Fast CNN Surveillance Pipeline for Fine-Grained Vessel Classification and Detection in Maritime Scenarios

Fouad Bousetouane and Brendan Morris
Electrical and Computer Engineering Department
University of Nevada, Las Vegas, NV 89154, USA
{fouad.bousetouane, brendan.morris}@unlv.edu

Abstract

Deep convolutional neural networks (CNNs) have proven very effective for many vision benchmarks in object detection and classification tasks. However, the computational complexity and object resolution requirements of CNNs limit their applicability in wide-view video surveillance settings where objects are small. This paper presents a CNN surveillance pipeline for vessel localization and classification in maritime video. The proposed pipeline is build upon the GPU implementation of Fast-R-CNN with three main steps:(1) Vessel filtering and regions proposal using low-cost weak object detectors based on hand-engineered features. (2) Deep CNN features of the candidates regions are computed with one feed-forward pass from the high-level layer of a fine-tuned VGG16 network. (3) Fine-grained classification is performed using CNN features and a support vector machine classifier with linear kernel for object verification. The performance of the proposed pipeline is compared with other popular CNN architectures with respect to detection accuracy and evaluation speed. The proposed approach mAP of 61.10% was the comparable with Fast-R-CNN but with a $10\times$ *speed up (on the order of Faster-R-CNN) on the new Annapolis Maritime Surveillance Dataset.*

1. Introduction

Early understanding and prediction of threats is critical for modern vision-based fleet security systems. However, analyzing complex maritime scenes of wide coverage areas in real-time is a challenging problem because of the large number of simultaneous vessel activities, waves, small vessel size, and occlusion. Further, actionable threat analysis and behavior recognition requires detailed understanding about vessel location, speed, context, and capabilities. In order to obtain these attributes, vision-based systems can be employed to detect, recognize, and track vessels. However, assigning identity to vessels and precisely estimating their positions in complex wide-area views is still a challenging vision task. In addition to the typical object detection challenges in terms of object description and localization, vessel identification is further complicated by fine-grained intra-class shape, appearance, and size variability in addition to distinct inter-class similarities. Further, the typical fleet security system should operate on video in real-time for early threat detection and warning.

Recent work using deep convolutional neural networks (CNNs) have been successful for fine-grained classification (FGC) as they are able to catch subtle inter-sub-class differences. In addition, newer frameworks have been developed using search space reduction techniques to provide dramatic speed-up in CNN evaluation [5, 1]. These new CNN frameworks, such as R-CNN [6], Fast-R-CNN [5], and Faster-R-CNN [11], utilize region proposal techniques to process sub-images as well as region-of-interest pooling for the computation speed-up required for video surveillance.

In spite of the spectacular detection accuracy of these frameworks, straight-forward application to surveillance is not possible. These frameworks were designed for lower dimensional images of non-crowded scenes such as in Pascal VOC ($\sim 469 \times 387$ pixels) or ImageNet ($\sim 482 \times 415$) images. Surveillance images in contrast cover wide-views with higher resolution images and have crowded scenes with many smaller objects of interest. This results in two main challenges: First, the number of candidate regions grows exponentially with image size. Second, the extraction of CNN features is complex and time consuming and should be limited to only likely regions of interest.

In this paper, a simple but powerful deep CNN-based FGC pipeline is proposed specifically for maritime surveillance. The proposed system is built on top of the Fast-R-CNN framework [5] with three main stages. It utilizes a weak HOG-based object detector for efficient region proposal in the first stage. The second stage utilizes the maritime fine-tuned VGG16 network to extract high-level CNN

IEEE AVSS 2016, August 2016, Colorado Springs, CO, USA

features for off-the-shelf classification using a linear SVM. The final stage provides region verification and assigns final class labels using a simple confidence scheme. The full system is optimized through multi-core parallel operation of region proposal and GPGPU computation for CNN evaluation for near real-time performance.

The main contributions of this paper are: i) the development of a deep CNN pipeline suitable for detection and fine-grained classification in surveillance. ii) Use of simple object detectors based on hand-engineered features for region proposal in wide-view images. iii) Evaluation of state-of-art sophisticated deep CNN pipelines for surveillance application on the new Annapolis Maritime Surveillance Dataset.

The overall structure of the paper is organized as follows: Section 2 is dedicated to description of background information and related work. Section 3., presents the proposed CNN off-the-shelf pipeline for vessels detection. Experimental comparison on the Annapolis with various state-of-the-art CNN architectures is provided in Section 4 with a discussion of results. Finally, concluding remarks are provided in Section 5.

2. Background and Related Work

Over last 10 years, researchers have developed a wide spectrum of different hand-engineered descriptors which encode shape, structure and context. These descriptors were integrated into object detection approaches of various levels of sophistication for object localization in real-time. Popular examples of these approaches include the HOG detector [2] and deformable parts model (DPM) [4].

Recently, a sequence of results on challenging visual recognition benchmarks has demonstrated that deep CNN-based frameworks greatly outperform hand-engineering detectors [5][6]. Sermanet et al. [12] proposed a CNN based framework for object detection in which a sliding window is used to densely sweep the entire image at different scales. While effective, the sliding window approach generates a very large number of candidate regions to evaluate for which the feed-forward pass through the deep CNN architectures was computationally very expensive.

To improve the computation time and to utilize deep CNN features for object detection, a limited region-based paradigm was proposed [5][6]. In contrast to sliding window strategy, this paradigm generates a few hundred to thousands of regions which are likely to contain target objects [8]. The most notably region based CNN is R-CNN framework proposed by Girshick et al. [6]. This framework uses the selective search algorithm [14] as a low-level segmentation process in the region proposal stage. The resulting ~ 1000 regions are then input to a appropriately tuned CNN model. Other region proposal methods that are class agnostic have been proposed for CNN detection such as the deep bounding box network DeepMultiBox [3].

To further address the issue of computation time, computation sharing techniques have been proposed, e.g., SPPnet [7]. SPPnet accelerates the R-CNN by sharing computation at the high-level spatial pyramid pooling layer. The fixed convolutional layer of SPPnet limits the utilization of SPPnet framework on new datasets. This limitation was recently addressed by Fast-R-CNN[5] object detection framework. The Fast-R-CNN provides a new training algorithm to update all convolutional layers to speed up the R-CNN and SPPnet frameworks and improve detection accuracy. Fast-R-CNN framework trains the deep network models (e.g., VGG16) 9x faster than R-CNN and 3x faster than SPPnet. However, the major limitation of CNN-based frameworks for object detection (e.g., R-CNN, SPPnet, Fast-R-CNN, etc.) is in the region proposal stage. If the region proposal (e.g., selective search) does not provide a bounding box close to a real object location, then no matter how deep, the CNN will not be able to detect it [16].

The Region Proposal Network (RPN) was recently proposed by Shaoqing Ren et al. [11] to share full-image convolution features with the detection network. The Faster-R-CNN framework was formed by combining the RPN for region proposal with Fast-R-CNN for classification. Faster-R-CNN is able to generate just 300 hundred region proposals on Pascal VOC images, while achieving state-of-art object detection accuracy on PASCAL VOC challenge.

Despite all the recent advances and success in object detection with deep CNNs, these frameworks have only be utilized on images of smaller dimension (e.g., Pascal VOC and ImageNet). In addition, these datasets generally provide few objects per image and objects that occupy a large portion of the entire image. However, in surveillance setting, there can be many objects of interest in a crowded environment with objects representing only 1%-5% of the scene. In these situations, will the state-of-the-art CNN object detection frameworks be effective?

This paper provides response to this question through experimental study of state-of-the-art object detection frameworks for maritime vessel detection in high resolution surveillance images of Annapolis harbor. Furthermore, a simple but effective deep CNN-based pipeline for fine-grained classification and detection of vessels is proposed which outperforms the state-of-the-art methods.

3. Surveillance CNN Detection Pipeline

The proposed surveillance pipeline is based on three main stages as illustrated in Fig. 1: 1) Region proposal stage to generate object candidates, 2) fine-grained classification, and 3) Region verification to maintain only confident detections.

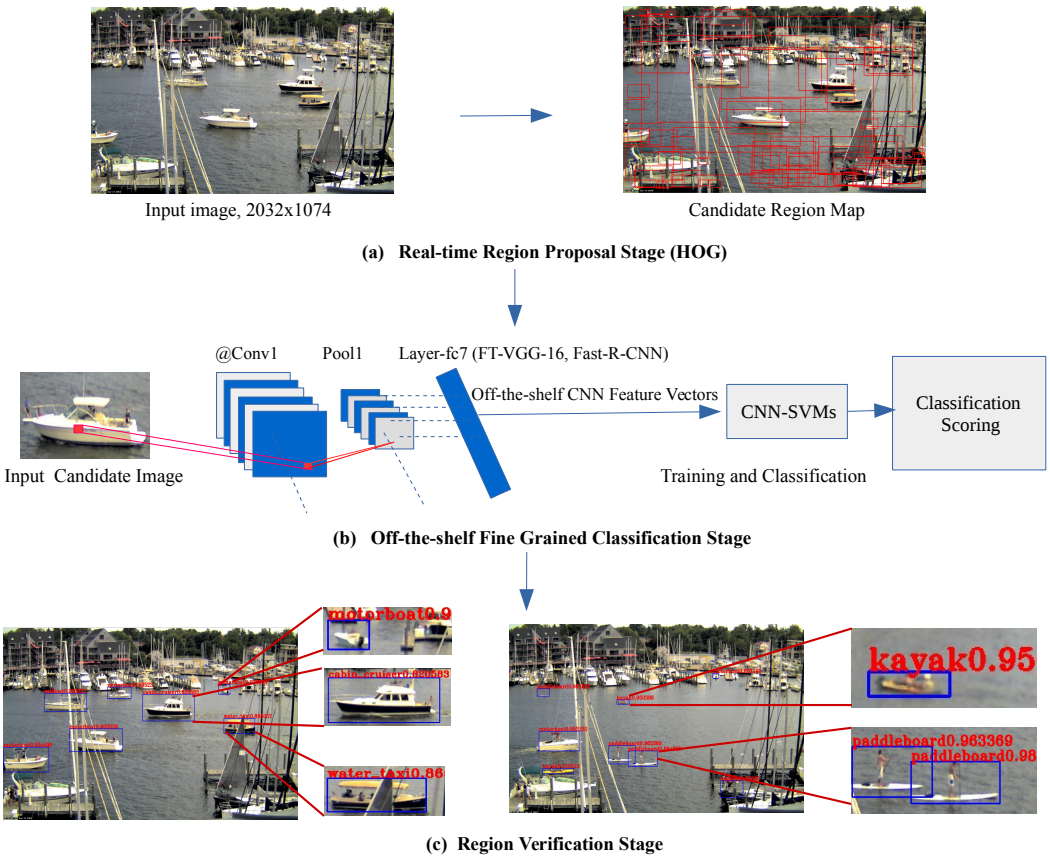


Figure 1: Overview of the proposed three stage CNN-based pipeline for Vessels classification and detection in wide-view Annapolis maritime surveillance images.

3.1. Real-Time Region Proposal

Although sliding window search based methods for object detection were popular for many years, exhaustive search becomes prohibitively expensive when using deep CNN features for region classification. To speed up this process, the selective search (SS) algorithm is used in recent CNN-based frameworks [5, 6, 7] to generate $\sim 1000-4000$ candidate regions per image of lower resolution (e.g., 469×387 pixels for Pascal VOC). However, the computation time of this algorithm grows exponentially relative to scale and image dimensions and since it is a bottom-up visual segmentation operation, it does not perform well in crowded scenes, wide-view imagery, or when objects are of small size as is the case for surveillance.

The proposed system utilizes a region proposal stage that is better suited for identifying likely object regions in wide-view surveillance images. The image is filtered and examine at different scales using weaker detectors trained with low-cost hand-engineered features (e.g. HOG, LBP, etc.). The classifiers can be tuned to have very high recall since errors are acceptable and expected to be pruned with a more complicated classification stage later. In fact, a set of detectors can be trained in a one-vs-all training scheme for object specific performance. The candidate regions from each weak-detector are fused together into a candidate map which should contain all objects of interest along with false positives.

In practice, the separate object detectors can be run in parallel on multiple cores (or GPU) to accelerate the evaluation time. In our experiments, only ~ 100 candidate regions were returned after non-maximal suppression from the weak-detector and passed on to the next stage for CNN feature extraction.

3.2. Off-the-Shelf Fine-Grained Classification

Fine grained classification (FGC) is the task of distinguishing sub-ordinate categories of the same class such as boats [1], aircraft, bird or car models [15]. The major challenge of FGC is object description with relevant and discriminative appearance features to differentiate fine details in appearance between visually very similar object class. Recently, a sequence of results has demonstrated that fea-

tures generated from deep CNNs have powerful inter-class discrimination [10, 1]. Discriminative features are needed in this stage to deal with vessel's intra-class shape variability to reach high recognition accuracy. In order to recognize the generated candidate regions, an off-the-shelf FGC scheme with two stages is developed as illustrated in Figure.1.

3.2.1 Deep CNN Features Extraction

In deep CNNs, different layers correspond to a hierarchy of features; earlier layers have more low-level features. CNN features are extracted at the highest level (or last convolution/pooling layer) to encode appearance relationships [1]. In this work, the VGG16 CNN model with `ROIPooling` layer implemented in the Fast-R-CNN framework is used as a black box deep CNN features extractor. The high-level CNN features were extracted from the first fully connected layer `fc7` and are of dimension 4096. The deep VGG16 CNN is pre-trained on ImageNet and is fine-tuned on the application specific data (see Section 4.1). The `ROIPooling` layer of the Fast-R-CNN framework is critical for surveillance operation since it dramatically improves computation time by performing only one feed-forward evaluation for the multiple region proposals that come from the previous stage.

3.2.2 Region Evaluation

Region evaluation is the operation of the assignment of subclass (label) to each candidate region. It is common to use the softmax function for classification problems using CNNs as it can be elegantly integrated as the last layer of the network to produce a probability distribution over the c classes through the minimization of the cross-entropy loss. However, recent work has shown improved classification results in CNN pipelines using margin-maximizing support vector machine (SVM) classifiers [13, 10]. A multi-class SVM with linear kernel is trained using the high-level CNN features to provide a class specific score for each of the c object classes. The SVM output is mapped to a value between $[0, 1]$ to approximate the class probability.

3.3. Region Verification

After the FGC stage, each candidate region is represented by probability distribution $p(c) \in [0, 1]$ over the c object sub-classes. The final region class is assigned as the sub-class with highest probability as long as it is over a threshold of acceptance

$$c^* =_c p(c) > Th^c. \quad (1)$$

In this work, the threshold is fixed at 0.5 but could be learned for each sub-class from training data. An example of region verification is displayed in the bottom of Fig. 1 where the original detections are labeled by the vessel subclass and false detections are filtered (see top of the Candidate Region Map) by the classification threshold. Note that in contrast to R-CNN and Fast-R-CNN object detection frameworks, no bounding-box regression is used in the proposed pipeline since localization is handled by the HOG region proposal.

4. Experimental Evaluation

The CNN surveillance pipeline is evaluated on the new wide-view Annapolis Maritime Surveillance Dataset. The experiments utilize the ImageNet pre-trained VGG16 CNN that is part of the Fast-R-CNN framework [5] with fine-tuning on the Annapolis data. The proposed system is compared with other popular object detection frameworks including R-CNN, Fast-R-CNN, and Faster-R-CNN. Experimental results illustrate the impact on performance when utilizing high-dimensional images with small objects of interest (surveillance setting).

4.1. Annapolis Maritime Surveillance Dataset Overview

The Annapolis Maritime Surveillance Dataset consists of a collection of wide field-of-view images captured in high-resolution (2032×1072). Images were obtained from a camera overlooking the Annapolis Harbour at 1 Hz over the course of a week from 19:40 Friday August 13, 2010 through 03:00 Saturday August 21, 2010. The dataset consists of nine recreational vessel classes: cabin cruiser, canoe, kayak, motorboat, paddle-board, raft, row-boat, sailboat and water taxi. Fig. 2 provides example images from the dataset along with examples from the nine vessel classes. Docked vessels in the harbor were not annotated because it is difficult to reliably identify the the sub-class of these vessels only based on the rear view.

The main challenge of this dataset is wide variability in appearance of vessels from the same class (e.g. cabin cruiser in Fig.2b). Furthermore, small vessels like kayak and paddle-board represent only $1\% - 5\%$ of the total image resolution. Finally, the harbor scene is crowded and complicated (waves) which makes visual segmentation techniques for region proposal inaccurate.

4.2. Implementation Details

A HOG-based vessel detector is trained for large and small vessels separately to generate region proposals. This detector is based on a cascade of HOG classifiers. Vessels were divided by size since the visual features are more consistent for similarly sized objects. The small vessels detector was trained on a size of 58×24 and contained the five vessel classes of kayak, paddle board, raft, rowboat, and

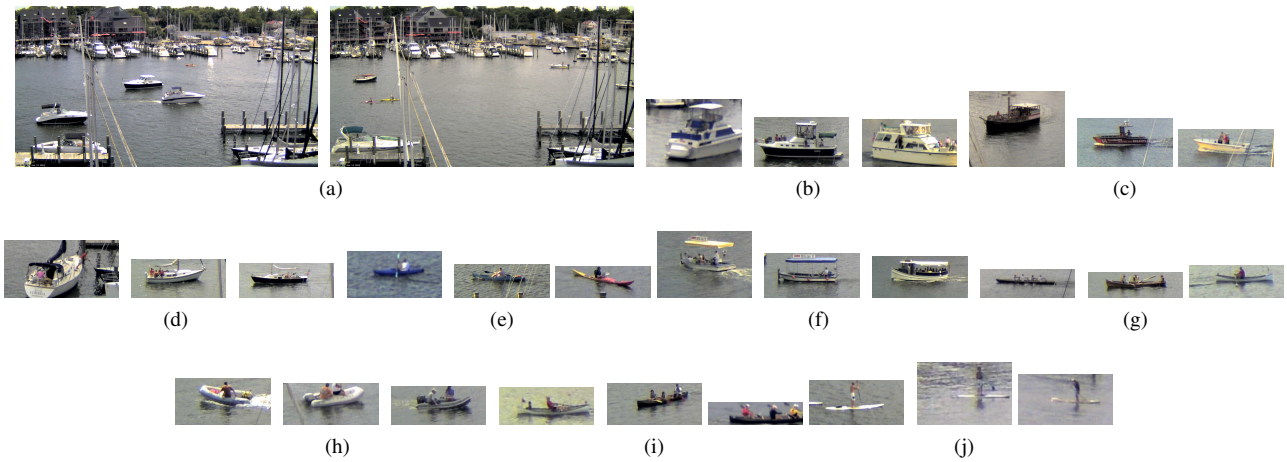


Figure 2: Annapolis Maritime Vessel Dataset. (a) Example of wide view of Annapolis Harbor. (b)-(j) Examples of images of annotated Annapolis Maritime vessel categories. (b) Cabin Cruiser, (c) Motorboat, (d) Sailboat, (e) Kayak, (f) Water Taxi, (g) Rowboat, (h) Raft, (i) Canoe, (j) Paddle Board

canoe. The big vessel detector is trained on sailboat, motorboat, cabin-cruiser and water-taxi at a size of 110×32. The detectors operate in parallel on multiple CPU cores in full sliding windows at multiple scales in the region proposal stage.

The full image and vessel-like regions and used as input to the VGG16 CNN from the Fast-R-CNN code. The dataset was enriched by performing basic transformations, like rotation, for fine-tuning. The CNN implementation is derived from the publicly available GPU-version of the Caffe toolbox [9] and Fast-R-CNN code [5]. A 4096 dimension feature vector is obtained for each vessel image 4096 and normalized using the L2-norm and used to train a multi-class linear SVM using the LibSVM toolbox.

The available online open source codes of R-CNN, Fast-R-CNN and Faster-R-CNN frameworks are used for comparison with the proposed pipeline. The CNN network of Fast-R-CNN and the RPN network of the Faster-R-CNN framework were both pre-trained on ImageNet dataset and fine-tuned on Annapolis training data.

4.3. Results and Discussion

The detection performance is evaluated using the Pascal VOC protocol based on the precision-recall curves to obtain the average precision (AP) for each vessel class. The AP results as well as the computation time for the different detection techniques is provided in Table 1. Note that for R-CNN and Fast-R-CNN two variants for region proposal were examined, the standard selective search (SS) and a sliding window (SW) approach to obtain and upperbound on performance.

4.4. Vessel Detection Results

Overall, the mean average precision (mAP) 61.55% is obtained for SW+Fast-RCNN. The proposed HOG+Fast-R-CNN pipeline has comparable performance with the SW approaches at mAP=61.10%. However, the performance is significantly greater than the standard SS region proposal variants. In addition, the computation time of the whole detection process (including region proposal) of 0.67 s/frame is comparable with faster-R-CNN with its region proposal network. The dramatic speedup with HOG+Fast-R-CNN comes from the reduced number of regions to evaluate. To catch very small vessels (e.g. kayak or paddle board), the selective search process was initialized with very small segment size (10x10 pixels) which results in more than $30,000$ region proposals generated per large 2032×1072 pixel Annapolis image. In contrast, the RPN of Faster-R-CNN generated ~ 2000 regions and only ~ 100 from the HOG detectors. The main bottleneck was the segmentation and region proposal from selective search.

4.5. Surveillance-Size Results

Further exploration highlighted the impact of vessel size and classification accuracy (Fig. 4). In this experiment, each vessel is cropped from the full-size image with five different padding settings. The resulting image was designed to have the vessel occupy 10%, 20%, 40%, 50% or 80% of the total image size and was used as the input image into the CNN detection pipelines. Not surprisingly, the smaller vessels had poorer performance since there are fewer visual cues to distinguish. In Fig 4a, the effect of the ratio of object to image ratio is illustrated. In the case the object is larger than 40% of the image size, the detection accuracy is high

Table 1: Average Precision (%) and Timing (s/image) Results on K80 GPU

Method	# Regions	Cabin Cruiser	Motorboat	Sailboat	Canoe	Paddle Board	Kayak	Raft	Rowboat	Water Taxi	mAP	Time
SS+R-CNN	30879	13.25	21.50	34.10	6.94	7.75	10.24	9.14	11.24	51.78	18.43	12.954
SW+R-CNN	-	83.50	87.15	75.06	**30.57**	38.73	51.32	42.02	**45.78**	87.17	60.14	21.500
SS+Fast-R-CNN	30200	15.75	27.24	37.87	8.02	7.98	13.56	9.35	12.04	52.23	20.44	6.023
SW+Fast-R-CNN	-	84.32	**87.97**	79.98	29.32	**40.87**	**53.84**	43.14	45.25	**89.32**	**61.55**	10.350
RPN+Fast-R-CNN	2000	45.30	38.96	41.32	17.71	18.62	26.67	16.35	18.89	56.24	31.17	**0.660**
HOG+Fast-R-CNN	100	**85.45**	87.32	**81.54**	29.84	39.80	49.56	**45.35**	44.16	86.95	61.10	0.670

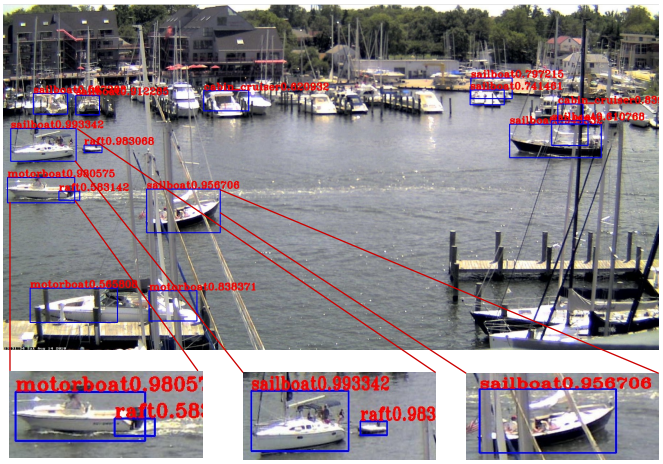
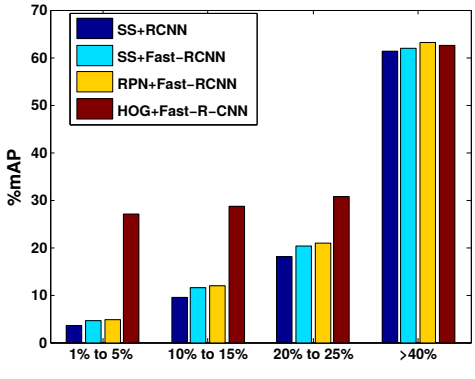
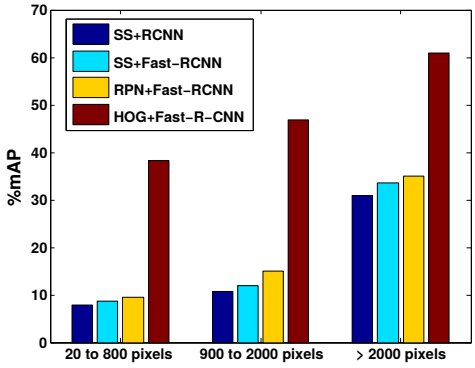


Figure 3: Examples of Annapolis vessel detection and classification results using the proposed HOG+Fast-R-CNN pipeline. Note correct classifications in bottom have confidence sore close to 1 while the "raft error" has low confidence. The system is also able to correctly detect and classify even with occlusion.



(a)



(b)

Figure 4: CNN frameworks for surveillance: (a) Impact of vessel/image ratio on detection accuracy. (b) Effect of Annapolis vessel size on CNN detection.

(mAP $>$ 60%) for R-CNN variants. This case corresponds to the scenario for most detection challenges in vision such as VOC or ILSVRC. However, when the object occupies a smaller fraction of the image, there is a dramatic drop-off in performance. The HOG+Fast-R-CNN in contrast does not suffer with size as much. The size-performance relationship is further illustrated in Fig. 4b on evaluation of a full image. In this plot, it is clear that the performance of HOG+Fast-R-CNN is much less dependent on the size of the vessel than the traditional R-CNN variants. Even when the object is large ($>$ 2000 pixels), the performance is not more than 40% since this is still a small fraction of the total image size.

5. Conclusion

In this work, a powerful CNN-based pipeline for detection in wide-area surveillance is developed. The pipeline makes a simple replacement of selective search with a HOG detector for region proposal in the Fast-R-CNN system. The resulting pipeline has high accuracy on a complex maritime vessel dataset while simultaneously increasing the evaluation speed 10$\times$ over the original Fast-R-CNN. The main conclusion from this work is that state-of-the-art region proposal techniques designed for traditional detection challenges are not well suited for wide-view surveillance settings where objects are very small in comparison with image size.

Acknowledgments

Thanks to ONR 311 and NRL for supporting this research.

References

[1] F. Bousetouane and B. Morris. Off-the-shelf cnn features for fine-grained classification of vessels in a maritime environment. In *Advances in Visual Computing*, pages 379–388. Springer, 2015.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[3] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2155–2162. IEEE, 2014.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[5] R. Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer Vision–ECCV 2014*, pages 346–361. Springer, 2014.

[8] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *arXiv preprint arXiv:1502.05082*, 2015.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[10] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[13] Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[14] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[15] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *arXiv preprint arXiv:1411.6447*, 2014.

[16] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee. Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. *arXiv preprint arXiv:1504.03293*, 2015.