A Three-Level Computational Attention Model

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

[1] Faculty of Engineering, Mons (FPMs), TCTS Lab
31, Bd. Dolez, 7000, Mons, Belgium
{matei.mancas, 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 a low-level step which aims in fastly highlighting important areas and a middle-level step which analyses the behaviour of the detected areas. Their application on both still images and videos provide results to be used by the third high-level step.

Keywords: computational attention, visual importance, saliency, rarity

1 Introduction

The human visual system (HVS) is a topic of increasing importance in computer vision research since Hubel's work [1] and the comprehension of the basics of biological vision. Mimicking some of the processes done by our visual system may help to improve the existing computer vision systems. Visual attention takes part to one of the most important tasks of the HVS, which is to extract relevant features from the surrounding images in order to react in a relevant manner for our survival.

In this article, we describe a biologically-motivated three-level visual attention model for both still images and video. One of its main advantages is to efficiently handle spatial and temporal textures leading to noise reduction in the case of moving trees or flickering lights in video sequences for example. The goal of this article is to provide a first approximation of how humans perceive their environment depending on the importance they award to the different regions of their visual field.

The general idea of our visual attention model is described in the next section. Part three and four provide attention mechanism application to still images and video sequences. The final section will conclude the work and discuss our approach.

2 Visual Attention (VA)

Pre-attentive visual attention is reflex-based and it occurs faster than 200 milliseconds for humans. The pre-attentive interest areas detection is a "parallel" fast process as

Proceedings of the ICVS Workshop on Computational Attention & Applications - WCAA 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other ICVS contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

opposed to saccade-based image analysis which is a "serial" and slower process. Treisman and Gelade [2] defined a set of features which can be detected pre-attentively by using several basic experiences. An interesting conclusion of this study is the clear separation between pre-attentive and attentive vision. This separation is one of the basis of our three-level computational attention model.

2.1 Biological background

The Superior Colliculus (SC) is a brain structure which directly communicates with the eye motor command in charge of eye's orientation. One of its tasks is to direct the eyes onto the "important" areas of the surrounding space: studying the SC afferent and efferent paths can provide important clues about visual attention.

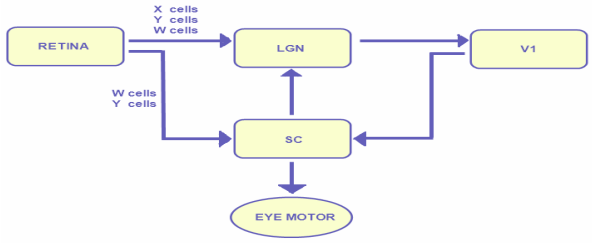

Fig. 1. Pre-attentive visual system

A first simplification of the SC interactions with other brain structures can be summarized as displayed in **Fig. 1**. There are two afferent pathways for the SC, one direct path from the retina, and one indirect path crossing the Lateral Geniculate Nucleus (LGN) and the primary cortex area (V1) before coming back to the SC. There are two efferent paths, one to the eye motor area, and the other one to the LGN.

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. Neither the SC afferent W cells or Y cells have colour or directional capacities.

2.2 Attention modelling

Many methods may be found in the literature about visual attention and image saliency. Some of them attempt to mimic the biological knowledge as Itti and Koch's method [4]. They define a multi-resolution and multi-feature based system which models the visual search in primates. Le Meur et al. [5] suggested a global architecture close to the Itti's one, but using a smart combination between the different feature maps. Instead of combining simply normalized feature maps, they use some coefficients coming from biological studies and which provide more or less importance to the different features into the final saliency map.
In these approaches only local processing mimicking different cells is used.

Walker et al. [6], Mudge et al. [7], Stentiford [8] and Boiman and Irani [9] base their saliency maps on the idea that important areas are unusual in the image. The

saliency of a configuration of pixels is inversely related to their occurrence frequency. These techniques use comparisons between neighbourhoods of different shapes and at different scales to assign an attention score to a region. Itti and Baldi [10] also published a paper describing a probabilistic approach of surprise as an attention factor. An integration of this new theory into the initial Itti's framework is currently done. These methods have a more global approach and are all based on the similarity quantification inside an image or a database.

We think that the local processing done by the cells is somehow globally integrated (possibly inside the SC structure). Our definition will be based on the **rarity** concept which is necessarily a global concept integrating the local processing of different cells. We noticed that our vision can be attracted by homogeneous areas into a heterogeneous scene, but also by heterogeneous areas into a homogenous scene.

The concept we highlight here is that our visual attention is not driven by a specific feature as some models could assess. Heterogeneous or homogeneous, dark or bright, symmetric or asymmetric, fast moving or low moving objects can all attract our visual attention. The HVS is attracted by the features which are in minority in an image. That is why we can say that the visual attention is based on observing things which are **rare** in a scene. We also could note that the pairs of features we mentioned are opposite features describing the order and the disorder at several scales, in space and time. The HVS describes rare, therefore anomalous, objects as "interesting".

Based on a global rarity approach and the Treisman studies, we propose a three-level approach of visual attention. This model is summarized in **Fig. 2.**

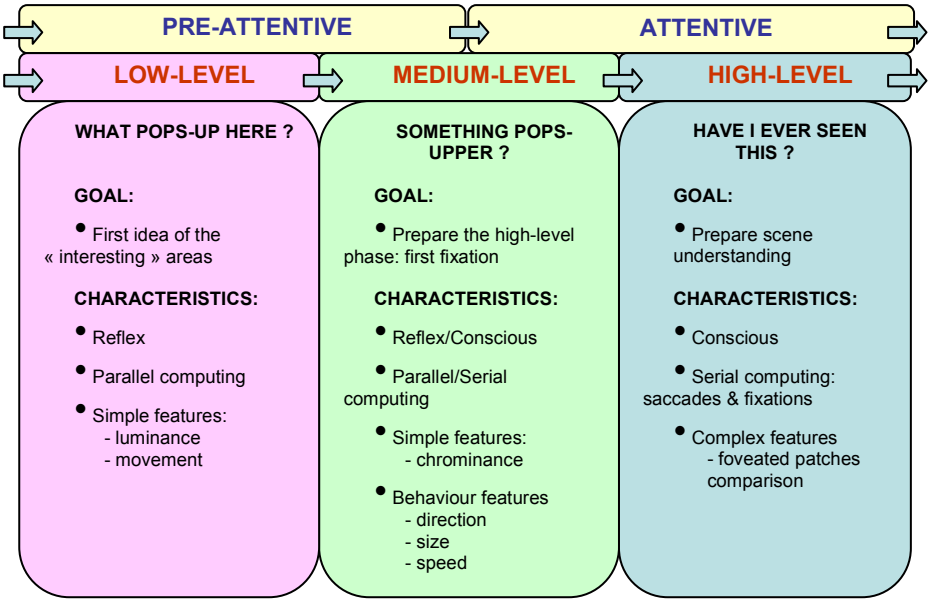

Fig. 2. Our three-level model proposition

The attention mechanism 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

generally on the number of medium-level features. The low-level approach could be directly carried inside the SC where only luminance and movement cells are available and the medium-level approach could be achieved within the second pathway where the LGN could provide chrominance information and the V1 cortical area could provide directional cells. In this article, we shall only address the low-level and medium-level pre-attentive processes of visual attention.

2.3 Rarity quantification

A pre-attentive analysis is achieved by humans in less than 200 milliseconds. How to model rarity in a simple and fast manner?

The most basic operation is to count similar areas in the image so to use the histogram. Within the context of information theory, this simple 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:

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

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 replacing each message m_i by its corresponding self-information $I(m_i)$. The self-information is also known to describe the amount of surprise of a message inside its message set: rare messages are surprising, hence they attract our attention.

We estimate $p(m_i)$ as a two-terms combination:

$$p(m_i) = A(m_i) \times B(m_i) \tag{2}$$

The $A(m_i)$ term is the direct use of the histogram to compute the occurrence probability:

$$A(m_i) = \frac{H(m_i)}{Card(M)} \tag{3}$$

Where $H(m_i)$ is the value of the histogram H for message m_i and $Card(M)$ the cardinality of M. The M set quantification provides the sensibility of $A(m_i)$: a smaller quantification value will let messages which are not the same but quite close to be seen as the same. $B(m_i)$ quantifies the global contrast of a message:

$$B(m_i) = 1 - \frac{\sum_{j=1}^{Card(M)} \left| m_i - m_j \right|}{Card(M) \times Max(M)} \tag{4}$$

If a message is very different from all the others, $B(m_i)$ will be low so the occurrence likelihood $p(m_i)$ will be lower and the message attention will be higher. $B(m_i)$ was introduced to avoid the cases where two messages have the same occurrence value, hence the same attention value using $A(m_i)$ but in fact one of the two is very different from the others while the other one is just a little different.

3 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 at a given time as shown in **Fig. 3**. If we consider as a message the pixel with the coordinates (2,2, t_0) we have $m_i=11$ and $M=\{25, 2, 16, 200, 11, 12, 200, 150, 12\}$.

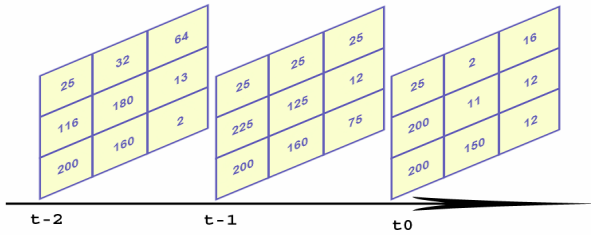

Fig. 3. Example of m_i and M on a three frame 3x3 image.

3.1 Low-level spatial attention

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 ($A(m_i)$ in **Eq. 3**). Than we compute the distance between the pixel neighbourhood mean and the others to get $B(m_i)$ as in **Eq. 4**.

Contours and statistically smaller areas get higher attention scores on the VA map (**Fig. 4,** top row, second image). If we consider only local computations as, for example, the local standard deviation or the local entropy (**Fig. 4,** top row, third and fourth image), contours are also highlighted but there are also some differences like the camera fixation system or the cameraman's trousers. The local entropy seems to provide better results but the textured grass area has a too high score.

This difference is even more important on textured images. As it contains repeating patterns, its rarity score will be lower.

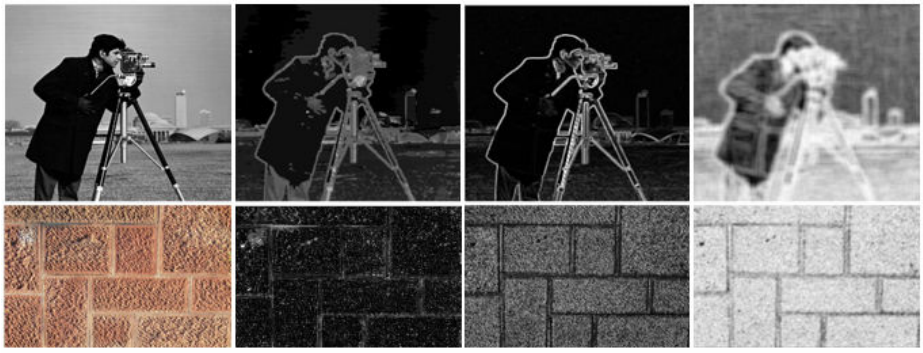

Fig. 4. From left to right: initial image, our VA model, local standard deviation, local entropy

More regular a texture is, less surprising it is and less important the attention score will be [12]. Local computations have a uniform high response for this textured image (**Fig. 4**, bottom row, third and fourth image). In the case of our VA map (**Fig 4.**, bottom row, second image), the response is important only for clearer or darker colours which are rare and which consequently attract human attention. Most of the vertical and horizontal separation lines between the bricks are also well highilighted.

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.

3.2 Medium-level spatial attention

The purpose of the medium-level attention system is to prepare 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 after a low-level analysis. In this case a fast analysis using other criteria than luminance has to be achieved. Three features are used: one is simple (colour information) and two others are more complex (size and direction).

In order to integrate colour we simply compute the VA map of the two remaining components of the opponent colour system which is the colour system used by the HVS (the red-green opposition and the blue-yellow opposition). Afterwards we combine these two VA maps with the previously computed VA map of the luminance component by using the maximum operator: if a feature has a high attention score in at least one of the three maps it will attract our final attention.

After the colour step, a VA map integrating the simple features of luminance and colours is achieved. We then apply the Otsu [13] thresholding method to select the most salient areas of the VA map which pop-out in the image.

After getting a list with the size of each area weighted by its low-level importance, **Eq. 1** is applied on the probability to find an unlikely or rare size. **Fig. 5** shows that in this case the final area selected is the area which on the low-level and colour map contained already the maximum of importance. Some vertical and horizontal separation lines also got a high score, but several repeating little areas due to texture were excluded as their size importance is very low.

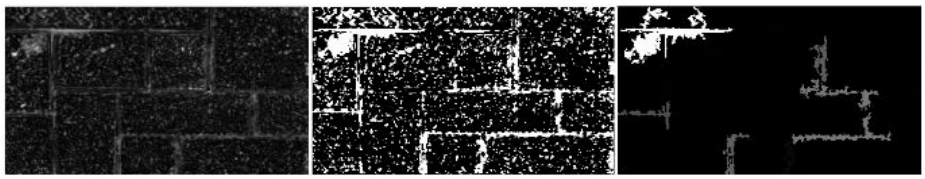

Fig. 5. From left to right: low-level and colour VA map, pop-out areas, higher attention areas

Fig. 6 shows a case where the medium-level step is very important. The low-level VA map highlights all the white rectangles from the initial image. In reality we "see" first the bigger rectangle because all the others have approximately the same size. The bigger rectangle is here emphasized on the top row of the third image of **Fig. 6** which is the size-based medium-level attention map.

The direction feature is computed by comparing the length of the projection of the selected area on several directions (4 in our implementation). **Eq. 1** is then applied to the set of directions of the pop-out areas: rare directions will thus provide a high attention score. **Fig. 6** on bottom row shows a case where after the low-level VA map and the size VA map (bottom row, second and third image) there is no decisive way to choose the first fixation, even if we intuitively look first to the rectangle which has another direction than all the others. The fourth image, bottom row of **Fig. 6** shows the final VA map after using size and direction considerations. As we can see, the rectangle which does not have the same direction as the others is highlighted.

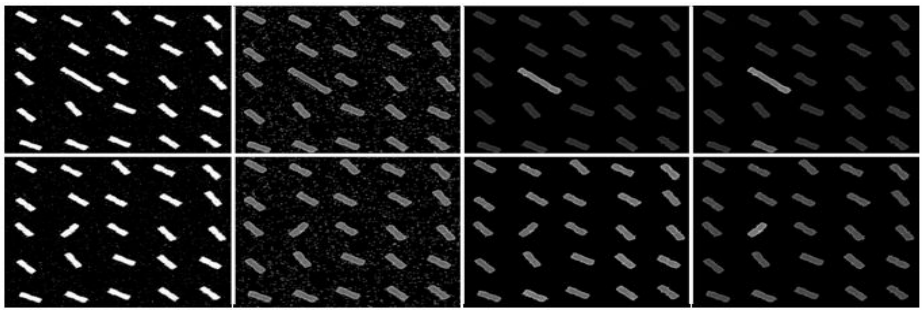

Fig. 6. From left to right: original image, low-level and colour VA map, size areas VA map, size and directions areas VA map

4 Temporal Visual Attention

4.1 Low-level temporal attention

Y cells, which are responsible of the motion analysis, have a high temporal resolution but a low spatial one [1]. Thus, the image spatial resolution is reduced and a 3x3 window mean filtering is applied on the resulting image. As Y cells are not sensitive to colour, only the luminance is used.

Message *m* is here the grey-level of a pixel at a given spatial location and message set *M* is the history of all grey-levels the pixel had over time. For example, the pixel with the coordinates (2,2, t_0) in **Fig. 3** has *m=11* and *M={180, 125, 11}*.

However, if at each frame, the whole pixel history is needed, this will lead to overloaded memory problems. Hopefully, our ability to forget lets us specify a history size and to take into account only recent frames providing a limit to the set *M*.

As motion is generally rare in an image where most pixels are quite the same from one frame to another, moving objects will be naturally well highlighted. On the top-left of **Fig. 7**, a video frame was annotated with two regions. Region 1 is a flickering light (regular time texture). The second region is a walking person. In the middle row of **Fig. 7**, the VA map was computed on a 200-frame history (on the left) and a motion estimation map which is obtained by the subtraction of the current frame from a 200-frame estimated background using a Gaussian Model (GM) method [14] (located on the right). The two thresholded maps (second image for our VA map and fourth image for GM) show that the region 2 is detected by both approaches.

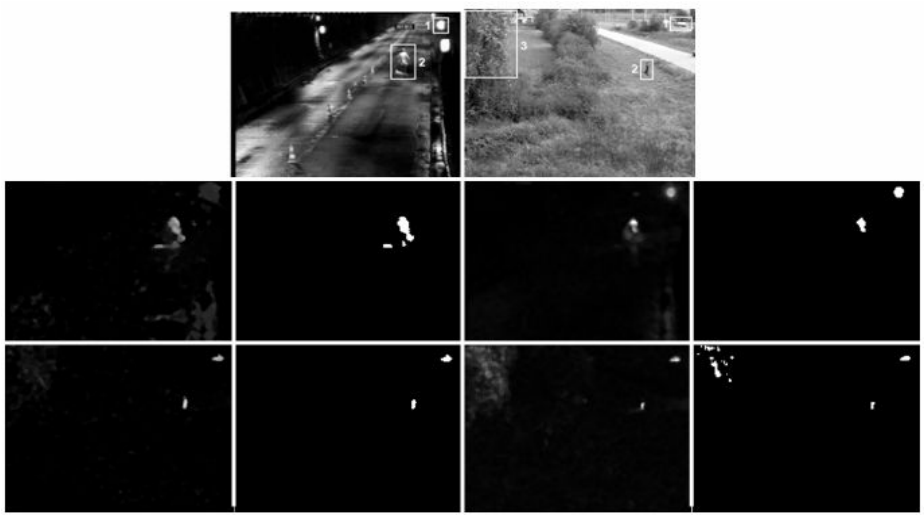

Fig. 7. Annotated frames from two video sequences on top. Middle row: motion estimation for the top-left sequence: VA map, thresholded VA map, GM-based motion map, thresholded GM-based motion map. Bottom row: motion estimation for the top-right sequence: VA map, thresholded VA map, GM-based motion map, thresholded GM-based motion map.

Our model seems to detect more largely the walking person which is underestimated by the GM method, but it also detects a little part of its shadow. The most noticeable difference is in the region 1. Our VA model awards little attention score to the flickering light as it has a higher frequency and thus is a less rare event.

Fig. 7 also provides the results on another video sequence whose frame is on the top-right and the VA map and GM-based motion estimation on the bottom row. Both methods correctly detected regions 1 and 2 (a moving car and a walking person). Moreover our method reacted with a very low attention score on region 3 (a tree moving because of the wind). These two examples show that the same behaviour is obtained for temporal or spatial attention: textures, in space or in time, are considered as less important areas. This consideration seems natural to us: first we pay attention to the texture but as soon as we understand that it can be described by a repeating pattern, our reaction is to inhibit these areas and to look elsewhere.

Proceedings of the ICVS Workshop on Computational Attention & Applications - WCAA 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other ICVS contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

4.2 Medium-level temporal attention

Within the temporal sequences the medium-level step is even more relevant than within spatial images. One can find pop-out areas with similar VA score each time that at least two objects are in relative motion. The region of interest (ROI) temporal behaviour can be described by its direction (temporal equivalent to the spatial direction) and by its speed (temporal equivalent to the spatial size).

The speed is computed by measuring the distance between the centroids of a ROI from a frame to another one. We apply **Eq. 1** to compute the attention of a ROI using its speed against the speed of the other ROIs in the current frame and also against the speed of all ROIs on a limited history computed on the previous frames. In **Fig. 8** one can see on bottom-left a simulated low-level VA map on the initial frame containing four moving rectangles. Then, on its right, successive frames where the low-level information is mixed with the medium-level information are shown. We can see that the third rectangle which is faster than the others is better highlighted.

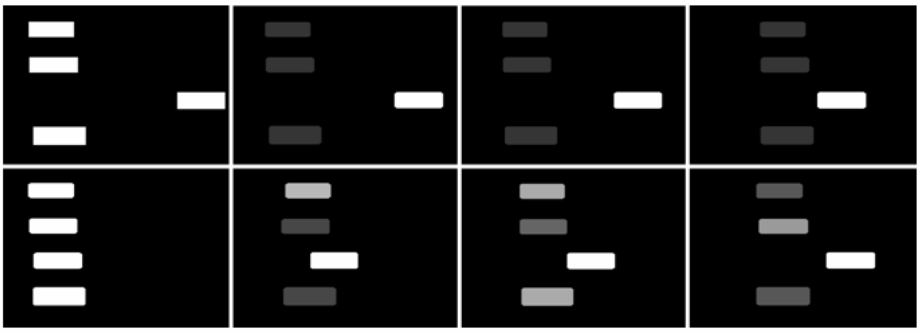

Fig. 8. From left to right: first frame (low-level VA map), frame 2, frame 3, frame 4

The second ROI motion behaviour is its direction. The direction is also computed from the ROI's centroid coordinates between two consecutive frames. A dot product is achieved between the motion coordinates and a set of vectors having 4 directions and the higher dot product result is chosen as the motion direction. Then, **Eq. 1** is used to get its final VA score. The ROI's direction is here compared to the other ROIs movement direction in the current frame and to all ROIs movement direction within the history extracted from the previous frames.

On **Fig. 8** one can see on top-left the low-level VA map on the initial frame. Then, on its right, successive frames where the low-level information is mixed with the medium-level information are shown. We can see that the third rectangle which moves in the opposite direction of the three others is better highlighted.

5 Discussion and Conclusion

We presented a bio-inspired rarity-based visual attention model working on both still images and video sequences. As one-dimensional signals can be handled by **Eq. 1,** its application to audio, tactile, smell or taste signals seems natural: the use of the

self-information as a global saliency measure appears to be universal.

We also presented a three-level computational attention architecture which progressively reduces the signal information and classifies it by perceived importance. A difference with existing models is that features are not computed in parallel but sequentially (medium-level features use the results of low-level features). In this way the image complexity decreases at each step selecting finer and finer ROIs until the high-level analysis will focus only on pre-selected areas. It speeds up computation time needed for scene comprehension, improving the species survival chances.

Some issues have still to be solved as how to correctly mix static and temporal VA maps which are intimately linked. Our serial three-level model simplifies the feature attention maps normalisation, but this problem is not completely discarded. Finally a crucial issue is in computational model validation. Eye-tracking detection is not perfect and calibration issues can disturb the results, moreover the experiments should be done on freely available databases to enable result comparisons.

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. Otsu, N., "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, 1979.
14. Wren, C.R., Azarbayejani, A., Darrell, T., and Pentland, A.P. “Pfinder: Real-time tracking of the human body”, IEEE PAMI 19:780–785, 1997

Proceedings of the ICVS Workshop on Computational Attention & Applications - WCAA 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other ICVS contributions archived and available at: http://biecoll.ub.uni-bielefeld.de