Comparison of Object Detection Algorithms on Maritime Vessels

Mark Chua[1], David W. Aha[2], Bryan Auslander[3], Kalyan Gupta[3], and Brendan Morris[1]

[1]Department of Electrical and Computer Engineering;
University of Nevada, Las Vegas; Las Vegas, NV 89154
[2]Adaptive Systems Section;
Naval Research Laboratory, Code 5514;
Washington, DC 20375
[3]Knexus Research Corporation; Springfield, VA 22153

Abstract

This manuscript conducts a comparison on modern object detection systems in their ability to detect multiple maritime vessel classes. Three highly scoring algorithms from the Pascal VOC Challenge, Histogram of Oriented Gradients by Dalal and Triggs, Exemplar-SVM by Malisiewicz, and Latent-SVM with Deformable Part Models by Felzenszwalb, were compared to determine performance of recognition within a specific category rather than the general classes from the original challenge. In all cases, the histogram of oriented edges was used as the feature set and support vector machines were used for classification. A summary and comparison of the learning algorithms is presented and a new image corpus of maritime vessels was collected. Precision-recall results show improved recognition performance is achieved when accounting for vessel pose. In particular, the deformable part model has the best performance when considering the various components of a maritime vessel.

1. Introduction

Detection and classification of objects within images is an actively pursued field in computer vision. To foster growth, PASCAL created the Visual Object Classes (VOC) Challenge to serve as a benchmark. Competing teams were tasked to detect or classify a number of object categories. This challenge took place annually during 2005-2012. For each competition, top researchers in the field of computer vision could submit their respective systems to classify objects, which led to the development of many state-of-the-art algorithms.

However, object placement in categories for the VOC can be considered to be generic; for example, all maritime vessels of varying sizes were under the category label 'boat'. There has been no formal evaluation for the multiple categories within this broad category. Obtaining high classification accuracy of maritime vessels given images with low inter-class variation could prove extremely useful in maritime threat assessment. Accurate systems could be developed that rely on visual data rather than electronic data exchange between vessels or parameter based behavior analysis.

In this survey, we selected three high-scoring algorithms that were submitted to VOC competitions to determine their relative performance when given more specific labels. They are *Histogram of Oriented Gradients* (HOG) [1], *Exemplar-SVM* (ESVM) [2][6], and *Latent-SVM with Deformable Part Models* (LSVM) [3][5]. These algorithms performed extremely well in their given submission years. The HOG/SIFT representation has several advantages. It captures edge or gradient structure that is very characteristic

of local shape, and it does so in a local representation with an easily controllable degree of invariance to local geometric and photometric transformations [1]. *E-SVM* simplifies the HoG learning task by treating every instance as a separate object of fixed orientation. *L-SVM* develops a flexible representation that learns subparts of an object to provide robustness to appearance change.

In Section 2, we briefly review these algorithms. We then describe the maritime corpus we used to compare them in Section 3. In Section 4, we describe our empirical methodology, and report the results in Section 5. We end with a discussion in Section 6.

Our main finding was that state-of-the-art object detection systems, especially *Latent-SVM*, were accurate in identifying boats within images. However, boat categories whose appearances are small and have little visual cues such as kayaks and canoe generally did worse than boats that were large and had unique visual features such as the water taxis (canopied tops) and sailboats (mast and sails). These visual features were most important in the case of *L-SVM*. The main strength of *L-SVM* is to extract latent features using deformable part models and use these part models to correctly detect an object. We believe this is why *L-SVM* achieved good results in vessels that have a unique visual feature.

2. Description of Algorithms

Here we briefly describe the selected object detection algorithms.

2.1 Histogram of Oriented Gradients

Dalal and Triggs (2005) introduced a feature descriptor they named *Histogram of Oriented Gradients* (HOG). Their empirical study showed that *HOG* can be used to derive feature sets that can be used to attain high precision and recall on some pedestrian detection tasks. *HOG*, when paired with a linear SVM, gave essentially perfect results for the MIT pedestrian test set, which prompted INRIA to create the more difficult INRIA pedestrian test set.

The concept behind HOG is that local object appearances and shapes can be represented by the distribution of gradient intensities and orientation of edges. An image is divided into overlapping cells that create a histogram of gradients and edge directions of the pixels within the cell. Groups of cells are then grouped together into blocks, which are formally called *Histogram of Oriented Gradients*. Figure 1 shows an example of a water taxi with its corresponding HOG. (Other HOG examples can be seen in Figures 3 and 4).

Figure 1. In this image, the water taxi surrounded by the bounding box is taken as a positive instance for the water_taxi category (left). The confined area is extracted and converted to a HOG feature set for training (right).

2.2 Exemplar SVM

Object detection in *Exemplar-SVM* [2] relies on a conceptually simple but effective approach. Every positive instance of the object in the training set can be selected as an exemplar and trained individually. A single positive exemplar is trained against (potentially) millions of negative examples using HOG as a feature descriptor. This results in multiple, feature-specific SVM models. Detection is then performed with a nearest-neighbor approach to find the closest exemplar model that matches the current detection window.

E-SVM eschews generality and instead focuses on specificity. A single classifier no longer has to represent objects in general pose (all orientations and appearance variations), which may be quite difficult in complex cases. Each *E-SVM* is able to accurately recognize an object at a specific pose, operating like an appearance-based nearest neighbor operation (see Figure 2).

Exemplar-SVM can become computationally expensive due to the training of multiple models; each exemplar requires a separate model. This is expanded with the number of negative examples needed to fine-tune the deciding hyper plane for each exemplar's model. It is also unclear what is the optimal way to manage the exemplar database – to ensure all relevant poses and appearances are covered without requiring too many exemplars.

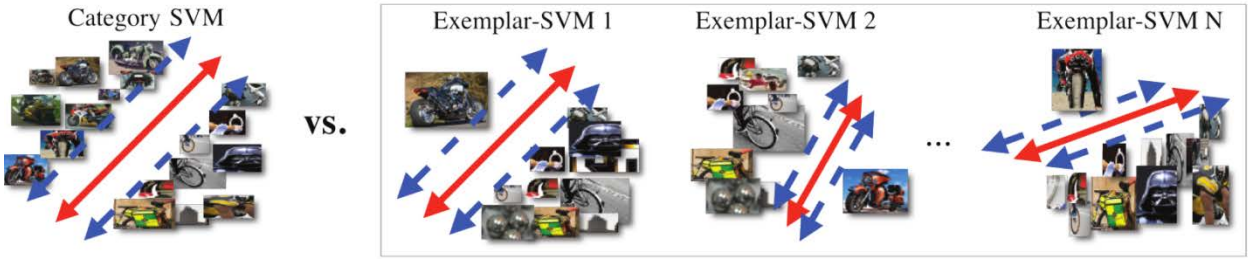
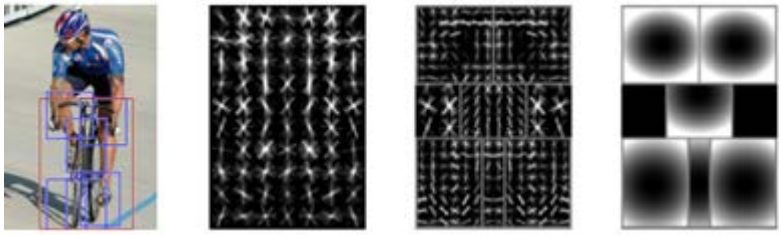


Figure 2. This shows the main difference between a general SVM and *Exemplar-SVM*. In a general SVM, multiple instances of the object are trained together as positives while in *E-SVM*, only a single instance of the object (usually of a certain pose) is trained. The three positives shown in *Exemplar-SVM* represent different view profiles of the motorcycle.

2.3 Latent SVM with Deformable Part Models

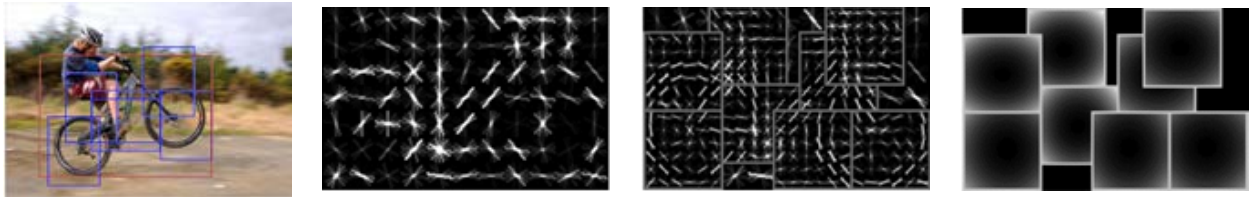
Latent SVM is an extension of the HOG model put forth by Dalal and Trigg. A HOG feature descriptor is created of the object named the root filter. The object is then expanded to twice the spatial resolution to generate part models. A part model consists of a spatial model and a part filter. The spatial model is a set of possible placements of the part model relative to the position of the root filter along with a cost for displacement from the original position, as seen in Figure 3. The overall score of the detection window is the score of the root filter along with the sum of the part filters. The star-structure of parts provides an enriched HOG model. Object categories are also represented as mixtures of these star models in order to represent various object appearances (Figure 4a,b).

The final classification score is a combination of the root filter along with optimal placement of the parts which provides robustness to noise and occlusion. However, this does come at the cost of significantly more involved and time-consuming learning process.



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

Figure 3. Example of Latent-SVM with Deformable Part Models being used on an image with a bicycle object (a). Image (b) shows the HOG representation of the root filter while (c) and (e) shows part models being generated at twice the spatial resolution of the root filter.



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

Figure 4. The same object shown in Figure 3 but with a different pose with parts displaced to match various orientations. (a) The part models deformed to match the "wheelie" being done by the bicycle. (b) HOG (c) (e) part models.

3. Dataset

To compare the performance of the selected algorithms, we needed a large image corpus of maritime vessels. We chose the Annapolis Harbor Dataset used by Morris et al. [4] in an earlier experiment for our training and testing data. Its images were obtained from a public streaming webcam hosted by the Annapolis Yacht Club. Images from the “Spa Creek” webcam were saved in one second intervals between the hours of 19:40 Friday August 13, 2010 through 03:00 Saturday, August 21, 2010; a total of 58365 images were obtained from 180 hours of video.

Marine vessels in the Annapolis dataset were divided based on visual distinctions, which resulted in 9 classes. The classes are *cabin_cruiser, canoe, kayak, motorboat, paddleboard, raft, rowboat, sailboat,* and *water_taxi*. Classifications of vessels were left to the labeler’s discretion based on visible cues. For example, to be classified as a water taxi, a vessel must have a canopied top regardless of size or color while sailboats are vessels that are equipped with sails. There may be multiple boats within a single image, thus resulting in multiple annotations.

4. Experiment Design

We believe that object detection systems are still able to perform well given a more detailed set of categories. We also believe that objected detection systems that are more complex in their approach will generally perform better than systems that are less intensive in generating detection models.

4.1 Annotations

The Annapolis Harbor corpus was annotated with the following fields:

- Vessel Type {cabin_cruiser, canoe, kayak, motorboat, paddleboard, raft, rowboat, sailboat, water_taxi}
- Bounding Box [x, y, width, height]
- Occlusion {none, masts, partial, full}

The Occlusion field was included to gauge the difficulty in detecting the object. *None* indicates that there is no to little occlusion. *Masts* indicate that masts of other ships, usually docked at the foreground of the view, occlude our annotated vessel. *Partial* is when the vessel is 25%–60% occluded, while *full* occlusion means that the vessel is almost completely covered.

4.2 Separating Training and Testing Data

The Annapolis Harbor corpus consists of 180 folders representing the 180 hours of monitoring. Each folder contains all saved images within its respective hour. Due to time constraints, only hour sets 11-20, 41, 46, 70, 94, 118, 142, and 166 were annotated. Hour sets 11-17 were set aside for training while the remaining hours were used for testing.

4.3 Training

Positive examples were selected based on the algorithm and only retrieved for hours 11-17. The statistics from the training data is presented in Table 1 [be sure to link properly]. In total, 10,000 frames were annotated with a vessel for a total of 12,157 labeled “boat” instances. Of these examples, 7741 were completely without occlusion.

All positive instances of the vessel with occlusion type *none* were selected as positive examples for the L-SVM and HOG. E-SVM requires only a single positive instance for training each of its models. We manually selected multiple positive instances that best represent the different orientations of each marine vessel. Each exemplar was individually trained to generate an independent model, and combining these models can be used for vessel classification.

Negative examples were obtained from the SUN Database and the Annapolis Harbor corpus (i.e., images from it that contained no boats). The full negative dataset consists of 17,425 images from the SUN database and 3,000 negative Annapolis images for a total of 20,425 negative images.

Table 1: Training Data Statistics

Class	Frames	Instances	Occlusion			
			None	Masts	Partial	Full
Cabin_cruiser	876	907	466	95	340	6
Canoe	229	229	121	36	49	23
Kayak	1979	3028	1792	190	710	336
Motorboat	3695	4274	2641	412	1163	58
Paddleboard	292	346	278	12	28	28
Raft	1198	1256	864	49	252	91
Rowboat	202	210	143	15	43	9
Sailboat	1534	1615	1265	65	281	4
Water_taxi	285	292	171	29	92	0
Total	10290	12157	7741	903	2958	555

4.4 Testing

The test set consisted of annotated Annapolis pictures of the hour sets 18-20, 41, 46, 70, 94, 118, 142, and 166. Every 10th frame of the test set was annotated for a total of 3455 images with 5048 labeled instances of vessels. Only images that contained a positive instance of the maritime vessel were retrieved for testing. Any vessels whose occlusion was considered *full* was withheld from testing due to inherent difficulty in these settings in accordance with the VOC testing procedure. The test set contains examples from different days of capture and a range of weather and lighting conditions (between dusk and dawn and rain).

Table 2: Testing Data Statistics

Class	Frames	Instances	Occlusion			
			None	Masts	Partial	Full
Cabin_cruiser	103	103	68	6	29	0
Canoe	10	10	121	9	1	0
Kayak	384	585	1792	408	110	38
Motorboat	1340	1986	2641	1326	500	19
Paddleboard	64	80	278	52	15	8
Raft	213	227	864	169	28	18
Rowboat	18	18	143	13	5	0
Sailboat	1202	1702	1265	1165	427	19
Water_taxi	334	337	171	221	89	1
Total	3668	5048	7741	3431	1204	103

5. Results

This section presents the performance results of the of object detection algorithms.

5.1 Evaluation Criteria

The standard for the Pascal VOC Challenge to compare object detection systems was to compare their precision-recall curve. We will be using the same standard to compare the three algorithms. Precision is the fraction of retrieved instances that were relevant. Recall is the fraction of relevant instances that were retrieved.

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}} \qquad \text{Recall} = \frac{\text{tp}}{\text{tp+fn}} \qquad (1)$$

The precision and recall are defined in equation (1). A true positive (tp) is a relevant instance correctly classified as relevant whereas a false positive (fp) is an irrelevant instance that was incorrectly classified as relevant. A false negative (fn) is a relevant instance incorrectly classified as irrelevant. The detection system relies on two bounding boxes to determine a hit on the image. The detection bounding box is the area of the image that was reported to contain an object. The ground truth bounding box is the area of the image where an object is located. The detection system will classify detection as positive if the intersection of the two boxes divided by the union of the two boxes is greater than 50%.

$$a = \frac{area(Bp \cap Bgt)}{area(Bp \cup Bgt)} \qquad (2)$$

where B_p is the detection bounding box and B_{gt} is the ground truth bounding box.

5.2 Annapolis Harbor Performance

The results of the performances for our algorithms are displayed in Table 3. The table displays the average precision for each vessel and type of occlusion. The final column displays the mean average precision of all maritime vessels for all rows. Figure 5 displays the graphical view of the data compiled in Table 3; due to the poor performance of *HOG*, its graph was not included in Figure 5.

L-SVM was unable to generate models for the *canoe*, *raft*, and *paddleboard* categories and because of this, these categories were not included in our initial comparison. The reason that *L-SVM* was unable to generate models for these specific classes in still unknown and may be determined in future work to show a full comparison of all nine maritime vessel categories. One opinion we put forward is that due to the small size of the boats within the images, *L-SVM* was unable to properly create a model using deformable part models leading to the algorithm failing. This is purely speculation and due to time constraints, we decided to remove these three categories from comparison.

As seen in both Table 3 and Figure 5, the best performing category of maritime vessel is *water_taxi*. In the case of *E-SVM*, it is the only category whose average precision is comparable to results achieved by Malisiewicz et al. [2]. However, we noticed that most results for *L-SVM* were on par with results achieved by Felzenszwalb et al. [5] in their submission of *L-SVM* for the VOC Challenge.

To compare the three object detection algorithms, we plot the results of two categories that give high average precision within their respective algorithm (*water_taxi* and *cabin_cruiser*). *L-SVM* outperforms the other two as seen in Figure 5. In fact, we observed that the results of *L-SVM* across all maritime vessel categories and occlusion types were higher than either *HOG* and *E-SVM*. The only exception is the

rowboat category with occlusion type *masts*, where *L-SVM* was unable to detect any boats within the set of 13 images of rowboats obstructed by masts. We concluded that given the results, *L-SVM* is the best performing object detection algorithm out of the three we selected.

Table 3: Average Precision Results

Approach	Occlusion	Cabin cruiser	Kayak	Motor-boat	Rowboat	Sailboat	Water taxi	Mean AP
HOG	None	0.030	0.000	0.009	0.000	0.045	0.119	0.034
	Masts	0.000	0.000	0.005	0.000	0.091	0.109	0.034
	Partial	0.015	0.000	0.002	0.000	0.091	0.091	0.033
E-SVM	None	0.027	0.025	0.012	0.007	0.038	0.544	0.108
	Masts	0.023	0.023	0.011	0.007	0.038	0.489	0.099
	Partial	0.016	0.018	0.007	0.012	0.025	0.346	0.071
L-SVM	None	0.439	0.091	0.520	0.364	0.503	0.802	0.453
	Masts	0.182	0.045	0.421	0.000	0.496	0.709	0.309
	Partial	0.361	0.091	0.244	0.143	0.335	0.441	0.269

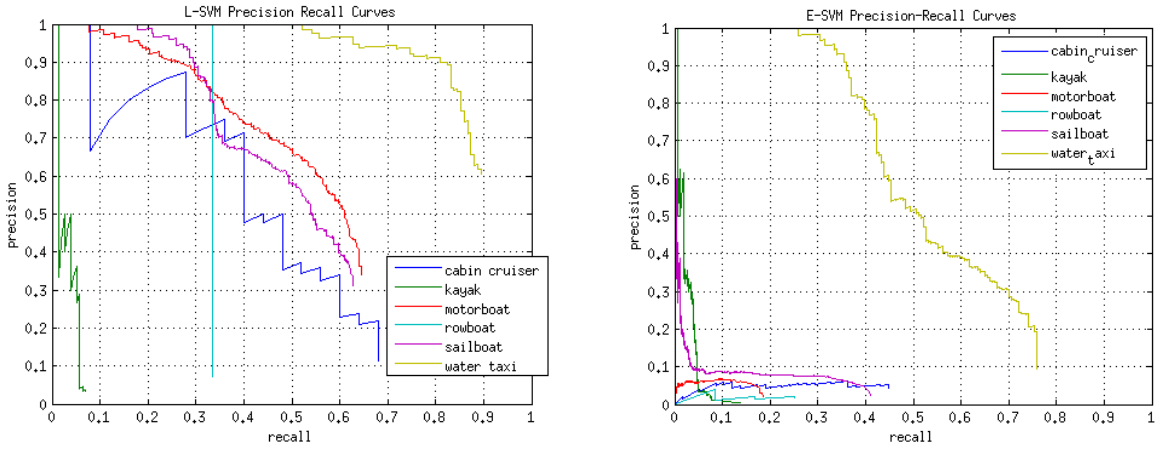


Figure 5. Precision-Recall performance graphs for *L-SVM* and *E-SVM*. Category *water_taxi* is the best performing maritime vessel category for all three object detection algorithms. Smaller class vessels such as *kayak* performed poorly.

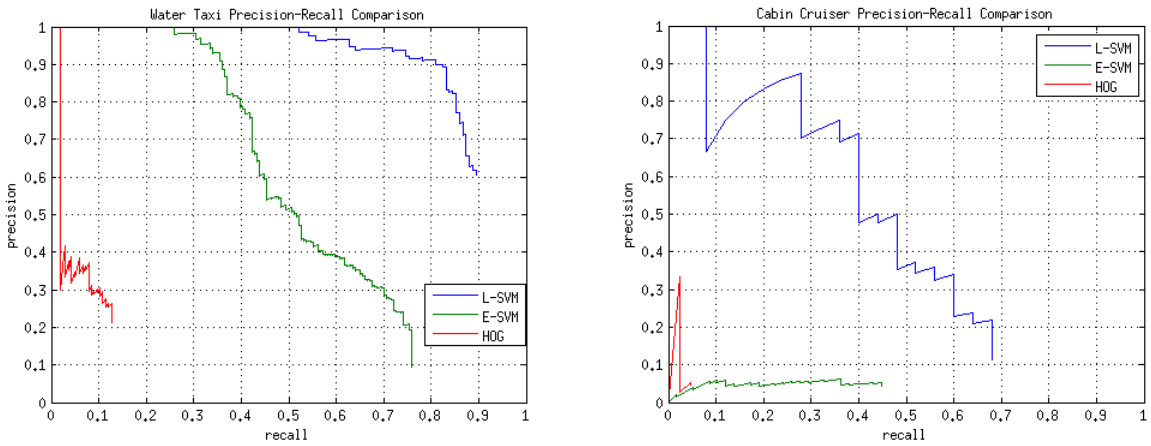


Figure 6. Performance comparison of the three object detection algorithms of two maritime vessel categories (*water_taxi* and *cabin_cruiser*). *L-SVM* is the best performing algorithm with the highest average precision rating on all six categories.

To view how occlusion factors in the detection rate, we plot the average precision of all three occlusion types in Figure 7. We plotted two categories (*water_taxi* and *motorboat*) using the results we received from *L-SVM*. This was because *L-SVM* provided the best results and variations between occlusions. This allowed us to easily see how occlusion affects the rate of detection. We correctly guessed that maritime vessels with little to no occlusion were more detectable than vessels that suffered from obstruction. Boats that were partially obstructed performed the poorest.

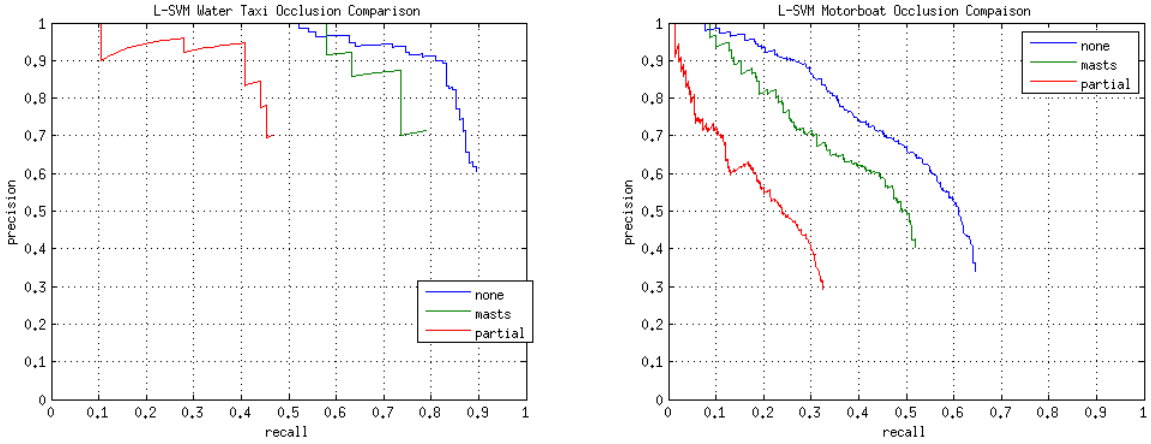


Figure 7. As expected, maritime vessels with the least occlusion are more easily detectable than vessels that suffer from obstruction.

6. Discussion

The results from this comparison show that given a more specific labeling scheme within a general object category, the algorithms chosen were able to detect objects within the images. Objects that had a

unique visual feature performed better than object categories that vary little in visual appearances. The category *water_taxi* is the most detected object from the image corpus with the highest average precision value throughout all three algorithms. We speculate this is due to the water taxi having a large, unique visual feature, specifically a canopied top. This allowed the object detection systems to better differentiate the water taxi from other boats.

Categories that included small vessels such as *kayak* and *rowboat* generally performed worse than larger vessels. Given the small size of the vessels, HOG was not able to create a feature descriptor that allowed the detection algorithm to distinguish the vessel from other areas of the images. A common false positive we noticed were waves coming from the water. When the area containing the waves is converted into a HOG feature descriptor, the SVM classifier cannot properly differentiate the HOG descriptor of a small vessel and feature descriptor of water waves. This might be improved by adding more images of waves into the negative feature set to help improve the SVM classifier and will be looked into further in future works.

Other attributes that increase object detection is the amount of occlusion on the maritime vessel. Boats that are easily visible (none to very little occlusion) are detected more often as seen in Figure 6. As mentioned in the testing section, boats that were mostly obstructed were excluded from testing. We believe that even though it was not tested, the occlusion field *full* would perform poorer than those tested.

Furthermore, object detection systems that relied on more complex models generally performed better. *Latent-SVM* is by far the best performing algorithm out of the three, achieving higher average precision scores in all six categories. *Histogram of Oriented Gradients*, being the least complex, performed the poorest. We believe that relying on deformable part models allows a greater flexibility in detection compared to a strict HOG feature descriptor.

Future work would include adding more negative images that contained only ocean waves. This could improve performance by reducing the amount of false positive detections on water. An expansion of the comparison could also be performed to include other state-of-the-art object detection systems that also performed well on the VOC Challenge. It will be important to compare other descriptors than those that use HOG to help determine how to best identify boats. Since they are rigid bodies, edge-based descriptors may be helpful for removing internal “noise” and focus on the ship itself.

Acknowledgements

Thanks to ONR for funding this research. Mark Chua’s visit to NRL during the summer of 2013, during which time he conducted the research described in this paper, was funded by the ONR NREIP Program.

References

[1] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection,” *CVPR*, pages I: 886-893, 2005.

[2] T. Malisiewicz, A. Gupta, and A. A. Efros, “Ensemble of exemplar-SVMs for object detection and beyond,” in *Proc. IEEE Inter. Conf. Computer Vision*, 2010, pp. 89-96.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. “Object Detection with Discriminatively Trained Part Based Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Sep 2010, Vol. 32, No. 9.

[4] B. Morris, D.W. Aha, B. Auslander, and K. Gupta. “Learning and leveraging context for maritime threat analysis: Vessel classification using Exemplar-SVM,” 2012.

[5] P. Felzenszwalb, D. McAllester, D. Ramanan. "A Discriminatively Trained, Multiscale, Deformable Part Model," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[6] T. Malisiewicz, A. Shrivastava, A. Gupta, and A. A. Efros. "Exemplar-SVMs for Visual Object Detection, Label Transfer and Image Retrieval," ICML, 2012.