VENUS: A System for Novelty Detection in Video Streams with Learning

Roger S. Gaborski, Vishal S. Vaingankar, Vineet S. Chaoji, Ankur M. Teredesai

Laboratory for Applied Computing,
Rochester Institute of Technology,
102 Lomb Memorial Drive, Rochester, NY 14623
{rsg, vsv8846, vsc2002, amt }@cs.rit.edu

Abstract

Novelty detection in video is a rapidly developing application domain within computer vision. The motivation behind this paper is a learning based framework for detecting novelty within video. Since, humans have a general understanding about their environment and possess a sense of distinction between what is normal and abnormal about the environment based on our prior experience; any aspect of the scene that does not fit into this definition of normalcy tends to be labeled as a novel event. In this paper, we propose a computational learning based framework for novelty detection and provide the experimental evidence to describe the results obtained by this framework. To begin with the framework extracts low-level features from scenes, based on the focus of attention theory and then combines unsupervised learning techniques such as clustering with habituation theory to emulate the cognitive aspect of learning.

Introduction

Novelty detection, also referred as event detection, in video has received widespread attention in the past several years within the computer vision and AI domains. Learning to detect novelty from a given video sequence is a very challenging task. In this paper we propose a novelty detection framework based on the low-level features extracted from a video sequence and a clustering based learning mechanism that incorporates habituation theory. Since the overall area of video processing is also referred to as video exploitation, we have termed our framework as the VENUS: Video Exploitation and Novelty Understanding in Scenes. Figure 1 provides a simple example to facilitate the understanding of such a novelty detection framework. The sequence of frames in the figure shows people walking in our laboratory. The system processes the incoming video data, extracts the features and learns the motion and still aspects over time. Initially any form of event in the scene is flagged as novel. Over time, as the system learns the events it tends to consider this as normal behavior and 'habituates'.

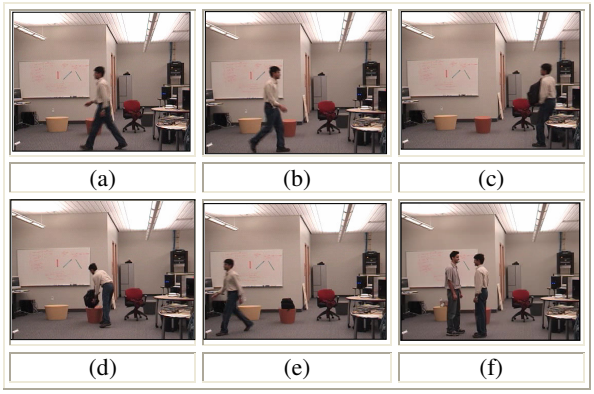


Figure 1: Selected frames a-f of an example video sequence describing normal vs. novel behavior

Subsequent frames process this information and learn that such motion is normal. In the frame shown in 1(d) the placement of a bag is recorded as novel and in the frame shown in 1(f) the fact that two people stop walking and stand in front of each other is also flagged as novel by the system. These precisely are the expectations of a novelty detection framework. The system considers any activity in the scene as an event, such as people walking, cars entering and leaving an area. These events are classified as novel events if they have not been witnessed before in the scene
This paper is organized as follows: In the next section we briefly overview the related work done in this domain. The subsequent sections describe the key components of the VENUS system followed by the experiments and results.

Related Work

Recent advances in video sequence understanding and video exploitation were motivated by an engineering perspective to develop state-of-the-practice video surveillance systems (Medioni et al. 2001). Research by Stauffer et al (2000) proposed detecting events in real-time by learning the general patterns of activity within a scene. This learnt information is subsequently used for activity classification and event detection in the videos. Prior to that, semantic event detection by Haering, Qian and Sezan (1999) successfully tracked and detected events in wild life hunt videos. Recently, Tentler et al., (2003) proposed an

event detection framework based on the use of low-level features. The proposed VENUS framework uses the low-level features obtained from the focus of attention theory and combines it with habituation based clustering.

Habituation is an effect by which a system ceases to respond after repeated presentations of the same stimulus (Siddle, Kuiack and Kroese 1983). Computational modeling of habituation has been applied in mobile robots by Marsland, Nehmzow and Shapiro (1999). Their work models habituation as an exponential function that lends to describing the short-term and long-term memory aspects of learning. In a related previous work we described the primitives of the learning aspect as inspired by biological theories such as habituation (Vaingankar et al. 2003). Initial experimental results on the video sequence described in Figure 1 are discussed in (Gaborski et al. 2004).

VENUS' System Framework

The event detection model described in this paper consists of two major components. First, a focus of attention component that generates the low level features - intensity contrast, color, orientation and directional motion. Second, a learning component that handles novelty detection. The following sub-sections describe the system components.

Focus of attention

Figure 2 shows the block diagram of VENUS' novelty detection framework. The first part of our framework is the focus of attention system. The motivation for using the focus of attention theory is provided by Koch and Ullman (1985). Given the enormous amount of visual information available in a scene, we, as humans, process only a subset of it. We tend to focus on the interesting aspects of the scene ignoring the uninteresting ones. The attention system in our framework is based on the selective attention theory initially modeled by Itti and Koch (2001), where a saliency map topographically represents the object's saliency with respect to its surrounding. Attention allows us to focus on the relevant regions in the scene and thus reduces the amount of information needed for further processing as verified in Gaborski, Vaingankar and Canosa (2003). Objects that are highly salient in the scene are further tracked for possible novel events. The video sequences are processed in the still and motion saliency channels. The still saliency channel processes every frame individually and generates topographical saliency maps. Consider an airport scene where someone leaves an object in a restricted area and walks away. The still saliency channel detects this object as a salient item. Since this object was not part of the original scene, the introduction of the object fires a novel event, which is a feature of the still learning & novelty detection module. The motion saliency channel detects the salient moving objects of the scene, in this case the motion of the person who brought the object.

Still Processing: The still saliency channel processes every frame of the video sequence and extracts the low-level

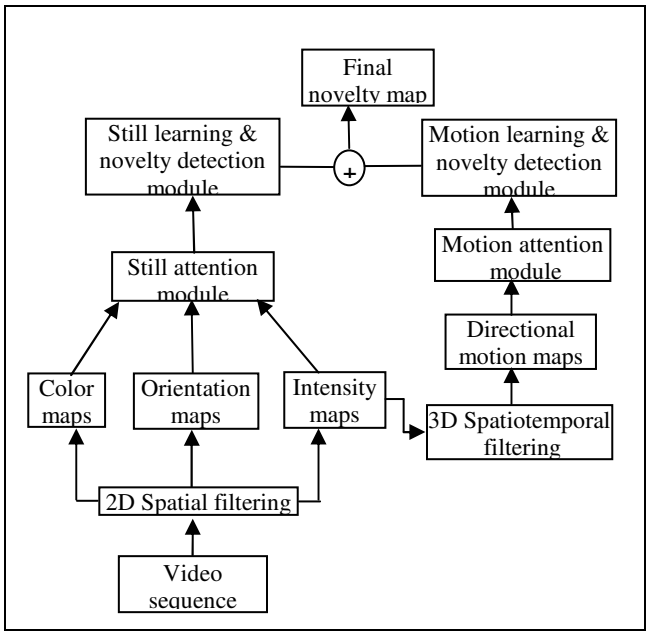


Figure 2: The VENUS' Novelty Detection Framework

features. The information from the video frames is extracted using multi-resolution orientation, color, and intensity contrast filters. This form of processing is called the bottom-up focus of attention (Itti and Koch 2001), since objects evoke attention based on the low-level feature conspicuity.

The 2D multiresolution spatial filters are convolved with the input image to obtain the topographical feature maps. Intensity contrast is extracted using difference of Gaussian filters. The intensity contrast filtering simulates the function of the retinal ganglion cells which posses the centre-surround mechanism. The color information is extracted using the color opponent filters.

The orientation processing employs Gabor orientation filters to extracts edges of 0^{o}, 45^{o}, 90^{o}, and 135^{o} orientations. The sine and cosine Gabor filters oriented in the spatial axis is modeled based on the receptive field properties of the orientation tuned simple cells of the early visual cortex. Due to the centre surround antagonistic nature of the feature extraction filters, the topographical maps obtained are called the saliency maps. The still attention module combines the multi-resolution feature saliency maps within the respective feature channels to form the final orientation, color, and intensity contrast feature saliency maps which are inputs to the still learning & novelty detection module.

Motion Processing: Motion detection in our system is achieved by using the 3D Gaussian derivative spatiotemporal filters tuned to respond to moving stimuli (Young, Lesperance and Meyer 2001). This method of motion filtering is similar to quadrature sine and cosine pair spatiotemporal energy filters developed by Adelson and Bergen (1985). The first step in the motion detection process is convolving the multi-resolution intensity saliency maps with a 3D band-pass spatiotemporal filter.

This filter gives optimal response to regions of motion in the scene irrespective of the direction of motion. Theoretically, such a filter would give the maximum response after an object which was not there before, now appears. The result of the above convolution is bounded (by a threshold) to produce a mask for the motion in an image. In order to extract directional motion, the multi-resolution intensity saliency maps are then convolved with a set of four direction-selective spatiotemporal filters. The direction selective filters are similar to the band-pass filters except that the directional filters are oriented in both space-time (x,t) axes. The orientation of the Gaussian derivative lobes in the (x, t) axis determines the direction to which the filter is tuned. Suppose an (x,t) orientation of 45 degree detects right moving objects, then an (x,t) orientation of 135 degree would be tuned to detecting left moving objects. The result of this directional filtering is a set of multi-resolution directional motion maps (up, down, right, left). The degree of space-time orientation tunes the filters to varying speeds of moving objects. The previously obtained motion mask is then applied to the directional motion maps, to mask out any spurious response generated for stationary objects. The motion attention module combines the multi-resolution direction maps within each direction channel to give final four directional motion saliency maps. The still saliency and the motion saliency maps are the input into the novelty detection modules. The saliency maps as input to the learning & novelty detection modules causes only the salient regions of the scene to be tracked for potential novel events.

Novelty Detection using Learning

The novelty detection framework in VENUS has two components – "still learning & novelty detection module" and "motion learning & novelty detection module". The still learning & novelty detection module in this system is based on Tentler et. al.(2003). Still aspects of the scenes are learned based on simple topographical averaging of feature values in the still feature saliency maps. This averaged map is called the still habituation map which is representative of changes to non-motion aspects of the scene over time. The still novelty is calculated by taking a difference between the current frame's still saliency map and the still habituation map. An event can be novel by virtue of any of its low-level features or a combination of them.

On presentation of a never-seen-before motion feature value, the 'Motion learning & novelty detection module' classifies it as being novel. If the same feature value is observed repeatedly over time, the system habituates to it and stops flagging it as a novel event. On the contrary, lack of additional occurrences of the same event causes the system to recover its original sensitivity for the feature, i.e. the habituation effect decreases. This concept is based on Kohonen's theory (1988) of novelty detection filters with a forgetting effect. The theory states that the system can only memorize patterns when it is frequently exposed to them. The memorized pattern tends to be forgotten if it is not reinforced repeatedly over time. The forgetting term is similar to the dis-habituation effect described by Wang (1995).

Novelty detection in this system is region based where a region is an 8-by-8 pixel area on a frame of the video. A new event in any region is captured by the creation of a cluster in that region. An event can be intuitively considered to be any activity depicted in the motion maps. Each region has a pool of clusters that represent unique events observed by the system in that region. A new cluster formed is compared with clusters in the pool to see if it can be merged with any of them. Merging of clusters indicates occurrence of an event that is similar to a previously learnt event. The clustering algorithm treats the data as a Gaussian mixture. Each cluster obtained represents a Gaussian in the mixture. Novelty detection is thus reduced to identifying novel clusters in every region. The algorithm does not limit the number of clusters per region since the number of novel event cannot be predicted ahead of time. Each cluster follows a sigmoidal-like habituation curve (Vaingankar et al. 2003) to model learning. As per habituation theory, an event is not instantaneously learnt. It takes some number of occurrences before a system gets completely habituated. Novelty is inversely related to the degree of habituation the cluster has attained. Higher the habituation value, the lower is its features' novelty and vice versa. The novel events gathered from each motion direction map are combined with still novelty map to form a final novelty map.

Experiments and Results

Analysis 1:

The goal of the experiments was to quantify the amount of novelty and analyze the effect of the habituation model on learning. We define a novelty index which ranges from zero to one. The novelty index measures the amount of novelty as compared to the motion activity in the scene and is computed as the ratio of the number of novel regions to the number of motion activity regions in a frame. This ratio is calculated for every frame of the video. Higher the ratio, greater is the novelty in the frame. We also define degree of habituation as an index which describes how well a region has learnt a particular event. The experiments are conducted for two scenes. Scene 'A' shows a sequence of two people (P1 and P2) walking from right to left one after the other. Here, P1 and P2 walking are independent events which together form a single sequence. This sequence is repeated thrice without any delay between the sequences. The objective of this setup is to gauge the learning capability of the system to repeated occurrences of people walking in the scene. In scene 'B', the same scene is used but with considerable time delay (frames 130 to 440) between the second and third sequence. The objective for setting up this scenario is to test whether the system forgets a previously learnt event. Figure 3 and 4 show the variations in the novelty index over the complete video, for

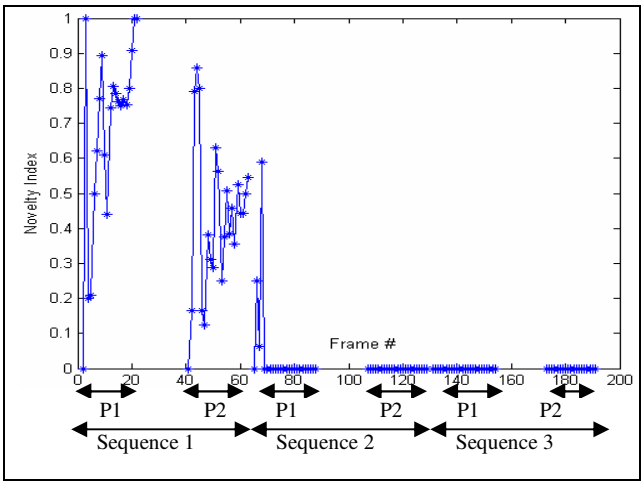


Figure 3: Novelty Index for Scene A

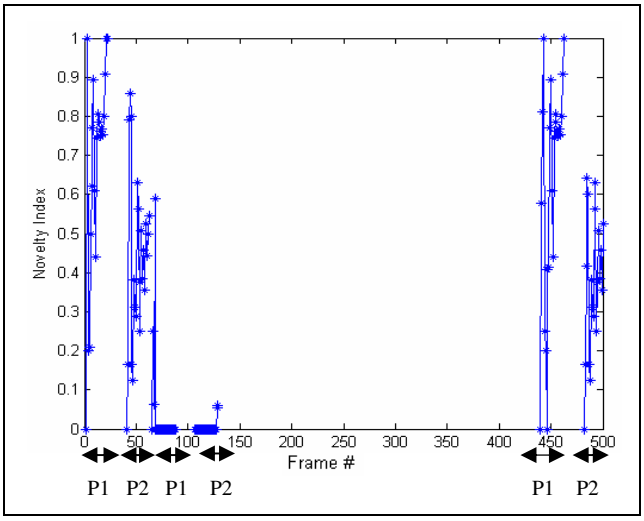


Figure 4: Novelty Index for Scene B

the two scenes respectively. It can be observed in figure 3 for sequence 1 (frames 1 to 60), that the novelty index for event P2 is lower than that of P1 indicating that the system is getting habituated to the event. On the second sequence (P1, P2 frames 60 to130), the novelty index drops to zero. This effect is observed since the regions in the scene have already learnt the motion sequence that occurred in sequence 1. The system continues to invoke zero novelty (novelty index zero) for sequence 3 (frames 130 to 190). Thus with regular reinforcement of similar motion values the clusters do not fire novelty.

Figure 4, shows the corresponding results for scene B. The delay between sequence 2 and 3 causes the system to completely forget the events which were learnt previously. This can be observed by the sudden rise in novelty index for event P1 in sequence 3. Again the system starts to learn this novel event which is visible in the reduced novelty index of the following P2.

Graphs in figures 5 and 6 illustrate the habituation curve for a cluster in a particular region over all the frames in the

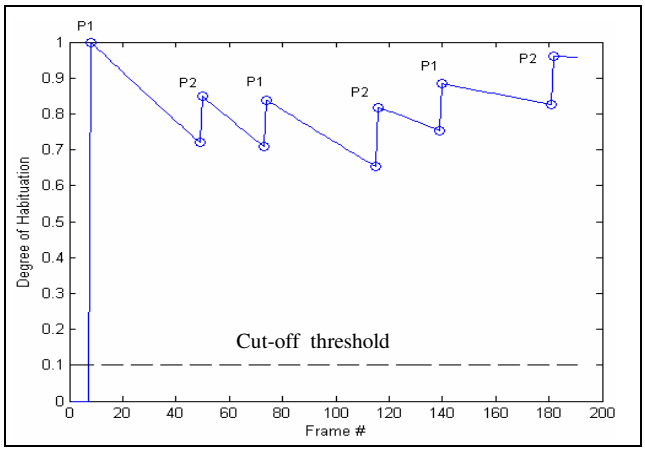


Figure 5: Habituation curve for one cluster observed within a 8X8 region (Scene A).

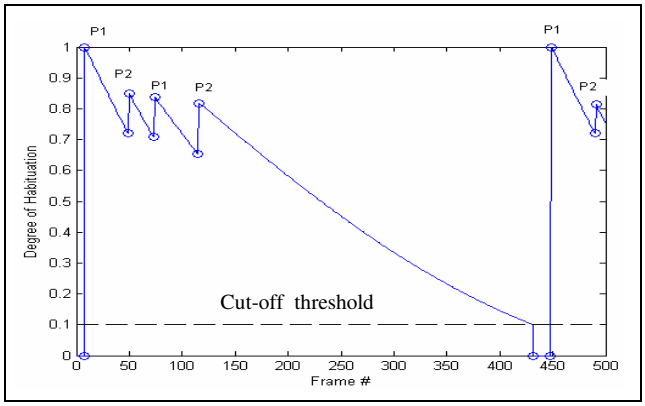


Figure 6: Habituation curve for one cluster observed within a 8X8 region (Scene B).

scene. The y axis shows the degree to which the cluster is habituated to that particular event.

Figure 5 shows the habituation curve for scene A. On seeing the event P1, a cluster is created for that region and an initial habituation value of 1 is assigned. Overtime the habituation value decays at a pre-assigned decay rate. This is seen in the smooth drop in the degree of habituation for the cluster. At frame 40, when the same region is activated by event P2, the cluster recovers from the drop in habituation. This is seen in sudden upward spike in the curve. Similar behavior is observed for the remaining events. As seen from the graph the cluster never reaches the cut-off threshold, indicating that the cluster was regularly reinforced, thereby retaining the event for a longer duration. After the cluster is reinforced with a repeated event, the decay rate is updated such that the system forgets the event at a slower rate. The decreasing slope of the habituation curve after each recovery confirms this effect.

Figure 6 shows the corresponding habituation curve for scene B. During the delay after sequence 2, the habituation value reaches the cut-off threshold. This threshold symbolizes the loss of this event from the system's memory. On frame 450 the event P1 reoccurs. Since the

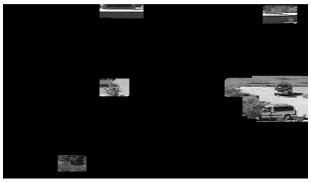
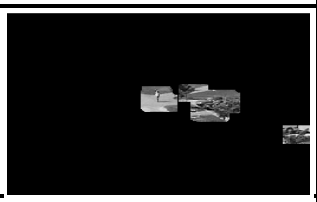
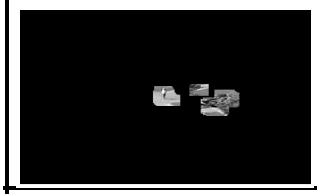
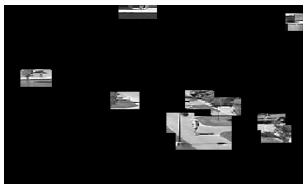
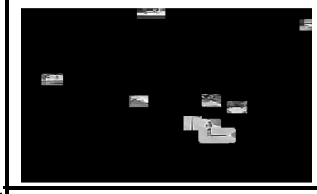
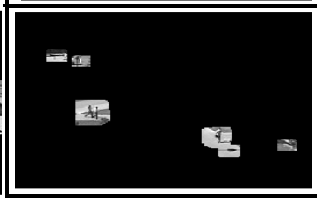
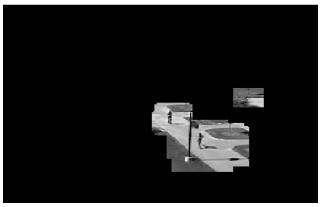
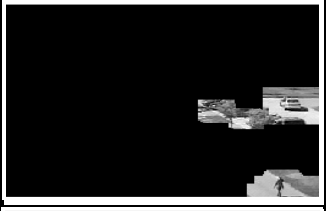
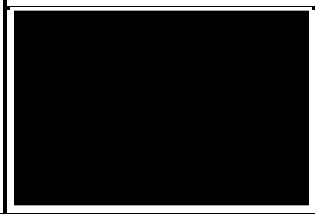
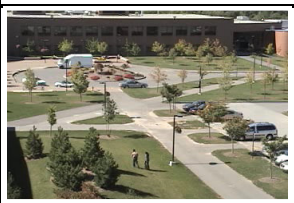
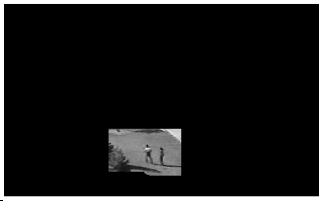
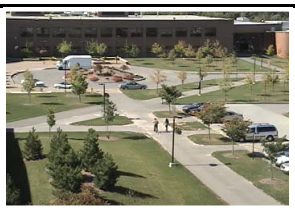
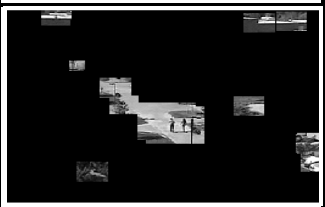
Frame #	Image from video	Motion mask	Motion Novelty mask	Analysis
25 Car entering the parking lot.				Car seen for the first time and was detected by the novelty mask
133 Person walking down the pathway.				The person walking briskly is detected by the novelty mask.
200 Person walking down the pathway.				The system still considers this novel since it is not completely habituated yet.
364 People walking and car leaving the parking lot.				Moving trees stop firing novelty as their swaying motion is learnt.
628 People walking on the already habituated pathway.				People walking are not detected as novel events.
723 People and cars on the learnt regions of the scene.				People and car are not detected as novel events.
957 People walking on the lawn				Novelty detected since this is first instance of people walking on the lawn
1248 People back on the pathway.				People not detected as novel since the pathway is already habituated to motion

Figure 7: Images from video, Motion Mask and the Motion Novelty mask

system has already forgotten this event, it fires a novelty for it and again assigns an initial habituation value.

Analysis 2: The system successfully detects and tracks regional events. Figure 7 shows scenes of videos recorded on a university campus. This video sequence shows people walking on the pathway and cars entering and leaving the parking lot. The figure shows the actual image, the motion mask and the motion novelty mask during various frames of the video. Moving objects in the frame are shown within bounded box in the motion mask. The regions within the motion mask that depict a novel event are shown in the novelty mask. Initially the system fires novelty for each motion activity (event) until it learns the motion in the regions. Consider the example of motion detected for the swaying trees. Initially the system considers this tree motion as novel but gradually habituates to it and no more shows it as a novel event. This can be seen in novelty mask of frames 25, 133, 200 that detects the tree motion as novel, but in frames 364 and further the tree regions do not show a novel event. The system has been successfully tested on outdoor as well as indoor video scenes.

Future Work and Discussion

In the current framework we do not make use of any high level object descriptors. This can be one of the avenues to explore in future for content analysis. In this paper we described a system for novelty detection on natural scenes. We termed this system as VENUS. The central contribution of this work was to combine the recent advances in Computer Vision (saliency and focus of attention) with Data Mining (mining high speed data streams). The described system successfully employs the theory of habituation for learning novel events over time in a video data stream. We did not place any constraints or assumptions on the type of video data our system can process. As long as low-level features can be extracted and motion maps can be generated to capture motion, the system's learning component will detect the novel events. This combination, of the focus of attention theory with habituation theory; is the real strength of the VENUS system.

References

Adelson, E.H.; and Bergen, J.R. 1985. Spatiotemporal energy models for the perception of motion. *Journal of Optical Society of America.* 2:284-321.

Gaborski, R.; Vaingankar V.S.; and Canosa, R.L. 2003. Goal Directed Visual Search Based on Color Cues: Co-operative Effects of Top-down & Bottom-up Visual Attention. In *Proceedings of ANNIE.* Rolla, Missouri.

Gaborski R. S., Vaingankar V. S., Teredesai A., Chaoji V., Tentler A. 2004. Detection of inconsistent regions in video streams. *IS&T/SPIE Electronic Imaging: Human Vision and Electronic Imaging IX,* San Jose.

Haering, N. C., Qian, R.J., and Sezan, M. I. 1999. A Semantic Event Detection Approach and Its Application to Detecting Hunts in Wildlife Video. In *IEEE Transactions on Circuits and Systems for Video technology*, 10:857—868.

Itti L., and C. Koch. 2001. Computational modeling of visual attention. *Nature Neuroscience Review.*, 2(3):194-203.

Koch, C., and S. Ullman. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*. 4:219-227.

Kohonen, T. eds. 1988. Self-Organization and Associative Memory. New York: Springer-Verlag.

Marsland, S.; Nehmzow, U.; and Shapiro, J. 1999. A model of habituation applied to mobile robots. In *Proceedings of Towards Intelligent Mobile Robots*. Bristol, UK.

Medioni,G.; Cohen, I.; Brmond, F.; Hongeng, S.; and Nevatia R. 2001. Event Detection and Analysis from Video Streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23: 8, 873-889.

Siddle, D. A. T.; Kuiack, M.; and Kroese, S. B. 1983. The Orienting reflex. (pp. 149-170). In *Physiological Correlates of Human Behaviour.* Edited by Gale A. and Edwards, J. Academic Press: London.

Stauffer, C.; and Grimson, E. 2000. Learning Patterns of Activity Using Real-Time Tracking. In *IEEE Transactions* on *Pattern Analysis and Machine Intelligence.* 22(8):747-757.

Tentler A.; Vaingankar V.; Gaborski R.; and Teredesai A. 2003. Event Detection in Video Sequences of Natural Scenes. In *IEEE Western New York Image Processing Workshop,* Rochester, New York.

Vaingankar V.S, Chaoji V, Gaborski R S, Teredesai A M. 2003. "Cognitively motivated habituation for novelty detection in video", *NIPS Workshop on Open Challenges in Cognitive Vision*. Whistler, Canada.

Wang D.L. 1995. Habituation. Arbib M.A. (ed.), *The Handbook of Brain Theory and Neural Networks.* 441-444, MIT Press.

Young, R. A; Lesperance, R. M., and Meyer, W. W. 2001 The Gaussian Derivative model for spatialtemporal vision: I. Cortical Model. *Spatial Vision,* 14(3,4):261-319.