A cognitively motivated video detection system

Roger S. Gaborski
Rochester Institute of Technology
rsg@cs.rit.edu

Jeremy Paskali
Rochester Institute of Technology
stealth582@gmail.com

ABSTRACT

The Cognitively Motivated Video Detection (CMVD) system models our ability to detect interesting video events by first learning commonly occurring events in a particular environment. New events are then compared to previously learned events. If the event has not been seen before, the event is noted as interesting. Spatial-based events are described in terms of cognitively motivated low-level features including color, orientation, contrast, and motion. A schedule is used to represent temporal event patterns. The CMVD system has been successfully tested on video acquired from stationary cameras. The system accurately models human behavior.

INDEX TERMS

Cognitive, Habituation, Interesting Events, Low-level features, Surveillance, Video

I. INTRODUCTION

An important aspect of our cognitive skills is our ability to detect interesting events in the environment while ignoring uninteresting events. This capability allows humans to function in a complex environment with large amounts of data being sensed by our sensory systems. The Cognitively Motivated Video Detection (CMVD) system is composed of two subsystems. The first subsystem, VENUS (Video Exploitation and Novelty Understanding in Streams), detects interesting stationary and non-stationary spatial events [1], and the second subsystem, IVEE (Interesting Video Event Extraction), detects interesting temporal events. Consider the following situation and how a human observer might react: an observer sees a red vehicle slowly proceeding down a road traveling east to west. The first time this happens our observer's attention will be drawn to the vehicle, and he or she will unconsciously note the passage of the vehicle and such details as the vehicle's color, relative speed, and direction of travel. The second time a similar vehicle of the same color, speed, and direction of travel passes, the observer will be less interested because this event is no longer unusual. Furthermore, with sequential passing of the vehicle, the observer will be even less interested. From a cognitive science point of view, the observer has been habituated to this event and no longer finds it interesting. However, if the color of the vehicle, speed, or direction of travel is different, the observer again finds this event interesting. This is the type of spatially interesting event that is detected by the VENUS subsystem. Now consider the situation where the vehicle travels on a schedule, say once an hour. Our observer will note the schedule and realize the vehicle is absent at the expected time. The vehicle's appearance at a different time will also be interesting. This is the temporal type of interesting event the IVEE subsystem is designed to detect. The two subsystems are combined to identify interesting spatial and temporal events. It should be noted that both subsystems operate on low-level features, and in the current framework, higher-level objects are not recognized.

II. BACKGROUND

A survey of the current video surveillance literature reveals that current systems lack several of the key capabilities needed to simulate human behavior. This is especially true in complex environments with respect to discovering interesting events. First, current systems do not calculate degrees of interest of the stimuli but simply report if the stimulus is unusual

or abnormal [1][2][3][4]. Second, some systems require training with hand-labeled examples of abnormal and normal events in order to operate successfully [3][5]. Third, systems lack domain knowledge of time and are not able to represent how people operate on a twenty-four-hour day and a seven-day-a-week schedule. Last, the lack of domain knowledge of time leads to the system's inability to discover event patterns of occurrences and thus to not have any expectations of the event's next appearance.

A. *What describes an "interesting event"?*

An interesting event is defined as any event that is out of the ordinary. To function effectively in an environment, we require a sensory system to sense the environment, a learning mechanism to learn events, and short- and long-term memory mechanisms to store interesting events along with their schedule and the ability to recall similar events.

In this paper we develop a cognitively motivated framework that models the human's capability to detect interesting events in a scene using visual and temporal information. The described system automatically learns the normal visual environment and detects and displays events of interest.

III. VENUS

The current implementation of the VENUS subsystem is illustrated in Figure 1. The video data from a stationary camera is processed by 2D spatial filters that extract color, orientation, and intensity information. A pool of Gaussian distributions is used to model the stationary background and detect motion. Video containing interesting stationary and non-stationary (motion) events are outputted by VENUS

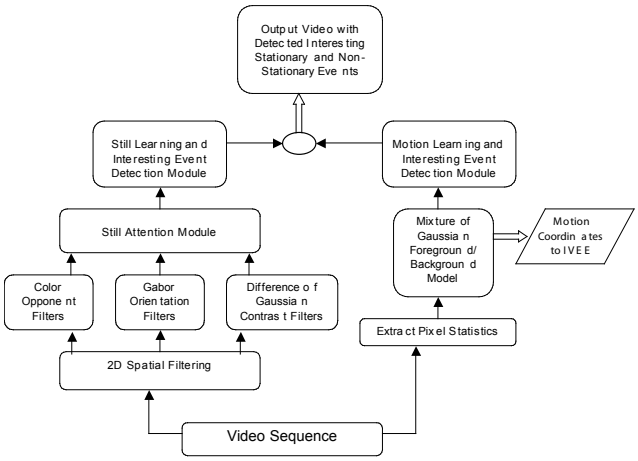

Figure 1. The VENUS subsystem

A. *VENUS Saliency and Focus of Attention*

The low-level features detected by the 2D spatial filtering and the motion information derived from the Gaussian models are combined to form a saliency map. This saliency map is used as the foundation for the focus-of-attention process. The motivation for using the focus-of-attention approach is provided by Koch and Ullman [6]. Given the large amount of visual information in a scene, we focus on the interesting aspects of the scene while ignoring the uninteresting regions. In the same manner, the VENUS subsystem minimizes the processing of information by focusing on the salient regions of the video. The actual attention algorithm in the VENUS subsystem is based on the selective attention theory initially modeled by Itti and Koch [7], where a saliency map topographically represents objects' saliency with respect to the surroundings.

In the VENUS subsystem, we separate the information content in the video into two channels: 1) the information contained within the still image, and 2) the information obtained from moving objects in a sequence of video frames. This approach results in a focus-of-attention system comprised of two sub-components, the Still Saliency Channel and the Motion Saliency Channel.

B. *VENUS Learning and Habituation*

The interesting event detection framework in VENUS has two components: a still- interesting event detection module and a motion-interesting event detection module. The video frame is divided into a mosaic of 8x8 pixel sub windows. For each sub window, the still-interesting event detection algorithm creates a model of the stationary objects in the scene based on color, edge orientation, and contrast using Gaussian probability distribution models. The Gaussian models are updated each time a new video frame is processed. Similar distributions may be merged, or new distributions may be created depending on the new features detected. Objects that are detected by the motion detection model are excluded from the stationary model. Each new frame is compared to the stationary model after objects in motion have been excluded. A never-seen-before feature value is classified as being inconsistent or interesting. The motion-interesting event- detection algorithm works in a similar manner, except objects that are in motion are compared to a motion-feature map. A never-seen-before motion feature is flagged as interesting.

Both the stationary and motion models are updated when a interesting event is detected. The system habituates to the same feature value observed repeatedly and stops flagging it as an interesting event. An event can be interesting by virtue of a low-level feature 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) [8] of novelty detection filters with a forgetting effect. The theory states that the system can memorize patterns only when it is frequently exposed to them. The memorized pattern tends to be forgotten if it is not reinforced repeatedly. The forgetting term is similar to the dis-habituation effect described by Wang (1995) [9]. In summary, the VENUS subsystem detects interesting stationary and non-stationary objects, tracks motion, records motion coordinates, assigns degrees of interest to events, and stores interesting video segments in a database with metadata for easy querying but does not incorporate a sense of time. The Interesting Video Event Extraction (IVEE) system was developed to incorporate a sense of time.

IV. IVEE

The Interesting Video Event Extraction (IVEE) (Figure 2) subsystem is part of the CMVD system. The inputs to this subsystem are the original video stream and motion coordinate data from the VENUS subsystem. The IVEE subsystem learns the temporal schedule of events that occur in the video environment. Events that are not in temporal agreement with previously determined schedules are labeled as interesting temporal events. A color cue is used to indicate the calculated degree of interest and the coordinates of the interesting events. The IVEE subsystem also outputs indications of missed expected events to the output video sequence. The IVEE subsystem separately calculates degrees of interest for events in motion from stationary events. By learning the normal events, the IVEE subsystem is able to discover interesting events, which include interesting and missed expected events.

A. IVEE Temporal Event Patterns

In the IVEE subsystem, the interesting event is based on changes to a learned schedule. As can be seen in Figure 3, events are learned on several time scales.

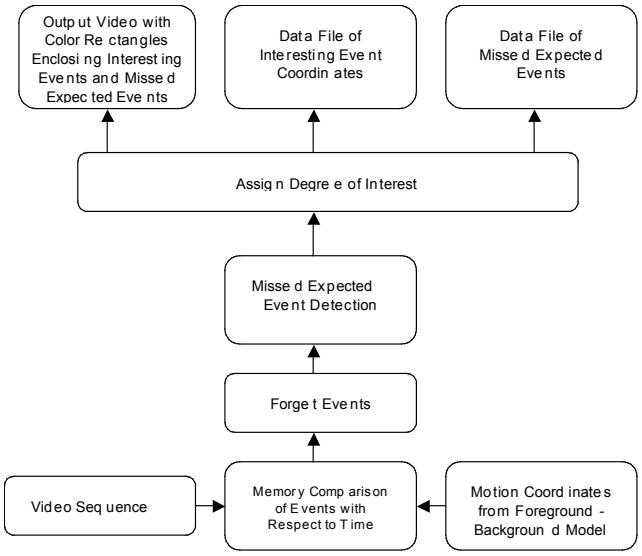

Figure 2. The IVEE subsystem

An important aspect of the IVEE subsystem is the discovery of event patterns. The subsystem is able to determine that certain patterns occur on an hourly, daily, or weekly schedule. The schedule enables the prediction of an event's next appearance.

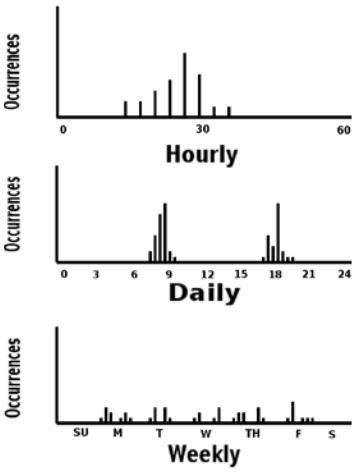

Figure 3. A specific event might occur on a given schedule.

B. IVEE Learning and Habituation

The learning system used in the IVEE subsystem is more appropriate for learning temporal events than the original learning system used in VENUS. When the IVEE subsystem detects an event, it calculates the center of the spatial event, which determines the 8x8 pixel sub window for comparison of the low-level event features. Each sub window contains membership groups, which are representations of a set of similar events.

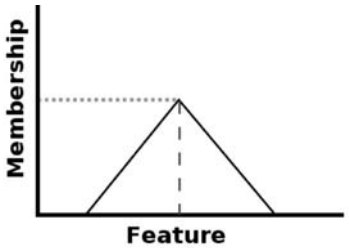

Figure 4. The vertical dashed line represents the mean of the group feature, plotted against the membership level of the group.

The low-level event features, mainly color, speed, direction, size, and time of the event, are compared against membership groups previously represented in the region. As Figure 4 depicts, each feature of the membership group is plotted against the membership level, which is the IVEE subsystem's familiarity with the set of events that the group represents. The higher the membership level, the less interesting the represented set of events will be to the IVEE subsystem. Two line segments are formed by plotting the membership level against the mean of the low-level feature. The slopes of the line segments are calculated from the width of the feature at the zero membership level. The user defines a similarity confidence value from zero to one, which determines how easily an event matches to a membership group. In order to determine if the event matches the membership group by at least the similarity confidence value, the IVEE subsystem linearly averages the similarity of the event features against the membership group features. Figure 5 displays one feature of the witnessed event being compared against the feature mean of the membership group. The average linear value is divided by the membership level in order for groups with large and small membership levels to compete fairly to discover if they match with the witnessed event. If the event features are exactly equal to the group feature means, then the average linear value will be equal to one.

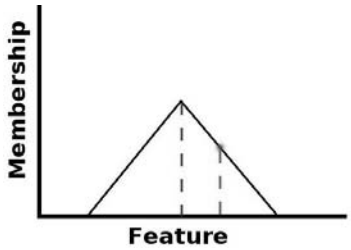

Figure 5. The dashed line on the right represents the event low-level feature, which is being compared against the mean and width of the feature in the membership group.

When an event is matched to a membership group, the IVEE subsystem performs three modifications to the membership group. First, the membership level of the group is increased, which allows IVEE to habituate the events represented by the group. Second, the feature means of the membership group are slightly adjusted toward the event features. Finally, the group feature widths are updated depending on where the event features fell on the line segments. As shown in Figure 6, the user defines a zone for contracting and expanding the width. If the event feature falls within a specified area around the membership group feature mean, the feature width contracts to improve the feature range representation. Similarly, if the event feature falls far from the membership group feature mean, the feature width expands. The width remains unchanged if the event feature does not fall in either the contraction or expansion zones.

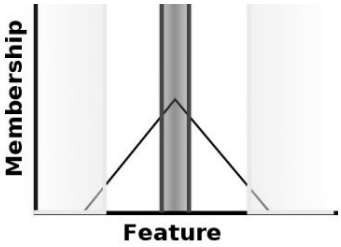

Figure 6. The dark area represents the contraction zone. The light area represents the expansion zone of the feature width.

The initial widths of the membership group will most likely not represent the event correctly after the first occurrence of an event. This is because not enough data exists to set the width properly. Thus, a similar set of events may create multiple membership groups, which eventually will cover similar feature widths and means as the groups adjust to represent the similar set of events. The IVEE subsystem merges similar membership groups together based on two rules:

1. Membership group A has a greater membership level in group B than in group A.
2. Group A is similar to group B.

The IVEE subsystem forgets, or dis-habituates, via one of three user-selectable approaches. The first ap-

proach is the incremental method. The user defines a number of days and a degree of membership level to forget. After a membership group has not been updated for the defined number of days, the membership level drops by the defined amount instantaneously. The second approach for forgetting is the linear method. The user defines the slope, which is the membership level over the number of days. After the processing of each input video frame, the time difference between frames is applied to the linear equation to determine the amount of membership level subtracted from each membership group. The last approach for forgetting is the quadratic method, where the slope is defined as the amount of membership level decrease over the number of squared days. The quadratic method allows the IVEE subsystem to slowly forget about the membership group that was recently updated and quickly forget membership groups that were not updated recently.

When a witnessed event does not pass the similarity confidence value, a new membership group must be created. The IVEE subsystem assigns a maximum degree of interest of one to the event. When the event is matched to an existing membership group, IVEE assigns a degree of interest to an event by taking a linear average across all membership group features. The IVEE subsystem calculates the linear average by comparing the low-level event features against the membership group feature means and widths. The interest value ranges from zero to one. It is calculated by taking the absolute difference of the features and multiplying by the negative membership level over the feature width, then adding the membership level.

C. IVEE Degree of Interest Display

The IVEE subsystem displays the degree of interest as a color in the output video by using the Hue-Saturation-Value (HSV) color space. The saturation and value planes are set to a value of one, while the hue plane is equal to one minus the interest value. This causes the represented color to start at red and transition to orange, yellow, green, blue, and purple. This color representation of the interest value allows people to understand visually the degree of interest assigned to the event by the IVEE subsystem. The IVEE subsystem builds knowledge of when events occur during the week. If a similar event continues to occur at the same time of the week, the event status changes from interesting to normal and, finally, to expected. An event is declared as expected when the membership level of the group surpasses a defined threshold. As Figure 7 depicts, the IVEE subsystem determines if the expected event occurred within the appropriate window of time by first determining if the expected window of time has expired by crossing the zero membership level. If the expected event did not occur within the appropriate time window, then the event is declared as a missed expected event. The IVEE subsystem draws a black and white rectangle of the expected event size and location to the output video.

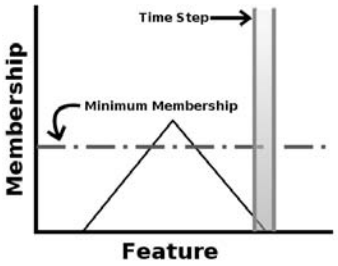

Figure 7. A missed expected event must meet a minimum membership level requirement, and the time width must cross the zero membership level during the current time step, which is usually the time between frames of the input video stream.

V. RESULTS

The subsystem was successfully tested on several video sequences.

The IVEE subsystem learns about events over short and long time periods. Figure 8 shows four frames taken from the same video, which was shown to the IVEE subsystem at five minutes intervals. The IVEE subsystem first shows extreme interest in the events, indicated with the red rectangles about events. The IVEE subsystem shows moderate interest in the same events five minutes later with orange rectangles. The IVEE subsystem shows minimal interest in the same events five minutes after with yellow rectangles. Finally, the IVEE subsystem considers the events normal with the fourth viewing of the same events. Figure 8 shows the same behavior, but with events shown to the IVEE subsystem at the same time of the week over four weeks. Figure 9 shows how the IVEE subsystem reacts to missed expected events.

Figure 8. The IVEE subsystem habituates similar events that occur over a short period, such as five minutes apart.

Figure 9. The IVEE subsystem habituates similar events that occur at the same time each week.

Figure 10. The IVEE subsystem draws black and white rectangles, which indicate where missed expected events have occurred.

VI. CONCLUSION

The integration of the VENUS and IVEE subsystems results in a system that is capable of detecting both spatially and temporally interesting events. In the current framework, we do not make use of any high-level object descriptors. Use of descriptors, such as, cars, buses, trucks, and people can be one of the avenues to explore in the future for content analysis. Object recognition algorithms are currently being developed to categorize objects that are identified as being interesting. The categorization of interesting objects will allow a higher-level description of the scene.

VII. ACKNOWLEDGEMENT

Funding was provided in part by the Intelligence Technology Innovation Center.

VIII. REFERENCES

[1] R. Gaborski, A. Vaingankar, V. Chaoji, and A. Teredesai, "Venus: A system for novelty detection in video streams with learning," in *Proceedings of the 17th International FLAIRS Conference,* South Beach, FL, May 2004, pp. 1-5.

[2] S.-W. Ban and M. Lee, "Selective attention-based novelty scene detection in dynamic environments," *Neurocomputing*, 69, pp. 1723-1727, 2006.

[3] M. Dahmane and J. Meunier, "Real-time video surveillance with self-organizing maps," in *Proceedings of the Second Canadian Conference on Computer and Robot Vision,* 2005, pp. 136-143.

[4] H. V. Neto and U. Nehmzow, "Automated exploration and inspection: comparing two visual novelty detectors," in *Advanced Robotic Systems International,* 2(4), pp. 355-362, 2005.

[5] C. Clavel, T. Ehrette and G. Richard, "Events detection for an audio-based surveillance system," in *Proceedings of the IEEE Int. Conf. on Multimedia and Expo* (ICME 2005), Amsterdam, 2005, pp. 6-8.

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

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

[8] T. Kohonen, eds. *Self-Organization and Associative Memory,* Springer-Verlag, New York, 1988.

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

ROGER GABORSKI is a professor in the computer science department at Rochester Institute of Technology. He received his Ph.D. degree in Electrical Engineering from the University of Maryland and is a member of Sigma Xi, Eta Kappa Nu, and a senior member of the IEEE. He has published more than fifty conference and journal articles and has been awarded 19 patents. Dr. Gaborski's current research interests include biologically inspired intelligent systems and pattern recognition. Please direct correspondence regarding this paper to rsg@cs.rit.edu.

JEREMY C. PASKALI is with the College of Applied Science and Technology, Rochester Institute of Technology, Rochester, NY 14623 USA, where he obtained his M.S. in Computer Science.