Image Surveillance Assistant Architecture: Status and Planned Extensions

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

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

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

Abstract

We describe our research on integrating deep learning with artificial intelligence techniques in the context of an imagery surveillance prototype designed to automatically identify imagery of interest to a user. In particular, we briefly overview the Image Surveillance Assistant (ISA) architecture, present a study focusing on ISA_1, ISA's preliminary implementation, and discuss our plans for future extensions. Our plans include two foci: the more expressive semantic and perceptual representations and processing we will use, and user interaction processes.

1 Introduction

Watchstanders are tasked with monitoring their environment for potential threats. However, watchstanders are constrained by information overload and fatigue - humans are capable of monitoring only a fixed-width input, and only for a certain amount of time before fatigue degrades performance. As such, watchstanders benefit from tools that can reduce information overload and fatigue by partially automating the task of surveillance.

We previously introduced the Image Surveillance Assistant (ISA) architecture [Maynord *et al.*, 2016], an architecture which culls from an input stream input which is of no interest to the operator, and thus eases the constraints of information overload and fatigue. It does this by matching input against *context specifications* - operator provided descriptions of the situations (or contexts) for which the operator wishes to receive notification.

ISA is composed of a hierarchy of modules, which correspond to a hierarchy of representations of increasing levels of abstraction. Towards the bottom, raw input is fed through detectors and through a caption generation module. Towards the top, the operator configures the system through a GUI using high level concepts. ISA stands in contrast to architectures engineered for a specific task in that it is highly adaptable and quickly deployable.

Additionally, we previously introduced ISA_1 [Maynord *et al.*, 2016], a proof of concept implementation of the ISA architecture. Towards the bottom of ISA_1, objects are detected using simple SVM detectors, and captions are produced using a Long-term Recurrent Convolutional Network (LRCN) [Donahue *et al.*, 2014]. Towards the top, contexts are defined using a decision boundary over object detections, and a set of exemplar captions.

ISA_1 effectively demonstrated the feasibility and utility of the ISA architecture. However, ISA_1 is a preliminary implementation. We present in this paper future extensions to ISA_1. These extensions are centered around two foci: internal representations and processing; and user interaction. The representations which we here introduce greatly expand the capabilities of ISA_1 to detect context specifications with greater accuracy and nuance. The alterations to user interaction which we present improve the fluidity of user interaction, which is needed particularly as the internal representations and processes of ISA_1 become more sophisticated.

This paper extends our work in [Maynord *et al.*, 2016] - the overview presented on the ISA architecture in Section 3, and on the first implementation, ISA_1, in Section 4, as well as the illustration of use given in Section 5, is a summarization of the descriptions provided in that paper. In Section 2 we

cover related work. In section 6 we cover future extensions, with 6.1 detailing extentions to internal representations and processes, and 6.2 detailing extensions and modifications to user interaction. Finally, we conclude in Section 7.

2 Related Work

Work on ISA is contextualized within a desire to improve detection of threats to maritime assets - this is of great interest to the Department of Defense [DoD, 2012]. A variety of maritime surveillance systems exist; they vary in multiple ways [Auslander *et al.*, 2011], e.g., type of coverage (aerial vs. ground based sensors), and model category (modeling of movements across the globe vs. within a single harbor). Two common types of systems are those that trigger warnings when a perimeter is breached [Lipton *et al.*, 2002], and those that monitor chokepoints [McArthur, 2015], or limited areas such as harbors [DoD, 2012]. However, unlike ISA, these systems do not allow the operator to dynamically configure the system to be sensitive to contexts of interest.

This paper is an extension of our work in [Maynord *et al.*, 2016], which is in turn an extension of our proposal in [Smith *et al.*, 2015]. Additionally, our group has studied artificial intelligence (AI) methods for maritime threat assessment, including probabilistic graphical models [Auslander *et al.*, 2012a], and plan recognition [Auslander *et al.*, 2012b]. However, [Maynord *et al.*, 2016] was the first point at which we merged Deep Learning (DL) and AI techniques for computer vision. To our knowledge, we are the only group studying a unified approach spanning both DL and AI for the purposes of maritime surveillance.

On the more general topic of using DL in support of symbolic AI tasks including inference, there has been increasing interest. [Doshi *et al.*, 2015] makes use of Convolutional Neural Networks (CNNs) in the construction of episodic memories of video scenes which are then used in generating future predictions (such as objects that will appear). Additionally, there has been substantial work recently on caption generation for images [Donahue *et al.*, 2014], and to a lesser extent for video [Donahue *et al.*, 2014], [Venugopalan *et al.*, 2015]. There has been interest in more tightly integrating processes of vision and reasoning [Aditya *et al.*, 2015], [Wang and Yeung, 2016], [Maslan *et al.*, 2015], with a recent area of particular interest being visual question answering [Antol *et al.*, 2015], [Yang *et al.*, 2015], [Zhang *et al.*, 2015].

3 ISA Conceptual Architecture

The full ISA architecture is depicted in Figure 1. Information in ISA flows in two directions: top-down, and bottom-up. In the top-down direction, context specifications - definitions of scenarios of interest provided by the operator - are used in biasing the ISA's perceptual pipeline - the series of modules through which imagery data is processed. In the bottom-up direction, representations for imagery of successively higher levels of abstraction are constructed until the level at which the operator interacts with ISA is reached.

The watchstander starts the top-down process either by selecting from a set of pre-defined context specifications (which are stored in Context Specifications) through interacting with the Context Elicitor via the GUI, or by constructing and storing new context specifications, with the aid of the Context Elicotor, in accordance with the constraints contained within the Environment Model.

The Translator is provided those context specifications for which the operator has indicated an interest. The role of the Translator is to "translate" context specifications into system parameterizations - increasing the sensitivity of the Pattern Interpreter to imagery properties of relevance to the provided context specifications (for example, if certain objects are of particular concern in the provided contexts, the precision/recall balance of detections for these objects can be altered - increasing recall at the cost of precision).

In the bottom-up process input is abstracted by a Feature Extractor, such as a CNN, to extract a set of features that are useful for tasks such as detection of objects, attributes, and general image properties (e.g., is the provided input of a reflection of a night scene, rather than a day scene?).

Input is also fed through a Caption Generator, which produces English descriptions of the input. English image captions are powerful in their expressiveness - using captions allows us to represent a large range of image properties (such as relations between entities) without the need to employ more formal structured representations. However, the expressivity of automatically generated image captions comes at the cost of precision.

The pattern interpreter "interprets" the features produced by the Feature Extractor and the captions produced by the Caption Generator in accordance with the parameterizations specified by the Translator, and passes this interpretation to the Context Recognizer. The role of the Context Recognizer is to come to a determination of which, if any, of the contexts in which the operator indicated an interest are active, given the interpretation provided by the Pattern Interpreter, and the list of context specifications of interest as provided by the Context Elicitor. The determinations of the Context Recognizer are then provided to the operator via the GUI.

4 ISA_1 Prototype

ISA_1 is a proof of concept implementation of the ISA architecture, which includes most of the components and functions of the full ISA architecture. ISA_1 provides the watchstander with the ability to select one or more of four pre-defined context specifications, as well as the ability to define and refine novel context specifications. The choice of the four predefined contexts was a practical decision based on the availability of images in the SUN Image Corpus [Xiao *et al.*, 2010] that correspond to those contexts, as well as the observation that the 80 object categories in the Microsoft COCO dataset [Lin *et al.*, 2014] can be used to distinguish these contexts.

The context specifications of ISA_1 includes a set of exemplar captions per context, and logistic regression (LR) models trained over object detection vectors produced by 80 SVMs corresponding to the 80 object categories of MSCOCO. The Pattern Interpreter takes as input the features computed by a PCA module, and an image caption produced by an LRCN. It then applies the 80 SVM object detectors, and compares semantic distances between the generated caption and exemplar

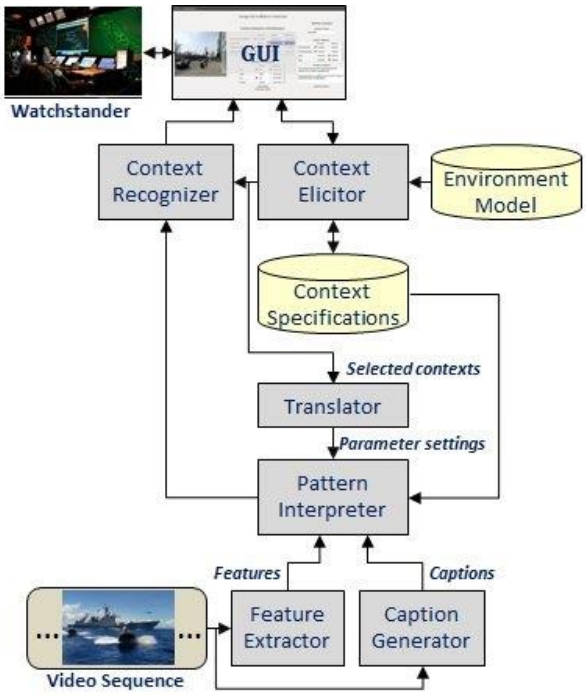

Figure 1: Conceptual Architecture for the Image Surveillance Assistant (ISA)

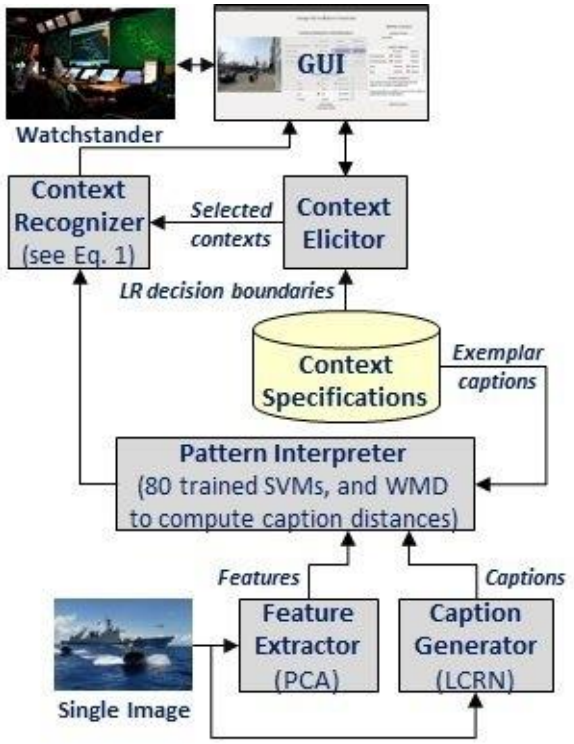

Figure 2: ISA_1, an Initial Implementation of the ISA architecture

captions of each context specification using Word Movers Distance (WMD) [Kusner *et al.*, 2015] as a distance metric. The WMD is good choice since it captures semantic distance between two sentences. The Pattern Interpreter then passes the resulting detection vectors and caption distances to the Context Recognizer.

The Context Recognizer applies the LR models to the object detection vectors, and applies a nearest neighbor (NN) classifier to the caption distances, to predict which contexts are active in the input. ISA_1 uses a weighted sum that combines and balances the predictions of the LR models and NN applied to captions to compute an activation level for each context. This activation level is then compared to a fixed threshold, which if exceeded indicates that the context is active.

See [Maynord *et al.*, 2016] for an evaluation of performance.

5 Example

The GUI of ISA_1 is shown in Figure 3 and has three columns. The first column shows the image, which is being processed, the middle takes user input and provides output, and the third allows the user to define contexts. A watchstander defines a new context specification by interacting with widgets to provide a context name, selecting which objects are present and absent in the context, and by entering exemplar sentences for the new context. A new image is loaded and evaluated by interacting with the middle column. Under results, the detected objects and detected contexts are presented.

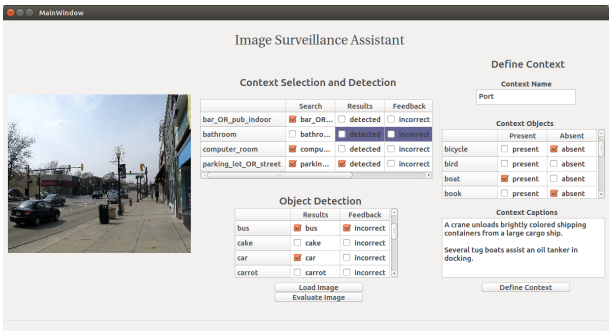

Figure 3: A Screenshot of ISA_1's GUI

6 Future Extensions

ISA_1 is a simple initial implemention of our vision for a comprehensive surveillance support system. We plan several extensions, and discuss two types here: (1) more comprehensive representations, and associated processes, for expressing user contexts and (2) more expansive user interaction processes.

6.1 Perceptual and Semantic Representations and Processes

Our definition of "context" is complex. The representations and processes we use in ISA_1 are a first step towards capturing a context, but are limited. Extending ISA_1 with a wider variety of representations and processes will enable processing contexts with finer granularity.

Consider that many contexts may be differentiated from similar situations by differences that the object detections of ISA_1 are not able to capture and which ISA_1's caption based comparisons could miss. For examples, a "danger of fire" context could be defined by the operator to be any situation where flammable material is in close proximity to a flame. To competently handle this context the representations internal to ISA_1 would need to be expanded to include a representation for flame, a representations for common flammable objects, and a representation for proximity relations.

With this objective in mind, we will extend ISA_1 as follows:

Objects We will extend beyond the initial 80 MS COCO objects we have used, as well as introduce simple attributes for describing them.

Scenes We will use probabilistic graphical models to represent scenes as a set of objects, their attributes, and (temporal and spatial) relations among them.

Context evaluation over detections We will test whether methods for learning probabilistic classifiers over vectors of detections, other than via logistic regression (e.g., SVMs), are preferable for distinguishing contexts.

Caption matching We will replace our nearest neighbor algorithm for caption matching with a Gaussian Mixture Model (GMM) approach.

We discuss the latter three in more detail in the following subsections.

Scenes: Spatial Relations

We represent objects and their attributes as atoms. Symbols can be attached to these atoms and manipulated by symbolic methods, and as discrete entities we can learn statistical relations among them. These relations could be used in the lower to mid-levels of ISA to help interpret and match imagery with contexts. (Currently, relations among entities are not explicitly represented in ISA_1.)

Probabilistic graphical models [Koller and Friedman, 2009] can be used to explicitly represent relations among low and mid-level entities, such as co-occurrence. For example, a conditional random field (CRF) can more robustly capture and represent co-occurence relations (e.g., among a knife, a fork, and a spoon) than relations learned implicitly using a neural network. These relations can then be used to regularize the results of detections, and improve detection accuracy - CRFs are not infrequently used in a regularizing role [Russell *et al.*, 2009], [Mann and McCallum, 2010].

We will also consider the use of scene graphs [Johnson *et al.*, 2015], which are designed to represent an image's objects, attributes, and their relations. To provide a grounding for full and partial scene graphs over images, we will use a CRF formulation that maximizes the likelihood over possible groundings [Koller and Friedman, 2009]. We could map a context specification to a scene graph representation, and using this mechanism to evaluate the match of an image with a context's scene graph. This will then constitute one more interpretation mechanism - in addition to decision boundaries over detection vectors and semantic comparison of produced and exemplar captions - which ISA can leverage for input evaluation.

Introducing probabilistic graphical models into the mid-level of ISA is, while not necessarily easy, in principle straightforward, and we will pursue this. Use of scene graphs is more difficult, and more consideration of how best to integrate them is needed before determining whether to include them.

Scenes: Temporal Relations

In addition to spatial relations, knowledge of temporal relations can help detect contexts in input. To illustrate, consider the following temporal relation of observations: person A is carrying backpack B; A is sitting on a bench next to B; person C is sitting next to A and B; C is carrying B. In isolation, each observation is innocuous, but when put in temporal relation to each other, the significance of these collective observations becomes apparent. Temporal relations can be captured using probabilistic graphical models. Hidden Markov Models (HMMs) [Rabiner and Juang, 1986] are well-suited for expressing temporal relations among observations. This use of temporal relations as features for context evaluation will make transitioning input from individual images to image sequences (or videos) more tractable. As ISA_1 is expanded to operate over input that extends through time, we intend to employ HMMs.

Context Evaluation Over Detections

Currently, ISA_1 uses boundaries learned using logistic regression (LR) to match object detection vectors with context specifications (where each specification is associated with a distinct LR model). This has the advantage that logistic regressors, unlike many classifiers, are probabilistic classifiers - they provide an associated confidence with their classification prediction. This confidence is important for ISA_1, as it is this confidence, rather than the binary classification, that ISA_1 uses for context matching. However, LR is limited in that its decision boundary is linear. Classifiers other than logistic regression can provide information on classification confidence [Wu *et al.*, 2004] (e.g., SVMs [Platt and others, 1999]). We will test them in ISA_1, where we expect they will provide an advantage over LR models.

Caption Matching

ISA_1 compares generated captions against context exemplar captions using a semantic sentence distance metric. ISA_1 then uses a nearest neighbor algorithm (1-NN) to evaluate which context specification is active in the given input. The advantage of this approach is that 1-NN is straightforward to implement and often performs reasonably well in comparison to more sophisticated methods. However, it assumes that each input belongs to (precisely) only one context. Additionally, 1-NN generates a classification, but not an associated confidence, which would be valuable.

A more nuanced method for caption matching should produce degrees of activity associated with each context specification for a given input. 1-NN could be extended to provide a measure of confidence [Cheetham, 2000] through a function of the ratio of distances between the nearest example of

the predicted class and the next-nearest example of a different class. However, because of the imprecision of captions there is significant overlap between the distribution of captions of different contexts, and so the confidences produced by such a method may not be reliable (the more the overlap, the less consistent the ratio of the distances of the examples of the closest two classes becomes). A more stable caption to context metric is needed. One complicating factor is the potential need for context-specific similarity functions (consider that some contexts may be more "tightly" defined than others) or even asymmetric similarity functions (e.g., a change in one "semantic direction" may be associated with a larger change in context similarity than a change in another semantic direction).

Thus, the semantic clusters associated with each context specification may not be of similar sizes or symmetric. One potential solution to this challenge is to match a Gaussian Mixture Model (GMM) [Bilmes and others, 1998] over all context labeled captions, where each Gaussian distribution within that model is associated with a single context specification. This will permit calculating the degree to which an automatically-generated caption is associated with each context specification, which then counts as evidence towards the activity of each context in the input.

However, GMMs operate in Euclidean space. The present 1-NN implementation in ISA_1 compares generated and exemplar captions directly, and relies on an imprecise distance metric. As such, ISA_1 does not contain a Euclidean space in which captions are represented as single points, nor are the distances between captions likely to be precisely representable in a Euclidean space. However, there exist semantic embeddings for sentences, such as [Socher *et al.*, 2014] (constructed, in particular, for sentences with "visually grounded meaning"). This represents entire sentences (as opposed to individual words, for which many semantic spaces are constructed) as a single point in a high dimensional Euclidean space, where the location of the point in that space carries a semantic meaning that approximately matches the semantic meaning of the sentence. The GMM may be constructed on captions represented in this semantic space, which will allow comparision with automatically-generated captions with the Gaussian distributions of the context specifications.

Given the advantages which GMMs in a Euclidean semantic space may provide, we intend to pursue them.

6.2 User Interaction

In Section 6.1 we detailed how more expressive and more precise representations are important for extending ISA_1. However, finer granularity is not the only desired property for an extended ISA_1 implementation. Our objective is for ISA to reduce operator burden by partially automating some surveillance tasks, given an operator-provided context specification. In order for ISA to meet this objective it must be capable of interpreting operator provided specifications in terms of its internal representations - this puts constraints on the form which context specifications can take. As ISA_1's internal representations become more sophisticated, the more sophisticated the context specifications become, and the more important is the process through which ISA_1 guides the operator in selecting, defining, or refining context specifications.

In this section we detail extensions to ISA_1 which aid the operator in more effectively interacting with ISA_1, particularly as the sophistication of the system increases. We cover the following:

GUI extensions More information will be displayed to the operator, and the operator will have the option of providing more precise constraints. This requires a modification to the nature of the GUI to maintain ease of use.

Interpretation feedback As the nuance of the interpretations of ISA expands, so does the utility of human oversight. We allow the operator to aid ISA in its interpretations, while maintaining ease of use.

Defining novel context specifications With more sophisticated context specifications, comes the need for an alternative method of novel context specification definition - we outline such a method.

Active refinement of context specifications We introduce an iterative approach to defining context specifications, allowing gradual refinement.

Online refinement of context specifications We allow context specifications to be updated on the fly, outside of the more formal active refinement process.

GUI Extensions

The current interface is depicted in Figure 3. Object detections are presented to the user via a series of check-boxes. We will change this such that object detections are shown on the image itself, using labeled bounding boxes, where the color of the bounding box corresponds to the confidence of the object detection. The GUI will also display boxes for more objects, as well as attributes and relations.

A desirable property for an automated surveillance tool such as ISA to possess is *transparency*; if ISA determines that a given context is active in the input, it should be made clear to the operator what factors underly that determination. Often, at least some principal factors concerning a context specification's match can be displayed using annotations overlaid on the input, in a manner similar to bounding-box annotations for detections (of objects, attributes, and relations). However, not every relevant factor may be easy displayed as an overlay on an image. For example, the relations based on reasoning over generated captions are not straight-forward to elucidate through input annotations, and will require a different presentation format for the operator.

Operator Feedback on Context Predictions

In addition to providing additional feedback to the operator, the extended GUI will permit the operator to refine or correct ISA interpretations. This includes feedback on detections, their bounding boxes, and associated confidences. Inevitably, the methods on which ISA rellies for detections will produce errors (i.e., both false positives and false negatives), and this will decrease context detection accuracy. Allowing the operator to correct errors will increase accuracy, though a balance must be identified between operator burden and expectations of operator input.

Defining Novel Context Specifications
We will provide users a more comprehensive ability to define context specifications in our extension of ISA_1. They will be able to refer to the additional objects, attributes, and relations mentioned earlier, although how best to elicit such specifications from the operator may require additional analysis, and will depend on the nature of relations that ISA can accurately perceive.

More radically, we will extend ISA_1 such that novel context specifications can be "seeded" with a small set of positive and negative examples (i.e., imagery). Generating approximate context specifications from a small set of examples is in principle straight-forward. Vectors of detections derived from those examples can be used as training instances for supervised learning to produce a decision boundary for context evaluation. Exemplar captions can be generated by selecting prototypical captions from a set of automatically generated captions.

Active Refinement of Context Specifications
Defining context specifications by providing examples can ease the operator's burden, particularly as the form of the ISA context specification is extended to be more sophisticated. However, because the size of the set of examples provided by an operator will by necessity be small, the context specifications derived from that set will be approximate. This motivates the need for a process that can be used to refine an initial context specification.

Future ISA prototypes will allow an operator to engage in iterative refinement of context specifications, which are currently defined by the operator in such a way that the operator must infer how ISA will interpret them. Presently, for the operator to evaluate how ISA interprets a context specification the operator must select input and observe ISA's behavior. This requires the operator to infer which input will be informative for ISA interpretation. This is problematic in that it places a new burden on an operator and it depends on an operator's understanding (or more likely intuition) of how ISA behaves.

To address this, we will enable ISA to automatically select input that is informative with regards to the validity of its interpretation of the novel context specification. This can be achieved by selecting from an input library inputs that are near the decision boundary for an "active" determination of the novel context specification. That is, ISA will present to the operator scenarios in which it has low certainty of whether the operator intends the context to be considered active. The operator will then be invited to provide feedback on whether ISA's prediction is correct and, if not, provide feedback as to *why* ISA's prediction is incorrect. This feedback can be used to refine the context specification.

Online Refinement of Context Specifications
Similar to how input interpretations can be modified online according to user feedback, context specifications can also be modified online. This process will be similar to iteratively refining context specifications during definition except that the input will be taken from the input stream, rather than selected by ISA. Note that input from a small time window will likely only be representative of a small portion of the input space over which the operator desires the context specification to perform well. To deal with this, the weight that ISA gives to operator feedback provided online will be lesser than that given to feedback provided during the process of active refinement. Online refinement is useful, but not as powerful as active refinement.

7 Conclusion

We provided an overview of the Image Surveillance Assistant architecture, an architecture for reducing the constraints of information overload and fatigue which confront watchstanders. ISA is highly adaptable and quickly deployable. We provided an overview of ISA_1, a proof of concept implementation, and discussed at length future extensions to ISA_1. These extensions centered around more sophisticated representations and processes, and more fluid operator interaction. Representation and process extensions will include representations for more objects, attributes, and their relations, as well as probabilistic graphical models to constrain the relations between these, and more sophisticated mid-level interpretation mechanisms, including interpretation of detection vectors and image captions. Operator interaction extensions include an altered and expanded GUI, the enabling of feedback for input interpretation, an alternate mechanism to define context specifications based on a "seed" of examples, and iterative and online refinement of context specifications.

References

[Aditya *et al.*, 2015] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. Visual common-sense for scene understanding using perception, semantic parsing and reasoning. In *2015 AAAI Spring Symposium Series*, 2015.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[Auslander *et al.*, 2011] Bryan Auslander, Kalyan Moy Gupta, and David 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.

[Auslander *et al.*, 2012a] Bryan Auslander, Kalyan M Gupta, and David W Aha. Maritime threat detection using probabilistic graphical models. In *Proceedings of the Twenty-Fifth Florida Artificial Intelligence Research Society Conference*, 2012.

[Auslander *et al.*, 2012b] Bryan Auslander, Kalyan Moy Gupta, and David W Aha. Maritime threat detection using plan recognition. In *Proceedings of the Conference on Technologies for Homeland Security*, pages 249–254. IEEE Press, 2012.

[Bilmes and others, 1998] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter

estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[Cheetham, 2000] William Cheetham. Case-based reasoning with confidence. In *Advances in case-based reasoning*, pages 15–25. Springer, 2000.

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

[Donahue *et al.*, 2014] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[Doshi *et al.*, 2015] Jigar Doshi, Zsolt Kira, and Alan 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.

[Johnson *et al.*, 2015] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3668–3678. IEEE, 2015.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[Lipton *et al.*, 2002] A.J. Lipton, C.H. 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.

[Mann and McCallum, 2010] Gideon S Mann and Andrew McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *The Journal of Machine Learning Research*, 11:955–984, 2010.

[Maslan *et al.*, 2015] Nicole Maslan, Melissa Roemmele, and Andrew S Gordon. An integrated evaluation of perception, interpretation, and narration. 2015.

[Maynord *et al.*, 2016] Michael Maynord, Sambit Bhattacharya, and David Aha. Image surveillance assistant. In *Computer Vision Applications in Surveillance and Transportation: Papers from the WACV-16 Workshop*, 2016.

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

[Platt and others, 1999] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[Rabiner and Juang, 1986] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[Russell *et al.*, 2009] Chris Russell, Pushmeet Kohli, Philip HS Torr, et al. Associative hierarchical crfs for object class image segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 739–746. IEEE, 2009.

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

[Socher *et al.*, 2014] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[Venugopalan *et al.*, 2015] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence–video to text. *arXiv preprint arXiv:1505.00487*, 2015.

[Wang and Yeung, 2016] Hao Wang and Dit-Yan Yeung. Towards bayesian deep learning: A survey. *arXiv preprint arXiv:1604.01662*, 2016.

[Wu *et al.*, 2004] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.

[Xiao *et al.*, 2010] Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, Antonio 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.

[Yang *et al.*, 2015] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Neural self talk: Image understanding via continuous questioning and answering. *arXiv preprint arXiv:1512.03460*, 2015.

[Zhang *et al.*, 2015] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. *arXiv preprint arXiv:1511.05099*, 2015.