Smith, L.N., Bonanno, D., Doster, T., & Aha, D.W. (2015). Video surveillance autopilot. In J. Hays, A. Khosla, S. Savarese, & J. Xiao (Eds) *Scene Understanding: Papers from the CVPR Workshop*. Boston, MA.

Video Surveillance Autopilot

Leslie N. Smith[1], David Bonanno[1], Tim Doster[2], and David W. Aha[1]
[1]Naval Research Laboratory; Washington, DC
[2]NRC Postdoctoral Researcher; Naval Research Laboratory; Washington, DC
{leslie.smith, david.bonanno, tim.doster, david.aha}@nrl.navy.mil

1. Introduction

Navy security watchstanders can be overwhelmed with the task of monitoring the content of multiple video streams for rare or important events. For example, if stationed on a ship in a busy port location, identifying those activities that require additional attention may benefit from familiarity with normal port operations. However, many activities may take place nearby (e.g., refueling operations, water taxi transportation, and leisure activities), and the constant monitoring of videos (and other sensors) can be taxing.

We describe our plans for developing a *Video Surveillance Autopilot* (VSA), a software tool for automating surveillance tasks by providing watchstanders with tailored notifications and summaries of video content. Tremendous progress in deep learning and cognitive systems are making it possible to create such a system.

We describe our vision of a VSA system. Given a video stream, the VSA (Figure 1) will consist of three primary components: (1) a text Annotator, (2) a Scene Interpreter, and (3) a User Interface (UI). Their developments pose substantial research challenges.

2. Approach

The VSA should infer and communicate the presence or absence of unexpected or otherwise important entities, activities, or relations of interest to a watchstander. To do this, it will decode videos into constrained natural language (NL) text descriptions of the scene's entities, activities, and their relations. The following sections describe the components of our approach.

Text annotation: Several researchers have proposed processes for automated caption generation from images or video [2, 6, 8, 9]. Most use deep convolutional neural networks (CNNs) to encode image data and some type of recurrent neural network (RNN) to decode this into an NL sentence. For example, long-short-term memory (LSTM) RNNs have performed well on this decoding task. These approaches have been applied to images and videos, but typically to generate *unconstrained* text annotations.

We seek to constrain the generated text so that it highlights content of pre-specified interest to the watchstander. That is, we wish to focus the input on only relevant parts of the video and *bias* the output of annotation. For example, if the watchstander needs to be informed of specific threats (or threat types) as they evolve, or be notified that particular activities have been completed, then these constraints need to be provided to the system at an appropriate level of abstraction.

To support this capability in the Scene Interpreter, the deep CNN encoding and LSTM decoding need to be trained on a large corpus of examples of the types of (normal and abnormal) activities that need to be brought to the watchstander's attention. In addition, the Annotator needs to focus on a vocabulary for the types of entities, activities, and other scene elements relevant to the watchstander's task. This bias can reduce the input to the VSA because most surveillance video content is background (both spatially and temporally) and is not relevant to the watchstander's task. Thus, we are investigating methods that incorporate these biases to focus the VSA's attention on the watchstander's information needs.

Scene interpretation: The Scene Interpreter must process the generated text in the context of the watchstander's information needs. For example, it must assess whether highlighted activities are abnormal or constitute a threat, and generate notifications accordingly. In addition, it should process a constrained set of queries from the watchstander and reply appropriately, provide explanations for alerts when requested, and recommend response actions.

This requires models of the watchstander's information needs along with entities and activities of interest. Thus, the VSA's World Model will contain models of expected (e.g., procedures for ship defense during refueling operations) and known unexpected entities and activities (e.g.,

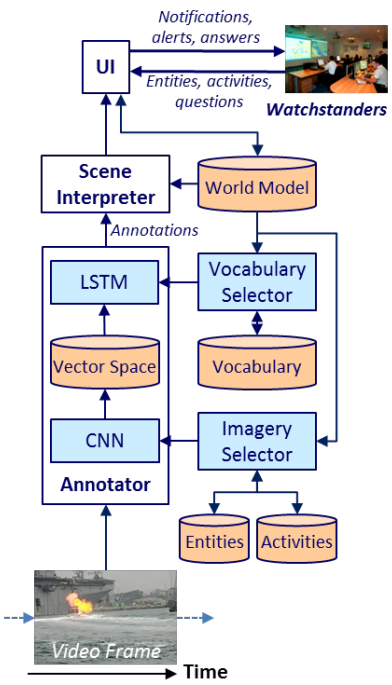

Figure 1: VSA Conceptual Architecture

violations of port navigation rules), along with their expectation probabilities as a function of environment conditions (e.g., location, time). We will also model some known abnormal or suspicious entities and activities (e.g., broaching of security perimeters around Navy assets), annotated such that these models can be used to generate explanations to the watchstander should they be observed.

Our initial representations will include manually-defined frames and related symbolic representations (e.g., scripts, plots, narratives) [3], which encode entities, activities, and the possible intentions of interacting agents. This will require training our Annotator with imagery of these entities and activities, and mapping them to selected vocabulary. These representations will be amenable to inference; we will use them for activity recognition, to confirm expected normal events, and to identify suspicious events. We will also include methods to predict the effects of recognized activities so that the VSA can assess those effects along with the watchstander's information needs, and react accordingly (e.g., by providing early warnings). In later work, we will attempt to learn these representations using unsupervised methods.

User interface: The UI will allow a watchstander to identify the types of communications they seek from the VSA. The watchstander will be able to request notifications on the status and completion of specific activities (or activity types), alerts on abnormalities, corresponding explanations, and direct the VSA to answer specific questions. For example, these questions may pertain to static attributes of when an entity was detected, prior occurrences of an activity, why an activity is deemed abnormal, the system's confidence in a classification of a specific predicted activity, and the effects of an ongoing activity once complete. The watchstander will be presented with a continuous interpretation of the scene and be able to point and click to answer questions.

3. Status

Our current focus is on translating surveillance video input into text descriptions. Among other tasks, we are investigating how to automatically determine the relevant parts of the video and how to train each component with a bias to detect and recognize entities and activities of interest.

We chose Caffe for our CNN implementation; we are using both the 16 layer VGG architecture and the GoogleNet architecture that were top competitors in the ImageNet 2014 competition because higher-performing classification architectures improve the results of image description generation. In addition, in our exploratory research, we found that increasing and decreasing learning rates can increase classification performance.

The training of the CNN, LSTM, and World Model is important for recognizing entities and activities of interest. Initial training of the CNN with ImageNet data provides a good initialization of the lower layers' weights and we are fine tuning the weights on relevant datasets (e.g., VIRAT[1]). In addition, we are using several other datasets (e.g., generated from movie Descriptive Video Service (DVS) [4, 9]) to train the integrated CNN-LSTM networks to maximize recognition accuracies.

We are also investigating the use of attention [8, 9] and preprocessing to spatially and temporally reduce the size of the CNN's inputs so as to reduce its computational requirements and improve the VSA's performance.

Since we are focused on activities from video, we have adopted the state-of-the-art two-stream convolutional network approach [5]. In this approach one CNN detects entities in select frames and another one, trained on multi-frame dense optical flow, classifies short term motion. On top of this we are investigating use of late fusion and an LSTM for longer term motion and activity recognition.

Although we choose an LSTM implementation of an RNN, teams from DeepMind and Facebook recently published innovations in RNN architectures that outperform LSTMs [1, 7]. In the future we plan to assess these other architectures for how reliably each can be biased to recognize activities of interest.

References

[1] Graves, A., Wayne, G., and Danihelka, I. Neural Turing machines, arXiv preprint, arXiv:1410.5401 (2014).

[2] Karpathy, A., and Li, F.-F. Deep visual-semantic alignments for generating image descriptions. arXiv preprint arXiv:1412.2306 (2014).

[3] Mataeas, M., and Sengers, P. Narrative intelligence. Proceedings of the AAAI Fall Symposium on Narrative Intelligence (pp. 1-10) (1999).

[4] Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B.. A dataset for movie description. arXiv preprint arXiv:1501.02530 (2015).

[5] Simonyan, K., and Zisserman, A. Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems (2014).

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

[7] Weston, J., Chopra, S., and Bordes, A. "Memory Networks", arXiv preprint, arXiv:1410.3916 (2014).

[8] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015).

[9] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., and Courville, A. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. arXiv preprint arXiv:1502.08029 (2015).

[1] http://www.viratdata.org/