The 2016 AAAI Fall Symposium Series:
Artificial Intelligence for Human-Robot Interaction
Technical Report FS-16-01

# Automated Surveillance from a Mobile Robot

**Wallace Lawson, Keith Sullivan, Esube Bekele,**
**Laura M. Hiatt, Robert Goring, J. Gregory Trafton**
Naval Research Laboratory, Washington DC

## Abstract

In this paper, we propose to augment an existing video surveillance system with a mobile robot. This robot acts as a collaborator with a human who together monitor an environment to both look for objects that are out of the ordinary as well as to to describe people found near these objects. To find anomalies, our robot must first build a dictionary describing the things that are normally seen while navigating through each environment. We use a computational cognitive model, ACT-R/E to learn which dictionary elements are normal for each environment. Finally, the robot makes note of people seen in the environment and builds a human understandable description of each individual. When an anomaly is seen, the robot can then report people recently seen, as they may be potential witnesses or people of interest. Our system operates in real-time, and we demonstrate operation on several examples.

Video surveillance is a simple and effective way to monitor large environments for potential threats. In such systems, a person monitors a number of cameras mounted around a facility. They make note of anybody present, while also looking for suspicious actions or objects. Unfortunately, the effectiveness of these systems can be limited (Smith 2002). Monitoring a large environment requires a lot of cameras, and it is difficult to watch all of the video feeds at the same time. Many of these video feeds contain nothing of interest, and such tedious, repetitive tasks can be difficult for a human to perform effectively (Hockey 2013). Further, if something suspicious is seen, it is not always straightforward to fully understand what is happening from a 2D video screen.

*Automated surveillance* can improve the effectiveness of video surveillance systems by adding additional support for detecting and understanding events. For example, such systems can look for abandoned objects anywhere in the view of the camera and notify an operator when this event has occurred. This alleviates the cognitive load on the human, and ensures that high priority events are not missed. In this paper, we propose to augment such systems with an automated surveillance robot, Octavia (see figure 1 (Breazeal et al. 2008)). Octavia can support the human's surveillance efforts, alerting the human when events of interest occur. The mobile nature of a robot also has the secondary advantage of being able to move about freely and monitor any ad-hoc area as need. This allows it to monitor areas that may not be well-covered by cameras, as well as inspecting areas around suspicious events with much greater scrutiny than what might be possible using fixed cameras.



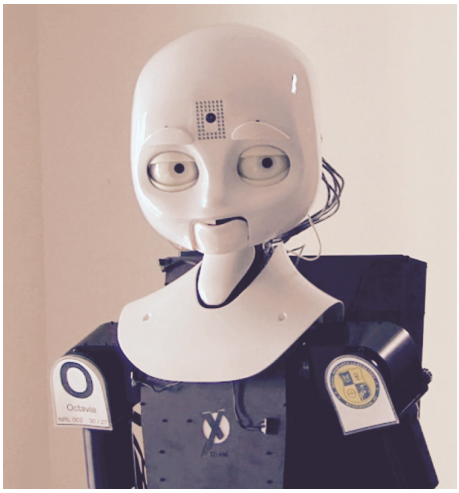

Figure 1: Our MDS Robot, Octavia, operates as PatrolBot

In addition to her patrol duties, Octavia should interact in several different ways with the people around it. She should accept instructions from a human teammate or supervisor, and report back with a status report after its task is complete. As a key part of this, the robot needs to be able to describe the people it near anomalous events in a way that a human teammate can understand.

Artificial intelligence techniques are essential to accomplishing both of Octavia's functions: detecting anomalies on a building-wide scale, and describing the people she sees in a human-like way. In this paper, we first describe our approach for learning about the environment and reporting anything that is out of the ordinary. The robot first learns a normative model of the environment by navigating through it and extracting features from small regions in each image. As our features, we cluster deep features from the AlexNet convolutional neural network architecture (Krizhevsky, Sutskever, and Hinton 2012), which allows her to learn its normative model in an unsupervised fashion. At the same time, the

images are also annotated with their location (Zhou et al. 2014). When detecting anomalies, this location is used to provide context to anomaly detection. Context is used both distinguish between similar objects as well as locate things that might be normal for one environment whereas they are anomalous in another. For example, while a bag might be normal in an office, it may be seen as highly suspicious in a corridor or outside of a building.

We next move to describing how Octavia utilizes work in artificial intelligence to enable natural, human-like descriptions of people she encounters during her patrol. To accomplish this, we again leverage the AI technique of deep learning to extract ten different biometric features of human appearance, including gender, clothing and hairstyle. The primary challenge in training a deep network to provide such description is how to make use of data that is highly biased and not large enough to train network that is very deep. We resolve this issue by building "shortcuts" into our networks which bypass multiple layers in the network (He et al. 2015), effectively permitting the network to learn better features using less data. We also have a customized loss function, which allows us to tolerate highly biased data.

Together, these two capabilities, built upon common artificial intelligence techniques, greatly enhance Octavia's capabilities to patrol and ability to interact with a human supervisor. We support this claim by describing a working demonstration of Octavia as PatrolBot in action. We then conclude with a brief discussion and future extensions.

## Anomaly Detection

Our overall approach to anomaly detection is to build a model of what is normal for each environment. During patrol, this model is queried to compare what is seen against what is normal. Anomaly detection proceeds as follows (Lawson, Hiatt, and Sullivan 2016): to properly determine context, we first must determine the location of the robot using the PlacesCNN (Zhou et al. 2014). This CNN includes a wide range of places such as "kitchen", "conference room", and "corridor". Figure 2 includes several examples of labeled scenes from the places database.

In parallel, the robot continuously analyzes the environment using patches extracted from the camera images. The patches are of a fixed size $NxN$ with a step-size of $N/4$ (in our experiments, $N = 256$ for an image of size $640x480$), meaning that there is an intentional overlap in order to both properly handle objects that straddle the boundary of multiple patches as well as to see objects that are larger than a single patch. Each patch is represented by features from a deep convolutional neural network; a combination of a cognitive model and per-environment dictionary of common items is then used to determine if the image patch is anomalous of not.

During training, the robot patrols each environment collecting data to construct a dictionary of patches that normally appear within each environment. In practice, this results in an enormous amount of data (in our experience, we generate 1.5 million patches each time the robot goes through all environments), so we use a streaming clustering algorithm which requires only a single pass through the

Figure 2: Examples of locations encoutered from the robot. In this case the PlacesCNN classified the locations as top row:"waiting room", "kitchenette" and 'bottom row:"corridor".

data and can add new clusters during runtime. Each patch $p$ is evaluated by a deep neural network (using the "AlexNet" architecture, fully trained on ImageNet) to generate a feature vector $v$ for the patch from the last layer in the network. This feature vector is then compared to the current clusters $C$, and, if the patch is sufficiently close (using Euclidean distance) to an existing cluster, it is added to that cluster. If the patch $v$ is not similar to an existing cluster, a new cluster is created. Using this approach, PatrolBot constructs a dictionary for each environment, and new features are added as they are encountered.

Based on this dictionary, we construct a normative model of context from the computational cognitive architecture ACT-R/E (Trafton et al. 2013). Context in ACT-R/E takes the form of associations between concepts (here, the dictionary of features and environment location): as the robot traverses the world, associations between environments and features are strengthened based on how frequently a feature appears within the environment. After training, features that are typical for an environment have strong associations while features that are atypical for an environment have weak associations.

To determine the correct dictionary element, the robot takes both the distance to the patch and the location into consideration. We weight each observed patch by the association strength from the cognitive model $c_{ka}$ for each dictionary element $k$ in scene $a$:

$$p_k = \exp(-d_k/\sigma) \tag{1}$$

$$k(a) = argmin_k(p_k c_{ka}) \tag{2}$$

Figure 3: Sample images used to train the convolutional network for attribute recognition.

where $d_k$ is the distance to cluster $K$, and $\sigma$ is a parameter that can be estimated from the expected distribution of the data. A patch is marked as anomalous when $k(a)$ falls above a threshold.

## Person Attribute Detection

When the robot encounters a potential anomaly, it first looks up to check if the anomaly is related to a person (e.g., a person is standing nearby). If a person is seen, the robot greets them and generates a description for later use.

Our person attribute detection system uses a modified residual deep convolutional neural network with an optimization (i.e., gradient loss function) to account for imbalances in the training data. In deep learning, researchers noticed that adding additional layers to the their networks increased accuracy, especially in difficult datasets. However, as the network depth increased, the learning process became more difficult, as the error gradients within the network started to either vanish or explode, resulting in the inability of the network to converge. While several techniques sought to solve this problem, another problem arose: degradation of training accuracy. The introduction of residual blocks solved this degradation problem by making deep networks as easy to solve as shallower networks. An upshot of this approach is a decreased amount of training data (i.e., no need for data augmentation tricks) leading to decreased training time.

One of the biggest problem in training a deep network to learn pedestrian attributes is the highly unbalanced nature of the data. Datasets will have hundreds, perhaps thousands, of negative examples and only a few positive examples. For example, an attribute as simple as "wearing a red shirt" may only be present in 5% of the data. When presented with an image, a naive classifier can then just always predict "False" and produce an accurate of 95%! To solve this problem, we need to assign more weight to those examples. To do this, we changed the loss function to incorporate a weighting of positive and negative samples differently so that predicting the rare example wrong is penalized with higher cost. Finally, attribute recognition is inherently a multi-label learning problem. Multiple pedestrian attributes may be present in an example image and there is inherent relationships among attributes (Zhu et al. 2015).

In person attribute detection, a single image typically contains multiple attributes, thus, a multi-label loss function is needed to learn the relationship between attributes. Assume we are looking for $M$ attributes to describe a person, then the loss function is

$$TotalLoss = \sum_{m=1}^{M} \gamma_m loss_m \tag{3}$$

where $loss_m$ is the contribution of the $m^{th}$ attribute to the total loss and $\gamma_m$ controls the contribution of attribute $m$ to the total loss. This is useful for situations where particular attributes are more important than others (e.g., if we find determining gender is more important than recognizing shirt color). While this can be tuned towards human prefers, in our work, we assume all attributes are equally important, so set $\gamma_m = \frac{1}{M}\ \forall m$.

The individual losses are computed using a sigmoid binary cross-entropy for each attribute (see Eqn. 4). This loss function accounts for imbalances of positive and negative examples within an attribute using the weight $w_m$.

$$loss_m = -\frac{1}{N}\sum_{i=1}^{N} w_m(binary\_cross(y_m^i, p(x_m^i))) \tag{4}$$

The weight $w_m$ is computed from the ratio of the positive and negative examples as shown in Eqn. 5 where $p_m$ is the number of positive examples in the $m^{th}$ attribute. $\sigma$ is a control parameter. This parameter can be used to control the number of true positives and hence control the recall in effect.

$$w_m = \begin{cases} exp((1-p_m)/\sigma^2) & if\ y_m = 1 \\ exp(p_m/\sigma^2) & else \end{cases} \tag{5}$$

Using the residual deep network with this custom loss function we detect ten attributes including as gender, hair length, shirt color, and the presence of a jacket.

## Demonstration

PatrolBot works as follows: a human could tell the robot to investigate an area of the environment via simple natural language. The robot looks for keywords in verbal communications, and then drives towards the indicated area. As the robot moves through the environment, it stops whenever it encounters either anomalous objects or people. When it sees a person, it will greet them while simultaneously extracting desriptive information. In the case of an anomalous object, it will stop and point at the object.

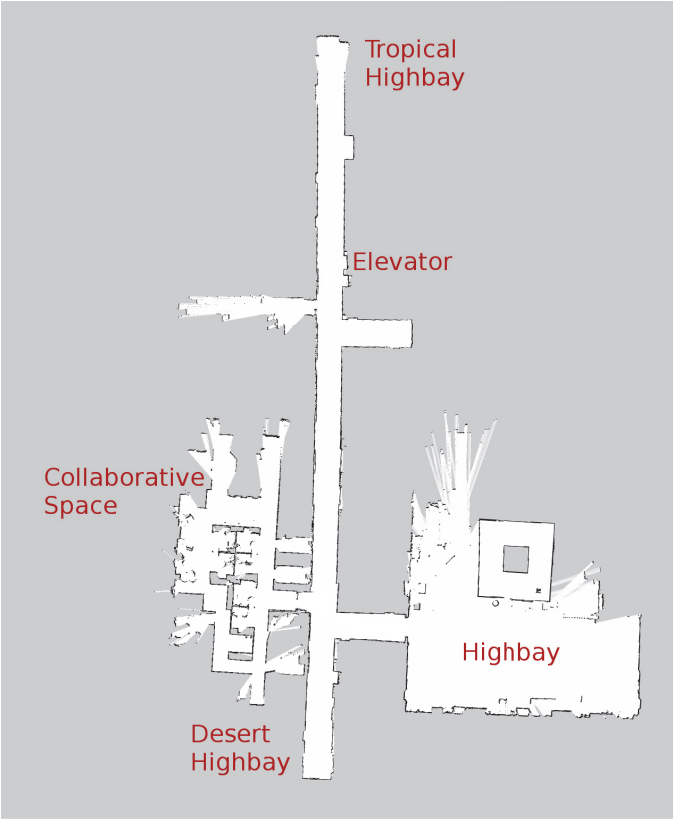


Figure 4: Map of LASR

For navigation, we use the Robot Operating System (ROS) navigation stack (Quigley et al. 2009). This framework provides the capabilities needed to generate a map of its environment, and navigate while performing object avoidance. The primary sensor used for this navigation is a Hokuyo UTM-30LX scanning range finder. To generate the map (example shown in Figure 4) a ROS wrapper is used for OpenSlam's Gmapping (Grisetti, Stachniss, and Burgard 2007). This allows for the robot to simultaneously navigate and map-build in an unknown environment. When operating in a familiar environment, Adaptive Monte Carlo Localization (AMCL) is used to reduce computational load on the system.

We have labelled waypoints on the map and can instruct PatrolBot to move towards any of these waypoints as described in the preceding paragraphs. To train our anomaly detection algorithm, initially PatrolBot navigates between each waypoint to build a dictionary and to learn the associated model. Once trained, she can begin patrolling. During evaluation, she can process 390 patches per second using an NVIDIA GTX-980, which permits her to move about the environment and locate anomalies in real-time. When using a similar GPU, she can determine pedestrian attributes at a rate of 271 frames per second, meaning that she need only look at a person for a fraction of a second before extracting a description.

In the first case, the robot is in the waypoint "Highbay"

Figure 5: People that PatrolBot saw during patrol. The person on the left was descibed as "male, red shirt", the person on the right was described as "male, patterned light shirt"

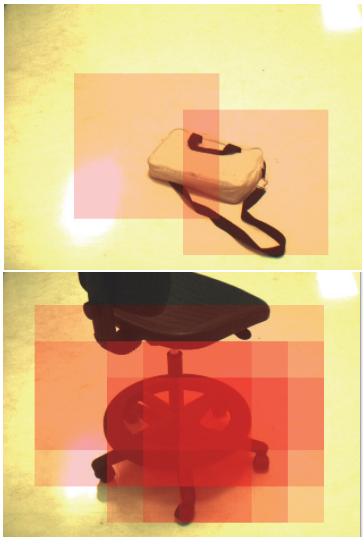

Figure 6: Anomalies that were seen during Patrol. In the first case, a bag was abandoned in the hallway and in the second a chair was in an unusual location.

and is instructed to move towards the waypoint "Desert highbay". During this operation, the robot encounters several people (see figure 5, and to each she looks in the eye and says "Excuse me, I am patrolling". She also collects information about each person. In the case of the first person, she saw a male wearing a red shirt. In the case of the second, she saw a male wearing a light colored, patterned shirt.

In the second case, the robot is at the waypoint "Desert Highbay" and is instructed to move back towards the waypoint "Elevator". She does not encounter people this time, but rather sees several anomalous objects (see figure 6). In the images, each red rectangle represents a position in the image that was seen as anomalous.

## Discussion/Future Work

As we have shown, our PatrolBot is able to accurately locate anomalies and describe people in order to act as a reliable and effective surveillance partner. In our experience, it is easy to train, and our experimental validation has shown a low false positive rate. Another strength of the approach is

that it is able to build a normative model for different environments by leveraging the power of a computational cognitive model. For example, PatrolBot might have a dictionary element for "briefcase" which she knows that is normal for an office, but would be abnormal in the hallway.

Anomaly detection on a moving robot is something that can potentially be extremely advantageous in crowded environments, but it most likely will need a greatly expanded dictionary to describe a number of atypical, yet not anomalous events. Some examples that we saw during our evaluation included things marked anomalous when they should not have been. For example, PatrolBot noticed that a plant had been turned during watering and during training it was leaning in one direction and in testing it was leaning towards the other direction. Although noticing such subtle differences can be a powerful advantage, in this case, PatrolBot needed to update its dictionary to reflect this as another definition of normal.

This leads to one potential area for future work, which is improved learning. Currently, she has a distinctive training and operation mode. It would be beneficial for a human observer to correct her in cases where she erroneously identifies something as anomalous. Although the cognitive model is capable if updating in such cases, it can be troublesome of the patch does not match something that has already been seen in the past. In this case, she needs to also update her dictionary to add this new information. In the previous example of the plant, a human collaborator can review the information and inform PatrolBot that this is normal.

In the future, we also plan to incorporate biometric re-identification. Here, PatrolBot would patrol an environment looking for people that matched a description (e.g., "look for a man with short hair and a red shirt").

## Acknowledgements

Wallace Lawson and J. Gregory Trafton were supported by the Office of Naval Research, Keith Sullivan was supported by the Naval Research Laboratory under a Karles Fellowship. Laura Hiatt was supported by the Office of Naval Research and the Office of the Secretary of Defense. Esube Bekele was supported by the National Research Counsel and the Office of Naval Research.

## References

Breazeal, C.; Siegel, M.; Berlin, M.; Gray, J.; Grupen, R.; Deegan, P.; Weber, J.; Narendran, K.; and McBean, J. 2008. Mobile, dexterous, social robots for mobile manipulation and human-robot interaction. In *ACM SIGGRAPH 2008 new tech demos*, 27. ACM.

Grisetti, G.; Stachniss, C.; and Burgard, W. 2007. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Transactions on Robotics* 23(1).

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Hockey, R. 2013. *The psychology of fatigue: work, effort and control*. Cambridge University Press.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks.

Lawson, W.; Hiatt, L.; and Sullivan, K. 2016. Detecting anomalous objects on mobile platforms. In *Proceedings of Moving Cameras Meet Video Surveillance: From Body-Borne Cameras to Drones Workshop at CVPR*.

Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; and Ng, A. 2009. Ros: an open-source robot operating system. *ICRA Workshop on Open Source Software*.

Smith, G. J. 2002. Behind the screens: Examining constructions of deviance and informal practices among CCTV control room operators in the UK. *Surveillance & Society* 2(2/3).

Trafton, J. G.; Hiatt, L. M.; Harrison, A. M.; Tamborello, II, F. P.; Khemlani, S. S.; and Schultz, A. C. 2013. ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction* 2(1):30–55.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems (NIPS)*.

Zhu, J.; Liao, S.; Yi, D.; Lei, Z.; and Li, S. Z. 2015. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *Biometrics (ICB), 2015 International Conference on*, 535–540. IEEE.