Image Surveillance Assistant

Michael Maynord
Computer Science Dept.
University of Maryland
College Park, MD 20742
maynord@umd.edu

Sambit Bhattacharya
Dept. of Math & Computer Science
Fayetteville State University
Fayetteville, NC 28301
sbhattac@uncfsu.edu

David W. Aha
Navy Center for Applied Research in AI
Naval Research Laboratory, Code 5514
Washington, DC 20375
david.aha@nrl.navy.mil

Abstract

Security watchstanders who monitor multiple videos over long periods of time can be susceptible to information overload and fatigue. To address this, we present a configurable perception pipeline architecture, called the Image Surveillance Assistant (ISA), for assisting watchstanders with video surveillance tasks. We also present ISA_1, an initial implementation that can be configured with a set of context specifications *which watchstanders can select or provide to indicate what imagery should generate notifications. ISA_1's inputs include (1) an image and (2) context specifications, which contain English sentences and a decision boundary defined over object detection vectors. ISA_1 assesses the match of the image with the contexts by comparing (1) detected versus specified objects and (2) automatically-generated versus specified captions. Finally, we present a study to assess the utility of using captions in ISA_1, and found that they substantially improve the performance of image context detection.*

1. Introduction

Maritime watchstanders perform critical security tasks: they monitor ship movements and suspicious activity for potential threats, conduct communication checks, and perform a variety of related duties [7]. They can benefit from some video surveillance tools, but these usually require monitoring multiple streams for long durations. This risks information overload and fatigue. This is particularly true if the tools have a high incidence of false alarms, or are otherwise not well-aligned with the watchstander's objectives.

Ideally, an intelligent decision aid would allow a watchstander to provide (and dynamically revise, as needed) a specification that describes the contexts for which they seek notification. For example, this could include the entities/objects, relations, activities, and other scene elements of interest, as well as the contextual conditions under which these are of interest (e.g., time of day). The decision aid, having been trained to recognize these elements, and having been given (or learned) models for interpreting these conditions, would then operate on video streams selected by the watchstander.

We present an abstract perception pipeline architecture for such a tool, called the *Image Surveillance Assistant* (ISA). ISA culls uninteresting images from the input stream, and presents imagery to a watchstander only when it is of predicted interest. It does this by comparing input imagery against predefined, user-selected contexts. For example, a watchstander may wish to be notified of mechanical failures, security breaches, or the appearance of certain objects in a monitored scene.

ISA combines top-down and bottom-up processes using a hierarchy of components that operate on data structures at different abstraction levels. It accepts as input a watchstander's specifications, and uses these to modify the configuration of its mid-level components, biasing them to be sensitive to patterns that are relevant to the watchstander's selected contexts. ISA's bottom-up processing of image content employs deep learning (DL) techniques, whose output is then interpreted for context matching. However, while this assists with constraining bottom-up perception, many options for this interpretation task exist.

In this paper, we assess how well automatically-generated image captions assist with image interpretation

for an initial ISA implementation named ISA_1. To do this, we compare ISA_1's performance when it uses, to a varying degree, (1) a set of trained Support Vector Machines (SVMs) for object detection and (2) a Long-term Recurrent Convolutional Network (LRCN) [5] for caption generation, combined with a semantic distance metric. While the first is an obvious choice, the second is less so, and leverages recent progress on caption generation. Our evaluation suggests that these captions can be valuable for automated detection of contexts in image surveillance tasks.

We briefly summarize related work in Section 2. Section 3 describes our conceptual approach. Section 4 then details ISA_1's implementation. We present an example of its use in Section 5 and our empirical study in Section 6. Finally, we discuss future work plans in Section 7 before concluding.

2. Related Work

Waterfront security concerns actions taken to defend against threats to maritime assets (e.g., facilities, vessels). it is of great interest to the DoD, which is encouraging a unified approach to protect maritime assets [4]. For example, SPAWAR developed the Electronic Harbor Security System (EHSS), which integrates electronic sensors and video systems to detect, assess, track, and archive capabilities for waterside surface and subsurface threats. EHSS includes the use of video surveillance, but does not address issues of watchstander information overload.

Many maritime surveillance systems exist. They vary according to several dimensions [1], such as their type of coverage (e.g., some provide aerial coverage using satellite photography, whereas we focus on ground-based sensors) and model category (e.g., some track vessels worldwide, whereas we focus on a local area, such as a harbor). Perhaps the most successful are those that perform perimeter defense; they trigger alerts when a perimeter is breached [14]. Others focus on specific types of surveillance tasks, such as chokepoint surveillance [15]. However, they are not designed to allow watchstanders to dynamically specify their contexts of interest, or use current-generation techniques for image processing and interactive interpretation.

This paper is an extension of our earlier proposal [18]. Here we describe a broader conceptual architecture, an initial implementation, and an empirical study on the utility of automatically generated captions for context prediction. Our group has also studied the use of several artificial intelligence (AI) techniques for maritime threat assessment, such as probablistic graphical models [3] and plan recognition [2]. However, this is our first use of DL techniques for scene recognition, and on the task of using contexts as a focus for triggering watchstander notifications.

We are using DL techniques to support symbolic inferencing tasks. Several other researchers are likewise examining integrations of DL and AI techniques for computer vision tasks. For example, Doshi et al. [6] are using Convolutional Neural Networks (CNNs) to help create episodic memories of video scenes that are matched to previous episodes to generate predictions (e.g., of objects that will appear in the near future). As another example, one among several that use Long-Short Term Memories (LSTMs) to automatically generate captions, Venugopalan et al. [19] describe a system which operates on video input and learns languge models that can generate natural, grammatical sentences. However, we are not aware of other groups that are studying the use of integrated DL and AI techniques for maritime surveillance tasks.

3. ISA Conceptual Architecture

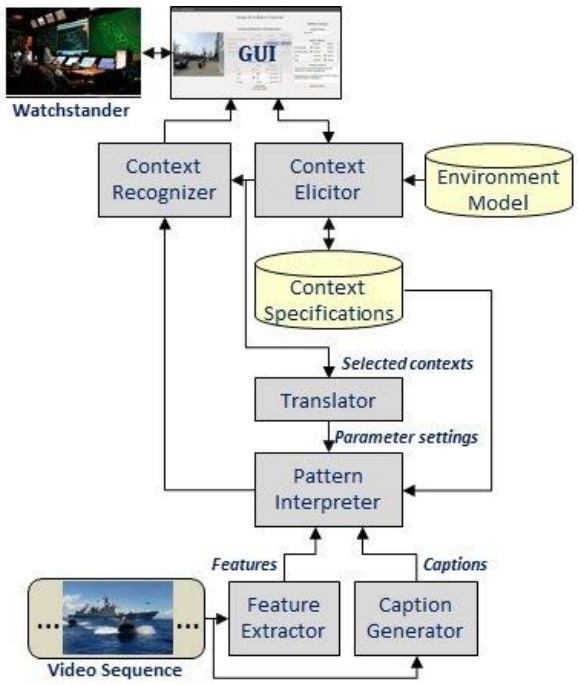

Figure 1. Conceptual Architecture for the Image Surveillance Assistant (ISA)

Figure 1 displays our vision for ISA's architecture, including its modules, data sources, and their relation. ISA conducts a top-down process that constrains the pipeline in accordance with specifications elicited from the watchstander, and a bottom-up process for imagery interpretation. The following paragraphs describe this abstract architecture, which we instantiate in Section 4.

The top-down process is intended to begin with the Context Elicitor prompting the watchstander (via a GUI) with a set of pre-defined contexts (whose encodings are stored in Context Specifications). The watchstander could select a subset of these, and also construct (with the assistance of the Environment Model, which could contain a variety

of semantic information) and store encodings for additional contexts. Selected contexts are made available to the Translator, which can modify parameter settings used by the Pattern Interpreter that are relevant to those contexts. This allows the Pattern Interpreter to be sensitive to patterns that are relevant to detecting watchstander-specified contexts.

We envision that a Context Specification will include at least two data structures. First, it will contain (e.g., feature) descriptions for each context. These could be used by the Context Recognizer to test whether the input imagery is a good match for a given context. Second, a Context Specification will include a set of exemplar captions for each context. These could be used by the Pattern Interpreter to match with a caption generated from the input imagery.

The bottom-up process includes two initial steps. First, a Feature Extractor will extract a set of features for the Pattern Interpreter, which could use them to assess whether they predict the appearance of an object, action, or other scene element in the input imagery. Second, a Caption Generator will automatically produce a sentence annotation that, in part, may describe valuable relations among scene elements. The Pattern Interpreter could compare this with a context's exemplar captions (mentioned above) to assess the degree to which they match.

The results of these interpretations will then be provided to the Context Recognizer, which is also provided with information on the contexts selected by the watchstander, and their encodings. Given these inputs, its task is to predict which (if any) contexts are *active* in the current imagery, and communicate this to the watchstander via display/notification in the GUI.

4. Prototype

Section 3 described the full ISA architecture. We now describe an initial ISA implementation, ISA_1 (Figure 2). ISA_1 includes most but not all of the full architecture's components and functions.

Using ISA_1's GUI, watchstanders can select one or more of four pre-defined contexts, or define new contexts. The four contexts we predefined and encoded into ISA_1 are:

1. bar_OR_pub_indoor
2. bathroom
3. computer_room
4. parking_lot_OR_street

We selected these contexts because we were able to obtain images that correspond to them from the SUN Image Corpus [21]. Also, these contexts can be distinguished using the 80 object categories defined in the Microsoft COCO dataset [13]. Finally, the choice of four contexts suffices for our initial study (Section 6).

ISA_1's Context Recognizer leverages a set of logistic regression (LR) models that are included in the Context Specifications. We trained these LR models over object detection vectors produced by SVMs. ISA_1's Context Specifications also include a set of exemplar captions per context. (See Section 6.2 for details.)

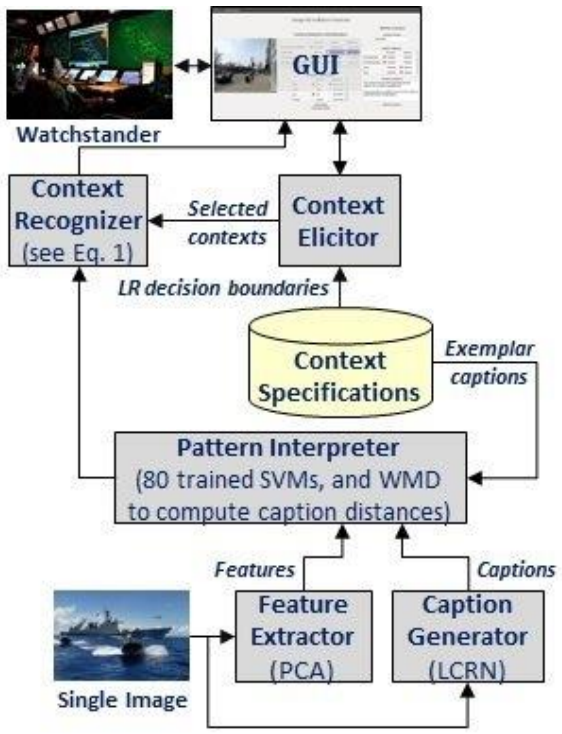

Figure 2. ISA_1, an Initial Implementation of the ISA architecture

The LR models for the four contexts are made available to the Context Recognizer while the exemplar captions are made available to the Pattern Interpreter.

ISA_1 takes as input individual images, rather than video. These images are fed through two modules: a PCA compression module that performs feature extraction, and a LRCN to generate captions. We chose to use PCA here because it is fast and required little time to integrate for this first implementation of ISA.

ISA_1's Pattern Interpreter takes as input the output of the PCA and LRCN modules. It applies the set of 80 trained SVMs on the features extracted by PCA, and sends their predictions to the Context Recognizer.

This Pattern Interpreter also computes the Word Mover's Distance (WMD) [12] of the LRCN-generated caption to each exemplar caption of each context, and passes these distances to the Context Recognizer. WMD is analogous to Earth Mover's Distance [16], which determines the similarity between two probability distributions by calculating the effort involved in moving the "earth" of one distribution such that it matches the other distribution. In WMD sentences are represented as a set of points in a Euclidean space semantic embedding. WMD defines the similarity of

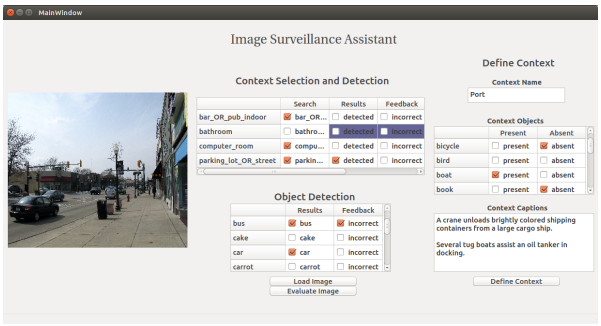

Figure 3. A Screenshot of ISA_1's GUI

two sentences as the effort involved in transporting the set of points associated with one sentence onto the set of points associated with the other sentence. WMD is desirable in this case as it captures semantic distance.

The Context Recognizer takes as input the output of the Pattern Interpreter's object class detections (using the trained SVMs) and the computed caption distances. It applies the LR models provided by the Context Elicitor to the object class detections, and a nearest neighbor classifier (1-NN) to the caption distances, to predict which (if any) contexts are active in the input image. Finally, the Context Recognizer combines these predictions using Equation 1, where $al_i(\vec{o}, C)$ is ISA_1's predicted activity level for context i given object detection vector $\vec{o}$ and caption distances C. $b(\vec{o})$ denotes the distance of $\vec{o}$ from the LR model's decision boundary, $e(C)$ denotes whether 1-NN's predicted context is i, and $\alpha \in [0, 1]$ is a tunable parameter. We define a context to be *active* when $al_i(\vec{o}, C) > 0.5$.

$$al_i(\vec{o}, C) = \alpha \times b(\vec{o}) + (1 - \alpha) \times e(C) \quad (1)$$

This is a simple and constrained first implementation of ISA. We discuss future extensions of it (e.g., alternatives to using PCA for feature extraction) in Section 7.

5. Example

In this section we describe an example use of ISA_1, whose GUI (Figure 3) includes three columns. The first contains the image on which it is operating, the middle takes user input and presents output, and the third allows defining of contexts.

A watchstander/user can define a new context specification as follows. The name for the new context is provided in the text field labeled *Context Name*. Within *Context Objects* is a list of the 80 objects over which a linear decision boundary is defined. For each object, the user can specify whether the object's presence or absence is associated with the context, for which they can check the *Present* or *Absent* boxes, respectively (or check neither). The user can also insert sentences into *Context Captions* that exemplify the context being defined. Note that in defining a new context the user does not provide any example images of that context - the user provides only a description of the context, one of a form that is easy for humans to understand and interact with. In Figure 3, a user is defining a new context called *Port*. The user specified that the presence of a *boat* object indicates that the context is active in an image, whereas the presence of *bicycle* and *book* indicate that the context is not active, and the presence of *bird* is neutral. The remaining 76 objects become visible when scrolling. The user has provided two example sentences for the Port context: "A crane unloads brightly colored shipping containers from a large cargo ship", and "Several tug boats assist an oil tanker in docking".

The middle column contains two buttons. The Load Image button loads a new image, while the Evaluate Image button runs ISA_1's evaluation process over the curent image and displays the results. Within *Object Detection* is a list of the 80 COCO objects, and a column for *Results*; its boxes are checked for those objects that ISA_1 detected in the image. *Context Selection and Detection* contains a list of all pre- and user-defined context specifications, and the *Search* and *Result* columns. (The Feedback columns in this column are notional, as ISA_1 does not adjust its configuration based on correcting feedback.) The user can check a Search column's box to indicate a context of interest, in which case ISA_1 should evaluate input with respect to this context specification (to determine whether it is active). When the user presses *Evaluate Image* the cells of the Result column will be in one of three states: grayed out, ungrayed and unchecked, or ungrayed and checked. A cell is grayed out if it is associated with a context specification for which the user indicated no interest, by leaving that context specification's Search checkbox unchecked. An ungrayed unchecked box indicates that ISA_1 evaluated the image with respect to the associated context specification, and predicted the context is not active. Similarly, for an ungrayed and checked box, ISA_1 predicted the context to be active.

Figure 3 shows that the user selected an image of a street, and indicated interest in all but the bathroom context. ISA_1 predicted that, among the contexts the user selected, only the parking_lot_OR_street context is active in the image. All context detections are correct. Among the first four objects listed in *Object Detection*, both *bus* and *car* were detected.

6. Evaluation

6.1. Hypothesis

In this section we evaluate the hypothesis that, for at least our study, ISA_1's use of automatically generated captions can increase its performance, where the task is to predict the active context of a given image, the image belongs to

exactly one context, and our metric is the F-score.

6.2. Datasets and Training

ISA_1's LR models were produced as follows. First, we collected 2910 images from the bar, pub indoor, bathroom, computer room, parking lot, and street scene categories of the SUN Corpus. After grouping into our four context categories, we randomly selected 40% of the images from these context categories to serve as a test set (totaling 1165 images), while the remaining images serve as a training set (totaling 1745 images). We then trained a separate SVM for each of COCO's 80 object categories using COCO images (i.e., because COCO's images are labeled according to these object categories whereas the SUN corpus does not have object labels), where we used PCA to extract features that we then standardized such that each feature had $\mu = 0$ and $\sigma = 1$. We similarly standardized the vector of SVM predictions. Next, for each context we trained a LR model on object vectors obtained by applying the SVMs to positive and negative examples selected from the training set we derived from the SUN corpus.

We generated a set of ten exemplar captions per context as follows. (1) We trained the LRCN model of [5] on COCO images. (2) We then applied it to the 1745 images of our SUN Corpus training set to generate 500 captions per context. An LRCN combines a trained CNN [11] with a Long-Short Term Memory (LSTM) [8], which is a recurrent neural network that can represent temporal patterns and produce captions for images or image sequences. (3) For each context's image captions, we then applied WMD to each pair of captions to produce a distance matrix among captions. For each caption in that matrix, we computed its sum of squared distances to the other 499 captions. (4) Finally, we selected, per context, the ten captions with the smallest sum of squared distances among them. These were included in ISA_1's Context Specifications. Table 1 presents a subset of the ten exemplar captions employed for each context.

6.3. Evaluation Method

To evaluate the extent to which the context exemplar caption method impacts context prediction, we recorded whether ISA_1's context predictions (on the 1165 images of the test set) were correct. Our independent variable was α in Equation 1, which we varied from 0 to 1 by increments of 0.05. When $\alpha = 1$, ISA_1 uses only the SVM's learned decision boundaries to predict/detect an image's context. In contrast, all lower values for α in our experiment correspond to increasing reliance on comparing a context's captions with the image's predicted caption.

For each of the four contexts, and each value of alpha, we applied the LR models, SVMs, and LRCN of ISA_1 to the test images of that context, determined context detections using Equation 1, computed precision and recall, and then combined these into an F-score. We averaged the F-scores across all four contexts, and report this below.

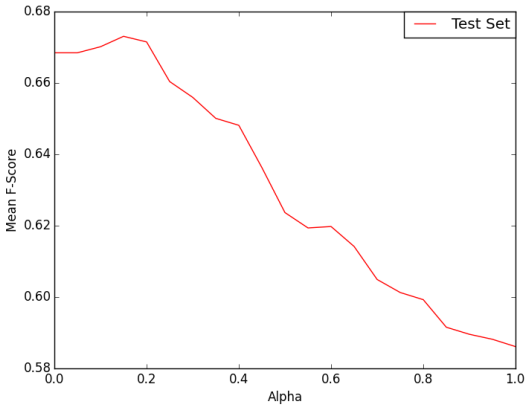

Figure 4. Mean F-scores as a function of α. When ($\alpha = 0$), only the generated captions are used for prediction, whereas when ($\alpha = 1$), only object detection predictions are used. These results indicate that using the generated captions can increase performance.

Table 2. Context Confusion Matrix for Caption Matching

		Predicted Context				
		1	2	3	4	ALL
True Context	1	198	57	178	67	500
	2	41	398	57	4	500
	3	117	78	287	18	500
	4	47	24	19	410	500
	ALL	403	557	541	499	2000

6.4. Results

ISA_1's context prediction performance is summarized in Figure 4. For our study, we found that its optimal performance is attained when $\alpha \approx 0.2$, where the predictions of the SVMs' and 1-NN (on caption matching) were employed. ISA_1 also recorded a comparatively high mean F-score when $\alpha = 0$, indicating that using only caption matching compares well to a joint method. Furthermore, its performance was much higher than when using only object detections for context prediction (i.e., when $\alpha = 1$). Thus, these results support our hypothesis (Section 6.1).

Table 2 displays a confusion matrix for our caption comparison method. As shown, the performance of caption matching varies with the context, and some contexts (e.g., a computer room and a bar or pub) are more easily confused than others. We conjecture that the computer_room and bar_OR_pub_indoor contexts are easily confused as they both involve indoor scenes that do not consistently contain the unique identifying objects that a bathroom scene contains.

Table 1. Example captions used for each of the four contexts used in our empirical study.

Context	Subset of 10 Exemplar Captions (Per Context)
bar_OR_pub_indoor	"A stove with stove and pans around it." "Large sun lights on display as a restaurant decorated."
bathroom	"A toilet that is in a bathroom with a curtain." "A tiny bathroom scene with a sink and toilet."
computer_room	"This is a woman sitting at a computer desk with two laptops." "A man plays a game with nintendo wii at television."
parking_lot_OR_street	"Cars and cars driving along a city street." "There is a view of a busy intersection of the city street."

7. Future Work

As mentioned, ISA_1 is a simple first implementation of ISA. We highlight a few future research directions here.

User Interaction: Future ISA versions will extend ISA_1's ability to dynamically define a new context, or alter an existing one, at any point during system operation. This is particularly important because a watchstander may want to modify the system's behavior as the monitored situation unfolds. We plan to extend this to scene elements as well, where the watchstander could provide imagery and additional information (e.g., annotations, features) to teach ISA new scene elements through a process of iterative refinement, where the system could use active learning techniques to prompt the watchstander for their feedback on predictions of the presence of these newly defined elements in new scenes. More generally, ISA should leverage watchstander feedback (e.g., highlighting false positives and negatives) on system performance. This could be used to automatically modify the system's configuration (e.g., models used in the Pattern Interpreter or Context Recognizer).

Feature Extraction: We will replace the object detectors of ISA_1, consisting of a set of SVMs applied to PCA compressed images, with a Regional CNN object detection method [9]. This should produce state-of-the-art image features. This approach is also better suited to detect objects in images that contain many objects per scene, and it will extract additional information from the input (e.g., object positions, spatial relations). Alternatives to LRCN for caption generation will be explored, such as [20], and [17] which functions over video.

Imagery: While ISA_1 is limited to single images, we will extend its scope to work with video. To do this, we will incorporate recent advances in processing video to extract features (e.g., [10]) and generate captions (e.g., [22]).

Pattern Interpretation: ISA_1's Pattern Interpreter uses SVMs. Future versions will instead leverage more sophisticated techniques, such as HMMs, GMMs, and scripts to represent temporal relations. The Pattern Interpreter will also take input from a Translator module to assist with mapping context specifications to these representations. This could assist watchstanders with monitoring processes of interest.

Caption Evaluation: ISA_1's method for evaluating captions to determine context detections assumes that each image belongs to exactly one context. We will relax this assumption, and are considering an extension that applies a GMM model to the space of captions, or applies maxium likelihood estimates to context-specific caption distance distributions. This will require ensuring the training set includes images that belong to multiple contexts.

Tuning: We will include methods to automatically tune system parameters, such as the number of exemplars to be included in a context specification's exemplar set and the threshold value used in Equation 1.

User Study: Finally, we will evaluate the effectiveness with which users are able to define and detect novel contexts using ISA_1.

8. Conclusion

In this paper we introduced a novel video surveillance architecture, the Image Surveillance Assistant (ISA), which is intended to assist watchstanders with identifying imagery that is of interest to them. This may be particularly important in the conditions of information overload or fatigue. This architecture embodies three key characteristics: incorporation of both connectionist and symbolic representations, a compositional hierarchy, and top-down information flow. In addition to the architecture we introduced a limited proof-of-concept implementation, ISA_1, and described its evaluation to assess whether automatically generated image captions can improve context detection performance. Our results provide some support for this hypothesis. However, this is an initial study, several topics remain to be addressed in future work, and we discussed some of these in Section 7. Finally, we are collaborating with Navy reservists to obtain their feedback on this system, and to demonstrate it to potential users.

References

[1] B. Auslander, K. M. Gupta, and D. W. Aha. A comparative evaluation of anomaly detection algorithms for maritime video surveillance. In *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2011.

[2] B. Auslander, K. M. Gupta, and D. W. Aha. Maritime threat detection using plan recognition. In *Proceedings of the Conference on Technologies for Homeland Security*, pages 249–254. IEEE Press, 2012.

[3] B. Auslander, K. M. Gupta, and D. W. Aha. Maritime threat detection using probabilistic graphical models. In *Proceedings of the Twenty-Fifth Florida Artificial Intelligence Research Society Conference*, 2012.

[4] DoD. Security engineering: Waterfront security. Technical Report UFC 4-025-01, Department of Defense, Washington, DC, 2012.

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[6] J. Doshi, Z. Kira, and A. Wagner. From deep learning to episodic memories: Creating categories of visual experiences. In *Proceedings of the Third Annual Conference on Advances in Cognitive Systems*, 2015.

[7] L. Guenther, D. Sine, and A. Sine. Sector command center watchstander structure front end analysis (fea) report. Technical Report SURVIAC-TR-2006-185, Survivability/Vulnerability Information Analysis Center, Wright-Patterson AFB, Ohio, 2006.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[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 International Conference on Multimedia*, pages 675–678. ACM, 2014.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. IEEE, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[12] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *ICML*, 2015.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[14] A. Lipton, C. Heartwell, N. Haering, and D. Madden. Critical asset protection, perimeter monitoring, and threat detection using automated video surveillance. In *Proceedings of the Thirty-Sixth Annual International Carnahan Conference on Security Technology*. IEEE Press, 2002.

[15] B. A. McArthur. A system concept for persistent, unmanned, local-area arctic surveillance. In *SPIE Security+ Defence*. International Society for Optics and Photonics, 2015.

[16] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*, pages 59–66. IEEE, 1998.

[17] R. Shetty and J. Laaksonen. Video captioning with recurrent networks based on frame-and video-level features and visual content classification. *arXiv preprint arXiv:1512.02949*, 2015.

[18] L. Smith, D. Bonanno, T. Doster, and D. W. Aha. Video surveillance autopilot. In *CVPR Scene Understanding Workshop*, 2015.

[19] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence–video to text. *arXiv preprint arXiv:1505.00487*, 2015.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[21] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the Conference on Computer vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[22] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and F. Li. Every moment counts: Dense detailed labeling of actions in complex videos. *CoRR*, 2015.