
11. Cognitively Motivated Novelty Detection in Video Data Streams

James M. Kang, Muhammad Aurangzeb Ahmad, Ankur Teredesai,
and Roger Gaborski

Summary. Automatically detecting novel events in video data streams is an extremely challenging task. In recent years, machine-based parametric learning systems have been quite successful in exhaustively capturing novelty in video if the novelty filters are well-defined in constrained environments. Some important questions however remain: How close are such systems to human perception? Can results derived from comparing human perception with machine novelty help tasks such as storing (indexing) and retrieval of novel events in large video repositories? In this chapter a quantitative experimental evaluation of human-based vs. machine-based novelty systems is canvassed. A machine-based system for detecting novel events in video data streams is first described. The issues of designing an indexing-strategy or “Manga” (comic-book representation is termed as “manga” in Japanese) to effectively determine the “most-representative” novel frames for a video sequence are then discussed. The evaluation of human-based vs. machine-based novelty is quantified by metrics based on location of novel events, number of novel events, etc. Low-level image features were used for machine-based novelty detection and do not include any semantic processing such as object detection to keep the computational load to a minimum.

11.1 Introduction

Extracting novelty from video streams is gaining attention because of the ready availability of large amounts of video being collected and due to insufficient means of automatically extracting important details from such media. Different ways to summarize video based on novel or important aspects of the video are being explored by a wide range of industries [9, 17, 24]. Businesses that use video conferencing are interested in ways to capture important sections of meetings and make an outline of each meeting available for future reference. Likewise, security/surveillance-based industries are looking for ways to detect novel events in huge streams of seemingly unimportant video data.

We explore interesting ways to generate a cluster index of video frames, based on image features within the frames. Human novelty detection is then compared against a machine-based novelty detection technique. An example of such comparison is shown in Figure 11.1. The frames in the figure are the “representative novel frames” of a cluster found for both human and the machine. Differences and similarities between



Fig. 11.1. Human vs. machine novelty. The top image is the original video frame depicting an office scene where an employee is typing things into a computer. The image on the lower left depicts the novelty component found by a human subject in the study using an eye tracker. The image on the lower right is the novelty as determined automatically by a machine vision system we developed (termed VENUS). The two novelty components for the same video frame show that both the human and machine can find similar parts to be novel.

the humans and machines detected novelties are explored with metrics based on region and location. A framework to cluster the results from the two techniques and show a comparison metric between the two will be discussed and analyzed.

We term this particular framework for indexing, retrieval and human comparison of video novelty detection as VENUS (Video Exploitation and Novelty Understanding in Streams). VENUS is a computational learning-based framework for novelty 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. VENUS uses a simple habituation technique for “remembering” novelties in order to compensate for recurring events within a scene. The eye-tracking system used in the experiments detects novelty items based on certain aspects of human eye tracks such as fixation duration and saccade velocity. However, it can be extended to incorporate many more features.

In this chapter, we first go over related work in different fields and how this work compares with other novelty detection systems. The VENUS framework is then described, followed by a description of how data for human novelty detection

was obtained. We then describe how novel clusters were obtained and indexed. The process of selection of representative frame for creating manga is then described. Lastly we compare the results obtained from novelty detection by humans and the machine system.

11.2 Related Work

11.2.1 Video Streams

In the past, a number of systems have explored novelty in video streams. Video surveillance has been a major concern especially since the September 11 attacks. Diehl and Hampshire [6] examined novelty that occurs within a video and classification of new objects based on previously labeled objects. They initially classified each image with a label, and then classification is done on a sequence of images. There are a number of differences between their system and VENUS. Their system uses a motion detection camera for novelty detection; hence, they assume that all the video consists of motion. In VENUS, on the other hand, still frames can also be considered to be novel. Also it is not apparent what features other than motion were used as a basis for novelty detection in their approach. One can say that the Diehl approach is a comparison of new images with a preclassified set of images to find novel events within a video.

Work by Medioni et al. [19], 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, et al. [11] successfully tracked and detected events in wild-life hunt videos. Research by Stauffer et al. [25] 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. [28] 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 cognitive apparatus of humans and animals is gauged to detect novel changes of the ordinary changes in their environment. This observation has been applied in robotics by Crook et al. [4] and also by Marsland et al. [18]. The former used images taken by a camera for robot navigation while the later uses such images in conjunction with sonar for the same purpose. There are certain similarities and affinities between these two systems and VENUS. Hence the feature set used by VENUS is quite similar to the feature set used by Crooks et al. [4]. Both the systems use color, intensity, and orientation as features. According to Marsland et al. [18], the base concept for their system design is same as VENUS, namely Habituation. Habituation is the idea that as the frequency of repetition of an event increase, the less novel the event becomes.

Other related work deals with extracting features from a video stream using the background of the video [2]. The assumption made by these approaches is that the background will be the same throughout the video. Consequently any changes to

the background causes a novel event. However, if the background changes, then everything would be considered to be novel. Where in the VENUS system the actual content in the video does not affect the novelty found but changes dynamically as described in Section 11.3.1.

11.2.2 Image Novelty

Even though the topic primarily focuses on video streams, image detection is an important step in reaching the goal of analyzing video streams and detecting corresponding novel events. The following applications discussed here use low-level features, similar to VENUS, and also illustrate the utility of this approach.

Consider the case of breast cancer diagnosis. Detecting breast cancer efficiently has always been a problem. According to Tarassenko [27], there are about 26,000 new cases in the United Kingdom each year. On average there needs to be at least two analysts to review an x-ray image to diagnose breast cancer, implying the great need to reduce the time required for diagnosis. Tarassenko [27] describes a tool for analysts to focus on areas with larger mass regions in certain parts of the images, although the process is not fully automated. This allows an expert to look at areas that are most important.

Novelty Detection is done using features of shape, texture, boundary (edges), and contour. These features form the basis of the function that distinguishes between normality and abnormality. A density function based on the feature vector is used to detect novelty. If the density function gives a value that is below a predefined threshold, then the frame is considered to be novel.

11.2.3 Clustering Novelty in Video Streams

Clustering is an effective technique for grouping elements together that are similar to each other. Clustering was an important component in implementing the VENUS framework. Instead of showing every novel event, VENUS shows a summary of novel events that are representative of the whole cluster. Video Manga [30] explains how a video can be summarized and viewed as a comic book or Manga. Manga is a Japanese term that refers to comics. A summary of a video is generated through several steps. Initially, they compressed on all similar regions and then used many hours of office meetings videos for novelty detection. Normally, when a person is talking in a meeting, only the person's mouth and hands are moving. All other body parts remain still. These images are compressed into a single image. This is done for all the events. However, the details of such image comparison were not given.

The algorithm we use is based on low-level features like the pixel count of red, green, and blue, and uses the concept of major colors. If the pixel count of a certain color is greater than a threshold, the image is then considered to have this major color. This is done for all three colors. Once this is completed for all the images, each image will then have a binary representation, where "101" means that they have the Major Colors of Red and Blue. These images are then grouped together with similar binary values. The VENUS clustering method is a modified form of this approach.

11.2.4 Event vs. Novelty Clustering

Novelty clustering should not be confused with Event clustering within a video stream. Any set of frames that are found to be out of normality in a video stream can be considered to be Event clusters since the goal of type of detection is to determine specific periods of time that are different from other time periods. For example, an event could be students walking to a bus or cars on a highway. These can be an event that occurred within a video.

Novelty clustering refers not only to anything that is different from normality, but also to anything that can be deemed unusual. Students walking to a bus or cars driving from left to right are not considered unusual. A car crash can be considered an event and to be novel. The VENUS framework is based on Novel events rather than on any event within a video.

Novelty clustering should also not be confused with temporal event clustering. Even though the novel frames generated from the VENUS system also include the frame number that denotes the time the novel frame occurred, it does not have any effect on the clustering procedure. Similar frames within a video are clustered together. Within temporal event clustering, the time each image is taken is used to determine the event it belongs to [3]. With this said, the novel series cannot be represented as a time series. Lin and Keogh [16] suggest that clustering of time series data is meaningless, but since novelty series is itself a sampling of the raw video consisting of complete description of what was found novel, clustering novelty does produce meaningful results unlike other time series clustering using a sliding window.

11.3 Implementation

11.3.1 Machine-Based Process

11.3.1.1 The VENUS System

Figure 11.2 shows that a block diagram of the VENUS novelty detection model consists of two major components: a focus of attention component that generates the low-level features, and the learning component that handles novelty detection. Since the amount of visual information available within a scene is enormous, we humans “process” only a subset of the entire scene.

Humans tend to focus on the interesting aspects of the scene, ignoring the uninteresting ones. The attention system in our framework is based on a topographically saliency map that represents an object’s saliency with respect to its surrounding. VENUS filters out noninteresting events thereby greatly reducing the amount of information to be analyzed. These interesting events are termed as novel or inconsistent events. The event detection model described in this chapter consists of the following two major components:

- A focus of attention component that generates the low-level features.
- The learning component that handles novelty detection. In this section we describe each of these components in detail.

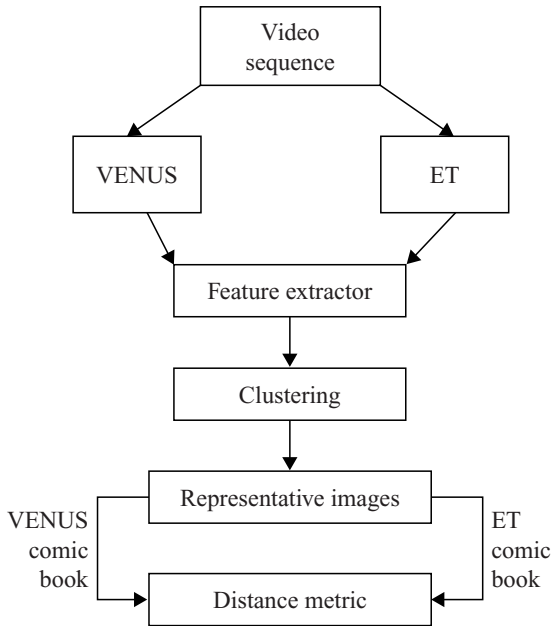


Fig. 11.2. VENOM system diagram.

The first phase of the project focused on detecting novel events. Consider the following example. If a casual observer is positioned on a freeway overpass where the vehicles below are traveling at about the same speed, after a short period of time the observer will generally ignore the individual cars (consistent events), but if a particular vehicle is traveling much slower or faster than the average speed of the other vehicles it is the subject of the observer's attention as a novel or inconsistent event. The VENUS system behaves in a manner similar to a human observer. It will first learn the normal, consistent events in a visual scene, and then detect novel or inconsistent events in the scene. A key point in the system is not programmed to detect fast or slow moving cars, but learns that they are novel in this environment. A common approach in prior work is to first manually define and then to store descriptions of inconsistent events in a database or they are defined using a predefined grammar. Events are then compared to stored events to determine their novelty. The VENUS system thus has the significant advantage of not requiring events to be detected to be predefined, but automatically learns what is normal and detects events that differ from normalcy.¹

¹ The VENUS project is spearheaded at the Center for Advancing the Study of CyberInfrastructure (<http://www.lac.rit.edu>).

11.3.1.2 The Attention System

The VENUS framework is based on the selective attention theory initially modeled by Itti and Koch [13], where a saliency map topographically represents the objects 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 et al. [8]. The 2D spatial filters used in the system are modeled after biological vision principles simulating underlying 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 intensity contrast filtering simulates the function of the retinal ganglion cells that possess the center-surround mechanism. The color information is extracted using the color opponent filters. 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.

Motion information in video sequences is extracted using the 3D spatiotemporal filters tuned to respond to moving stimuli [31]. Motion detection in our system is achieved by using a set of difference of offset Gaussian spatiotemporal filters. Hence 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 and 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.

11.3.1.3 Feature Extraction

The Feature Extraction takes place once the novel frames are generated from VENUS or the eye-tracker shown in Figure 11.2. Low-level features are extracted from a set of images using a color extractor that works on top of the HSV color space. HSV was chosen over RGB because of its similarity with the way in which humans perceive color. Feature sets are created based off of a set of query colors. The extractor scans images or regions of images for each query color and returns a hue score (proximity from the query color) for each pixel within the scanned region. The mean value of all the hue scores for a given query color is calculated and saved as that region's total score. During execution, multiple color features are used in extraction and their results saved as feature sets. These feature sets are used for indexing and clustering the images for later comparison and retrieval.

11.3.1.4 Machine Novelty Detection

Figure 11.3 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

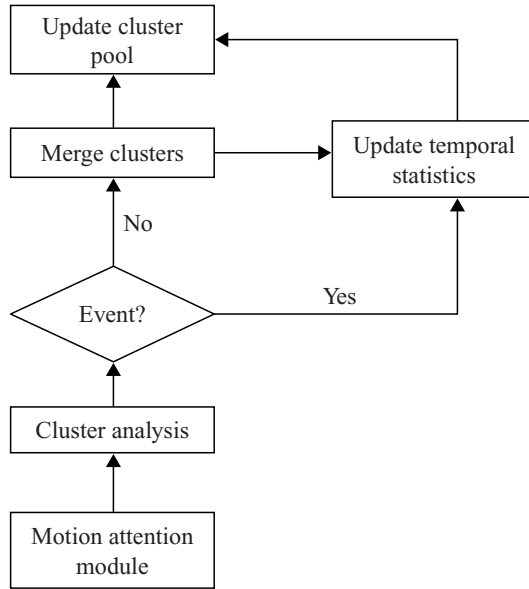


Fig. 11.3. Motion learning and novelty detection module.

area on a frame of the video. A direction map encodes motion values for the direction of motion. The regions of the direction maps that detect motion get excited if there is a change in the direction of motion in successive frames. These direction maps are input to the motion learning and event detection module. Within each region, a Gaussian mixture model represents 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.

The following example illustrates how VENUS novelty detects novelty in video streams. 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 (within 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 11.3. 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. If a similar cluster is already found in the pool, then this implies that a similar event had occurred in the past. Referring back to the example, when multiple people walk at 5 mph over a region, clusters representing their speeds are merged. This indicates

that people walking is not a novel event anymore. 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 [14]. The slower the decay rate the longer is the retention period for the event. The habituation function for a cluster is given by $H(t) = 1 - [1/(1 + e^{-a})]$, where $H(t)$ is the habituation value after t frames 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. A cluster with habituation value below the cutoff 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 1. Value of 0 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 is merged with similar clusters. e/f term indicates the reinforcement (cluster merging) rate. Higher the reinforcement rate, closer the new decay rate to 0. Smaller the reinforcement rate, closer the new decay rate will be to 1.

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 11.3. 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.

11.3.2 Human-Based System

11.3.2.1 Capturing the Eye Track

The Eye tracker is a system that captures eye tracks of how humans observe their environment. The Eye tracker is thus representative human system that is compared to the machine system shown in Figure 11.2. Human eye tracks are recorded while the subject watches the video to be processed. The experimental setup for the Eye

Table 11.1. Eye track example format. This is an example of the type of data that is extracted from the eye track images.

HR	MN	Sec	Total Secs	VPOS	HPOS
14	39	35.617	52775.617	-6.270	-1.645
14	39	35.633	52775.633	-6.245	-1.615
14	39	35.650	52775.650	-6.195	-1.545
14	39	35.667	52775.667	-6.205	-1.545
14	39	35.683	52775.683	-6.250	-1.525
14	39	35.700	52775.700	-6.295	-1.515
14	39	35.717	52775.717	-6.325	-1.500
14	39	35.733	52775.733	-6.385	-1.510

Tracker is as follows:

- The Eye Tracker is first calibrated to align the laser to the person's eye.
- It takes from 5 to 10 min of test video to confirm the calibration of the system.
- The actual test video is shown to the user.
- The system then reads the eye movement information as fixations within the video.

The eye track data are then filtered and modified for calibration and readable representation. This results in an easily parsable format for the attention finder algorithm. The format of this data file (Table 11.1) may vary depending on implementation; however, VENUS requires at least X and Y coordinates across a time series.

11.3.2.2 Extract Scan-Path

The data file is read in and the linear scan-path that the users' eye followed during his or her session is cleaned and saved. The path structure is later analyzed to group fixations together into "Attention Areas."

11.3.2.3 Segment Eye-Tracking Data

Due to the temporal nature of eye track data and the fact that there will almost always be at least one fixation per frame of the video, the data are segmented by a user specified amount. What this does is group together multiple fixations for a certain time period that is later analyzed for groups of fixations that may or may not correspond with Attention Areas (Figure 11.4).

11.3.2.4 Cluster Fixations

The clustering technique used in VENUS is simple, though effective. It can be replaced by a more robust method. For each segment, VENUS determines the centroids of groups of nearby fixations. VENUS uses a simple threshold measure to determine if fixations are "nearby" and thus belong to a particular cluster.

Centroids (Figure 11.5) are created to determine the most likely center point of groups of fixations that will later be used to create the mask that will highlight

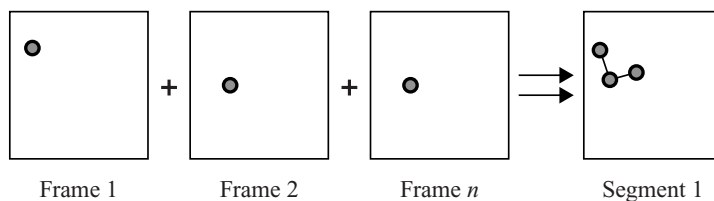


Fig. 11.4. The circle represents single fixation. Each frame has at least one fixation made by the subject. Based on the number of frames that is predefined beforehand, the frames are combined and clustered as Segment 1.

the Attention Areas and ultimately produce the novel image analogous to the Venus output.

11.3.2.5 Novelty Images and Novelty Video

For each centroid in a segment, the Attention Area is determined and a binary mask is created. This binary mask is applied to the corresponding frame from the video and is saved. The effect of combining the binary mask with a frame is to “black out” all areas that are not found as being novel. Novelty frames are then stitched together to create an AVI video of the sequence of novelty frames (Figure 11.6). (This last step is optional but is a good visualization of the progression of detected novelty.)

The novelty images constitute the desired output (Figure 11.5) of this process. The black areas correspond to the mask while the visible regions are the Attention Areas for each of the segments processed. VENUS uses a simple rule (size of area corresponds directly to the number of fixations used to determine the centroid) for determining the actual Attention Area; the rule can be changed should the need arise to use a more sophisticated measure in the future. Presently VENUS uses the first frame of the segment being processed to apply the mask to. It is, however, not known if this is the best way to representation purposes. However, this can easily be modified to use any frame number desired.

11.3.2.6 Experiment Setup and Usability Testing

The eye tracks were captured using the ASL head mounted ES501 system. A magnetic head tracker was not used in the experiments. Subjects were first calibrated and

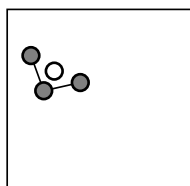


Fig. 11.5. The lighter circle represents a centroid. In a cluster made by the frames in Figure 11.4, a centroid is founded where this is the main fixation made for this set of frames.

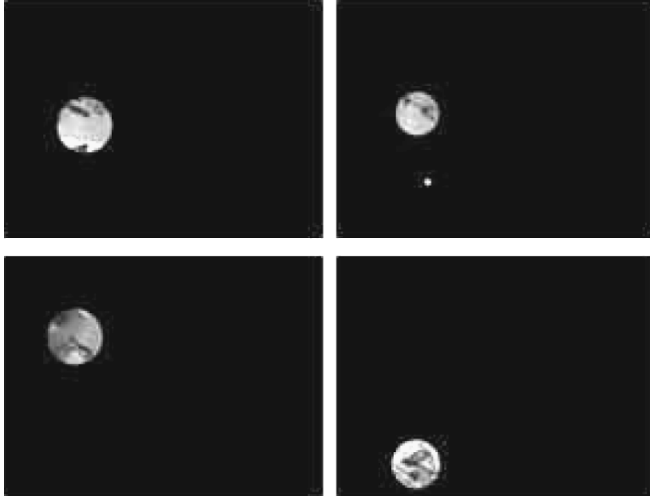


Fig. 11.6. Fixation Example Output. The visible part is the region of attention that is determined by centroids in a manner similar to Figure 11.3. The black areas are the mask that is very similar to how VENUS shows its novel frames.

then were instructed to watch a series of videos and to keep their heads relatively still. Subjects were shown four videos that were mainly from security or office cameras. These videos were chosen for their relatively small amount of motion and consistent viewing angle. After the subject finished watching all four videos, their data were saved and converted to ASCII format for later processing. The subjects who volunteered for our study came from a wide range of ethnic groups which ensured that the results were not skewed because of the person's background.

11.3.3 Indexing and Clustering of Novelty

Video data usually contain a large amount of novel events, although the frequency of novel events varies. For example, surveillance video of a basement warehouse will most likely contain less novel events than say an action movie. When many novel frames are extracted from a video, it is helpful to have an index of novelty frames with which to browse the novelty set. For this reason, VENUS creates this index, using an algorithm described by Mukherjea [20] shown in the system diagram in Figure 11.2.

11.3.3.1 Total Clustering

VENUS creates clusters of similar novelty images based on the feature set. In the current design, only low-level colors are used to generate these clusters. Table 11.2 shows an example of a set of features taken from a video. Here a tuple represents a score for each color in the novel frame. A threshold is created by taking the averages of each color's score. The algorithm (Figure 11.7) is based on the concept that major

Table 11.2. Example of Features. This is an example of a feature set take from novel frames from either VENUS or the Eye-tracker.

Red	Green	Blue
85	66	100
25	63	44
10	12	90
85	30	98

colors and pixel counts can be used to determine the most prevalent color of the image. In VENUS's implementation color score was substituted by the pixel count. Instead of using just the average as the threshold, a confidence interval is needed to ensure that features close to the average are included as part of the cluster. Without this confidence interval, very similar frames may be in different clusters. The confidence interval estimation is defined as: $\bar{x} \pm Z * (S/\sqrt{n})$, where \bar{x} is the threshold, Z is the interval coefficient, S is the standard deviation of a feature within the population, and n is the size of the population. A confidence interval of 95% ($z = 1.96$ for normal distribution) was expected for the current implementation. The standard deviation of each feature set is expressed by $s_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$, where s_N is the standard deviation of N , N is the size of the population, x_i is each feature value, and \bar{x} is the average of the feature population. For each color in Table 11.2, a 1 or a 0 is assigned if that color meets the threshold \pm confidence interval. Hence each frame will then be described by a binary number. Similar combinations are then clustered together. The algorithm for this process is shown in the following figure.

```

1 For each image
2   Calculate the sum of each color
3 end for
4 Find the average of each color and
5   store as threshold
6 For each image
7   For each color
8     If color is greater than threshold
9       Set "1" to color
10    Else
11      Set "0" to color
12    end for
13 end for
14 Group Images based on Binary Values

```

Fig. 11.7. Total clustering algorithm. The word *total* emphasizes that the algorithm does not pay attention to time or the frame number. It clusters the whole set of frames for its similarity.

11.3.3.2 Sequence Clustering

An interesting method of creating clusters is to identify the start and end points of each of the novelty that occurs within a video, instead of comparing all the novel frames within a video to find similar frames. The sequence of the novel frames that are similar is clustered. Figure 11.7 shows the sequence clustering algorithm. For example, suppose that there were five novel frames in a video, each occurring right after the other. If the distance between frames 1 and 2 is under the predefined threshold, then they are put within the same cluster. On the other hand, if the distance between frames 3 and 2 is greater than the threshold, then a new cluster is generated. This will create sections within the video stream of when a new novel event occurs. There are numerous forms of distance metrics that can be used such as Euclidean, Mahalanobis, Minkowski, Block Row, and Chebychev. Euclidean distance was used for clustering in the present case for simplicity. Each image can then be represented as a feature vector consisting of low-level features. Thresholds are created on the basis of the average or median of all the distances.

The total distances are not in a form of a matrix since the distance was based on the previous and following frame. Once a cluster has been created, a representative image of the cluster needs to be found to facilitate presentation in a comic book format.

Although the above algorithm describes the process of clustering together visually similar images, it does not prescribe a particularly good way of representing each cluster. To solve this problem VENUS uses pixel scores to get mean values for each color. Next, a distance is calculated for the features of each image with respect to the mean. The image whose features have the smallest distance with respect to the mean is used as the representative image. Representative images are then arranged on the

1	For each image
2	Calculate Euclidean distance between
3	image and the next image
4	end for
5	Find the average of all distances and store
6	as threshold
7	Find Confidence Interval (\bar{x})
8	For each image and distance
9	If image and next image distance is less than
10	threshold $\pm \bar{x}$
11	Cluster images together
12	Else
13	Create new cluster
14	end for

Fig. 11.8. Sequence clustering algorithm. Sequence means that time is a factor within this algorithm. A frame can be clustered only if it is similar enough to the previous frame based on a threshold.

```

1 For each cluster
2   Calculate total amount of color pixels
3   Add into the cluster average of pixels
4   For each image
5     Calculate distance from image to average
6   end for
7   Find shortest distance between image and average
8   Assign image as representative cluster
9 end for

```

Fig. 11.9. Representative image algorithm. Finds the centroid of the novel frame clusters for both the VENUS and Eye-tracker systems.

basis of the time when the respective frame occurs in the video. The images are then laid out in a comic book-like format (Figure 11.10).

11.3.4 Distance Metrics

11.3.4.1 Location Similarity

The first metric used by VENUS is location similarity. Novel frames usually contain only a small amount of novel area. The task here is to extract the actual location on the image where a certain novelty is present and to compare that with the location of novelty from a corresponding frame in order to determine whether or not the novelty detected by both systems captures the same location.

Novel location regions are extracted from the images using a recursive dissection technique. The algorithm recursively breaks up the image into subregions until a specified depth using quad-trees. Not only are novel regions quickly located, but a hierarchy of such regions is also created. This hierarchy can then be used later for further feature comparison, as in the case of the feature similarity metric. Figure 11.11 shows the comparison of similar regions between two novel images and the corresponding scores.



Fig. 11.10. Example of VENUS Manga. This is the represented frames generated by the algorithm in Figure 11.8. The frames are organized by time or frame number that represents a comic book or a summary of novel events within a video.

11.3.4.2 Feature Similarity

The second metric is feature similarity. Feature similarity is an abstract concept and can therefore be used for comparison of any type of comparable feature. In the case of VENUS, pixel color is the comparable feature. Regions of an image are compared to regions of other images on the basis of the mean hue scores generated by the feature extractor. The regions to be compared are identified via the location similarity metric detailed above. The feature similarity metric is able to identify regions across images that are similar to each other by using the hierarchy of similar regions. The user can specify how detailed of a comparison to perform, which facilitates fast comparison across very different images as well as detailed comparison between images that are very similar. Figure 11.11 shows scores for the feature similarity metric.

Figure 11.11 shows a visualization of the metrics VENUS uses. The second image is compared to the first, while the third image shows the novel regions as well as the intersection of those regions. Scores are calculated based off of the two metrics and printed on the middle image. As one can see, although the first image has a 0% location similarity to the second, their features are still 98% similar. Likewise, the features of the second comparison are 99% similar, with a location similarity of 15%.

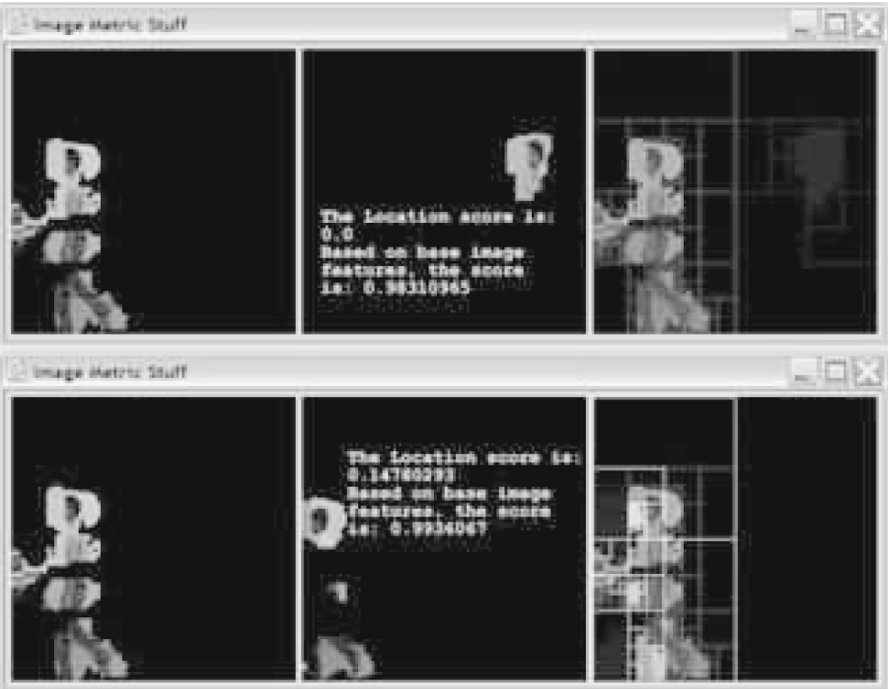


Fig. 11.11. Location and feature similarity metrics. The top image shows that the location feature is large since the two novel images are not within the same region. The image below shows that the distance feature is small since the two novel images overlap each other.

These image metrics enable VENUS to compare entire sets of novel images across different novelty detection systems shown in Figure 11.2.

11.4 Results

11.4.1 Clustering and Indexing of Novelty

11.4.1.1 Total Clustering Approach

VENUS' motion maps alone can give good visual results, although these results are not sufficient for identifying novelty. Figures 11.12 and 11.13 show a comparison between using motion versus all three measures.

VENUS' and the Eye-Tracking novel images were clustered using the Total Clustering algorithm discussed in Section 11.2.3. The results from this process were good. Images with similar features were grouped together into clusters that were later linked via their representative images. Figure 11.14 shows a graph of clusters obtained for a set of the Venus data. Since low-level features were used to cluster and index, images within clusters are visually very similar. New features can easily be incorporated to improve these results. The human-based novelty frames were also processed in the same manner and are also separated into visually similar groups.

11.4.1.2 Sequence Clustering Approach

Figure 11.14 also shows two types of distance equations within the graph. The first measures the distance from an image within a cluster to the representative frame and can be expressed as $D(C_i, I_j)$, where C_i is the centroid of the cluster and I_j is the image within that cluster. This distance is in terms of the difference in value of features in the base image and the query image (base image being Venus and

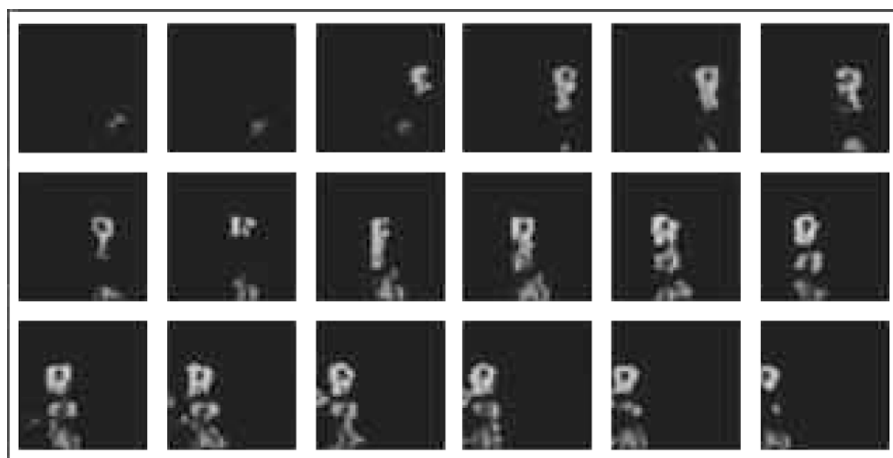


Fig. 11.12. Three measures: motion, still, and color.

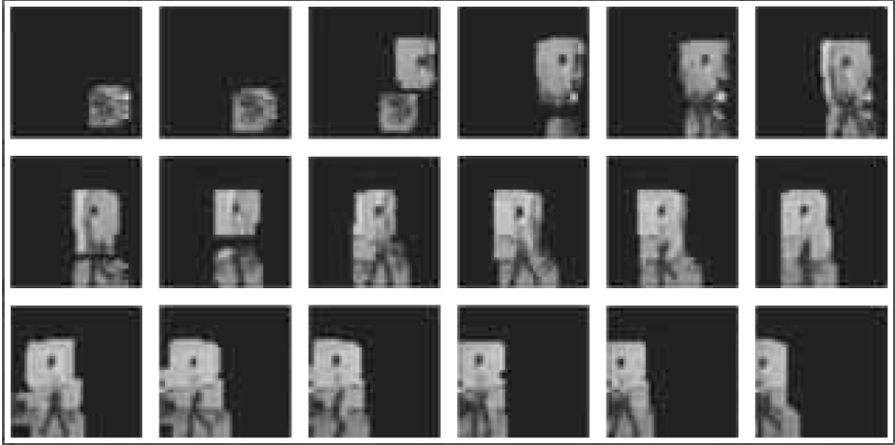


Fig. 11.13. One measure: Motion. VENUS using only the Motion measure. Since the goal of our project is to compare novelties between two different systems, the most visible novel areas should be used. This is why we chose to use just motion instead of all three shown in Figure 11.12.

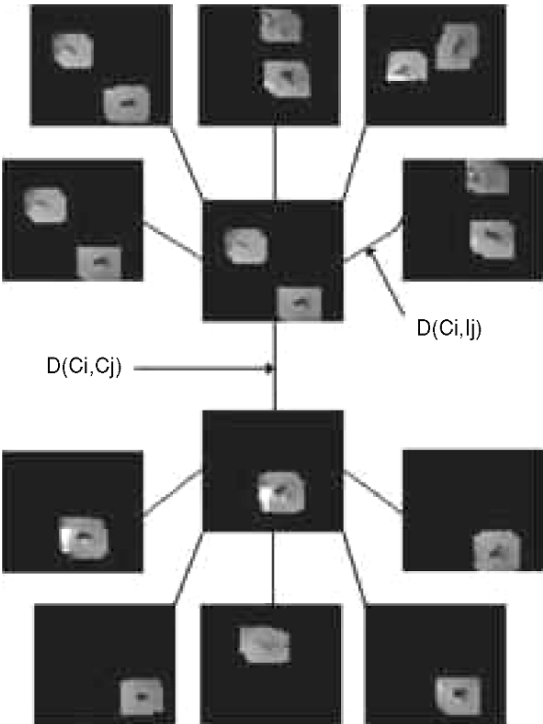


Fig. 11.14. Clustering example. This is an example of two clusters. The links within the cluster show the Euclidean distance between the cluster elements and the centroid. The distance between the two centroids are based on time or frame number.

Table 11.3. Human vs. machine total clustering. This is a comparison to show that the number of clusters generated by the Total Clustering algorithm, for both machines and humans is very similar. The average number of human clusters is based on 11 testers with each having 10 novel frames.

Video	Total number of novel frames	Machine clusters	Average human clusters
1	195	7	3
2	181	6	4
3	227	6	3
4	205	5	3

query image being the E.T. novel image). The second measures the distance between representative images within a video and can be expressed as $D(C_i, C_j)$, where C_i is the centroid of each cluster. Here, distance is expressed in terms of time, or frame number, and is used for the layout of the final comic book strip. After clusters are temporally linked together, each cluster's representative image is displayed and the comic book is formed.

The novel frames generated from Venus and the Eye Tracker were clustered together with this interesting clustering approach. The representing frames were generated in the same manner as the Total Clustering approach. The clusters are very similar to the ones shown in Figure 11.14, but each novel frame within a cluster was found to be within the same sequence. The results from this approach was very different from the results in the Total Clustering approach. The clusters generated by the previously described algorithm exhibited more intracluster similarities as compared to the Sequence Clustering approach. Table 11.3 shows the average number of clusters generated for the human-based and the machine-based approach using the Total Clustering algorithm.

Within the Eye Tracker and the Venus system, there were a total of 4 videos and 10 users for the Eye Tracker that we experimented with. These values are based on the average of those results. Each of these videos had very similar time duration even though the contents of these videos came from different domains.

Figure 11.8 shows the same statistics obtained via the Sequence Clustering algorithm. These results show that there is a significant difference between the human and machine novel approaches. This also shows that we are still far from reaching human-like novelty detection level. The number of clusters differ greatly against the total clustering algorithm shown in Table 11.4. This shows that Machine novelty detection is still significantly different from Human novelty detection.

Table 11.4. Human vs. machine sequence clustering. This is a comparison between machine and human clusters using the sequence algorithm.

Video	Total number of novel frames	Machine clusters	Average human clusters
1	195	69	2.8
2	181	57	3.0
3	227	42	4.2
4	205	60	3.4

11.4.2 Human Novelty Detection

Human novelty detection proved to be much more subjective, which was expected. Many factors came together to influence a subject’s tests but all subjects’ results were very similar as far as Attention Areas were concerned. First time novelty within the video gained high attention and quickly dropped down as new novelty was introduced. Essentially, humans tend to focus upon novelties “one at a time” and for short periods. On the other hand, a machine-based approach like Venus is capable of detecting all novelties simultaneously within a video.

The comic strip based on human novelty detection that was created was very homogeneous. It was observed that low-level features were not sufficient in creating and using a distance metric for comparison of novel images. This conclusion is a result of the manner in which representative images were produced. Using the similarity average of low-level features and the similarity of shared locations resulted in the selection of the largest Attention Area found within the subject’s novel image set as the closest pairing to each of VENUS’ representative images. This shows us that higher level features need to be considered in order to get more accurate scores between a subject’s novelty and VENUS’ novelty.

11.4.3 Human vs. Machine

The most apparent difference that was observed between the human novelty and VENUS’ novelty was in the apparent habituation of incoming novel areas (see Table 11.5 and Table 11.6). Venus uses a habituation technique where new novel events are “remembered” for a while and their significance to the current novelty within the video gracefully degrades over time. An example of where this is useful would be a busy airport. Initially, dozens of people walking around the airport would be very novel and Venus will display them as novelty within the scene. Over time, VENUS becomes used to dozens of people walking around and will not register them as being novel any longer. The real use of the habituation is in a situation where a person in the same airport drops a red bag on a chair and walks away. VENUS will pick this red bag up as being very novel and therefore show it in its results.

It has been observed that humans tend to have this same type of habituation [7]. Humans tend to pick up novelty within the video very quickly and, if it was not a major event, return their attention to the previous Attention Area just as quickly. This

Table 11.5. Total clustering: The table shows the raw data of every subject’s novelty clusters for each video we tested against the number of clusters generated from VENUS-based novelty detection.

	Total number of novel frames	Machine clusters	Human clusters									
			1	2	3	4	5	6	7	8	9	10
Video 1	195	7	3	4	3	2	3	3	3	3	3	3
Video 2	181	6	7	4	2	2	3	4	2	3	2	3
Video 3	227	6	2	2	3	4	2	4	3	3	2	3
Video 4	205	5	4	2	3	4	2	2	3	2	3	4

Table 11.6. Sequence clustering: This table shows the raw data of every subject’s cluster for each video we tested against the number of clusters generated from VENUS based novelty detection.

	Total number of novel frames	Machine clusters	Human clusters									
			1	2	3	4	5	6	7	8	9	10
Video 1	195	69	2	3	4	1	3	2	1	1	4	4
Video 2	181	57	3	3	3	3	3	1	2	1	2	4
Video 3	227	42	4	4	4	4	3	4	3	4	4	2
Video 4	205	60	3	4	2	4	4	2	2	4	2	3

is an important aspect of the comparison between VENUS and a Human’s novelty detection mechanisms and will be discussed in the next session.

Human novelty detection using segmented video and low-level feature extraction does not produce the quantity of novel frames for a comprehensive comparison to a machine-based approach. This was shown by the homogeneity of the final comic Human strip as well as the similarity among scores of each human novelty image. There does not seem to be any way in which to increase the number of novel frames without either repeating much of the novelty information in each frame or losing the high-level Attention Areas to many small fixation areas that would not visually show any usable results.

11.5 Discussion

11.5.1 Issues and Ideas

11.5.1.1 Venus

At present, the VENUS system does not gather semantic information about how relevant incoming novelty may be. Consequently it is difficult to model VENUS’ habituation algorithm to focus on “important” novelties rather than all novelty, where important novelties are defined as what is considered novel by a human being. It may be however possible to tweak the habituation mechanism to be trainable by a human expert whose job it is to discern between “important” and “unimportant” novelty. A system such as this would have great impacts on security surveillance systems, gambling monitors, highway safety, and many other areas. How this type of system may be implemented is still an open question.

Video indexing has not received as much attention in the community as novelty detection. With regards to video indexing, it is important to note the work of Furht and Saksobhavit on video indexing [7], Sun, Majunath, and Divakarn, on using motion to detect different levels in a video to index [26], and Detyniecki, who uses color features within key frames of a video for indexing [5]. Figure 11.15 gives an example of novelty detected by VENUS over the course of a video.

Another issue for Venus is the preprocessing time to generate novel images. Currently VENUS can take only uncompressed images, which can take up quite a bit of space. For example, consider the case when a 4-min compressed. AVI was

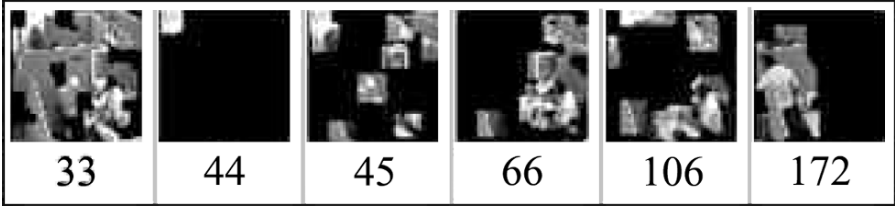


Fig. 11.15. Example of novelty detected in a Video with the respective frame numbers

used. When the file was uncompressed, the size of the decompressed file increased from 50 Mb to about 2 Gb. This .AVI movie had about 8000 frames and it took close to 30 s for each frame to be processed by Venus. This is a scalability issue and is an area of future development. By looking over the results from the Human novelty detection, it can be suggested that within the particular domain we experimented with (surveillance video) Human novelty detection most closely mirrored VENUS' motion images. This could be a way to speed up VENUS' processing time; i.e., by only processing the color and still maps when needed as opposed to every frame.

11.5.1.2 Eye Tracking

The issues concerning the eye-tracking experiments and setup mainly revolved around proper calibration of the subject's gaze. Without the support of a magnetic head tracker, each subject's gazes had some error associated with them. The error was reduced for many subjects by gradual tweaking. However, the error persisted for some subjects even after tweaking. The nature of the experiments however did not call for extremely precise measurements and can therefore receive little alarm.

For any real-time training of VENUS to take place by a Human, a noninvasive eye tracker would be needed. Remote eye trackers such as Tobii 1750 [5] or ASL's 504 HS model [6] would be ideal for this type of scenario. With these newer, noninvasive eye tracking systems many real-time applications of this sort would become possible.

11.5.1.3 Novelty Comparison

One of the goals of this chapter was to compare machine-based novelty and Human-precieved novelty from the eye tracker. It was observed that the number of novel clusters detected by the machine is far greater than the novelty clusters detected by a human. This is to be expected since VENUS exhaustively finds novel events in the video streams and as previously mentioned in Section 11.2.2.

Anything out of the ordinary is considered to be novel by VENUS. In this regard it can be said to perform better than the human. On the other hands, VENUS does not really distinguish between "important" or "unimportant" events that may be context dependent. This situation can be conceptualized as follows: Video novelty is just a subset of what is considered novel by a human being, e.g., if the video is that of a conversation between different people, in addition to visual changes in the video as being considered novel, any sudden change in the topic of discussion would also be

considered novel by a human. Such novelty is, however, not currently included in the scope of this project.

11.5.1.4 Future Work

A possible future project may involve using VENUS's framework to create graphs of novelty such as those described in Section 11.3.1. Mining these graphs may give a way to process difficult queries such as "Show me all of the events in videos X, Y and Z where this type of novelty occurred" or "Search for novel events that tend to lead up to a certain type of novel event" and others. To mine digital media in such a way, one needs a traditional structure in which to store these types of events. Future analysis of novelty, novelty clusters, novelty graphs, and representative images could lead to this type of data structure, on which more traditional search strategies can be applied.

11.5.2 Summary

A machine novelty detection system called VENUS was described in this chapter. VENUS successfully employs the theory of habituation for learning novel events over time in a video data stream. The utility of this approach was demonstrated by a series of experiments. VENUS does not use semantic information for novelty detection but rather uses low-level image features. An attempt was made to compare machine-based novelty detection scheme with the human-based novelty scheme gleaned from the eye-tracking system. It was, however, observed that the low level features for human based novelty detection were inadequate for such a comparison.

The data from the machine novelty system VENUS and novelty perceived by a human as recorded by the Eye Tracker were used to analyze different novelty detection strategies, compare them, and lay them out in a manga-like format. Important events in the video were described by the novelty detected in the video. In addition, indexing schemes for clustering and indexing novel frames was also implemented and discussed. A framework for comparison between the machine-based and human-perceived novelty was also implemented and tested. The results of the respective tests and experiments were listed and reviewed.

11.6 Acknowledgments

We thank the Vision and Data Mining Group for allowing us to use VENUS as the basis for novelty detection comparison. We also thank the Usability Group at RIT for allowing us to use their eye tracker. And lastly, we thank all of the volunteers who participated in the eye-tracking experiments for their time and patience.

References

1. Applied Science Laboratories [homepage on the Internet] Available from: <http://www.a-s-l.com>

2. Burl MC. Mining patterns of activity from video data. *SIAM Int. Conf. on Data Mining*, April 2004.
3. Cooper M, Foote J, Girgensohn A, Wilcox L. Temporal event clustering for digital photo collections. In: *Proceedings of the 11th ACM International Conference on Multimedia*, Berkeley, CA, 2003, pp. 364–373.
4. Crook P, Marsland S, Hayes G, Nehmzow U. A tale of two filters—On-line Novelty Detection. In: *Proceedings of International Conference on Robotics and Automations (ICRA'02)*, Washington, DC, 2002: 3894–3900.
5. Detyniecki M. Discovering indexing rules for video-news. In: *Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems—EUNITE'2002*, Algarve, Portugal, September, 2002: 44–6.
6. Diehl CP, Hampshire JP II. Real-time object classification and novelty detection for collaborative video surveillance. In: *Proceedings of the 2002 International Joint Conference on Neural Networks*, 3: 2620–2625.
7. Furht B, Saksobhavit P. A fast content-based video and image retrieval technique over communication channels. In: *Proc. of SPIE Symposium on Multimedia Storage and Archiving Systems*, Boston, MA, November 1998.
8. Gaborski R, Vaingankar V, Chaoji V, Teredesai A, Tentler T. VENUS: A system for novelty detection in video streams with learning. In: *Proceedings of the 17th International FLAIRS Conference*, South Beach, FL, 2004.
9. Journal of Net Centric Warfare [homepage on the Internet] Navy SEALs Using New Video Storage and Editing Laptop [cited 2005 February 14] Available from <http://www.isrjournal.com/story.php?F=658407>
10. Hayashi A, Nakashima R, Kanbara T, Suematsu N. Multi-object motion pattern classification for visual surveillance and sports video retrieval. In: *Proceedings of the 15th International Conference on Vision Interface*, Calgary, Canada, 2002.
11. Haering NC, Qian RJ, Sezan MI. A Semantic Event Detection Approach and Its Application to Detecting Hunts in Wildlife Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 1999;10:857–868.
12. Keim D, Sips M, Ankerst M. Visual data mining. In: *Visualization Handbook*, Eds. Johnson C.R., Hansen C.D., Academic Press, 2004.
13. Itti L, Koch C. Computational modeling of visual attention. *Nature Neuroscience Review*; 2001;2(3):194–203.
14. Kohonen T. *Self-Organization and Associative Memory*. New York: Springer-Verlag; 1988.
15. VirtualDub [homepage on the Internet]. Lee A. Available from: <http://www.virtualdub.org>
16. Lin J, Keogh E, Truppel W. Clustering of streaming time series is meaningless. In: *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003;56–65.
17. TechTrax [homepage on the Internet]. Holographic Video Storage. TechTrax; c2002-2005 [cited 2005 Dec 9] Available from: <http://pubs.logicaexpressions.com/Pub0009/LPMArticle.asp?ID=118>
18. Marsland S, Nehmzow U, Shapiro J. Detecting novel features of an environment using habituation. In: *Proceedings of Simulation of Adaptive Behavior*, MIT Press 2000; 189–198.
19. Medioni G, Cohen I, Brmond F, Hongeng S, Nevatia R. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2001;23(8):873–889.
20. Mukherjea S, Hirata K, Hara Y. 2000. Using clustering and visualization for refining the results of a WWW image search engine. In: *Proceedings of the CIKM 1998 Workshop on*

- New Paradigms in Information Visualization and Manipulation (NPIV 1998)*, Nov 3-7, 1998 ACM. 1998;29-35.
21. Nairac A, Corbett-Clark T, Ripley R, Townsend N, Tarassenko L. Choosing an appropriate model for novelty detection. In: *Proceedings of the 5th IEEE International Conference on Artificial Neural Networks*, Cambridge, 1997;227-232.
 22. Qiu G, Ye L, Feng X. Fast image indexing and visual guided browsing. In: *Third International Workshop on Content-Based Multimedia Indexing*, Sep 22-24, 2003 IRISA, Rennes, France.
 23. S. Singh, M. Markou. An approach to novelty detection applied to the classification of image regions. *IEEE Trans. Knowledge Data Eng.* 16(4);Apr, 2004; 396-407.
 24. Streamload [homepage on the Internet]. Available from: <http://www.streamload.com/>
 25. Stauffer C, Grimson E. Learning Patterns of Activity Using Real-Time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2000; 22(8):747-757.
 26. Sun X, Manjunath BS, Divakaran A. Representation of motion activity in hierarchical levels for video indexing and filtering. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Rochester, NY, Sep 2002: 149-152.
 27. Tarassenko L. Novelty detection for the identification of masses in mammograms. In: *Proceedings of the 4th IEE International Conference on Artificial Neural Networks*, Cambridge, UK, 1995, 4:442-447.
 28. Tentler A, Vaingankar V, Gaborski R, Teredesai A. Event Detection in Video Sequences of Natural Scenes. In : *Western New York Image Processing Workshop*, Rochester, New York, 2003.
 29. Tobii Technology. [homepage on the Internet]. Available from: <http://www.tobii.se>.
 30. Uchihashi S, Foote J, Girgensohn A, Boreczky J. 1999. Video Manga: Generating Semantically Meaningful Video Summaries. In: *Proceedings ACM Multimedia*, (Orlando, FL) ACM Press, October 30, 1999; 383-392.
 31. Young RA, Lesperance RM, Meyer WW, The Gaussian Derivative model for spatialtemporal vision: I. Cortical Model. *Spatial Vision*, 2001;14(3,4);261-319.
 32. Zhu X, Fan J, Elmagarmid AK, Wu X. Hierarchical video content description and summarization using unified semantic and visual similarity. *Multimedia Syst.* 2003;9(1):31-53.
 33. Zhu L, Rao A, Zhang A. Advanced feature extraction for Keyblock-based image retrieval. In: *Proceedings of the 2000 ACM Workshops on Multimedia*, Los Angeles, CA, 2000, pp. 179-183.