VRIJE UNIVERSITEIT BRUSSEL
FACULTY OF ENGINEERING
Department of Electronics and Informatics (ETRO)
Image Processing and Machine Vision Group (IRIS)

Visual Attention Framework: Application to Event Analysis

Thesis submitted in fulfilment of the requirements for the award of the degree of
Doctor in de ingenieurswetenschappen (Doctor in Engineering)
by

Geerinck Thomas

Examining committee
Prof. Hichem Sahli, Vrije Universiteit Brussel, promoter
Dr. Valentin Enescu, Vrije Universiteit Brussel, co-promoter
Prof. Ann Nowé, Vrije Universiteit Brussel
Prof. Rik Pintelon, Vrije Universiteit Brussel
Prof. Werner Verhelst, Vrije Universiteit Brussel
Prof. Eric Soetens, Vrije Universiteit Brussel
Prof. Jan Cornelis, Vrije Universiteit Brussel
Prof. Robert B. Fisher, University of Edinburgh
Dr. Lucas Paletta, Institute of Digital Image Processing

Address
Pleinlaan 2, B-1050 Brussels, Belgium
Tel: +32-2-6291300
Fax: +32-2-6292883
Email: tgeerinc@etro.vub.ac.be

Brussels, May 2009

Contents

List of Figures

List of Tables

Abstract

One of the monolithic goals of computer vision is to automatically interpret general digital images or videos of arbitrary scenes. However, the amount of visual information available to a vision system is enormous and in general it is computationally impossible to process all of this information bottom-up. To ensure that the process has tractable computational properties, visual attention plays a crucial role in terms of selection of visual information, allowing monitoring objects or regions of visual space and select information from them for report, recognition, etc. This dissertation discusses one small but critical slice of a cognitive computer vision system, that of visual attention. In contrast to the attention mechanisms used in most previous machine vision systems, which drive attention based on the spatial location hypothesis, in this work we propose a novel model of object-based visual attention, in which the mechanisms which direct visual attention are object-driven. Considering the temporal dynamics associated with attention-dependent motion, an attention-based visual motion framework is also proposed. Finally, since a vision system will always have a set of tasks that defines the purpose of the visual process, a top-down approach is proposed to define the competition of the visual attention occurring not only within an object but also between objects, and illustrated in the framework of a surveillance system.

Chapter 1

Spatiotemporal Attentional Selection of Active Salient Objects

1.1 Visual Event Detection in Computer Vision

In the context of video analysis, a visual event is commonly defined based on a moving object with constraints in its size, color, shape. Also, motion instances which are not accepted into the definition of normal, are regarded as events [7, 33]. Moreover, motion instances which haven't been seen before are considered novel events. In the literature, an important amount of work has been carried out addressing novelty detection as attentional event detection.

Subsequently, we start by extensively discussing the existing state of the art on novelty detection approaches. Special attention has been devoted to approaches applying data mining techniques to identify the time points at which changes (i.e. events, novelties) occur. In the literature, this has been called the change point detection problem. Of course, novelty detection is not the only point of view towards visual event detection in video. We present an overview of approaches to video event detection and video analysis. All these methods share visual attention principles as starting point with the purpose of the extraction of regions of interest in video at a later processing stage.

1.1.1 Novelty Detection

The ability to identify perceptions that were never experienced before, referred to as novelty detection, is an essential component of intelligent agents aspiring to operate in dynamic environments. Animals, for example, are able to quickly detect and focus their attention in unusual situations by using different sources of sensory information. Novelty detection mechanisms and, more generally, attention mechanisms are extremely important competencies, which maximize chances of survival, not only because they help to reduce threats and exploit opportunities, but also because they enables the animal to learn from experience.

Novelty detection is performed by means of a so called *novelty filter* [25]. The general term "novelty filter" refers to all learning mechanisms which acquire a model of normality from the environment and are able to use it to filter out abnormal inputs. One of the crucial issues is the selection of feature vectors,

which are simplified abstractions of the original visual aspects, are expected to describe relevant characteristics and eliminate unnecessary details. They reduce dimensionality of the data to be processed by the novelty filter while trying to preserve the ability to discriminate between different classes of features as much as possible.

Because a model of normality needs to be acquired prior to the use of the novelty filter in inspection tasks, the experimental procedure is divided into two phases. First, exploration of the environment takes place with learning enabled so that the model of normality can be acquired. During the exploration phase, performance of the learning mechanism can be evaluated. After the model of normality is acquired for a particular environment, the trained system can be used in inspection tasks to filter out any abnormal perceptions in that context.

Novelty detection has been used in problems where large amounts of data exist in which the result of the test is negative, and relatively few examples of the important features that have to be detected. It is therefore usually not possible to install or learn models of abnormality, because too little training data is available, if any (in some cases, one often does not know even what to look for). Instead, a model of normality is acquired and used to filter out any input stimulus that does not fit the learnt model [21].

The implementation of novelty detection systems is usually based on statistical approaches [19] or artificial neural networks [20]. In both cases, a model of normality is built and used to filter out any previously unobserved situation. Depending on the algorithm being used, learning can be performed in a supervised or unsupervised manner, either in batch (off-line) or continuous (on-line) mode. A new approach for novelty detection, based on a model of *habituation*, has been proposed in [22]. The use of habituation, a reversible response reduction to repeated stimuli, allows not only to detect new perceptions but also to quantify their degree of novelty.

Event or novelty detection in video sequences has been extensively researched from an engineering perspective. The VSAM system developed by Medioni et al. [24] is an example of the real time system processing for events. Semantic video detection approach by Haering et al. [11, 10] successfully tracked and detected events in wild life hunt videos. In the field of robotics, the concept of novelty is linked to comparisons between a pre-acquired memorized environment representation with current sensory data in order to detect deviations. For example, Marsland et al. [23] have presented an autonomous robot that senses an environment through 16 sonar sensors and produces a novelty measure for each scan relative to the model it has learned.

In [7], a system for novelty detection in video streams is presented based on the low-level features extracted from a video sequence and a clustering based learning mechanism that incorporates habituation theory. The system is named VENUS: Video Exploitations and Novelty Understanding in Scenes. Initially any form of event in the scene is flagged as novel. Over time, as the system learns events it tends to consider this as normal behavior and habituates. An event can be novel by virtue of any of its low-level features or a combination of them. In [34] the primitives of the learning aspect are inspired by biological theories such as habituation. In [1] a biologically motivated novelty scene detection model is proposed, in which an input scene is represented using a topology of a visual scan path obtained from

a scaled saliency map generated by a visual attention model. In order to indicate novelty of a current input scene, the topology of the current input scene is compared with that of a previously experienced scene. In addition, the energy signature with scale information in the corresponding visual scan path is also considered to decide novelty of the input scene.

More recently, Neto and Nehmzow [26] have combined an attention model with a novelty filter. Rather than processing novelty over the entire scene, the novelty filter only processes salient locations in order to improve performance.

In [29], novelty is divided in perceptual novelty, real novelty, partial novelty, contextual novelty, and semantic novelty; where perceptual novelty refers to the static saliency map. These different novelty types are then combined into a master novelty map.

1.1.2 Change Point Detection

The change point detection approaches apply data mining techniques to identify the time points at which the changes, i.e. events, occur. This has been discussed in several applications:fraud detection [4], rare event discovery [37], event/trend change detection [8], and activity monitoring [6].

In standard statistical approaches, change point detection has been made by (a) *apriori* determining the number of change-points to be discovered, and (b) deciding for the model to be used for fitting the subsequence between successive change-points [9, 12, 14].

A standard assumption in using data mining techniques to extract interesting patterns from temporal sequences is that the raw data collected from the sensor is somehow (pre)processed to generate a sequence of events. Considering learning-based approaches, for example, a prior model must exist that is both sophisticated enough to model the application and computationally tractable for deriving the posterior model. However, deriving an event sequence from raw sensor data, in absence of any knowledge of models or possible patterns and events, requires a more systematic approach in terms of processing.

Moreover, for the change point detection mechanism to be effective, the following requirements are stated:

- The detection process should be online. A change point should be detected as soon as possible, after it has appeared.
- The detection should be adaptive to non-stationary data sources. A change point should be detected even if the nature of the data source may vary over time.
- The detection is performed unsupervised. There is no first learning step of an underlying model of data-generation mechanisms.

In [36], a unified framework for detecting outliers and change points from non-stationary time series data is presented. Although in most works outlier detection and change point detection have not been related explicitly, [36] presents a unifying framework for dealing with both of them based on the theory of online learning of non-stationary time series. In this framework, a probabilistic model of the data source is incrementally learned using an online discounting learning algorithm, which can track the changing

data source adaptively by forgetting the effect of past data gradually. In order to handle non-stationary data sources, an autoregressive model has been introduced with time varying coefficients, i.e. parameter estimates are updated incrementally so that the effect of past examples is gradually discounted. Then a score is assigned to each data/each time point, with a higher score indicating a high possibility of being an outlier/a change point.

In [16], since novelty is always a relative concept with regard to our current knowledge, novelty should be defined in the context of a representation of our current knowledge. To each novel event, a value is associated characterizing how confident the judgement is. The online detection algorithm is developed using online support vector regression.

In [8] a method has been proposed for the detection of the appropriate set of number of points that minimizes the error in fitting a pre-decided function using maximum likelihood. There is no fixed number of change-points to be detected. Moreover, no constraints are imposed on the class of functions that will be fitted to the subsequences between successive change-points.

Two approaches have been proposed, the batch (offline) and the incremental (online). In the batch version, the entire data set is available, as in the case of 24-hour data from traffic sensors, from which the best set of change-points is determined. In the incremental version, the algorithm receives new data points one at a time, and determines if the new observation causes a new change-point to be discovered.

Following the notation in [8], let $y(t), (t = 1, ..., n)$ be the time series to be segmented, where t is the time variable. It is assumed that the time series can be modeled mathematically, where each model is characterized by a set of parameters. The problem of event detection becomes one of recognizing the change of parameters in the model, or perhaps even the change of the model itself, at unknown time(s). The change-points detection, is then formulated as finding a piecewise segmented model, given by

$$\begin{aligned} Y &= f_1(t, w_1) + e_1(t), (1 < t \leq \theta_1), \\ &= f_2(t, w_2) + e_2(t), (\theta_1 < t \leq \theta_2), \\ &= ..... \\ &= f_l(t, w_l) + e_l(t), (\theta_{l-1} < t \leq N). \end{aligned} \tag{1.1}$$

Where $f_i(t, w_i)$ is the function (with its vector of parameters w_i) that is fitted to the segment i. The θ_i's are the change-points between successive segments, and $e_i(t)$'s are error terms.

Maximum Likelihood Estimation

If all change points are specified a priori, and modeled with parameters w_i's and estimated standard deviations σ_i's found for each segment, then the statistical likelihood L, of the change points is proportional to, using the homoscedastic error model:

$$L = [\sum_{i=1}^{l} m_i \sigma_i^2]^{\frac{N}{2}} \tag{1.2}$$

Here, l is the number of change-points, m_i is the number of time points in segment i, and N the total number of time points. The homoscedastic error model specifies that $\sigma_1 = \sigma_2 = ... = \sigma_l$. In contrast the heteroscedastic error model doesn't impose this constraint.

Using the heteroscedastic error model, the statistical likelihood L, of the change-points is proportional to:

$$L = \prod_{i=1}^{l} \sigma_i^{m_i} \tag{1.3}$$

If the change points are not known, the maximum likelihood estimate (MLE) of the θ_i's can be found by maximizing the likelihood L over all possible sets of θ_i's, or equivalently, by minimizing $-2 \log L$. This is equivalent to, for the homoscedastic case:

$$-2 \log L = n \log (\sum_{i=1}^{l} m_i \sigma_i^2) \tag{1.4}$$

For the heteroscedastic error model, this gives:

$$-2 \log L = \sum_{i=1}^{l} m_i \log \sigma_i^2 \tag{1.5}$$

The term *likelihood criteria* will refer to the function $-2 \log L$, denoted as $\mathcal{L}$. Since log is a monotonically increasing function, an equivalent likelihood criteria of minimizing the function $\sum_{i=1}^{l} m_i \sigma_i^2$ is used, for the homoscedastic error case.

Model Selection

For each segment i, model estimation is the problem of finding the function $\hat{f}_i(t, w_i)$ that best approximates the data. The quality of an approximation is measured by the loss function $Loss(y(t), \hat{f}_i(t, w_i))$, where $\theta_{i-1} < t < \theta_i$. The expected value of loss is called risk functional $R_i(w_i) = E[Loss(y(t), \hat{f}_i(t, w_i))]$. Therefore, for each segment we have to find $\hat{f}_i(t, w_i)$ that minimizes $R(w_i)$.

Concerning the nature of the approximation functions $\hat{f}_i(t, w_i)$'s, in general it is impossible to determine the nature of these functions from domain knowledge. As such, several types of basis functions can be considered, e.g. algebraic polynomials, wavelet, Fourier, etc. [5].

Batch Algorithm

Here the assumption holds that the entire data set is collected before the analysis begins.

Assume that the best model that maintains time points $t_i, t_{i+1}, ..., t_j$ as a single segment has been selected. Let S be the residual sum of squares for this model. The number of points in the current segment is $m = j - i + 1$. Let $\mathcal{L}(i, j) = m \log S/m$ if a heteroscedastic error model is used, and $\mathcal{L}(i, j) = S$, if the error model is homoscedastic.

The key idea behind the proposed algorithm is that at every iteration, each segment is examined to see whether it can be split into two significantly different segments. The splitting procedure can be illustrated by a consideration of the first stage, since all subsequent stages consist of equivalent scaled-down problems.

Let the data set cover the time points $t_1, t_2, ..., t_n$. The change-points in the first stage is the j minimizing $\mathcal{L}(1, j) + \mathcal{L}(j + 1, n)$, say j^*. Here j^* is defined as:

$$\mathcal{L}(1, j^*) + \mathcal{L}(j^* + 1, n) = \min_{p \leq j \leq n-p} \mathcal{L}(1, j) + \mathcal{L}(j + 1, n)$$

The range of j depends on the fact that at least p points are needed for model fitting in each segment. Further, the model fitted in each segment is the best possible from the space described by the basis functions, according to the model selection method used.

At the second stage, each of the two segments is analyzed as described above, and the best candidate change-points c_1 and c_2 of each are located. The better of these candidates is then selected, yielding a division of the original sequence into three segments. Without loss of generality, we assume point c_1 is chosen. Now the likelihood criteria of the model becomes:

$$\mathcal{L} = (\mathcal{L}(1, c_1) + \mathcal{L}(c_1 + 1, j^*) + \mathcal{L}(j^* + 1, n)) < (\mathcal{L}(1, j^*) + \mathcal{L}(j^* + 1, c_2) + \mathcal{L}(c_2 + 1, n))$$

The above procedure is repeated until a stopping criterion is reached. Since the number of change-points is not known a priori, a stopping criterion must be used by the algorithm. Once the algorithm has detected all "real" change-points, adding any more change-points the likelihood will increase. Therefore, the algorithm should stop when the likelihood criteria becomes stable or starts to increase. Formally, if in iterations l and $l + 1$ the respective likelihood criteria values are $\mathcal{L}_l$ and $\mathcal{L}_{l+1}$, the algorithm should stop if

$$(\mathcal{L}_l - \mathcal{L}_{l+1})/\mathcal{L}_l < s$$

where s is a user-defined stability threshold. When stability threshold s is set to 0%, the algorithm stops only when the likelihood criteria starts increasing.

Incremental Algorithm

The batch algorithm is useful only when data collection precedes analysis. In some cases, change-point detection must proceed concurrently with data collection. Towards this an incremental version of the algorithm has been developed. The key idea is that if the next data point collected by the sensor reflects a significant change in phenomenon, then its likelihood criteria of being a change-point is going to be smaller then the likelihood criteria that it is not. However, if the difference in likelihoods is small, we cannot definitively conclude that a change did occur, since it may be the artifact of a large amount of noise in the data. Therefore a user-defined likelihood increase threshold is introduced.

$$(\mathcal{L}_{no\_change} - \mathcal{L}_{change})/\mathcal{L}_{no\_change} > \delta,$$

where δ is a user-defined likelihood increase threshold (see Experimental Evaluation - section **??**).

Suppose that the last change-point was detected at time t_{l-1}. At time t_l the algorithm starts by collecting enough data to fit the regression model. Suppose at time t_j a new data point is collected. The candidate change-point is found by determining t_i, with likelihood criterion $\mathcal{L}_{min}(l, j)$, such that

$$\mathcal{L}_{min}(l, j) = \min_{l<i\leq j} \mathcal{L}(l, i) + \mathcal{L}(i + 1, j).$$

If this minimum is significantly smaller than $\mathcal{L}(l, j)$, i.e. the likelihood criteria of no change-points from t_l to t_j, then t_i is a change-point. Otherwise, the process should continue with the next point, i.e. t_{j+1}.

In the incremental algorithm, execution time is a significant consideration. If enough information is stored, some of the calculations can be avoided. Thus, at time t_{j+1} to find likelihood criteria

$$\mathcal{L}_{min}(l, j + 1) = \min_{l<i\leq j} \mathcal{L}(l, i) + \mathcal{L}(i + 1, j + 1)$$

it is only necessary to calculate $\mathcal{L}(i+1, j+1)$, since (l, i) was calculated in the previous iteration.

It should be noted that if a change-point is not detected for a long time, the successive computations become increasingly expensive. A possible solution is to consider a sliding window of only the last q points.

1.1.3 Attentional Video Analysis Techniques

Extracting regions-of-interest in videos is very important for various applications ranging from video surveillance to video retrieval and video summarization. In order to ease explanation, a region-of-interest (ROI) in video is a portion of a frame that contains the key-concept or main subject of a visual scene and provides end users a more concise and informative representation of a document. Video surveillance systems seek to automatically identify events of interest in a variety of situations. Extracting a salient object is the most important step of a surveillance system. Similarly, prominent actions in video sequences are more likely to attract our first attention than surrounding neighbors.

Visual attention models have proved suitable for static scene processing. Extension of these models for treating video related processing tasks in a more efficient way, have been proposed in the literature. In this section, we highlight the most important approaches concerning attentional video analysis.

Ouerhani [27] aims at extending the saliency-based model of visual attention, described in [15], to consider also dynamic features, which gave rise to a model of dynamic visual attention [28]. The basic idea is to compute a conspicuity map related to motion which will be integrated with static conspicuity maps to compute the final saliency map. Motion estimation is done by hierarchical gradient-based optical flow estimation method.

A similar approach is presented in [31], where a visual attention framework is presented for the purpose of skin-based face detection. The visual attention architecture contains a motion channel to identify moving objects in the scene, using a multiresolution gradient-based approach [2] to estimate optical flow and generate a motion conspicuity map in the same manner as with static maps, calculated as in [15].

In several studies, image sequences are processed and analyzed in groups of two frames in order to infer the short-term objects' temporal evolution. Linking together the obtained results generates longer-term dynamics. However, the actual temporal dimension of the video is therefore disregarded. In [32] the extension of the visual attention scheme, described in [15], is proposed for volumetric data in spatiotemporal space. Under this framework, the video sequence is treated as a volume with temporal evolution (frame number) being the third dimension. The dimensions of width and height are the usual x- and y-axes of a frame. The third dimension is derived from layering frames of video data sequentially in time ($x - y - t$ space). Consequently, the movement of an object can be regarded as a volume carved out from the 3D space. Instead of feature maps, feature volumes are generated for each feature of interest (intensity, color, orientation). Each of them encodes a certain property of the video. Actually, every volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature. Interestingly, it is claimed that by exploiting the last consideration, motion estimation, needed to infer the dynamic nature of the video content, is avoided.

The same approach to spatiotemporal visual attention has been adopted in [30], for the purpose of video classification. The spatiotemporal visual attention model, that treats the temporal dimension of a video sequence as an intrinsic feature provides a unifying framework to analyze the spatial and temporal video organization. It is commonly believed that in order to achieve robust global classification, i.e. without prior object detection or recognition, it is crucial to select an appropriate set of visual descriptors. In [30] it is claimed that simple visual features bound to spatiotemporal salient regions will better represent the video content. Hence, feature vectors extracted from these regions are believed to enhance classifier performance.

Spatiotemporal-based visual attention detection in video sequences has also been adopted in [38]. A spatiotemporal video attention detection technique is presented for detecting the attended regions that correspond to both interesting objects and actions in video sequences. Both spatial and temporal saliency maps are constructed and further fused in a dynamic fashion to produce the overall spatiotemporal attention model. In the temporal attention model, motion contrast is computed based on the planar motions between images, estimated by point correspondences in the scene.

Extension of spatial attention to video sequences, where motion plays an important role, has also been researched in [3], where the problem of detecting irregularities in visual data is tackled. The term "irregular" depends on the context in which the "regular" or "valid" are defined. Yet, it is not realistic to expect explicit definition of all possible valid configurations for a given context. The problem of determining the validity of visual data is posed as a process of constructing a puzzle; i.e. a new observed image region or a new video segment ("the query") using chunks of data ("pieces of puzzle") extracted from previous visual examples ("the database"). Regions in the observed data which can be composed using large contiguous chunks of data from the database are considered very likely, whereas regions in the observed data which cannot be composed from the database (or can be composed, but only using small fragmented pieces) are regarded as unlikely/suspicious.

Cheng et al. [35] has incorporated the motion information in the attention model. The proposed motion attention model analyzes the magnitudes of image pixel motion in horizontal and vertical directions. As such, the presented framework determines automatically regions-of-interest in video sequences. A short video clip, called frame-segment, is used as the unit for conducting the video ROI analysis. On each frame-segment feature maps, a temporal median filter is applied, to ensure the general characteristics of a specific feature in the frame-segment.

Approaches for analyzing video attentions are also aiming at user attention models for video skimming and summarization. E.g. Ma et al. [17, 18] presented user attention models for video skimming and summarization, which utilized more audio-visual features of semantics, for example, motion, speech, camera operation, and lexical information. In their work, although the video features are shown to be effective in detecting temporal attentions, their interactions with spatial visual features are still unknown. Ho et al. [13] proposed a framework for video focus detection based on visual attention, which introduced a video-genre-based method for saliency map generation. That is, in different video categories, different parameter sets are elaborately optimized and accordingly assigned. The experiment shows impressive results, but the method is too highly domain dependent to be extended for general purpose.

Bibliography

[1] S.-W. Ban, W.-J. Won, and M. Lee. Novelty scene detection using scan path topology and energy signature in sclaed saliency map. *Neural Information Processing, Letters and Reviews*, 8(3):57–66, 2005.

[2] M. Black and P. Anandan. The robust estimationof multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. *International Journal of Computer Vision (IJCV)*, 74(1):17–31, 2007.

[4] P. Burge and J. Shaw-Taylor. Detecting cellular fraud using adaptive prototypes. In *Proc. of AI Approaches to Fraud Detection and Risk Management*, pages 9–13, 1997.

[5] V. Cherkassky and F. Mulier. *Learning from Data*. Wiley-Interscience, New York, N.Y., 1998.

[6] T. Fawcett and F. Provost. Activity monitoring: noticing interesting changes in behavior. In *Proc. of KDD-99*, pages 53–62, 1999.

[7] R. S. Gaborski, V. S. Vaingankar, V. S. Chaoji, and A. M. Teredesai. A system for novelty detection in video streams with learning. Technical report, Laboratory for Applied Computing, Rochester Institute of Technology, Rochester, NY, USA, 2004.

[8] V. Guralnik and J. Srivastava. Event detection from time series data. In *Proc. of KDD-99*, pages 33–42, 1999.

[9] S. B. Guthery. Partition regression. *Jr. Amer. Statist. Ass.*, 69:945–947, 1974.

[10] N. Haering, R. J. Qian, and M. I. Sezan. A semantic event-detection approach and its application to detecting hunts in wildlife video. In *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, volume 10, 2000.

[11] N. C. Haering, R. J. Qian, and M. I. Sezan. Detecting hunts in wildlife videos. *Multimedia Computing and Systems, International Conference on*, 1:9905, 1999.

[12] D. M. Hawkins. Point estimation of parameters of piecewise regression models. *Jr. of the Royal Statistical Society Series*, 25(1):51–57, 1976.

[13] C.-C. Ho, W.-H. Cheng, T.-J. Pan, and J.-L. Wu. A user-attention based focus detection framework and its applications. In *The Fourth IEEE Pacific-Rim Conference on Multimedia (PCM 2003)*, 15-18 December, , Singapore., 2003.

[14] M. Huskova. Nonparametric procedures for detecting a change in simple linear regression models. *Applied Change Point Problems in Statistics*, 1993.

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[16] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *In Proc. of the Ninth ACM SIGKDD*, pages 613–618. ACM Press, 2003.

[17] Y. Ma, L. Lu, H. Zhang, and M. Li. A user attention model for video summarization. In *Proc. ACM Multimedia 2002*, pages 533–542, 2002.

[18] Y.-F. Ma and H.-J. Zhang. A model of motion attention for video skimming. *ICIP*, 1:129–132, 2002.

[19] M. Markou and S. Singh. Novelty detection: A review - part 1: Statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

[20] M. Markou and S. Singh. Novelty detection: A review - part 2: Neural network based approaches. *Signal Processing*, 83:2499–2521, 2003.

[21] S. Marsland. Novelty detection in learning systems. *Neural Computing Surveys*, 3:157–195, 2003.

[22] S. Marsland, U. Nehmzow, and J. Shapiro. Detecting novel features of an environment using habituation. In *From Animals to Animats: Proceedings of the 6th International Conference on Simulation of Adaptive Behavior (SAB2000)*, pages 189–198, Paris, France, 2000. MIT Press.

[23] S. Marsland, U. Nehmzow, and J. Shapiro. Vision-based environmental novelty detection on a mobile robot. In *Proceedings of the International Conference on Neural Information Processing (ICONIP01)*, Shanghai, China, 2001.

[24] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:873–889, 2001.

[25] H. V. Neto. *Visual Novelty Detection for Autonomous Inspection Robots.* Phd thesis, University of Essex, Colchester, UK, 2006.

[26] H. V. Neto and U. Nehmzow. Visual novelty detection with automatic scale selection. *Robotics and Autonomous Systems*, 55(9):693–701, 2007.

[27] N. Ouerhani. *Visual Attention: From Bio-Inspired Modeling to Real-Time Implementation.* Phd. thesis, Universite de Neuchtel, 2003.

[28] N. Ouerhani and H. Hügli. A model of dynamic visual attention for object tracking in natural image sequences. *IWANN LNCS*, 1:702–709, 2003.

[29] C. Peters and D. Grandjean. A visual novelty detection component for virtual agents. In T. J. Paletta, L., editor, *Proceedings of the Fifth International Workshop on Attention and Performance in Computational Vision (WAPCV)*, pages 289–300, Santorini, Greece, May 2008.

[30] K. Rapantzikos, Y. Avrithis, and S. Kollias. On the use of spatiotemporal visual attention for video classification. In *Proc. of Int. Workshop on Very Low Bitrate Video Coding (VLBV '01)*, September 2005.

[31] K. Rapantzikos and N. Tsapatsoulis. Enhancing the robustness of skin-based face detection schemes through a visual attention architecture. In *ICIP05*, pages II: 1298–1301, 2005.

[32] K. Rapantzikos, N. Tsapatsoulis, and Y. Avrithis. Spatiotemporal visual attention architecture for video analysis. *Multimedia Signal Processing, 2004 IEEE 6th Workshop*, pages 83–86, 2004.

[33] A. Tentler, V. Vaingakar, R. Gaborski, and A. Teredesai. Event detection in video sequences of natural scenes. Technical report, Rochester Institute of Technology, Laboratory for Applied Computing, 2002.

[34] V. S. Vaingankar, V. S. Chaoji, R. S. Gaborski, and A. M. Teredesai. Cognitively motivated habituation for novelty detection in video. Technical report, Laboratory of Apllied Computing, Rochester Institute for Technology, Rochester, NY, USA, 2003.

[35] C. Wen-Huang, C. Wei-Ta, and W. Ja-Ling. A visual attention based region-of-interest determination framework for video sequences. In *IEICE TRANS. INF. & SYST.*, 2005.

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

[37] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixture with discounting learning algorithms. In *Proc. of KDD2000*, pages 250–254. ACM Press, 2000.

[38] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 815–824, New York, NY, USA, 2006. ACM.