VAIS: A Dataset for Recognizing Maritime Imagery in the Visible and Infrared Spectrums

Mabel M. Zhang[1], Jean Choi[2], Kostas Daniilidis[1], Michael T. Wolf[3], Christopher Kanan[3]
[1]University of Pennsylvania
{zmen@seas, kostas@cis}.upenn.edu
[2]Gwangju Institute of Science and Technology
pereus@gist.ac.kr
[3]Jet Propulsion Laboratory, California Institute of Technology
{wolf, kanan}@jpl.nasa.gov

Abstract

The development of fully autonomous seafaring vessels has enormous implications to the world's global supply chain and militaries. To obey international marine traffic regulations, these vessels must be equipped with machine vision systems that can classify other ships nearby during the day and night. In this paper, we address this problem by introducing VAIS, the world's first publicly available dataset of paired visible and infrared ship imagery. This dataset contains more than 1,000 paired RGB and infrared images among six ship categories - merchant, sailing, passenger, medium, tug, and small - which are salient for control and following maritime traffic regulations. We provide baseline results on this dataset using two off-the-shelf algorithms: gnostic fields and deep convolutional neural networks. Using these classifiers, we are able to achieve 87.4% mean per-class recognition accuracy during the day and 61.0% at night.

1. Introduction

Creating Autonomous sea Surface Vessels (ASVs) that can navigate the world's waterways with very little human intervention has enormous economic implications for the global supply chain. The seaborne cargo shipping industry moves over 9 billion tons of cargo per year, is worth $375 billion, and is responsible for 90 percent of world trade [18]. Crew costs are estimated to account for 44% of the operating expenses for seaborne cargo transport [2]. This expense could be eliminated by ASV cargo ships, with the added benefit of creating more room for cargo by eliminating life support systems. Large tanker and cargo ASVs could also operate at very slow speeds, known as super slow steaming, that would be unacceptable to a crew. Slow steaming can significantly increase fuel efficiency, which saves money and reduces pollution. Beyond shipping, ASVs have military applications such as surveillance.

Figure 1. An example autonomous ship recognizing another vessel observered by both its IR and visible cameras. To follow international maritime traffic regulations, ASVs need to classify other ships during the day and in low-light conditions.

One of the main obstacles to the development of ASVs is that they need to obey the International Regulations for Preventing Collisions at Sea 1972 (COLREGS). COLREGS governs when a vessel has the right of way over other vessels, and the rules depend on the kind of ship encountered. For example, a motorized vessel must give way to sailing vessels and vessels engaged in fishing, but they obey different rules when they encounter another motorized vessel. Therefore, to follow COLREGS, it is necessary for an ASV to categorize other vessels. An ASV may also need to classify ships for other reasons. For example, a military ASV may need to categorize hostile military vessels to determine how to best escape conflict. This recognition is best done visually using camera systems. Radar can be used to direct cameras where to look or analyze, but it alone is not suffi-

cient to differentiate among COLREGS relevant categories. Moreover, using radar would give away the vessel's position, which precludes its use in some military applications. While some ships broadcast information that includes the vessel's type and location, many ships do not, especially military and pleasure vessels.

To begin developing ASVs capable of obeying COLREGS, we created a dataset of synchronized and paired visible (RGB) and long-wavelength infrared (LWIR) photographs of vessels on the water. LWIR cameras produce images that measure thermal emissions in the environment, so they are capable of seeing ships during the day and night. Each pair of images was taken within the same full second, in the same scene, on a rigid mount. The images were collected over 9 days (10–15 hours per day) at 6 different piers, in the form of image sequences. Hundreds of individual ships were observed in this span, and we manually annotated multiple images of each ship.

Beyond the typical recognition challenges caused by changes in viewing angle, ships are especially difficult to recognize due to huge variations in lighting conditions and scale. Moreover, because the images were collected from piers, most of the images do not have open water backgrounds. Instead, almost all of them have clutter such as land, buildings, and larger docked ships. For instance, a tall crane on the land behind a cargo ship could be seen as a mast, potentially causing the cargo ship to be misclassified as a sailboat. Furthermore, the inherent structure of ships introduces intra-class similarities that require discriminating features to identify.

2. Related Work

There are a few small datasets that have been created for detecting and recognizing ships. In [24], the authors created a 136-image dataset of ships captured with an infrared camera. This dataset contained six different vessels, with the goal being to recognize the individual ships and not to generalize to ship categories. In [21], the authors created a dataset of 2205 infrared open water images, with 749 of them containing ships. Their goal was to determine whether the image contained a ship, clutter, or an irrelevant object. None of these datasets are publicly available.

Most other existing datasets of *paired* IR and visible imagery are of faces (*e.g.*, [9, 8, 6, 13, 19, 25, 16]). These datasets were created under tightly controlled conditions, *i.e.*, a static pose with no background clutter. The datasets in [13] and [25] have a small amount of pose variation, and [16] has more poses and lighting conditions, but still no background clutter. All of these datasets were acquired with cameras less than six meters from the subject's face. In contrast, the dataset we have constructed contains objects imaged at distances ranging from about ten to several hundred meters and beyond, resulting in enormous variation in image resolution. Additionally, most of these datasets do not contain multiple instances of an individual's face under differing conditions, making them ill-suited for testing multi-class recognition algorithms.

Non-face paired datasets exist, *e.g.* [5, 10], for pedestrians. The dataset provided in [5] consists of one or more moving pedestrians in a small enclosed space. While this dataset has more pose variations and occlusion, its IR imagery has little background clutter, and the cameras are still within ten meters from the subjects. Imagery from [10] has a longer distance from the subjects, but the subjects are not individually identified, which makes the dataset more suitable for tracking than multi-class recognition.

The CVC Multimodal Stereo Datasets are some of the only other non-human datasets to have paired IR and visible-light images [1, 4, 3]. They contain images of roads, buildings, and other urban areas. These datasets were created for studying stereo matching and creating depth maps, and they are not suitable for recognition.

3. The VAIS Dataset

3.1. Apparatus and Data Collection

Over nine days, we captured two terabytes of synchronized image sequences using a horizontal stereo rig, which is shown in Figure 2. The cameras are tightly mounted next to each other and checked to ensure no interference. The images captured were synchronized to retrieve one frame per second from each camera, by matching the two closest microsecond timestamps within each full second. The RGB global shutter camera was a ISVI IC-C25, which captures 5,056×5,056 bayered pixel images (25 MP). For infrared, we used a state-of-the-art Sofradir-EC Atom 1024 camera that captures 1024×768 pixel images, one of the highest resolution LWIR cameras available on the market. The infrared camera has a spectral range of 8–12 microns and uses uncooled ASi microbolometer infrared detector. The complete apparatus was approximately $40,000 (US) to purchase and assemble.

Prior to capturing data, we manually tuned the focus and exposure of the RGB camera to account for environmental conditions and the time of day. The infrared camera had no parameters that could be manually adjusted. Whenever ships appeared, we aimed the rig and recorded until the ships' sizes become insignificant. During the day, we used a monocular to find each ship's name. At night, we discovered ship names by using the MarineTraffic Android application, which shows the locations of nearby ships that are relaying Automatic Identification System (AIS) information in real time. Typically only large commercial vessels are equipped with AIS devices. The ship names are later used for recovering unique instances.

After collecting the synchronized image pairs, the RGB

Figure 2. Our multi-modal stereo camera rig collecting data.

images were debayered using MATLAB's "demosaic" function and a 3×3 median filter was applied to the image to remove pixel artifacts. No image resolution is lost in the debayering process.

3.2. Annotations

For each unique ship in the dataset, we manually drew bounding boxes in the images it appeared in, labeled the ship type, and assigned it the name recorded using the monocular or AIS data. In the rare case when we could not identify the ship's name, we assigned it a short description based on its appearance. Because the images were captured as image sequences at one frame per second, consecutive frames were near-duplicates, which is undesirable to use for classification tasks. To avoid having duplicates of the same ship instance, we do not label every frame that a ship appears in. Only three to five frames of each ship facing each discrete 45-degree orientation were selected. The 45-degree period comes from discretizing a 360-degree rotation into 8 orientations, all of which are possible directions that a ship could be facing. The discretization is done at the annotator's estimation. For most instances, only one orientation was captured; for a few, up to 5 to 7 orientations. This way, we avoid duplicates in the dataset, but still include as many orientations of any given instance as possible. Example bounding box images are shown in Table 1; example pairs are shown in Figure 3.

Bounding boxes with areas smaller than a reasonable threshold (200 pixels) were discarded from the dataset. Since a given IR image has a much lower resolution than its corresponding RGB image, smaller or farther objects may only satisfy the threshold in RGB. After discarding bounding boxes smaller than the threshold, a portion of paired bounding box images were left with only the RGB image, without its IR correspondence. We kept these singleton images in the dataset. All of the RGB images captured at night were discarded, leaving all night images to be IR-singletons.

3.3. Dataset Statistics

The dataset consists of 2865 images (1623 visible and 1242 IR), of which there are 1088 corresponding pairs. There are a total of 154 nighttime IR images. The dataset includes 264 uniquely named ships in 6 coarse-grained categories (or 15 fine-grained categories): merchant ships (26 cargo instances, 9 barge instances), sailing ships (41 small sails up, 21 small sails down, 3 large sails down), medium passenger ships (11 ferry, 4 tour boat), medium "other" ships (8 fishing, 14 medium other), 19 tugboats, and small boats (28 speedboat, 6 jetski, 25 smaller pleasure, 13 larger pleasure, 36 small). The area of the visible bounding boxes ranged from 644–6350890 pixels, with a mean of 181319 pixels and a median of 13064 pixels. The area of the IR bounding boxes ranged from 594–296510 pixels, with a mean of 12249 pixels and a median of 2272 pixels.

We partitioned the dataset into "official" train and test splits. Because we are interested in generalization, we used the names of the ships to ensure that each individual ship was assigned to either the testing or training sets. The number of unique instances in the dataset are counted by ship names as well. To create the train and test splits, we greedily assigned all images from each named ship to either partition, such that the number of images was roughly the same in each partition for all categories. This resulted in 539 image pairs and 334 singletons for training, and 549 image pairs and 358 singletons for testing. All nighttime imagery was assigned to testing, which enables us to measure how well information transfers when object representations are only learned from daytime imagery. This is important because ship traffic is much lower at night, so labeled data is more difficult to gather. All of the categories are represented in the nighttime data.

4. Classification Algorithms

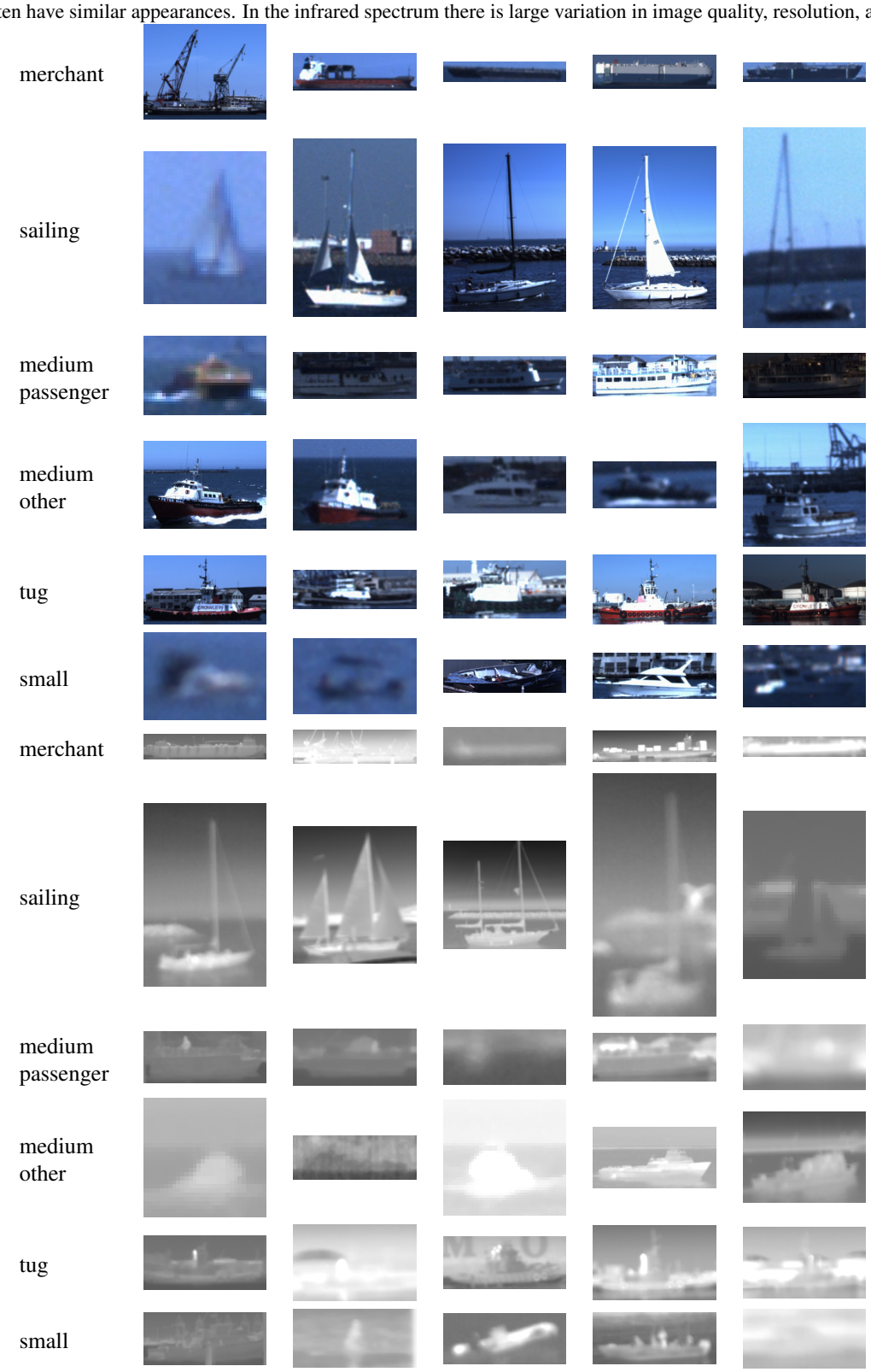
We use two classification algorithms with VAIS: deep convolutional neural networks (CNNs) and gnostic fields.

4.1. Deep Convolutional Neural Networks

Deep CNNs have recently come to dominate object recognition research due to their excellent performance on many challenging datasets harvested from the web, such as ImageNet. Training a large neural network on a small dataset, such as VAIS, would lead to overfitting. To overcome this issue, we use a CNN that has been pre-trained on ImageNet and use it to extract features from VAIS. This works because after being trained on millions of images from one thousand classes, the CNN learns features that are discriminative for object recognition in general.

We used the MatConvNet CNN MATLAB toolbox [23] with the 16-layer CNN from [20] to extract features from images, which achieved excellent results on Image Net

Table 1. Five visible (rows 1–6) and IR (rows 7–12) samples from each of the main categories. Medium passenger, medium other, and small boats often have similar appearances. In the infrared spectrum there is large variation in image quality, resolution, and heat ranges.



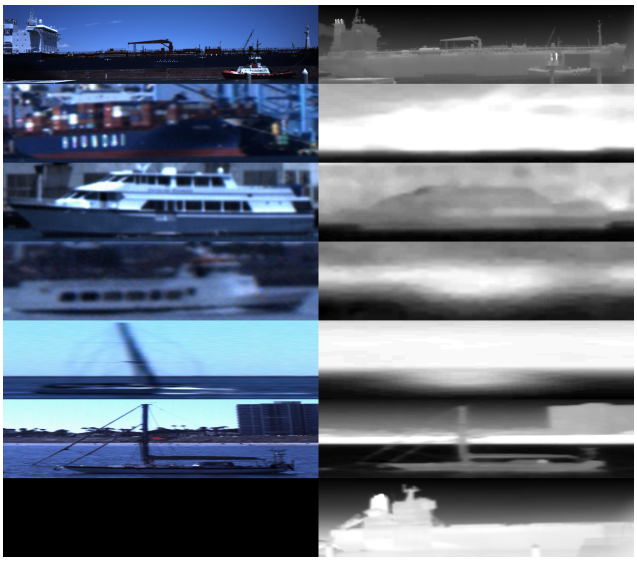


Figure 3. Example VAIS image pairs, with the left being in the visible and the right being infrared. The bottom pair of images are taken in the dark. We have stretched the images to have the same resolution to make them easier to view in this figure, although this distorts their aspect ratio. The quality of the infrared images is highly dependent on distance and differences in temperature.

ILSVRC-2012 dataset. We train a multi-class logistic regression classifier on the output of the 15th weight layer, with all negative values in these features set to zero (ReLU nonlinearity). The final layer of the CNN is specific to recognizing the 1000 ILSVRC categories, so it is not used. The CNN accepts 224×244 RGB images as input. For the VAIS image crops, we resize each crop to this size using bicubic interpolation. To handle the IR images, we simply duplicate the single IR channel three times to create faux RGB images. To train the multi-class logistic regression model, we use the LIBLINEAR toolbox [12].

In preliminary experiments we tried fine-tuning the CNN using backpropagation with VAIS's training images. While others have shown that this can improve results with other datasets such as birds (*e.g.*, [7]), we found that the model performed worse even when regularization techniques, such as drop-out, were used. This is likely because the CNN is very large compared to VAIS.

4.2. Gnostic Fields

To complement our CNN-based approach, we also train Gnostic Fields [15, 14] with SIFT features [17]. A Gnostic Field is a brain-inspired algorithm for recognizing images, and is one of the best non-CNN-based classifiers. While CNNs operate directly on pixels, Gnostic Fields need to operate on an intermediate representation such as SIFT descriptors. A Gnostic Field learns a representation for each category, called a gnostic set, using spherical k-means. When a gnostic field "sees" an image during run-time, each gnostic set compares every descriptor to its learned representation. The output of the gnostic sets is pushed through a competitive normalization function, followed by accumulating information across all of an image's descriptors. The final decision is made by a linear classifier, and we used multi-class logistic regression because it provides probabilities.

Our hypothesis is that gradient-based features, such as SIFT, will be especially useful in the infrared domain. For our infrared camera, the parameters at shooting time are automatically calibrated by its firmware and are not manually adjustable. Image values are automatically rescaled by the firmware from absolute temperatures to a range of $[0, 255]$. The camera automatically recalibrates itself a few times per hour. As a result, the values across IR images taken at different times of the day, or between different calibrations, are not directly comparable. Because of this variation in pixel values across IR images, it makes sense to use gradient-based features that are most sensitive to edge features. We used the dense SIFT implementation in the VLFeat toolbox [22]. For the RGB images, we first converted the images to grayscale prior to extracting SIFT descriptors.

We set SIFT to use 11×11 spatial bins with a step size of 5 pixels. Prior to extracting dense SIFT descriptors for a ship, we cropped out its bounding box image, and then resized it so that its shortest side is 60 pixels while the other side is resized proportionally to preserve the aspect ratio of the bounding box image. These settings produced about 50–700 128-dimensional feature vectors per image. We then augmented the descriptors with a 5-dimensional vector containing spatial location information. This was done by appending a vector $\hat{\ell}_{c,t} = \frac{\ell_{c,t}}{||\ell_{c,t}||}$ to the SIFT descriptor, where $\ell_{c,t} = \left[x_t, y_t, x_t^2, y_t^2, 1\right]^T$ and (x_t, y_t) is the spatial location of grid point $\mathbf{g}_{c,t}$ normalized by the image's dimensions (size) to be between -1 and 1. This yields n 133-dimensional feature vectors for a bounding box image, with n dependent on the size of the bounding box. Subsequently, we used whitened PCA to reduce the dimensionality to 80, which is the same setting used in previous gnostic field research (*e.g.*, [14]).

5. Experiments

All methods used the same fixed training and test sets. We assess Gnostic Fields and CNNs on the night and day data. We also combine the probabilistic outputs of classifiers operating on the IR and visible imagery by averaging them to see if the IR information can enhance recognition during the day. Likewise, we also average the probabilistic outputs of the Gnostic Field and CNN models (denoted Gnostic Field + CNN) to see if the algorithms complement each other. To do this, we use a weighted average with 0.8

Table 2. Daytime mean per-class accuracy on VAIS.

	IR	Visible	IR + Visible
Gnostic Field	58.7%	82.4%	82.4%
CNN	54.0%	81.9%	82.1%
Gnostic Field + CNN	56.8%	81.0%	87.4%

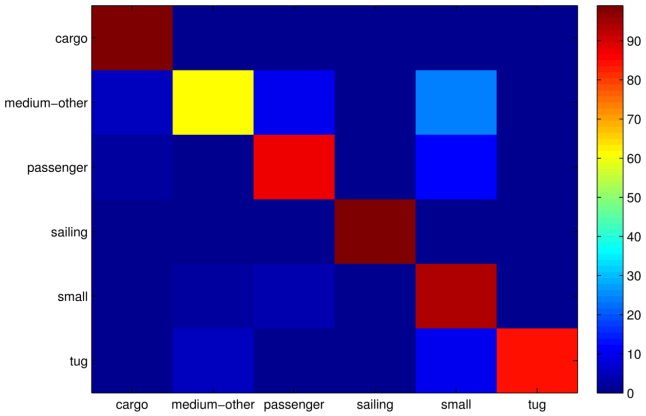


Figure 4. Daytime confusion matrix for the best performing classification model. All categories except for medium-other are above 85% accuracy. Medium-other achieves only 61.6% because it is often confused with passenger and small ships. Cargo and sailing ships are best discriminated with both over 98% accuracy.

Table 3. Nighttime mean per-class accuracy on VAIS.

	IR
Gnostic Field	51.9%
CNN	59.9%
Gnostic Field + CNN	61.0%

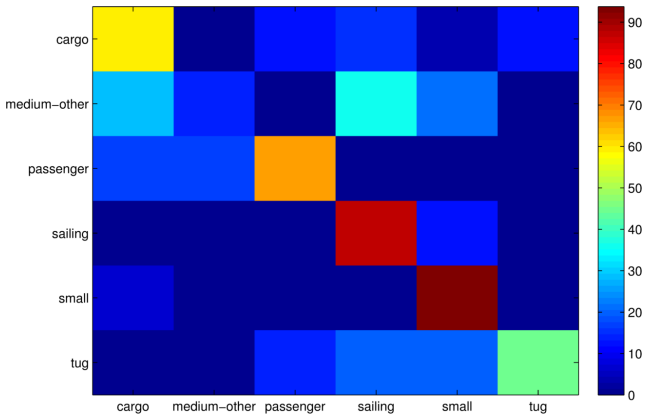


Figure 5. Nighttime confusion matrix for the best performing classification model. As with the daytime results, medium-other is the most confused category and accuracy is very low (14.3%). The most accurate categories are sailing (87.5%) and small (93.8%) ships.

applied to the CNN's output and 0.2 to the Gnostic Field's.

Daytime and nighttime images were tested separately to compare performance. Both daytime and nighttime were trained on the same data; the only difference is in the test data.

Daytime mean-per-class accuracy, *i.e.* the mean of the diagonal of the confusion matrix, for each method is shown in Table 2. As expected, IR performs worse overall compared to visible. This is likely due to the IR camera having significantly lower resolution than the visible camera. The gnostic field and CNN classifiers perform very similarly on the daytime data. Averaging IR with the visible outputs provides little improvement in accuracy for a particular classifier; however, the best performing approach was when all four models were combined (gnostic field on IR, gnostic field on visible, CNN on IR, and CNN on visible), which yielded 87.4% accuracy. The confusion matrix for this model is shown in Figure 4.

Nighttime accuracy for the IR camera is shown in Table 3. The confusion matrix for the best performing method is shown in Figure 5. Accuracy is similar to the daytime IR results.

6. Discussion

We described the VAIS dataset and took important first steps toward making ASVs that can recognize ships and comply with COLREGS a reality. Our results indicate that with a larger dataset and improved algorithms, ship recognition during the day is within reach with current technology, assuming that the ships can be localized. Our IR results suggest that ships can also be recognized moderately well at night, but it is likely that the camera technology will need to be improved before the results are comparable to daytime recognition.

We were able to annotate 16 fine-grained categories in VAIS, but we did not evaluate them here. This was done for two reasons. First, these fine-grained distinctions are not relevant for COLREGs or control. Second, several categories, *e.g.* jetski, were severely underrepresented due to containing too few pixels to be annotated.

Even after nine days of gathering data, VAIS is still relatively small because we were only able to capture ships that were fortuitously traveling nearby. One way to improve this situation is to use transfer learning by augmenting our training data with ship imagery from the web. ImageNet [11] could be used for this purpose, and it already contains labeled images of ships at a fine grain. However, there are many errors in its ship annotations and the majority of images do not have bounding boxes. Moreover, the domain shift (difference) between ImageNet's web images and our dataset gathered "in the wild" is huge in terms of image quality because many images in VAIS are low resolution and exhibit glare or other artifacts. Images from the web often suffer from photographer bias, in that images with more

aesthetic appeal tend to be uploaded. One of our next steps is to try to annotate ImageNet with accurate bounding boxes and remove erroneous images, in order to use it for transfer learning with VAIS.

Ultimately, to make ASVs a reality it will be necessary to collect and annotate images from cameras mounted on moving ships. This poses many difficulties, since a ship's "bouncing" ego-motion means that images captured in this way will have properties somewhat different from VAIS.

Acknowledgments

We wish to thank David Zhu for help setting up the cameras. The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Funding for this work was provided by the Defense Advanced Research Projects Agency (DARPA), under contract #NNN13R513T. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement A: Approved for public release; distribution is unlimited.

References

[1] C. Aguilera, F. Barrera, F. Lumbreras, A. Sappa, and R. Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–12672, Sept. 2012. 2

[2] I. Arnsdorf. Rolls-royce drone ships challenge $375 billion industry: Freight. *Bloomberg*, Feb. 2014. 1

[3] F. Barrera, F. Lumbreras, and A. Sappa. Multimodal stereo vision system: 3D data extraction and algorithm evaluation. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):437–446, Sept. 2012. 2

[4] F. Barrera, F. Lumbreras, and A. Sappa. Multispectral piecewise planar stereo using Manhattan-world assumption. *Pattern Recognition Letters*, 34(1):52–61, Jan. 2013. 2

[5] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi. Thermal-visible registration of human silhouettes: a similarity measure performance evaluation. *Infrared Physics and Technology*, 64:79–86, May 2014. 2

[6] T. Bourlai, N. Kalka, A. Ross, B. Cukic, and L. Hornak. Cross-spectral face verification in the short wave infrared (SWIR) band. In *20th International Conference on Pattern Recognition (ICPR)*, pages 1343–1347, 2010. 2

[7] S. Branson, G. Van Horn, P. Perona, and S. Belongie. Improved bird species recognition using pose normalized deep convolutional nets. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. 5

[8] H. Chang, H. Harishwaran, M. Yi, A. Koschan, B. Abidi, and M. Abidi. An indoor and outdoor, multimodal, multispectral and multi-illuminant database for face recognition. In *IEEE CVPR Workshop, 2006*, pages 54–54, 2006. 2

[9] X. Chen, P. Flynn, and K. Bowyer. Visible-light and infrared face recognition. In *ACM Workshop on Multimodal User Authentication*, pages 48–55, 2003. 2

[10] J. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2–3):162–182, 2007. 2

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 6

[12] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *J Machine Learning Research*, 9:1871–1874, 2008. 5

[13] M. Grgic, K. Delac, and S. Grgic. SCface - surveillance cameras face database. *Multimed. Tools Appl. J.*, 51:863–879, 2011. 2

[14] C. Kanan. Fine-grained object recognition with gnostic fields. In *Proc. IEEE WACV*. 5

[15] C. Kanan. Recognizing sights, smells, and sounds with gnostic fields. *PLoS ONE*, e54088, 2013. 5

[16] S. G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. R. Abidi, A. Koschan, M. Yi, and M. A. Abidi. Multiscale fusion of visible and thermal ir images for illumination-invariant face recognition. *IJCV*, 71(2):215–233, Feb. 2007. 2

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 5

[18] United Nations Conference on Trade and Development. Review of maritime transport, 2013. 1

[19] Z. Pan, G. Healey, M. Prasad, and B. Tromberg. Face recognition in hyperspectral images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1552–1560, 2003. 2

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[21] M. Teutsch and W. Kruger. Classification of small boats in infrared images for maritime surveillance. In *Waterside Security Conference*, 2010. 2

[22] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008. 5

[23] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *CoRR*, abs/1412.4564, 2014. 3

[24] P. Withagen, K. Schutte, A. Vossepoel, and M. Breuers. Automatic classification of ships from infrared (FLIR) images. In *Signal Processing, Sensor Fusion, and Target Recognition VIII*, volume 3720, 1999. 2

[25] B. Zhang, L. Zhang, D. Zhang, and L. Shen. Directional binary code with application to polyu near-infrared face database. *Pattern Recognition Letters*, 31(14):2337–2344, 2010. 2