Detection of Inconsistent Regions in Video Streams

Roger S. Gaborski, Vishal S. Vaingankar, Vineet S. Chaoji, Ankur M. Teredesai, Aleksey Tentler
Laboratory for Applied Computing,
Rochester Institute of Technology,
102 Lomb Memorial Drive, Rochester, NY 14623
{rsg, vsv8846, vsc2002, amt, axt8828}@cs.rit.edu

ABSTRACT

Humans have a general understanding about their environment. We possess a sense of distinction between what is consistent and inconsistent 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 classified as a novel event. An example of this is a casual observer standing over a bridge on a freeway, tracking vehicle traffic, where the vehicles traveling at or around the same speed limit are generally ignored and a vehicle traveling at a much higher (or lower) speed is subject to one's immediate attention. In this paper, we extend a computational learning based framework for detecting inconsistent regions in a video sequence detection. The framework extracts low-level features from scenes, based on the focus of attention theory and combines unsupervised learning with habituation theory for learning these features. The paper presents results from our experiments on natural video streams for identifying inconsistency in velocity of moving objects and also the static changes in the scene.

Keywords: Attention, Spatiotemporal filtering, motion detection, Habituation, Learning, Clustering.

1. INTRODUCTION

In this paper we propose an inconsistency detection framework based on the low-level features captured from a video sequence and a clustering mechanism that combines habituation theory and k-means clustering. This framework forms a part of the Video Exploitation and Novelty Understanding in Scenes (VENUS) system. Recent advances in video sequence understanding and event detection have been approached from an engineering perspective to develop state-of-the-practice video surveillance systems. Work by Medioni et al.(2001), part of the Video Surveillance and Monitoring System project, is an example of a system for tracking and detecting events in videos collected from an unmanned airborne vehicles. Prior to that work, semantic event detection approach by Haering, Qian and Sezan (1999) successfully tracked and detected events in wild life hunt videos. 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. Recently, Tentler et al., (2003) proposed an event detection framework based on the use of low-level features. The proposed VENUS framework also uses the low-level features and then advances the state-of-the-art by combining the focus of attention theory and habituation based clustering.

The learning aspect in VENUS is inspired by biological theories such as habituation. Habituation is an effect by which a system ceases to respond after repeated presentations of the same stimulus (Siddle, Kuiack and Kroese 1983). Work by Paul Crook (2000) implements a neural network to habituate to the surrounding elements and classifies stimulus patterns as familiar or novel. Computational modeling of habituation has been proposed by Wang (1995) and recently 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. The inconsistency detection framework in VENUS differs from the past work in this area in the following ways. First, the system uses intensity contrast, edge, color and motion information as low-level features instead of commonly used high-level knowledge about the objects. As opposed to systems that store codebook information (Stauffer and Grimson 2000) about the tracked objects, this system learns patterns of activity in the scene from the extracted low-level features. This makes VENUS environment independent. Second, we use habituation theory for modeling the memory aspect of learning events. Finally, we use a relaxed version of k-means for grouping events to detect novelty. The system classifies any activity in the videos as an event, such as people walking, cars entering and leaving an area. These events are classified as inconsistent events if they have not been witnessed before in the scene. We also show results of inconsistency detection based on object velocity in the scene, and the static changes.

2. MODEL

The event detection model described in this paper consists of two major components:
a) A focus of attention component that generates the low level features, and
b) The learning component that handles inconsistency detection.
In this section we describe each of these components in detail.

2.1 Focus of attention

Figure 1 shows the block diagram of VENUS' inconsistency 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). The claim is that given the enormous amount of visual information available within a scene, we "process" only a subset of the entire scene. 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).

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 the situation where someone places an object in an empty corridor and leaves. 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 an inconsistent event, which is a feature of the still novelty detection module. The motion saliency channel detects the salient moving objects of the scene, such as the motion of the person who brought the object in the corridor. The outputs of the saliency channels form the inputs to the still and motion learning & novelty detection modules (SLID) and Motion Learning & Novelty detection module (MLID).

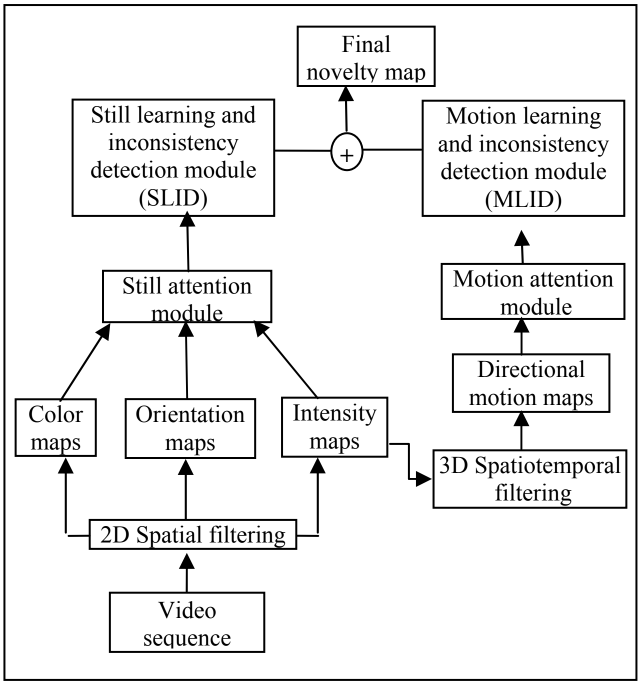


Figure 1: The VENUS' Novelty Detection Framework

2.1 Still Processing.

The still saliency channel processes every frame of the video sequence and extracts the low-level 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 spatial filters used in the system are modeled after the findings of the biological vision principles simulating the functioning of the retina, lateral geniculate nucleus and the early visual cortical areas. The spatial filters are convolved with the input image to obtain the topographical feature maps. Intensity contrast is extracted using Difference of Gaussian filters. The Difference of Gaussian filters used were on-centre-off-surround and off-center-on-surround behavior. 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. These filters simulate the functions of the color opponent cell responses in the LGN. The evidence for this arrives from the color opponent theory proposed by Hering (Palmer, 1999). It suggests that there exist a combination of colors like Red-Green and Blue-Yellow that have an opponent behavior to each other. That is, the On-centreRed-Off-surround-Green opponent cell gets excited by red and inhibited by green.
The orientation processing employs Gabor orientation filters to extracts edges of 0o, 45o, 90o, and 135o 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. They are termed as saliency maps because the output values obtained after applying these filters are dependent on the relative excitatory and inhibitory components of the centre surround filters. Thus, a region which is highly salient will have a low inhibitory component opposing the excitatory component. 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 and inconsistency detection module (SLID).

2.3 Spatiotemporal Motion Processing.

Motion information in video sequences is extracted using the 3D Gaussian derivative (GD) spatiotemporal filters tuned to respond to moving stimuli (Young & Lesperance, 1993) , (Young et al, 2001). Motion detection in our system is achieved by using a set of spatiotemporal first and second order Gaussian derivative filters. This method of motion filtering is similar to quadrature sine and cosine pair spatiotemporal energy filters developed by Adelson and Bergen (1985). Wherein the outputs of the sine and cosine filters are squared and summed to make the filter outputs phase independent. The similarity between the gabor model and Gaussian derivative model is explained in (Petkov & Westenberg, 2003). They use the energy error minimization between the two methods to analyze the similarity of the two models. Also (Young & Lesperance, 1993) explain the physiological fit of the GD filters with the visual cortex physiological data. Their research compares the GD model with the various receptive field profile models like Gabor model, hermite model and the Difference of Offset Gaussian model. The GD filters are oriented in both space and time axes. The spatial axis (x,y) orientation (θ) tunes the filter to the spatial orientation of the edge. The second is the space-time domain (x,t) orientation (ϕ), tunes the filter to speed of the objects. Suppose a ϕ of 45 degree detects right moving objects, then ϕ of 135 degree would be tuned to detecting left moving objects. By varying the ϕ from 0 deg to 90 deg we can tune the filters for varying speeds of motion. The GD filter with ϕ= 0 deg, will produce the maximum response for fast moving objects, ideally this filter behaves like a blink field. Theoretically, such a filter would give the maximum response after an object which was not there before, now appears. This blink filter gives optimal response to regions of motion in the scene irrespective of the direction of motion. The result of convolving the blink filed with an image is that regions of scene with motion are captured irrespective of the direction of motion which is used as a motion mask. This mask topographically represents only the motion regions in a scene.

The first step in the motion detection process is convolving the multi-resolution intensity saliency maps with 3D band-pass spatiotemporal filters. A set of velocity tuned directional filters are constructed to capture directional motion. A space-time orientation (ϕ) of 45 deg is chosen for the directional filters. The result of this directional filtering is a set of multi-resolution directional motion energy maps (up, down, right, left). 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 energy maps. The direction maps obtained represent the motion energy and not the motion velocity. Further processing is applied on the directional maps to obtain the velocity. A stationary filter is constructed that responds maximum for stationary objects of the scene and produces a reduced response for moving objects. The velocity calculation mechanism used in our system is similar to the one used by Adelson & Bergen (1986) for calculating velocity. Velocity is calculated by considering the opponent motion maps and the stationary map. For instance, the left velocity is derived by subtracting the right filter response from left filter response and taking a ratio of this difference with the stationary map response. The output obtained assigns maximum velocity value for high velocities and low values for slower moving objects. Thus we apply the same velocity calculation mechanism for all the four directional maps. The velocity maps obtained are referred as motion saliency maps, since the high velocities are represented with higher saliency than the slower moving objects. The motion saliency maps are the input into the motion learning and inconsistency detection modules (MLID). The saliency maps as input to the event detection and learning modules causes only the salient regions of the scene to be tracked for potential inconsistent events.

3. LEARNING & INCONSISTENCY DETECTION

The foundation of novelty detection derives its source from outlier detection, an area of active research in the statistical learning community. A comprehensive survey for novelty detection using statistical methods is provided by Markou and Singh (2003). With the focus on surveillance and monitoring applications increasing, there is considerable emphasis on novelty detection within the data mining community. Yamanishi and Takeuchi (2002) discuss novelty detection for non-stationary (adaptive) data while incorporating the effect of forgetting previous data. Recent work by Ma and Perkins (Ma and Perkins 2003) on online novelty detection uses support vector regression but requires substantial amount of memory to store previous data points. Oh, Lee and Kote (2003) discuss an algorithm for detecting motion on segments of frames and clustering these segments to identify normal events. Accordingly, anything not within these clusters forms an abnormal event. Their system lacks a comprehensive learning component for novelty detection, which is the focus of this paper.

The novelty detection framework in VENUS has two components - still novelty detection module and motion novelty detection module. The still novelty detection module in this system is based on Tentler et. al.(2003). In this section we discuss the novelty detection based on motion events. On presentation of a never-seen-before feature value, the MLID classifies it as being inconsistent. If the same feature value is observed repeatedly over time, the system habituates to it and stops flagging it as a novel event. An event can be novel by virtue of any of its low-level features or a combination of them. 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).

Figure 2 shows the working of the motion novelty detection module. Novelty detection and learning in this system is region based where a region is an 8-by-8 pixel area on a frame of the video. Depending on the direction of motion the relevant direction map will encode the motion values. The regions of the direction maps that detect motion get excited. These direction maps are input to the motion learning and event detection module. Within each region, a Gaussian mixture model can represent the values obtained from the directional maps over a period of time. Each distribution in the mixture is represented by a cluster resulting in a pool of clusters representing the entire distribution. Novelty detection is thus reduced to identifying novel clusters in every region.

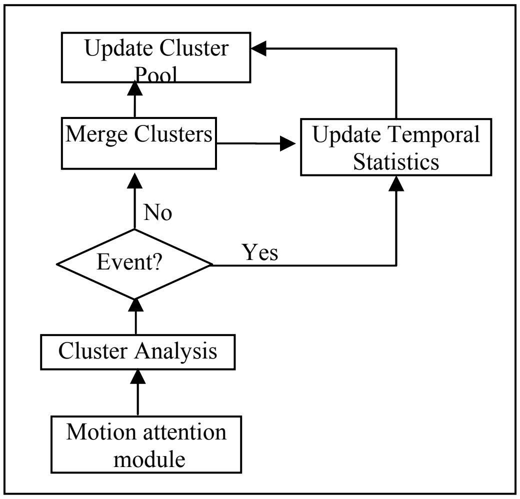


Figure 2: Motion learning and novelty detection module

The working of the system could be explained with the following example. Consider a video sequence in which people are walking from right to left at a speed of 5 mph. When a person passes over a region (during a group of contiguous frames), the left directional motion map gets invoked. The excited regions of the map provide motion values that correspond to the speed of the person walking. A single cluster representing a Gaussian distribution is formed from these values in the cluster analysis step in figure 2. . This cluster is compared with existing clusters in the pool of cluster. If this cluster is similar (in its distribution) to any cluster in the pool, it is merged with the cluster in the pool. Otherwise, if the cluster cannot be merged with any existing cluster, a new cluster is inserted into the pool. The similarity measure between two clusters is a function of their means and standard deviations. The fact that a similar cluster was already found in the pool indicates that a similar event had occurred in the past, while the insertion of a new cluster in the pool indicates the occurrence of a never-seen-before event. Going back to the example, when multiple people walk at 5mph over a region, clusters representing their speeds are merged. This indicates that people walking is not a novel event any more. Now, when a person runs at 15 mph from right to left, a new cluster for 15 mph is formed. This represents occurrence of a novel event. Similarly the above phenomenon will be observed if a person walks from left to right, thereby firing an event in the right directional map. This algorithm is incremental in nature, in that the clusters for a region are updated as events occur in the scene. The algorithm does not limit the number of clusters per region since the number of novel event cannot be predicted ahead of time.

New clusters added to the pool are assigned an initial habituation value and an initial decay rate that determine its temporal characteristics. The decay rate symbolizes the forgetting term described by Kohonen (1988). The slower the decay rate the longer is the retention period for the event. Each cluster follows a sigmoid-like habituation curve as shown in figure 3. The habituation function for a cluster is given by:

$$H(t) = 1 - [1/(1+e^{-at})]$$

where *H(t)* is the habituation value *t* frames after the creation of the cluster and *a* is the current decay rate of the cluster. When clusters are merged we update the decay rate for the older cluster. This indicates that the learnt event was reinforced resulting in increased retention. This is performed in the update temporal statistics step in figure 2. A cluster with habituation value below the cut-off threshold is considered completely decayed and is discarded from the pool of clusters. Effectively, the system has forgotten the event that the discarded cluster represented. Hence the forgotten event becomes novel once again. This models the concept of forgetting in habituation theory. The initial decay rate is set to zero which can go up to one. Value of zero indicates no decay (longer retention) while one indicates maximum decay (shorter retention). The decay rate for a cluster is adjusted as follows:

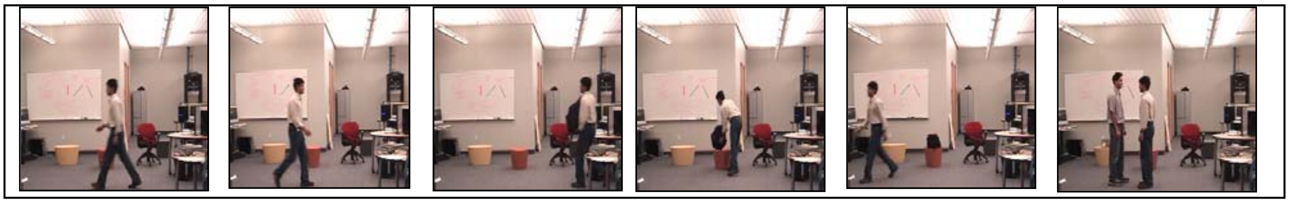
$$a_t = 1 - [e/f]$$

where a_t is the decay rate *t* frames after its creation, *f* is the number of frames passed since the creation of the cluster and *e* is the number of times the cluster merged with similar clusters. The e*/f* term indicates the reinforcement (cluster merging) rate. Higher the reinforcement rate, closer the new decay rate to zero. Smaller the reinforcement rate, closer the new decay rate will be to one. As per habituation theory, an event is not instantaneously learnt. It takes some number of occurrences before a system gets completely habituated. The recovery in degree of habituation prior to the system reaching complete habituation (also known as stable state) is lesser than the recovery after reaching complete habituation as seen in figure 3. 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.

4. Experimental Results

The system successfully detects and tracks regional events. Figures (5-9) shows scenes of video recorded in a laboratory where inconsistent events like people leaving objects in the room and sudden stopping and interactions of people are detected as inconsistent. The next set of results (figure 10-12) detects people briskly walking and running as inconsistent in the environment learnt for a stroll (walk) in the college building. These set of results can be seen as a easy extension for tracking threat situations against non-threat situations.

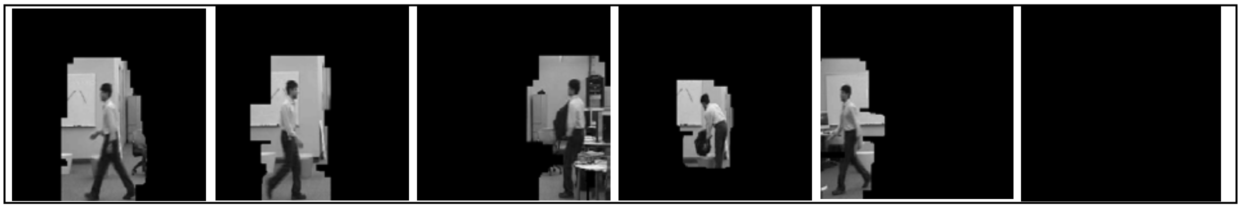
Scenario 1:



Frame 10 Frame 146 Frame 933 Frame 1071 Frame 1183 Frame 2300

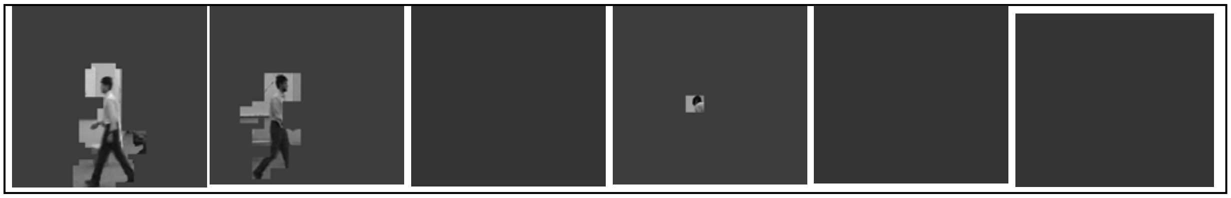
Figure 5: Original Frame Sequences

The scene description is as follows:
Frame 10: Person walking the first time from right to left in the lab.
Frame 146: Person walking the first time from left to right.
Frame 933: Person walking in the same area with a bag in his hand.
Frame 1071: Placing the bag on the table.
Frame 1183: Person walks away after placing the bag.
Frame 2300: People stop and talk.
The output maps of the various modules of the system are shown in the subsequent figures.



Frame 10 Frame 146 Frame 933 Frame 1071 Frame 1183 Frame 2300

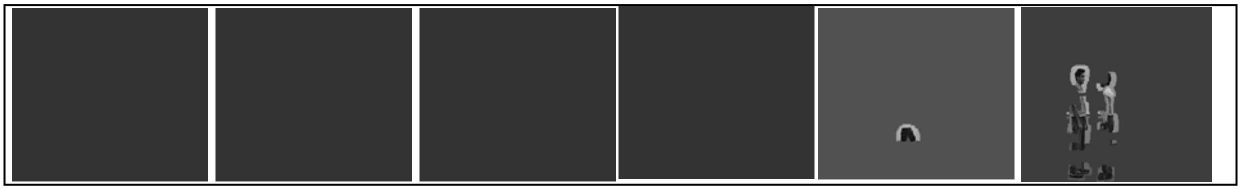
Figure 6: Motion Detection Outputs detects person moving in the scene represented within a bounding mask.



Frame 10 Frame 146 Frame 933 Frame 1071 Frame 1183 Frame 2300

Figure 7: Motion Novelty Map.

Analysis:
Frame 10, 146 : Motion of the person walking towards left and right is detected as novel in the initial frames when the system is still learning the motion areas of the scene.
Frame 933: Does not indicate the walking as a novel event since the system has learnt the walking action.
Frame 1071: detects person bending down motion as a novel motion
Frame 1183: Does not indicate the walking as a novel event since the system has learnt the walking action.
Frame 2300: People standing is not captured by the motion filters.

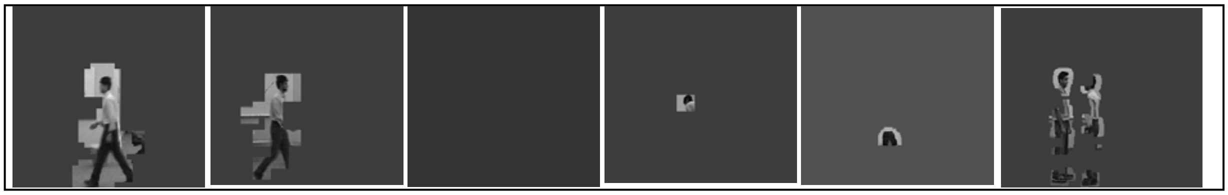


Frame 10 Frame 146 Frame 933 Frame 1071 Frame 1183 Frame 2300

Figure 8: Still inconsistency Map

Still inconsistency Analysis:

Frame 1183: Still Novelty detection map detects the bag placed on the table as a novel event since the bag was not part of the original scene. But this detected bag overtime loses its novelty due to system learning the bag at that position and stops being novel. Frame 2300: People standing in the room for the first time are captured as an inconsistent event. Overtime this event loses its importance when people continue to stand in that position.

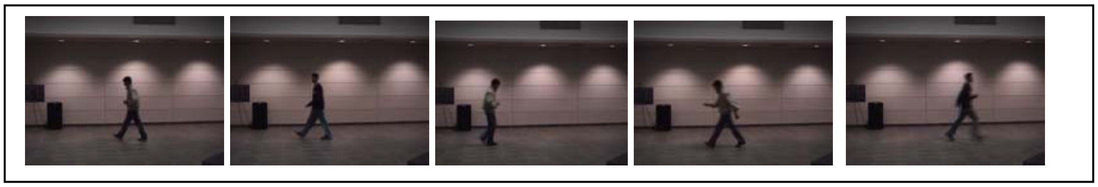


Frame 10 Frame 146 Frame 933 Frame 1071 Frame 1183 Frame 2300

Figure 9: Total inconsistency map is the combined inconsistency map of the still and motion inconsistency maps.

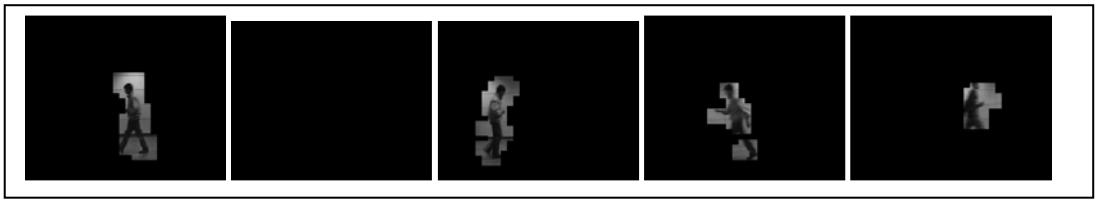
Scenario II: Inconsistency detection based on speed.

Scene description:
Frame 98: First instance of person strolling from right to left.
Frame 176: Second instance of person strolling in the same area from right to left.
Frame 335: first instance of person walking left to right.
Frame 671: person briskly walking with increased speed.
Frame 789: Running across from left to right.



Frame 98 Frame 176 Frame 335 Frame 671 Frame789

Figure 10: Original frame sequences



Frame 98 Frame 176 Frame 335 Frame 671 Frame789

Figure 11: Speed inconsistency detection maps

Speed inconsistency detection analysis:
Frame 98: person walking right to left the first time is indicated as inconsistent since the system is still learning this strolling speed values.
Frame 176: person not detected as inconsistent since strolling speed is already learnt from the previous frames of strolling sequences.
Frame 335 person walks left to right the first time is indicated as inconsistent
Frame 671: Brisk walk is detected as inconsistent since the speed is faster than the strolling speed.
Frame 789: person's run is detected as inconsistent since the speed was not encountered before.

5. 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. The learning component based on still-maps for novelty detection can also be improved using other unsupervised learning methods. The presently used clustering technique relies heavily on the distribution of data and relaxing this assumption would certainly result in a more generic habituation based learning strategy.

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. Several experiments, performed on captured video sequences, suggest the utility of our approach. We did not place any constraints or assumptions on the type of video data our system can process for novelty detection. 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.

6. 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.

Crook, P. 2000. Spotting Novelty: A Neural Network Model for Familiarity Discrimination.

Emerson, R. C., Bergen J. C., and Adelson, E. H. 1992. Directionally Selective Complex Cells and the Computation of Motion Energy in Cat Visual Cortex. *Vision Research.* 32(2):208-218.

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 the Artificial Neural Networks in Engineering, Rolla, Missouri.* Forthcoming.

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.

Ma, J., and Perkins, S. 2003. Online Novelty Detection on Temporal Sequences. In *Proc. of the Ninth ACM SIGKDD, ACM Press.* 613-618.

Markou, M.; and Singh, S. 2003. Novelty Detection: A Review, Part I: Statistical Approaches, *Signal Processing*. Forthcoming.

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.

Oh, J; Lee, J.; and Kote S. 2003. Real Time Video Data Mining for Surveillance Video Streams. In *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining,* 222-233. Seoul, Korea: Springer-Verlag

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*

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

Yamanishi, K., and Takeuchi, J. 2002. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proc. of the Eighth ACM SIGKDD, ACM Press.* 676-681.

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.

Vaingankar V. S, Chaoji V, Gaborski R S, Teredesai A M., "Cognitively Motivated Habituation for Novelty Detection in Video", *NIPS 2003 Workshop on 'Open Challenges in Cognitive Vision'. Whistler, BC, Canada, December, 2003*

Palmer S., Vision Science: Photons to Phenomenology, MIT press 1999.

Petkov N, Westenberg M A: Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, Vol. LNCS 2756, Springer-Verlag Berlin Heidelberg, pp. 762-769, 2003

Adelson, E. H., and Bergen, J. R. The extraction of Spatiotemporal energy in human and machine vision, *Proceedings of workshop on Motion: Representation and Analysis* (pp. 151-155), Charleston, SC, May 7-9 (1986).