable of dealing with occlusions and re-identifying objects after full occlusions. One of the drawbacks of this method is that a user interaction stage is required for every new video sequence during the learning phase. This suggests it is not deployable for real-world tasks that desire hard, real-time processing. Another example of a context based tracker can be seen in the work of [13]. This method learns scene context with a Bayesian probabilistic approach and handles the occlusion scenario in a rather novel way by modelling target births after occlusion events and the spatial layout of clutter. The result is a multi-object tracker improved by using scene context.

In addition, context is beneficial in classification tasks. A recent paper that exploits the use of context is [14], where they propose a Markov Random Field model called segDeepM. It allows each candidate box to choose a segment and scores how well they match up. This approach is notably more accurate than the context-free RCNN (Regions with Convolutional Neural Networks) baseline, making segDeepM at the top of the current PASCALs leaderboard. A paper by [15] uses context to find kerb ramps which are missing in images of street scenes. The model works as follows, it takes input images, masks the object (the kerb), to focus on learning only contextual information. Then it uses this contextual information to scan the images and generate a heat map showing where kerb ramps are likely to be in the image. The CNN can then be used to detect if kerb ramps are present. This contextual information helps the CNN localise kerbs and increases the likelihood of detection and recognition. In the work of [16], a CNN is trained to create contents of an image which are missing using its surrounding information. The method produces suitable results and shows another way context is vital to produce successful outcomes.

From the wealth of evidence supporting context in detection and recognition systems we will be using it to enhance an ATDR system developed in [3]. Our approach provides semantic context to deep learning classification approaches by combining CNN with learned probabilistic object location maps [17]. We use colour band imagery to generate the spatial context for the scene. A flow diagram of the set-up is shown in Figure 1.

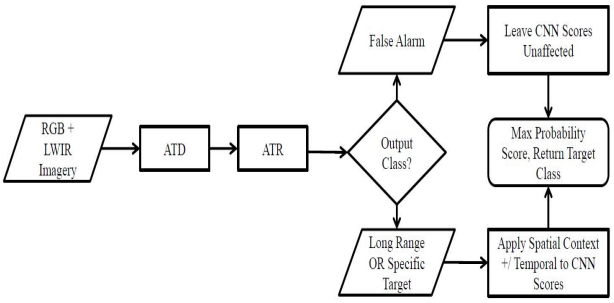

Fig. 1. Overview of the ATDR algorithm. When given multi-modal input data, candidate detections can be generated via an ATD process. These candidates are fed to the trained CNN, where the output class and score vector decides the next step. If the maximum class score is a false alarm, do nothing. If the target is a long-range class, remove FA and long range scores from CNN vector. Re-weight using spatial context. If real object class returned, re-weight CNN scores using spatial context.

Our contributions are:

1) Mobilisation of scene context via a Bayesian framework incorporating CNN output.
2) A semantic scene segmentation process to generate region and object priors, utilising multi-modal sensor information.

II. Classifying Objects in LWIR Imagery via CNNs

In work [3], a CNN is trained to classify objects in long wavelength infrared (LWIR) imagery. Please note the following definitions: short wavelength infrared includes wavelengths from $1.4-3\mu m$, mid-wavelength infrared is from $3-8\mu m$ and long wavelength infrared is from $8-15\mu m$. A long wavelength infrared camera, Catherine MP, is used as it is ideal for working with the objects for detection in this research; person, land-vehicle, helicopter, aeroplane, unmanned aerial vehicle (UAV), false alarm and long-range target, i.e. objects near room temperature. They classify objects at short, medium and long-range which have the following distances respectively: $0-200m$, $200-750m$ and $>750m$. The longest targets were at $2.5-3km$. This CNN achieves a test set accuracy of 95.7%. The system was evaluated over short to long-range surveillance sequences, which had been manually ground-truthed. The long-range surveillance video achieved an accuracy of 39.5%. The sequence of events in [3] is depicted in the flow diagram illustrated in Figure 2. Given LWIR input data, candidate detections can be generated via an ATD process. In other words, a bounding box is made around the suspected objects

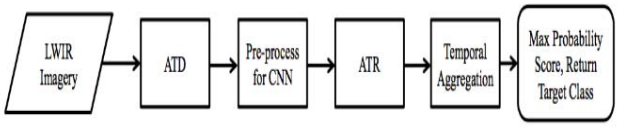

Fig. 2. General overview of the ATDR algorithm stages in [3].

in an image. These candidates are fed to the trained CNN (ATR), where the output score vector is temporally aggregated (explained in Methodology section) and re-weighted. The maximum probability returns the output target class.

When analysing the ground truth tables of this work, there is a high proportion of the results in the "long-range class" instead of the more accurate label "land vehicle class". This means that if a sufficient mechanism existed to switch the long-range class to the vehicle class, the overall accuracy would increase. In our work, we make use of scene context from RGB images to guide the CNN results and thus improve the accuracy of the system.

III. Methodology

As stated before, our aim is to develop an ATDR system which uses RGB contextual information to help classify long-range targets in LWIR imagery. We use a Thales thermal imager named Catherine MP which is sensitive to radiation at wavelengths $8\mu m$-$12\mu m$. The design of the systems has three key stages:

- The generation of targets via an ATD algorithm
- The passing of the detections into a trained classifier.
- The input of the probability scores from each detection into the contextual framework to affect the final classification.

A. ATD algorithm

We use a proprietary Thales algorithm for the detection of objects in the LWIR imagery. This algorithm can localise short-long range targets using hot spot detection. This choice of algorithm is not important to this work, as our main objective is to enhance object recognition.

B. ATR classifier

We use a trained CNN developed in [3] to classify the targets. The network is a scaled down version of Krizhevsky architecture [18], it preserves the overall depth and sequential structure, but removes the width as training used a much smaller dataset.

C. Incorporating spatial context

The following four components are needed to generate spatial context information:

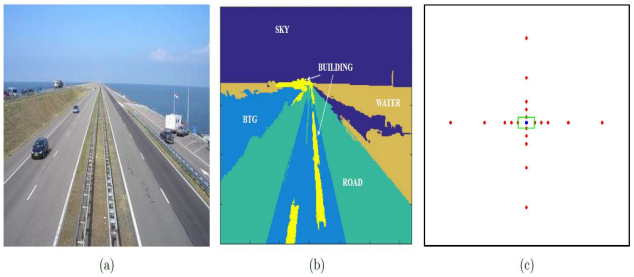

Fig. 3. A highway scene is shown in image (a) and is segmented and labelled with region class information in (b). Image (c) shows the sampling feature that is described in section...

(1) An algorithm for semantic segmentation
(2) A spatial sampling feature
(3) Prior scene knowledge for each object class
(4) The probability for each object to exist given its surrounding context.

1) Semantic segmentation: We use efficient graph-based segmentation to perform region segmentation. Colour intensity, textural information and spatial context are extracted from the different regions after segmentation. This information is then passed through the SVM to be assigned the correct label. In the following text, we will explain how this information is obtained.

Colour Features: The average and standard deviations of the HSV colour planes are used for the colour features for each region.

Texture features: Texture information is gained by using Gabor filters at 1 scale and 8 orientations and applying them to greyscale image regions. This emphasises edges and helps find region boundaries.

Spatial Context: Vertical position information is taken from the segmented image regions. Assuming the (x, y) image coordinate system is used, the vertical position feature is simply the average value of y-pixels per region.

The Stanford dataset [19] is used to compute the feature vectors which are used to train our SVM. The SVM can label five regions: sky, bush/tree/grass(BTG), road, water and building. An example of labelled segmentation is illustrated in Figure 3.

2) Spatial sampling feature: A sampling function is required to deduce what regions surround the candidate target. We use a spatial context feature developed in [17]. When a detection is made, we sample above, below, to the left and right of the centre of the bounding box at pixel locations, [1, 20, 40, 100, 200]. If a target is at the edge of an image then this sampling regime will fail. To rectify this, we pad the image by cloning the boundary pixels for a fixed length in each direction. This sampling feature can be observed in part c) of Figure 3.

3) Prior scene knowledge: Using 20 images for each class from ImageNet, a sampling scheme is used to learn the

prior probabilities $P_o(R_c|l_k)$. R_c is the expected region R for class c and l_k is the sample l at location k.

4) Probability given context: We use this prior scene knowledge to predict the object class given its context using the following equation:

$$P(O|C) = \frac{1}{n}\sum P_o(R_c|l_k) \tag{1}$$

The prior probabilities of each of the 20 locations are added up and divided by n, the total number of locations. For example, if we have a candidate detection with all location regions labelled sky, the equation for $P(aeroplane|context)$ would be:

$$P(aeroplane|context) = \frac{1}{20}(P(sky|location(1)) + ... + P(sky|location(20)) \tag{2}$$

The same process is completed for the other class objects and their values, $P(areoplane|context)$, $P(person|context)$, $P(helicopter|context)$, $P(UAV|context)$ and $P(landvehicle|context)$, make up the five element array $P(O|C)$. Thus, the five element array $(P(O|C)$ is populated by the probability for each object to exist given the context. It is this array we will eventually use to affect the CNN output scores.

D. Temporal aggregation

Even though we have built a robust system, classification errors still exist. A solution is to use temporal aggregation. This method tracks the objects over a period of time, removing the errors in the probabilities. As we are classifying long-range targets, we can achieve temporal aggregation without tracking as long-range targets move very little on the plane. To implement this, we create a circular buffer of detections and the corresponding CNN scores of length N. When a new detection is made, we find the closest previous detection location and find the Euclidean distance E_d between them. If this distance is below the threshold distance of $Thresh_{E_D}$, we aggregate the current CNN scores with the matched CNN scores. This process then moves onto the next detection and CNN score, moving the currently aggregated CNN scores into the buffer and dropping the last entry in the buffer, ensuring there are only N entries to match for any new detection. If the detection has no match below $Thresh_{E_D}$, we do nothing. By propagating through the detection sequence as so, the CNN scores are temporally aggregated and the errors are removed.

E. Overall framework

The overall framework for our work is shown in Figure 1. RGB and LWIR imagery is passed into an Automatic target detector (ATD) system and then into the trained CNN (labelled as ATR in Figure 1) created in [3]. The CNN will produce an output class for each target detected in the images. If this output class is a False alarm, the CNN result remains unaffected. However, if the output class is a long-range or specific target, spatial context is used to guide the system to the correct class. This stage is shown in Figure 1 as the block labelled "Apply Spatial Context+/Temporal to CNN Scores".

IV. Experiments

The final ATDR system is assessed using challenging, long-range multi-modal data sequences. The data is collected using a Thales Catherine LWIR thermal imager in a rural location on targets including; land-vehicle, helicopter and false alarms. All detections generated via the ATD algorithm are human ground-truthed to provide target classes. We evaluate over the same datasets as described in [3], in order for comparisons to be made.

Although the CNN is trained over 7 classes, the output probabilities after propagating through the contextual frame-work are a 5 element array. This is because the FA and long-range class are removed as described in Figure 1. We evaluate the CNN+Context and CNN+Context+Temporal Aggregation on a total of 8750 long range target candidates.

V. Discussion of results

Figure 4 shows the results. The performance using context is significantly better. Although, the images have only; helicopter, land-vehicles and false alarms present, the CNN we use from [3] was trained to identify 7 classes, so our confusion matrices must show all seven classes to highlight system error occurrence.

From Figure 4, we note the CNN and the CNN + Temporal aggregation cannot distinguish between the correct land vehicle classes due to the low signal over long ranges. We can see by mobilising contextual information to affect the CNN output scores, we can significantly improve the accuracy of our results. We achieve classification scores of around 92%, with the temporal function having slightly worse but negligible performance. The improvement from 39.5% to 92% in this system mainly comes from the switch from long range target to land vehicle, with slight improvements in the helicopter class as well. The results are used to obtain an $F_1 - Score$ (visualised in Figure 5) using the following equations:

$$F_1 = 2.\frac{precision.recall}{precision + recall} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

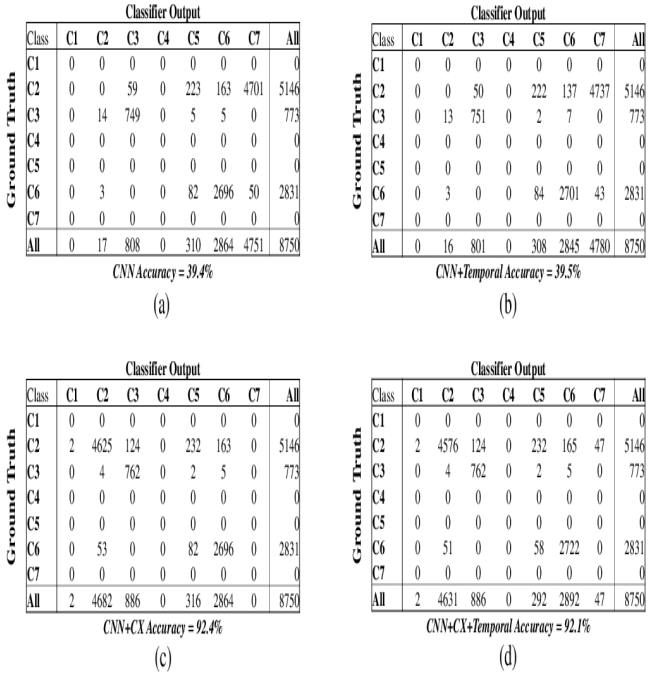

Fig. 4. Confusion matrices and overall accuracy results are presented for long range classification experiments. C1=person, C2=land vehicle, C3=helicopter, C4=aeroplane, C5=UAV, C6=false alarm and C7=long range target class. Matrix (a) is simply the trained CNN applied to ATD output target candidates, which does not perform well. This result is almost identical when the temporal aggregation is introduced in matrix (b), with only a negligible gain on offer. However, context has an overwhelmingly positive effect on the ATDR results as shown in matrices (c) and (d).

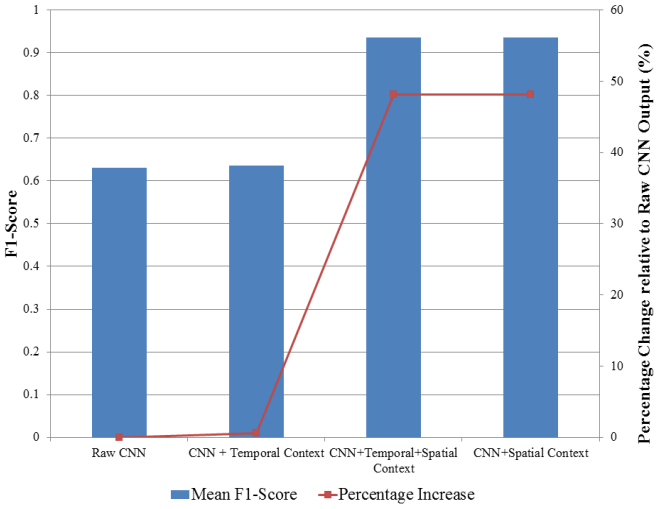

Fig. 5. The multi-axis plot shows mean F1 Scores for the different variants of classification algorithm in our final experiment. The F1 Score is a useful summary statistic in machine learning as it provides a weighted average of a classifier's precision and recall across classes. As we can see, there is a marked improvement gained from spatial context incorporation. This is highlighted by the red line showing the percentage increase in F1 Score relative to the raw CNN output

VI. Conclusion

We have successfully created an ATDR system for enhancing target recognition in long-range surveillance scenarios using multi-modal data. We have achieved this by using state of the art machine learning techniques in the form of a highly accurate CNN LWIR classifier. When used alone, this CNN is inadequate for challenging long-range scenarios. But our system improves it by adding context to infer accurate object classes.

Improvements could be made in the following aspects:

1) Inability to change incorrect false alarm cases, as the CNN output for this class is not passed through the contextual framework, as shown in Figure 1.
2) The system cannot change incorrect object class to false alarm. If the CNN outputs helicopter/land-vehicle/ aeroplane/person/UAV class then due to the structure of the system as described in the previous section, the false alarm class is removed from the list of options the contextual framework can choose from.

Overall, by incorporating a subtle amount of information from semantic segmentation and object priors as context, we have successfully classified challenging long-range targets in multi-modal surveillance data.

References

[1] T. P. Breckon, A. Gaszczak, J. Han, M. L. Eichner, and S. E. Barnes, "Multi-modal target detection for autonomous wide area search and surveillance," in *Proc. SPIE*, 2013, pp. 889 913–889 913–20.

[2] C. N. Dickson, A. M. Wallace, M. Kitchin, and B. Connor, "Vehicle detection using multimodal imaging sensors from a moving platform," in *Proc. SPIE*, vol. 8541, 2012, pp. 854 112–854 112–11. [Online]. Available: http://dx.doi.org/10.1117/12.971464

[3] I. Rodger, B. Connor, and N. M. Robertson, "Classifying objects in lwir imagery via cnns," in *Proc. SPIE*, 2016, pp. 99 870H–99 870H–14. [Online]. Available: http://dx.doi.org/10.1117/12.2241858

[4] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Multi-view automatic target recognition using joint sparse representation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 3, pp. 2481 – 2497, 2012.

[5] J. Sun, G. Fan, L. Yu, and X. Wu, "Concave-convex local binary features for automatic target recognition in infrared imagery," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 23, 2014. [Online]. Available: http://dx.doi.org/10.1186/1687-5281-2014-23

[6] T. M. Strat, "Employing contextual information in computer vision," in *In Proceedings of ARPA Image Understanding Workshop*, 1993, pp. 217–229.

[7] P. Harding and N. M. Robertson, "Visual saliency from image features with application to compression," *Cognitive Computation*, vol. 5, no. 1, pp. 76–98, 2013. [Online]. Available: http://dx.doi.org/10.1007/s12559-012-9150-7

[8] B. C. Hansen and E. A. Essock, "A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes," *Journal of Vision*, vol. 4, no. 12, p. 5, 2004. [Online]. Available: + http://dx.doi.org/10.1167/4.12.5

[9] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.

[10] A. Friedman, "Framing pictures: The role of knowledge in automatized encoding and memory for gist," *Journal for Experimental Psychology: General*, vol. 108, pp. 316–355, 1979.

[11] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, pp. 169–191, 2003.
[12] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Proc. IEEE CVPR 2012, Egocentric Vision workshop, Providence, Rhode Island*, Jun 2012, bu;td;mod;cv, pp. 1–8.
[13] E. Maggio and A. Cavallaro, "Learning scene context for multiple object tracking," *Trans. Img. Proc.*, vol. 18, no. 8, pp. 1873–1884, Aug. 2009. [Online]. Available: http://dx.doi.org/10.1109/TIP.2009.2019934
[14] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler, "segdeepm: Exploiting segmentation and context in deep neural networks for object detection," *CoRR*, vol. abs/1502.04275, 2015. [Online]. Available: http://arxiv.org/abs/1502.04275
[15] J. Sun and D. W. Jacobs, "Seeing What Is Not There: Learning Context to Determine Where Objects Are Missing," *ArXiv e-prints*, Feb. 2017.
[16] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.
[17] N. M. Robertson and J. Letham, "Contextual person detection in multi-modal outdoor surveillance," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 1930–1934.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
[19] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.