Image perception: Relative influence of bottom-up and top-down attention

Matei Mancas[1]

[1] Engineering Faculty of Mons (FPMs)
31, Bd. Dolez, 7000 Mons, Belgium
Matei.Mancas@fpms.ac.be

Abstract. Attention and memory are very closely related and their aim is to simplify the acquired data into an intelligent structured data set. Two main points are discussed in this paper. The first one is the presentation of a novel visual attention model for still images which includes both a bottom-up and a top-down approach. The bottom-up model is based on structures rarity within the image during the forgetting process. The top-down information uses mouse-tracking experiments to build models of a global behavior for a given kind of image. The second interesting point is that the relative importance of bottom-up and top-down information depends on the specificity of each image. For the three different sets of images within the database the importance of the top-down information is different. The proposed models assessment is achieved on a 91-image database.

Keywords: Visual attention, saliency, bottom-up, top-down, mouse-tracking

1 Introduction

The aim of computational attention is to automatically predict human attention on different kinds of data as sounds, images, video sequences, smell or taste, etc… This domain is of a crucial importance in artificial intelligence and its applications are numberless from signal coding to object recognition and going through image ergonomics and self-training machines. Intelligence is not due only to attention, but there is no intelligence without attention.

Attention is also very closely related to memory and cognition through a continuous competition between a bottom-up approach which uses the features of the acquired signal (here still images) and a top-down approach which uses observer's a priori knowledge about the observed signal. In this paper, a new model of bottom-up attention and a way to build top-down models of attention are proposed. An assessment of this approach is achieved which leads to a discussion on the relative importance of bottom-up and top-down influence.

In the next section, a state of the art in computational attention is achieved. The third section presents an original bottom-up computational attention which highlights the regions within an image which remain rare during the forgetting process.

The fourth section proposes a way of building top-down models which contain the mean behavior of the observers for specific images.

Section 5 achieves a computational attention algorithm assessment on a 91-image database. Finally, this section is followed by a discussion and a conclusion.

2 Computational attention: a state of the art

The result of attention algorithms is often an attention map or a saliency map of the input signal providing higher intensities for the most important areas. Attention maps and saliency maps are considered in this paper as synonyms.

The number of computational models has recently exploded as a confirmation of the maturity of the knowledge acquired within the biological, psychological and neuroscience domains. Several classifications of these methods are obviously possible, and most of them have similar philosophies, however it is possible to distinguish two main ideas. Attention may be due to:

- Local properties (a feature saliency depends on its neighborhood)
- Global properties (a feature saliency depends on the whole visual field)

If biological evidences supporting the local approaches are numerous, global approaches are for instance less well biologically motivated. This situation is normal as the local behavior of the cells on their classical receptive field (CRF) is obvious. Nevertheless, recent experiments in visual attention [1], [2] brought interesting confirmations for a global integration of features information all over the visual field. This is possible thanks to the impressive neuronal network which includes an important amount of "horizontal cells" which connect more or less directly the cells from the whole visual field.

2.1 Mostly local methods

In 1998, Itti et al. ([3], [4], [5]), set up the most well-known computational attention model. Based on the Koch and Ullman model [6], Itti proposed the extraction of three main features: luminance, chrominance, and orientation. These features are processed in parallel and then fused within a single saliency map.

Milanese et al. ([7], [8]) proposed an attention approach which is also based on the seminal architecture of Koch and Ullman. They added two more features which are contours amplitude and curvature. The normalization step is done by using Gaussian filtering and gradient descent-based relaxation before getting the mean of the maps.

Chauvin et al. [9] used the Koch and Ullman architecture, but they used Gabor filtering to get multi-resolution information. Additional computations reinforcing collinear and longer contours are also added. This model only deals with the luminance features, avoiding the difficult normalization step, but loosing important color information.

Petkov et al. [10] proposed a lateral inhibition technique to distinguish object contours from image texture. Le Meur et al. ([11], [12]) achieved a computational model of visual attention which is one of the closest to the local processing biological reality within the human visual system. Also based on the Koch and Ullman architecture, it integrates biological data for intermediate maps data fusion.

2.2 Mostly global methods

Mudge et al. [13] suggested as early as 1987 that object components saliency may be inversely proportional to their occurrence within the image. Osberger and Maeder [14] used a segmentation approach to separate the image into several homogenous areas. Five features were used in assigning a relative importance to the segmented areas. The problem of this kind of approaches is that errors within the segmentation may induce errors in the attention map.

Walker et al. [15] suggested that saliency may be related to the probability that a feature has to be misclassified with all the other features within an image.

Oliva et al. ([16], [17]) had a similar approach to Mudge et al. by stating that attention should be inversely proportional to the existence probability of a pixel. They modeled this probability with a Gaussian and used multi-resolution wavelet decomposition. Results seem similar to Itti's model as compared to eye tracking results. An interesting fact is that results are better than Itti's model if additional top-down information is used.

Bruce and Jernigan [18] integrated this idea by turning it into an information theory approach within the Koch and Ullman architecture. They afterwards [19] used ICA (Independent Component Analysis) to compare local features (local random patches of the image) in an image patches database obtained from the current image but also from other images.

Liu et al. [20] used image segmentation as Osberger and Maeder, but the mean shift [21] technique let it provide a more robust segmentation. They also assumed that centered regions may have higher attention scores. In section 4 of this paper it will be shown that this assumption is verified only in the case of natural scene images!

Itti and Baldi [22] also published a probabilistic approach of surprise based on the Kullback-Leibler divergence which is the energy of the so-called "net surprisal" within the information theory. The idea is that attention is due to a more or less important difference between what was expected to happen and the actual observation. This method has been integrated into Itti's model architecture and it provides better results compared to the original approach.

Stentiford [23] proposed a method related with Walker's ideas, but he defined no specific feature. Random pixel neighborhoods (forks) are directly compared and they are declared as matching if the distance between the two neighborhoods is below a threshold. If few matches are observed, the pixel is assigned with a high saliency score. The method provides very interesting results and its main advantage is to remain very general. It takes into account intensity, colors, directions and shapes mostly to smaller scales.

Boiman and Irani ([24], [25]) used comparisons between gradient-based patches of different sizes to define occurrence probabilities. One of the main originalities of this method is in the fact that not only different patches from the images are compared with other patches in the same image or in a database, but also the relative patches' positions were taken into account.

The author proposed ([26], [27], [28]) a global rarity approach of attention but not much local information was taken into account.

The approach proposed here is based on the fact that visual attention is the first filter which selects regions in an image which may be interesting to memorize. This

fact implies that during the forgetting process, rare regions are kept in mind while the others are forgotten. The proposed model will use both global rarity and local contrast information and thus it will have interesting properties from both local and global approaches.

3 Bottom-up attention: an unsupervised signal-based approach

This section describes a bottom-up attention approach which could also be seen as an unsupervised attention. Bottom-up attention uses the acquired image characteristics to predict its important regions and acts like a gate to memory. This model is somehow based on Edgar Alan Poe's proposition: "observing attentively is remembering clearly". Unsupervised attention is thus very important in remembering and it is able to keep in mind the details or rare regions within the image. Without attention, these important details are forgotten which implies a loss of crucial data.

When performing a remembering task about an already visited place for example, people remember a rough image about this place. The process of forgetting may be modeled by a low-pass filtering whose kernel size increases in time. Here, a set of six low-pass filters with increasing kernel sizes is used for each grey level of the image. The number of grey levels is reduced to 11 to speed up computation and avoid noise. The size of the largest low-pass filter kernel is chosen to be close to the half of the image. If the original image is larger or smaller than the largest filtering kernel size, it is resized to better fit the scale decomposition.

This idea is illustrated in Fig. 1 where the sky (big red upper rectangle) and the pool (small red rectangle in the middle) have the same grey level. At the higher resolution (top row on Fig. 1), two pixels (one in the middle of the sky and the other one in the middle of the pool) have the same global occurrence which is equal to the number of red pixels. When going from top images to bottom images in Fig. 1, low-pass filter kernels sizes (neighborhood sizes) are larger, thus the images are forgotten more and more. The occurrences of the two pixels have different behaviors (plots of the left column: sky; plots on the right column: pool). If the pixel within the sky has a slowly decreasing occurrence, the pool pixel's occurrence decreases very fast when larger and larger neighborhoods are taken into account (larger low-pass kernels). The pool pixel has an occurrence which gets rapidly very small while the sky pixel keeps a higher occurrence even when taking into account larger neighborhood sizes.

In order to quantify the behavior for each pixel, the sum on the scale space is used. This sum can be visualized in Fig. 1 on the right and left columns as the area behind the occurrence variation plots function of the neighborhood size. The occurrence probability of a pixel is obtained by the normalization of this sum and the self-information represents the attention score for the pixel (Eq. 1).

$$Attention(I_j) = -\log\left(\frac{1}{S \times Card(I_j)} \sum_{k=1}^{S} n_k\right) \quad \textbf{(1)}$$

In Equation 1, n_k is the occurrence value of the current pixel within the k^{th} resolution level. In the current implementation there are seven different resolutions and the one corresponding to k=1 means the grey level is unfiltered. S is a constant which is equal

to the total number of resolutions (here S=7). I_j is the j^{th} grey level of the image I and $Card(I_j)$ is its cardinality.

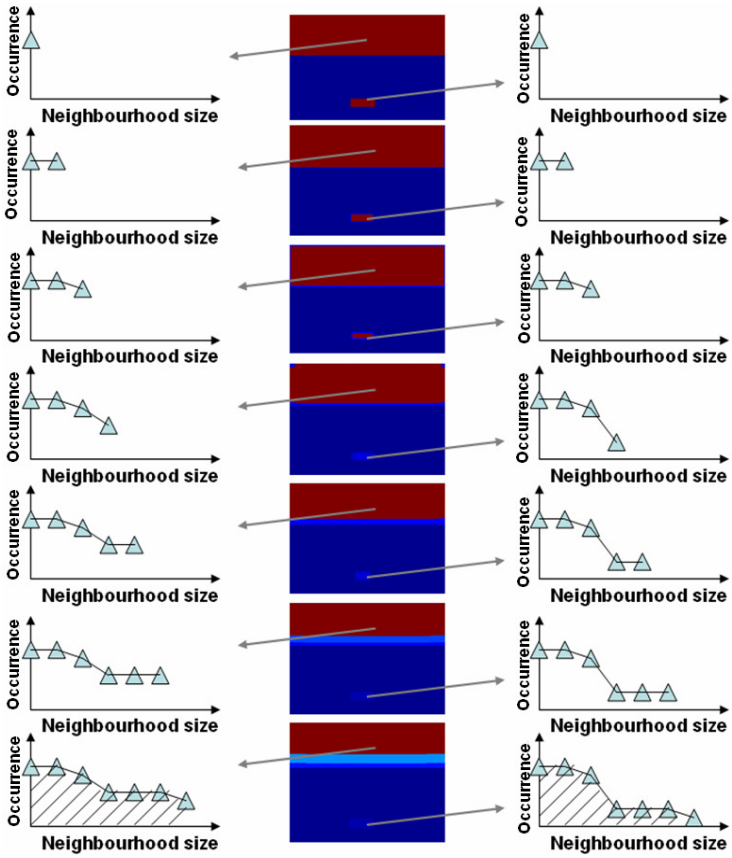

Fig. 1. From top to down: the decreasing resolution (increasing low-pass filtering kernel or neighborhood size) of an initial image grey level, From left to right: the occurrence behavior of a pixel in the upper red rectangle ("the sky") during the forgetting process; an image grey level; the occurrence behavior of a pixel in the lowest red rectangle ("the pool") during the forgetting process

In the current implementation, instead of simply using grey levels, their contrast maps are used. These maps are obtained as follows: for the grey level j (j is set between 1 and 11 in the current implementation), the pixels equal to j are assigned with the value 1 while pixels different from j are assigned with a value between 0 and 1. This value will be close to 1 if the pixel has a similar value with j and close to 0 if this pixel has a grey level, very different from j.

The color images are handled within an opposition color system. For each of the components (luminance, red-green opposition, blue-yellow opposition) a separate attention map is computed: the final map is obtained by adding the maps with a higher weight on the luminance which contains more information.

4 Top-down attention: a supervised application-driven approach

While a bottom-up approach uses signal characteristics to achieve attention computation, the top-down approach mainly uses feedbacks from the memory (a priori knowledge) and it depends on the task or the application to be achieved. Top-down attention can be seen as a supervised attention. In this section, a top-down approach for still images is proposed. The idea is to model the observers' behavior depending on the kind of images they look at.

Observers' behavior can be modeled by using eye-tracking or other alternative methods such as mouse-tracking to detect their gaze path. The mean of the gaze path of several observers is called a priority map and it highlights, for one image, the areas where the mean of a set of observers mostly looks as it is shown in Fig. 2.

Fig. 2. Left: original image, right: a priority map obtained by mouse-tracking

A top-down model can be achieved by using the mean of the priority maps obtained for a specific set of images (images with common meaning). Three sets of images [29] (set1, set2 and set3) were used within these tests to build three different top-down models:

- The first image set is made from 26 natural scene images.
- The second image set is made from 30 various advertisement images. Well-known trade marks were chosen as they have a huge advertising presence.
- The third image set is made from 35 various web sites. The websites of the 12 candidates to the French presidential election of 2007 are analyzed along with university and lab websites, institutional and government web sites. Some commercial websites have also been added. These media do not intend to provide the same informational content that is why it is interesting to see if there is a common attentional behavior to all these websites.

It is important to highlight the fact that a top-down model built in that way needs two main requirements to be meaningful:

- The first one is about the number of observers who provide their mouse track paths which should be high enough to get a realistic observer mean. Here, 40 to 60 observers' mouse paths per image were recorded. There were neither advertisement or web experts nor a specific age or gender class of observers: they can reasonably be considered as general public.
- The second requirement is about the homogeneity of the image set. The more the set of images is specific, the more the top-down model is accurate.

Fig. 3 displays the three top-down models from left to right: the sets presenting natural images, advertisements and websites. For the natural scene images, the mean priority map is mostly centered and it oddly looks like a centered Gaussian. The two other models are quite similar: high scores are detected in the top-left corner of the image decreasing towards the center. Nevertheless, the models used for advertisements and web sites also have some differences. Fig. 3 shows that the advertisement model is less selective than the web sites one: structures on its center are also quite well highlighted.

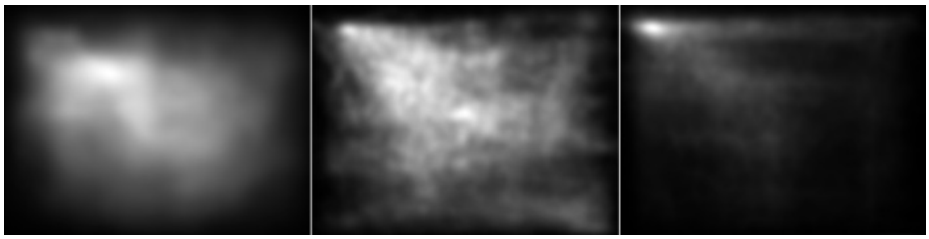

Fig. 3. Left to right: top-down models for a set of natural images, advertisements, web sites

The web sites model is a typical structured document model. The natural images model is typical of unknown unstructured images. The advertisements model seems to be a mix between these two extreme models. Its structure is close to a structured document one (human contribution is high), but it also covers the central areas of the image and the opposite corners where logos may often be found. This experiment shows how observers' attention behavior is different depending on the set of images.

5 Computational attention evaluation

The bottom-up attention model proposed in section 3 and the Itti's reference bottom-up saliency map [3] are both compared to the mouse-tracking results on a 91-image database. Both bottom-up methods are then added to three top-down models concerning the three categories of images within the database and compared again with the mouse-tracking data. A problem is that Itti's saliency maps and the attention maps computed with the model proposed in section 3 are of quite different nature. The proposed algorithm preserves the main structure of the image while this structure is no more recognizable with Itti's algorithm. Moreover, the priority maps are the results of several filterings, thus they are very smooth. In order for all compared algorithms and priority maps to have the same spatial characteristics (smooth areas) the algorithm proposed in section 3 is low-pass filtered.

5.1 Algorithms comparison: bottom-up information only

To obtain quantitative results, the classical linear correlation metric is used here. The correlation value goes from 0 (no similarity between the images) to 1 (there is a linear relationship between them).

Fig. 4 displays the correlation coefficient between the mouse-tracking priority maps and both Itti's bottom-up algorithm (dotted plot) and the proposed bottom-up algorithm (solid plot) for all images in the database. Very often, the tested images have better correlation coefficients for the proposed algorithm than for Itti's one.

Table 1 shows the mean and standard deviation of the correlation coefficients for the three sets of images.

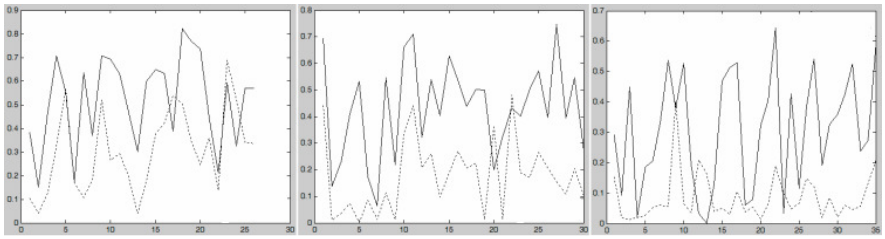

Fig. 4. Correlation coefficient between the priority maps and both the bottom-up algorithm of Itti (dotted plot) and the proposed bottom-up algorithm (solid plot). From left to right: natural scene database, advertising database, web sites database

Table 1. Bottom-up linear correlation mean (MEAN) and standard deviation (STD) results for the three sets of images

Image set	Itti MEAN	Itti STD	Mancas MEAN	Mancas STD
Natural Images	30%	18%	52%	19%
Advertisements	17%	14%	43%	18%
Web Sites	9%	10%	31%	19%

5.2 Algorithms comparison: both bottom-up and top-down information

In this section, the same top down models (those proposed in section 4 and displayed in Fig. 3) were added to Itti's and to the proposed (Mancas) bottom-up algorithms. High results improvements may be observed (Fig. 5) compared to the previous section where no top-down influence was taken into account. Moreover, even with the same top-down models, one may see that the bottom-up model remains very important as results of the proposed bottom-up method are better in terms of linear correlation than those of Itti's bottom-up model for 90 images on 91.

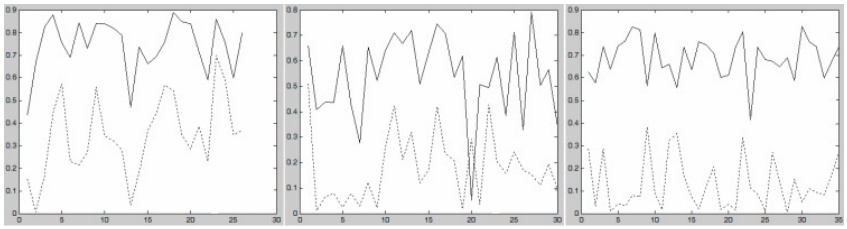

Fig. 5. Correlation coefficient between the priority maps and both the bottom-up and top-down attention algorithms based on Itti (dotted plot) and on Mancas (solid plot). From left to right: natural scene database, advertising database, web sites database

Table 2 summarizes the mean and standard deviation of the correlation coefficients for the three sets of images. A simple comparison with Table 1, where no top-down information was used shows the importance of the top-down step in attention.

Table 2. Bottom-up & top-down linear correlation mean (MEAN) and standard deviation (STD) results for the three sets of images

Image set	Itti MEAN	Itti STD	Mancas MEAN	Mancas STD
Natural Images	34%	17%	74%	12%
Advertisements	18%	14%	53%	16%
Web Sites	13%	11%	69%	9%

5.3 Algorithms comparison: a discussion

The bottom-up influence is higher for natural scene images than for websites images for example. For both Mancas and Itti methods, the bottom-up attention alone (Table 1) provides the best results for natural scene images, while this score decreases with the advertisement set and even more with the web sites set.

Moreover, the results of Table 2 show that the 74% of correlation for natural images (Mancas method) is due to 52% bottom-up and 22% top-down. On the other side, the result of 69% of correlation for the web site images is due to only 31% of bottom-up and to 38% top-down influence. A similar behavior can also be detected on the figures of the Itti method of Table 2.

A very interesting conclusion of these observations is that the more one knows about an image, the higher the top-down influence part will be. On the other side, for an unknown image, the bottom-up attention mechanism will be very important. Thus, if the role of top-down information in attention is always very important, its part in the attention process depends on the amount of knowledge that a mean observer may have on a given kind of images.

Nevertheless, advertisements score enhancement between Table1 and Table 2 is lower than expected: as advertisements are a mix between unstructured (natural scenes) and structured (web sites) documents, the influence of the top-down attention should be higher than for natural scene images. This is not the case because the top-down model used here only includes document structure and not faces and text which are also very powerful top-down stimuli. These stimuli proved to be very important for advertisements where faces and text are often very present.

The correlation results of Table 1 and Table 2 also need some remarks. The use of the linear correlation may not be the best metric to compare computational attention algorithms and other distances could be taken into account. Moreover, Itti's saliency map which focuses very highly on precise areas in an image may be penalized by the use of the linear correlation coefficient. It is thus quite difficult to precisely compare these two methods which have different behavior as there is no standard method for attention algorithms assessment. However, a precise analysis of both qualitative and quantitative results on the overall database shows that the proposed bottom-up algorithm outperforms the bottom-up algorithm proposed by Itti.

The purpose of this section was to show that the correlation coefficients between the mouse-tracking priority maps and the proposed algorithm become very interesting

and they can be considered as a quite good approximation of human attention. These correlation figures demonstrate that the use of attention to predict human gaze makes sense if both bottom-up and top-down information are used.

6 Conclusion

A bottom-up or unsupervised computational attention algorithm is presented which performs better than Itti's reference model on the test database. However, several improvements should be added to this algorithm, mainly to handle spatial orientations.

A top-down or supervised attention model based on the mean of the eye-tracking or mouse-tracking priority maps was also proposed. It proved to highly increase the results of the bottom-up algorithms and to finally provide a good approximation of human gaze. Other top-down influences should also be added to improve the results as face or text detection. Faces and text are known as very informative and that is why they represent very important top-down influences especially if there are few faces or few text within the images.

The encouraging results presented here confirm the more and more widely accepted idea that the automatic prediction of human attention for still images becomes quite accurate if both bottom-up and top-down information is used while bottom-up information alone remains insufficient. An issue in computational attention is the lack of a standard assessment method and database to really prove the pertinence of those approaches in predicting human attention.

An interesting point was found about the relative importance of bottom-up and top-down influences: the bottom-up mechanism is very important for new images, while for structured document where people are used with, the top-down influence is higher than the bottom-up one. These results which can be seen in sections 5.1 and 5.2 show that there is a complex relationship between bottom-up and top-down attention and that their roles are specific. On one side, bottom-up attention is oriented in learning which areas of an image are the most relevant mainly for new images and situations. On the other side, top-down attention uses already learnt situations to select some areas of the current image by inhibiting those where there are very few chances to find relevant information. Bottom-up and top-down attention interaction aims in optimizing reactions to both novel and already experienced situations.

Acknowledgement

This work was achieved in the framework of the Translogistic project which is funded by Région wallonne, Belgium.

References

1. Desimone, R., “Visual attention mediated by biased competition in extrastriate visual cortex”, Phil. Trans. R. Soc. Lond. B, 353, pp 1245 - 1255, 1998
2. Boynton, G.M., “Attention and visual perception”, Current Opinion in Neurobiology, 15:465-469, 2005
3. Itti, L., Koch, C., Niebur, E., “Model of saliency-based visual attention for rapid scene analysis”, IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(11):1254-1259, 1998
4. Itti, L., Koch, C., “A saliency-based search mechanism for overt and covert shifts of visual attention”, Vision Research, 40:1489-1506, 2000
5. Itti, L., Koch, C., “Computational modelling of visual attention”, Nature RevNeuroscience, 2(3) :194-203, 2001
6. Koch, C., Ullman, S., “Shifts in selective visual attention: towards the underlying neural circuitry”, Human Neurobiology, 4(4), 219-270, 1985
7. Milanese, R., Bost, J.M., Pun, T., “A bottom-up attention system for active vision”, ECAI92, 10th European Conference on Artificial Intelligence, pp. 808-810, 1992
8. Milanese, R., “Detecting salient regions in an image: from biological evidence to computer implementation”, PhD Thesis, University of Geneva, 1993
9. Chauvin, A., Herault, J., Marendaz, C., Peyrin, C., “Natural scene perception: visual attractors and image processing”, 7th Neural Computation and Psychology Workshop, 2000
10. Petkov, N., Westenberg, M. A., “Suppression of contour perception by band-limited noise and its relation to non-classical receptive field inhibition”, Biological Cybernetics, 88, 236-246, 2003
11. Le Meur, O., “Attention selective en visualisation d’images fixes et animees affichees sur ecran : Modeles et evaluation des performances - Applications”, PhD Thesis, University of Nantes, 2005
12. Le Meur, O., Le Callet, P., Barba, D., “A spatio-temporal model of bottom-up visual selective attention: description and assessment”, Vision Research, 2007
13. Mudge, T. N., Turney, J. L., Volz, “Automatic generation of salient features for the recognition of partially occluded parts”, Robotica, Vol 5, pp 117-127, 1987
14. Osberger, W., Maeder, A. J., “Automatic identification of perceptually important regions in an image”, 14th IEEE Int. Conference on Pattern Recognition, 1998
15. Walker, K. N., Cootes, T. F., Taylor, C. J., “Locating Salient Object Features”, British Machine Vision Conference, 1998
16. Oliva, A., Torralba, A., “Modeling the shape of the scene: a holistic representation of the spatial envelope”, International Journal of Computer Vision, 43(3):145-175, 2001
17. Oliva, A. Torralba, A., Castelhano, M. S., Henderson, J. M, “Top-down control of visual attention in object detection”, IEEE Inernational Conference on Image Processing, 2003
18. Bruce, N., Jernigan, E., “Evolutionary design of context-free attentional operators”, Proc. Of the IEEE International Conference on Image Processing, 2003
19. Bruce, N., Tsotsos, J.K., “Saliency Based on Information Maximization”, Proc. Of the the Neural Information Processing Systems, 2005
20. Liu, F., Gleicher, M., “Video Retargeting: Automating Pan-and-Scan”, ACM Multimedia, 2006
21. Comaniciu, D., Meer, P., “Mean Shift: A Robust Approach toward Feature Space Analysis”, IEEE Trans. Pattern Analysis Machine Intell., Vol. 24, No. 5, 603-619, 2002
22. Itti, L., Baldi, P., “Bayesian Surprise Attracts Human Attention”, Advances in Neural Information Processing Systems, Vol. 19 (NIPS 2005), pp. 1-8, Cambridge, MA:MIT Press, 2006
23. Stentiford, F.W.M., “An estimator for visual attention through competitive novelty with application to image compression”, Picture Coding Symposium, pp. 25-27, 2001

24. Boiman, O., Irani, M., “Detecting Irregularities in Images and in Video”, International Conference on Computer Vision (ICCV), 2005
25. Boiman, O., Irani, M., “Similarity by Composition”, Neural Information Processing Systems (NIPS), 2006
26. Mancas, M., Mancas-Thillou, C., Gosselin, B., Macq B., "A rarity-based visual attention map - application to texture description -", Proc. of IEEE International conference on Image Processing (ICIP), 2006
27. Mancas, M., Gosselin, B., Macq B., "A Three-Level Computational Attention Model", Proceedings of ICVS Workshop on Computational Attention & Applications (WCAA), 2007
28. Mancas, M., Gosselin, B., Macq B., "Perceptual Image Representation", EURASIP Journal of Image and Video Processing, Volume 2007, Article ID 98181, doi:10.1155/2007/98181, 2007
29. Validattention website: http://tcts.fpms.ac.be/~mousetrack/pageAccueil.php?langue=en