Computational Attention for Event Detection

Matei Mancas[1], Laurent Couvreur[1], Bernard Gosselin[1], Benoît Macq[2]

[1] Faculty of Engineering, Mons (FPMs), TCTS Lab
31, Bd. Dolez, 7000, Mons, Belgium
{matei.mancas, laurent.couvreur, bernard.gosselin}@fpms.ac.be
[2] Catholic University of Louvain (UcL), TELE Lab
2, Place du Levant, 1348 Louvain-la-Neuve, Belgium
Macq@tele.ucl.ac.be

Abstract. This article deals with a biologically-motivated three-level computational attention model architecture based on the rarity and the information theory framework. It mainly focuses on low-level and medium-level steps and their application in pre-attentive detection of tumours in CT scans and unusual events in audio recordings.

Keywords: computational attention, saliency, rarity, audio event, tumour

1 Introduction

The human visual system (HVS) is a topic of increasing importance in computer vision research since Hubel's work [1]. Mimicking some of the processes done by our visual system may help to improve the existing computer vision systems. Besides, there are evidences than some mechanisms that are involved in the human visual system can be generalized to other senses, *e.g.* the auditory human system.

In this article, we describe a biologically-motivated three-level computational attention model. Primarily designed for visual attention, we discover that it can be applied to handle audio stimulus as well. We apply mainly the low- and medium-level steps in event detection for medical images and audio ambiences.

Our model is described in the next section from a visual attention perspective. Parts three and four present the low- and medium-level attention mechanisms and parts five and six show two applications to tumour and audio event detection.

2 Visual Attention (VA)

In this article, we shall mainly address the low-level and medium-level pre-attentive processes of visual attention. Pre-attentive attention is reflex-based and it occurs faster than an eye saccade (eye movement) corresponding to 200 milliseconds for humans. The pre-attentive detection of areas of interest is a fast process by opposition with the saccade-based image analysis which is a "serial" and slower process [2].

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

2.1 Biological background

The Superior Colliculus (SC) is the brain structure which directly communicates with the motor command in charge of eye orientation. One of its tasks is to direct the eyes onto the "important" areas of the surrounding space. The study of the SC afferent paths definitively provides important clues about visual attention.

There are two afferent pathways for the SC, one direct path from the retina, and another indirect one crossing the Lateral Geniculate Nucleus (LGN) and the primary cortex area (V1) before coming back to the SC. Studies on afferent SC pathways [3] showed that the direct path from the retina is responsible for spatial (W cells) and temporal (Y cells) analysis and the indirect pathway is mainly responsible for spatial and motion direction and colour analysis. This architecture justifies a separation between non-colour and non-directional processing (low-level) and colour and behaviour (direction, speed, ...) information processing (medium-level).

2.2 Attention modelling

Many methods may be found in the literature about visual attention and saliency. They are mainly divided into two categories. The first one [4][5] uses locally computed salient features and the second one [6] [7] [8] [9] [10] uses similarity and comparisons all over the image in a global processing.

Our model of attention is based on the **rarity** concept that is necessarily a global concept integrating the local processing of different cells, thereby belonging to the second category. We claim that our vision is not attracted by specific features, but by features which are in minority wihtin an image. Likewise, attention can be defined for any other stimulus, *e.g.* our auditory attention is attracted by unusual audio features. Based on a global rarity idea, we propose a three-level approach of visual attention which is divided into three parts: a low-level approach which is exclusively pre-attentive, a high-level one which is exclusively attentive and a medium-level approach which can be either pre-attentive or attentive depending on the number of medium-level features.

2.3 Rarity quantification

A pre-attentive analysis is achieved by humans in less than 200 milliseconds, so the pre-attentive model should also be very fast. The fastest and most basic operation is to count similar areas in the image, hence to use the histogram. Within the context of information theory, this approach based on the histogram is close to the so-called self-information. Let us note m_i a message containing an amount of information. This message is part of a message set M. A message self-information $I(m_i)$ is defined as:

$$I(m_i) = -\log(p(m_i)) \quad \textbf{(1)}$$

where $p(m_i)$ is the probability that a message m_i is chosen from all possible choices in the message set M or the occurrence likelihood. We obtain an attention map by

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

replacing each message m_i by its corresponding self-information $I(m_i)$. We estimate $p(m_i)$ as a two-term product:

$$p(m_i) = \left(\frac{H(m_i)}{Card(M)} \right) \times \left(1 - \frac{\sum_{j=1}^{Card(M)} |m_i - m_j|}{Card(M) \times Max(M)} \right) \quad (2)$$

The first term is the direct use of the histogram: $H(m_i)$ is the value of the histogram H for message m_i and $Card(M)$ is the cardinality of the message set M.

The second term quantifies the distance between a message and the others. If a message is very different from the others, this term will be low, the occurrence likelihood $p(m_i)$ will be lower and the message attention will be higher.

3 Low-level Spatial Visual Attention

In an image we can consider in a first approximation that a message m_i is the grey-level of a pixel at a given space location and the message set M is the entire image. Nevertheless, comparing only isolated pixels is not efficient. In order to introduce a spatial relationship, areas surrounding each pixel should be considered.

Stanford [11] showed that the W-cells which are responsible of the spatial analysis inside the SC may be separated into two classes: the tonic W-cells (sustained response all over the stimulus) and the phasic W-cells (high responses at stimulus variations).

Our approach uses the mean and the variance of a pixel neighbourhood in order to describe its statistics and to model the action of tonic and phasic W-cells.

We compute the local mean and variance on a 3x3 sliding window as our experience showed that this parameter is not of primary importance. To find similar pixel neighbourhoods we count the neighbourhoods which have the same mean and variance (first term of **Eq. 2**). Then we compute the distance between the pixel neighbourhood mean and the others to get the second term of **Eq. 2**.

Contours and statistically smaller areas get higher attention scores on the rarity-based VA map. If we consider only local computations as, for example, the local standard deviation or the local entropy, contours are also highlighted but the textured areas have a too high score. In our method, more regular a texture is, less surprising it is and less important the attention score will be [12]. Achieved observations prove the importance of a global integration of the local processing made by the cells. Rarity or surprise, which obviously attracts our attention, cannot be computed only locally.

4 Medium-level spatial attention

The medium-level attention system prepares the high level image analysis by selecting the first eye fixation point. The low-level step already highlights some interesting areas from the image and sometimes, similar importance areas pop-out

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

after a low-level analysis. In this case an analysis using other criteria than luminance must be achieved. Three features are used for two-dimensional images: a simple one (colour information) and two more complex ones (size and direction).

In the case of one-dimensional signals, only one feature is used: the size of the detected events. Repetitive events with similar size have their global attention score decreasing while events with different, hence rare size will be awarded with higher attention scores.

5 Tumour Detection

5.1 Tumour detection using throat asymmetry

In this section we deal with tumour detection within head and neck areas in the scope of radiotherapy planning. The main modality we focus on is the CT Scan which can be directly used to plan the radiotherapy doses. The image volumes we consider go from the upper end of the lungs to the middle of the head.

If the human body bilateral symmetry is not respected, that is most of the time due to some abnormalities. For example, the symmetry measurement can aid in the detection of breast cancers [13] or neurological disorders [14]. Asymmetry was also used for brain tumours detection on MRI images [15].

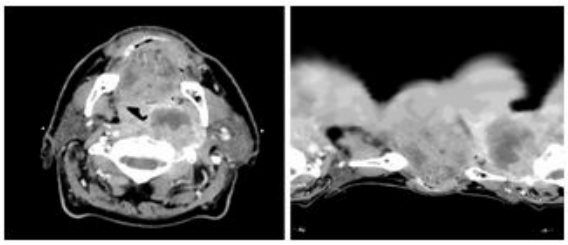

Fig. 1. Left: example of a slice form a CT scan volume of head and neck, Right: 360x360 Log-polar representation centred on the throat

In the neck area, cancers are due to smoke and tumours first develop themselves in the tissue which is directly related with this factor like the mouth, the tongue and the airways. The presence of a tumour close to the airways will introduce an asymmetry by pushing them in a given direction.

The cell repartition on the retina is concentrated on the fovea and that it exponentially decreases towards the retina boundaries [1]. We can model a medical doctor fast inspection of neck abnormalities as an eye fixation on the airways while different slices of the CT scan volume are displayed. A log-polar transform is used to mimic the eye fixation and to provide a first high-level analysis: the centre models the fixation point and the logarithmic transform models the retina cells repartition. **Fig. 1** shows on the right a result of a log-polar transformation from the initial image located on the left-side. This transform allows us focusing on the airways in the same way as humans do by fixating a position within the image.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

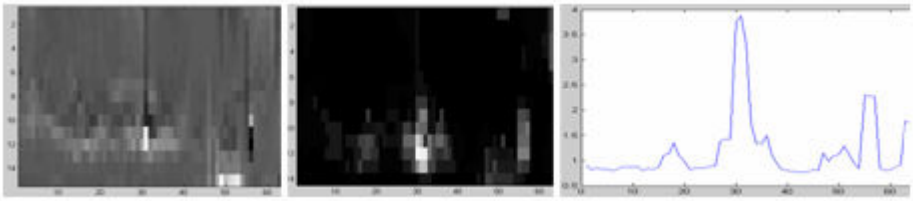

Fig. 2. Left: symmetry decomposition map for 15 grey-levels of a 65 slices CT-scan volume, Middle: low-level VA map of the first 14 grey-levels of the left image, Right: Projection of the VA map onto the horizontal axis i.e. the volume slice number

The symmetry decomposition is computed after resampling the log-polar image into 16 grey-levels. For each grey-level m_i into the M grey-level set the symmetry value is computed using the following equation:

$$S(m_i) = \left(\frac{Card_{total}(m_i)}{Card_{total}(M)} \right) \times \left(\left(\frac{Card_{right\_side}(m_i)}{Card_{right\_side}(M)} \right) - \left(\frac{Card_{left\_side}(m_i)}{Card_{left\_side}(M)} \right) \right) \quad \textbf{(3)}$$

Card(x) is the cardinality of *x*. The three terms represent the occurrence probability of each grey-level within the whole log-polar image or only within the right-side or the left-side. $S(m_i)$ can be positive or negative depending on which side of the image the grey-level is mainly located. Moreover the difference between the right-side and the left-side is weighted by the probability to find the grey-level in the whole log-polar image. As the logarithm provides increasing importance to the grey-levels located close to the centre point, **Eq. 3** means that if the difference between the right and the left side of the image is high and if this grey-level has a significant occurrence close to the airways, it will be very asymmetric, thus very likely to be an abnormality.

The left image from **Fig. 2** is the concatenation of the symmetry decomposition on 65 slices from a CT-scan volume. The grey-level m_0 has been eliminated as a tumour is never part of the image background or of the air, thus only 15 grey-levels are kept. Very clear values are strongly positive and the very dark ones are strongly negatives.

Now we have a map of the symmetry decomposition evolution within slices, we can directly apply the low-level VA map described in section 3 to highlight the most "important areas" which pop-out from the symmetry decomposition map (**Fig. 2**, middle). Finally, in order to quantify the rare slices, we project the VA map onto the slices axis and obtain the right-side image of **Fig. 2** which shows the slices which have a rare asymmetry.

5.2 Using an atlas: a priori knowledge integration

When inspecting the airways symmetry within a CT scan volume, a medical doctor not only compares the slices he sees but he also uses his experience in this domain about how symmetric the airways are in the neck area. In order to model the medical doctor experience we set a symmetry decomposition atlas built on 107 healthy slices of different patients. We can see this atlas on image I1 of **Fig. 3**.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

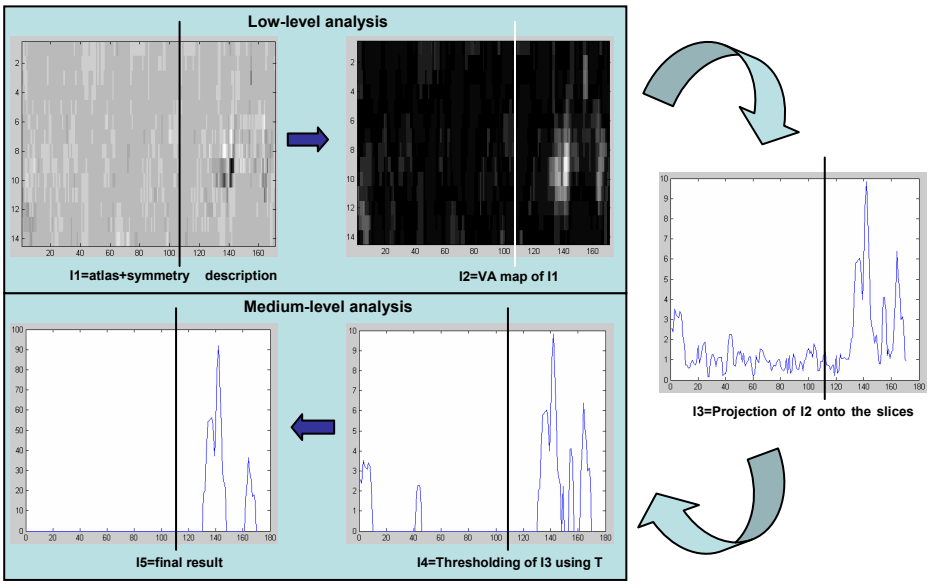

Fig. 3. From images I1 to I6: low-level to medium-level analysis

The image is obtained by concatenation of the atlas with a symmetry decomposition map of a CT scan volume we want to analyse. The black line is the border between the atlas (1 to 107) and the symmetry decomposition map (108 to 172). The low-level VA map can afterwards be computed on this new image. The result is shown in I2 (**Fig 3)**. The white line separates the atlas from the result of map we want to analyse.

If we compare this result with the VA map computed without an atlas (**Fig. 2**), we can say that the important parts are more highlighted which leads to less noisy results. If some values of symmetry are quite rare within the actual symmetry decomposition map but very numerous in the atlas, this means that this kind of values are in fact "normal". In this case, the occurrence probability will be higher than without the use of an atlas, therefore the importance will be lower. Generally, the abnormal values will be better highlighted and some normal symmetry values which appear as abnormal without the use of an atlas will become normal by using an atlas.

5.3 Medium-level analysis and results

After the low-level analysis we obtain the graph I3 (**Fig 3)** which represents the projection of I2 (**Fig 3)** onto the slices axis (horizontal). We will loose information about the the gray-levels which are responsible for the assymmetry, but this information is less important for detection purposes than for reconstruction.

We use the results obtained on the atlas (1 to 107) to extract normality features. Let us call $I3_{atlas}$ the vector of the first 107 values from **Fig 3**, image I3. Experimentally:

$$T = mean(I3_{atlas}) + std(I3_{atlas}) \quad \textbf{(4)}$$

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

T can be extracted from the atlas and used to threshold the I3 image (**Fig. 3**). If groups of values are higher than *T*, they will pop-out from the low-level step. The results are on image I4 (**Fig. 3)** where only the "interesting" structures were selected.

Finally, a medium-level step can be applied using the size (in slices) of the structures of image I4. **Eq. 1** is used to find rare structures. The most two interesting structures are shown in image I5 of **Fig. 3**.

We tested this method on 12 patients and obtained the following results:

Patient	Ground Truth	Detected	Rule-based	Precision	Recall
A3	14-30	**1-43** / 47- 65	**1-46**	37	100
A5	20-38	**20-43** / 51-58	**17-46**	63	100
A9	19-39	**9-32** / 61-64	**6-35**	57	81
A10	17-40	**18-42** / 57-64	**15-45**	77	100
A12	20-39	**24-38** / 55-62	**21-41**	90	95
A13	7-27	**8-22** / 47-56	**5-25**	90	90
B11	2-40	**13-40** / 45-64	**10-43**	91	79
B12	2-19	**5-16** / 19-24	**2-27**	69	100
B13	23-27	**3-11** / 25-27	**1-14** / 22-30	22	100
B15	10-32	12-18 / **20-28**	**9-32**	96	100
B16	1-21	1-7 / **9-22**	**1-25**	84	100
B17	8-13 / 22-37	8-16 / **23-42**	5-19 / **20-45**	54	100
Mean				**69 %**	**95 %**

Table 1. Results of visual attention tumour localisation

For each patient there is a "Ground Truth" column with the slices really containing tumours. For patient B17 there are two tumours, thus we have two sets of pathological slices. The "Detected" column shows the two more important detected areas after medium-level analysis. The bold sets are the most important. We then use two simple rules: we first dilate the set of detected slices by 3 on each side in order to include potential beginning or end of tumour which are more difficult to detect. This rule will also fuse some very close detected sets. The second rule eliminates all the detected sets which have pathological slices within 55 and 65. After slice 55 the airways are too irregular (as we arrive into the nose) or they disappear (as we arrive into the brain). The symmetry of the airways is no more relevant here. After using our two rules we have the results shown in the "Rule-based" column. The "Precision" and "Recall" columns show the results using the precision (ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved) and recall (ratio of the number of relevant records retrieved to the total number of relevant records in the database). The final "Mean" line provides the mean of precision and recall on the overall database. The recall results are very good and they have a low variance. This is not the case with the precision results which have a lower mean and higher variance. As this is a medical application, medical doctors prefer having more false alarms and detect all the pathological slices. In this case the recall accuracy is very important and the precision accuracy is not vital, therefore high percentage of precision is not mandatory.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

6 Audio event detection

Following recent tragic events, public authorities become more concerned with security issues for public areas and events. The common approach relies on video technology and the security operators have typically to monitor tens of displays in the control room. In order to help the security operators in assessing the situations, the video sensors are more and more often combined with other sensors, for instance microphones. Unlike for video streams, it is not possible to monitor simultaneously several audio streams because of the "transparent" nature of the audio data. Therefore, there is strong need for automatic audio processing techniques to detect unusual sound activities that may indicate critical situations, and focus the attention on them.

In the following, we present audio event detection algorithms that are tested on security-related sounds embedded in common audio ambiences. The audio events consist in gun shot, explosion, woman scream, car crash, glass breaking and siren sounds [16]. Since it is not practical to record these audio events in real conditions, we adopt a simulation approach where the audio events are mixed with ambience sounds, namely, train station, airport hall, sport stadium and street sidewalk [17]. Three ambience-to-event energy ratios are considered, namely –10, -5 and 5 dB corresponding to loud, weakly and very weakly audible audio events, respectively.

6.1 Low-level detection

The low-level computational attention algorithm can be either applied directly on the one-dimensional waveform (**Fig. 4** left) or on a two-dimensional representation of the audio signal. It is common to use the spectrogram in audio processing, which displays the Fourier power spectrum (vertical) as a function of time (horizontal).

In **Fig. 4** one can see on the second image the spectrogram of 15 seconds of audio signal containing a centre-located audio event. The third image displays the result of our low-level VA map applied to the spectrogram. This image is less noisy than the spectrogram and let us see more clearly the audio event. If we make a projection of this VA map onto the horizontal axis (time) we obtain the fourth image from **Fig. 4** where we can see that the attention is maximal in the same time as the audio event.

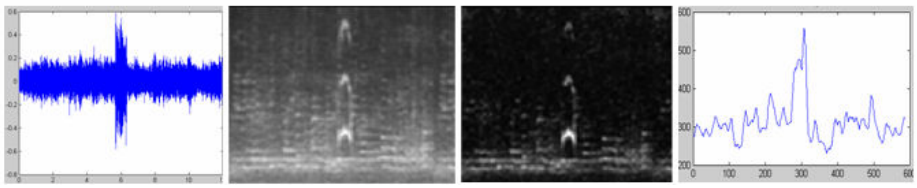

Fig. 4. Low-level audio signal analysis for a -10dB signal

Within our database for the -10 dB audio events the higher attention peak always match with the audio event. For -5 dB signals, the peak matching the event can be first or second. In order to solve this problem a medium-level approach is needed. For the 5 dB signals, the event is very weak, and, for some signals it is not possible to distinguish the attention peak of the event from the rest of the attention graph.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

6.2 Medium-level analysis and results

In order to get the pop-out structures after the low-level approach, we use a threshold equal to the mean of the importance graph (**Fig. 4**, right). As for tumours, once we have the pop-out structures we apply the **Eq. 1** on their duration (time size). We then select the two more important structures and finally keep only the one which contains the peak with the most important low-level attention score. **Table 2** presents the results for 6 events and 3 audio ambiences (side walk, airport hall, train station).

Event	-10 dB		- 5 dB	
	Prec.	Recall	Prec.	Recall
Car crash	99	96	99	97
Explosion	98	83	97	74
Glass breaking	100	62	96	58
Gun shot	100	74	100	69
Woman scream	95	98	96	97
Siren	100	29	99	34
Mean	**99 %**	**74 %**	**98 %**	**71 %**

Table 2. Results of visual attention audio events localisation for -10 dB and -5 dB

For the 5 dB ambience-to-event energy ratio we get 6 non-detections on 15 sounds. As there were non-detections, we did not compute their precision and recall which are measures of detection accuracy. Concerning the -10 and -5 dB ratios, the results are quite good in precision and even in recall which means that we may have a good estimation of the beginning and duration of the event. There were two cases working less well: first the "sport stadium" ambience where events were superimposed to an applause period. As applause occurs just once in 15 seconds they are considered as an event and there is no good distinction between the real event and the applause (the projection of the spectrogram on the time axis does not contain any spectral information), therefore we excluded this ambience where we cannot precisely say if the event or the applauses were detected. This kind of problem could be solved by using an atlas containing more applause. Secondly, the "siren" event is a temporal texture repeating itself. Our attention model will award it with low attention scores and only its beginning will be detected (high precision, low recall). This is coherent to human reaction: when a siren starts we pay attention but afterwards we get used to it.

7 Discussion and Conclusion

In this article, we used a computational attention technique to detect events into mono-dimensional signals as the symmetry variation within slices of a CT scan volume or audio signals. Initially used for still and video images, our computational attention model can also be used in mono-dimensional signals which shows the universality of the rarity concept and of our three-level approach.

In the medical imaging domain our method can be used to reduce the number of inspected slices. Moreover, a tumour reconstruction could be done by putting together

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

the structures close to the airways and containing the most asymmetric grey-values.

For audio signals, we could use more sophisticated representations than the spectrogram and even an atlas containing a "normal" audio signal for a given ambience. But the most important part of the work is to provide a real-time algorithm. This can be done by taking into account the history of each line (frequency) of the spectrogram and compute the novel pixel attention.

Finally one may want to obtain the entire spectral definition of the event in order to synthesise it and recognise it. An approximation of the spectral content of the event could be partially achieved by subtracting the spectral intensity just before or just after the event from the spectral intensity during the event.

References

1. Hubel, D.H., "Eye, brain and vision", New York: Scientific American Library, N°22, 1989
2. Treisman, A. M., and Gelade, G. "A feature-integration theory of attention", Cognitive Psychology, 12(1): 97-136, 1980
3. Crabtree, J.W., Spear, P.D., McCall, M.A., Jones, K.R., and Kornguth, S.E. "Contributions of Y- and W-cell pathways to response properties of cat superior colliculus neurons: comparison of antibody- and deprivation-induced alterations", J Neurophysiol., 56(4):1157-1173, 1986
4. Itti, L., and Koch, C. "A saliency-based search mechanism for overt and covert shifts of visual attention", Vision Research, 40:1489-1506, 2000
5. Le Meur, O., Le Callet, P., Barba, D., and Thoreau, D. "A coherent computational approach to model bottom-up visual attention", IEEE PAMI, 2005
6. Walker, K.N., Cootes, T.F., and Taylor, C.J. "Locating salient object features", Proc. of British Machine Vision Conference, 2:557-566, 1998
7. Mudge, T.N., Turney, J.L., and Volz, R.A. "Automatic generation of salient features for the recognition of partially occluded parts", Robotica, 5:117-127, 1987
8. Stentiford, F.W.M., "An estimator for visual attention through competitive novelty with application to image compression", Picture Coding Symposium, pp. 25-27, 2001
9. Boiman, O., and Irani, M., "Detecting irregularities in images and in video", Proceedings of Int. Conference on Computer Vision, 2005
10. Itti, L., and Baldi, P. "A principled approach to detecting surprising events in video", Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 631-637, 2005
11. Stanford, L.R. "W-cells in the cat retina: correlated morphological and physiological evidence for two distinct classes", J Neurophysiol., 57(1):218-244, 1987
12. Mancas, M., Mancas-Thillou, C., Gosselin, B., and Macq, B. "A rarity-based visual attention map -application to texture description -", Proc. IEEE ICIP, 2006
13. Alterson, R., and Plewes, D. B. "Bilateral symmetry analysis of breast MRI", Phys. Med. Biol. 48 3431-3443, 2003
14. Sharma1, R., and Sharma, A. "Physiological basis and image processing in functional magnetic resonance imaging: Neuronal and motor activity in brain", BioMedical Engineering OnLine 2004
15. Wang, Z., Hu, Q., Loe, K., Aziz, A., and Nowinski, W.L. "Rapid and Automatic Detection of Brain Tumors in MR images," Proc. of SPIE Medical Imaging, San Diego, 2004.
16. The BBC Sound Effect Library, http://www.bl.uk/collections/sound-archive/soundeffects.html
17. The AURORA-2 database, http://www.elda.org

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de