A SPATIOTEMPORAL SALIENCY MODEL OF VISUAL ATTENTION BASED ON MAXIMUM ENTROPY

Longsheng Wei, Nong Sang and Yuehuan Wang

Institute for Pattern Recognition and Artificial Intelligence
National Key Laboratory of Science & Technology on Multi-spectral Information Processing
Huazhong University of Science and Technology, Wuhan 430074, P.R.China
weilongsheng@163.com, nsang@hust.edu.cn, yuehwang@mail.hust.edu.cn

Commission III/5

KEY WORDS: Spatiotemporal saliency model, Visual attention, Maximum entropy, Saliency map

ABSTRACT:

This paper presents a spatiotemporal saliency visual attention model based on maximum entropy. A dynamic saliency map is created by calculating entropies in every corresponding local region between the current and some previous frames in a short video. In the same time, four low-level visual features including color contrast, intensity contrast, orientation and texture, are extracted and are combined into a static saliency map for the last frame in the video. Our proposed model decides salient regions based on a final saliency map which is generated by integration of the dynamic saliency map and the static saliency map. At last, the size of each salient region is obtained by maximizing entropy of the final saliency map. The key contribution of this paper is that the entropy value method is used to calculate dynamic saliency for some successive frames in a short video. Experimental results indicate that: when the moving objects do not belong to the salient regions, our proposed model is excellent to Ban's model.

1 INTRODUCTION

The human visual system can effortlessly detect an interesting region or object in natural scenes through the selective attention mechanism. Visual attention represents a fundamental mechanism for computer vision where similar speed up of the processing can be envisaged. The selective attention mechanism also allows the human vision system to more effectively process input visual scenes with a higher level of complexity. In a short video, motion has to be included based on the fact that people's attention is more easily directed to a motive stimulus in a static scene. The motion is clearly involved in visual attention, where rapid detection of moving objects is essential for adequate interaction with the environment. Therefore, the human visual system sequentially interprets not only a static input scene but also a dynamic input scene with the selective attention mechanism.

Most computational models (Tsotsos et al., 1995, Le Meur et al., 2006b)of visual attention are static and are inspired by the concept of feature integration theory (Treisman and Gelade, 1980). the most popular is the one proposed by L. Itti et al. (Itti et al., 1998) and it has become a standard model of static visual attention, in which salience according to primitive features such as intensity, orientation and color are computed independently. There are also many models (Ouerhani and Hugli, 2003, Veit et al., 2004, López et al., 2006, Le Meur et al., 2006a, Shi and Yang, 2007, Bur et al., 2007, Bogdanova et al., 2010) bringing dynamic saliency to visual attention mechanism. Ban et al. (Ban et al., 2008) propose a dynamic visual selective attention model. Firstly, a static saliency map is obtained by a frame in a video. Secondly, an optimal scale is calculated for each pixel location and for each static saliency map. Thirdly, those optimal scales and static saliency maps are used to calculate the entropy to form static entropy maps, then a static entropy map is obtained for every frame in the video. At last, all the entropy maps are used to calculate a new entropy map, which is called dynamic saliency map. However, when the moving objects do not belong to the salient regions, Ban's approach is very hard to find the moving regions.

In order to address the above problem, we propose a spatiotemporal saliency visual attention model based on maximum entropy in this paper. A dynamic saliency map is based on the successive frames including current and some previous frames in a short video. In our work, the dynamic saliency map is created by calculating entropies in every corresponding local region between the current and previous frames in the video. In the same time, four low-level visual features including color contrast, intensity contrast, orientation and texture, are extracted and are combined into a static saliency map for the last frame in the video. Our proposed model decides salient regions based on a final saliency map which is generated by integration of the dynamic saliency map and the static saliency map. At last, the size of each salient region is obtained by maximizing entropy of the final saliency map. Experimental results indicate that: when the moving objects do not belong to the salient regions, our proposed model is excellent to Ban's model. Our proposed model is shown in Figure 1.

This paper is organized as follows. Section two presents dynamic saliency map. Static saliency map is described in section three. Section four introduces how to acquire final saliency map and the size of salient region. Section five shows experimental results, and section six concludes this paper.

2 DYNAMIC SALIENCY MAP

Our proposed model is inspired by the human visual system from the retina cells to the complex cells of the primary visual cortex. The visual information goes through the retina preprocessing to the cortical-like filter decomposition (Massot and Hérault, 2008). The retina extracts two signals from each frame that correspond to the two main outputs of the retina (Beaudot, 1994). Each signal is then decomposed into elementary features by a bank of cortical-like filters. These filters are used to extract both dynamic and static information, according to their frequency selectivity, providing two saliency maps: a dynamic and a static one. Both

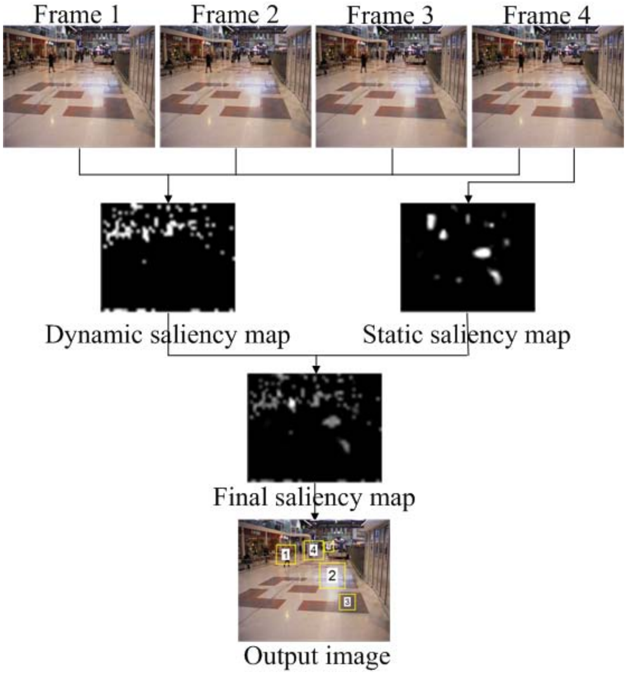

Figure 1: Our proposed model: Given four successive frames in a short video. Firstly, a dynamic saliency map is created by calculating entropies in every corresponding local region between these four frames. Secondly, four low-level visual features are extracted and are combined into a static saliency map for last frame. Thirdly, dynamic and static saliency map are fused into a final saliency map, which are guided human visual attention.

saliency maps are combined to obtain a final saliency map (Marat et al., 2009).

The retina, which has been described in detail in (Schwartz, 2004), has two outputs formed by different ganglion cells: magnocellular output and parvocellular output. Magnocellular output responds rapidly and provides global information which can be simulated by using lower spatial frequencies. Parvocellular output provides detailed information which can be simulated by extracting the high spatial frequencies of an image. This output enhances frame contrast, which attracts human gaze in static frame (Reinagel and Zador, 1999). Our proposed model decomposes the input short video into different frequency bands: a lower spatial frequency one to simulate the dynamic output and a high spatial frequency one to provide a static output.

For a given short video, motion has to be included based on the fact that people's attention is more easily directed to a motive stimulus in a static scene. The proposed dynamic saliency map model is based on the successive frames including current and some previous frames in the video. In our work, the dynamic saliency map is created by calculating entropies in every corresponding local region between the current and previous frames in the video. This map includes information from not only the current frame, but also the previous ones.

In order to understand better how the entropy is correlated with dynamic saliency, we convert the color successive frames into gray-scale successive frames. We work with 256 gray level successive frames and transform them to a lower number of levels n<256. Generally, good results are usually obtained with n = 8 levels in normal illumination indoor and outdoor scenes. A higher value rarely gives better results, whilst lower values (say, 2 or 4) may be used for night vision (Lópeza et al., 2006). In this paper, we choose the number of the input frames in this short video as the number of levels. Let the maximal pixel value of all frames is M, for each coordinate (x, y) at successive k frames in the video, we normal this pixel value divided by M to a fixed range $[0, 1]$. After dividing the $[0, 1]$ range into some k equal parts, we let the values in different parts be different integers, whose range is $[0, k-1]$. Those integers are defined by Equation (2). Figure 2 shows an example of transforming the input color frame (a) into gray frame (b), eight gray level bands (c) and four gray level bands (d).

$$f(x,y,k) = I(x,y,k)/M \tag{1}$$

$$g(x,y,k) = \begin{cases} 0 & 0 \leq f(x,y,k) \leq \frac{1}{k} \\ 1 & \frac{1}{k} < f(x,y,k) \leq \frac{2}{k} \\ & \cdots \\ k-1 & \frac{k-1}{k} < f(x,y,k) \leq 1 \end{cases} \tag{2}$$

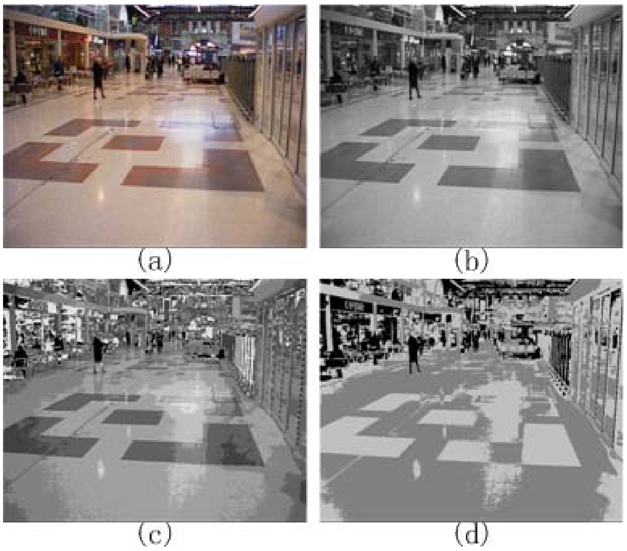

Figure 2: (a) Input color frame; (b) Gray frame; (c) Eight gray level bands; (d) Four gray level bands.

In order to be consistent with the size of static saliency map and to detect the moving region effectively, we reduce the size of every frame and calculate the dynamic saliency in a local region. Let $R(x, y)$ is a local region in coordinate (x, y). We calculate the corresponding entropy by the probability mass function, which is obtained by the histogram generated using this k integer values. The bigger of the entropy value, the more conspicuity of this point. The time-varying entropy $M_d(.)$ is calculated by Equation (3) and all the entropy values can form a dynamic saliency map.

$$M_d(x,y) = - \sum_{k\in\{0,1,\cdots,k-1\}} p_{g(x',y',k)} log_2 p_{g(x',y',k)} \tag{3}$$

where

$$(x', y') \in R(x,y) \tag{4}$$

$p_{g(x',y',k)}$ is the probability mass function, which is obtained by the histogram generated using those integer values at local region $R(x, y)$in all the successive frames.

3 STATIC SALIENCY MAP

Retinal input is processed in parallel by multi-scale low-level feature maps for every frame in a video, which detect local spatial discontinuities using simulated center-surround operations. Four low-level visual features including color contrast, intensity contrast, orientation and texture, are extracted and are combined into a static saliency map. Let r, g and b are three color channels of input image, four broadly-tuned color channels are created: $R = r - (g + b)/2$ for red, $G = g - (r + b)/2$ for green, $B = b - (r + g)/2$ for blue and $Y = (r + g)/2 - |r - g|/2 - b$ for yellow (negative values are set to zero). $RG = |R - G|$ is red/green contrast; $BY = |B - Y|$ is blue/yellow contrast. Therefore, color features are divided into red/green contrast and blue/yellow contrast two parties.We divided intensity into intensity on (light-on-dark) and intensity off (dark-on-light). We convert the color object image into gray-scale image to obtain an intensity image and let center/surround contrast be intensity on, surround/center contrast be intensity off. The reason is that the ganglion cells in the visual receptive fields of the human visual system are divided into two types: on-center cells respond excitatory to light at the center and inhibitory to light at the surround, whereas off-center cells respond inhibitory to light at the center and excitatory to light at the surround (Palmer and E., 1999). There are four orientations in our model: 0^o, 45^o, 90^o, 135^o. The orientations are computed by Gabor filters detecting bar-like features according to a specified orientation. Gabor filters, which are the product of a symmetric Gaussian with an oriented sinusoid, simulate the receptive field structure of orientation-selective neurons in primary visual cortex (Palmer and E., 1999). A Gabor filter centered at the 2-D frequency coordinates (U, V) has the general form of

$$h(x, y) = g(x', y')exp(2\pi i(Ux + Vy)) \tag{5}$$

where

$$\begin{aligned} (x', y') &= (xcos\phi + ysin\phi, -xsin\phi + ycos\phi), & (6) \\ g(x, y) &= \frac{1}{2\pi\sigma_x\sigma_y}exp(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}). & (7) \end{aligned}$$

σ_x and σ_y are the scale parameter, and the major axis of the Gaussian is oriented at angle ϕ relative to the axis and to the modulating sinewave gratings. In this paper, let the scale of Gabor filters equal to the scale of training object and let ϕ equal to 0^o, 45^o, 90^o and 135^o, respectively. For texture feature, we consider local binary pattern (LBP), which describes the local spatial structure of an image and has been widely used in explaining human perception of textures. Ojala et al. (Ojala et al., 1996) first introduced this operator and showed its high discriminative power for texture classification. At a given pixel position (x_c, y_c), LBP is defined as an ordered set of binary comparisons of pixel intensities between the center pixel and its eight surrounding pixels (Figure 3). The decimal form of the resulting 8-bit word (LBP code) can be expressed as follows:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{8}$$

where i_c corresponds to the gray value of the center pixel (x_c, y_c), i_n to the gray values of the 8 surrounding pixels, and function $s(x)$ is defined as:

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{9}$$

Two LBP operators are used in this paper, one is original LBP operator and the other is extended LBP operator with a circular neighborhood of different radius size. The extended LBP operator can keep size and rotation invariance and its pixel values are interpolated for points which are not in the center of a pixel. The two LBP operators are illustrated in Figure 4. Therefore, ten feature types are considered in this paper.

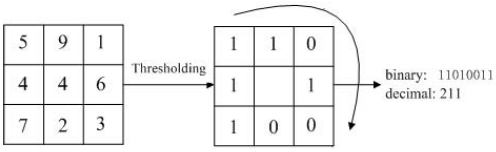

Figure 3: The LBP operator.

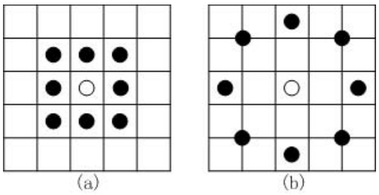

Figure 4: (a) The original LBP operator; (b) The extended LBP operator.

Center and surround scales are obtained using dyadic pyramids with nine scales (from scale 0, the original image, to scale 8, the image reduced by a factor 256). Center-surround differences are then computed as pointwise differences across pyramid scales, for combinations of three center scales ($c \in \{2, 3, 4\}$) and two center-surround scale differences ($\delta \in \{3, 4\}$); thus, six feature maps are computed for each of the ten features, yielding a total of sixty feature maps. Each feature map is endowed with internal dynamics that operate a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition. Resultingly, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings.

A region is salient if it is different from its neighbors. Thus, to strengthen the intermediate maps that have spatially distributed maxima, the method proposed by literature (Itti et al., 1998) is used. After being normalized in $[0, 1]$, each feature map (m_i $i = 1, \cdots, 60$) was multiplied by $(max(m_i) - \overline{m_i})^2$where $max(m_i)$ and $\overline{m_i}$ are its maximum and average respectively. Then, all values in each map that were smaller than 20% of its maximum were set to 0. All intermediate maps were added together to obtain a static saliency map M_s .

4 FINAL SALIENCY MAP AND THE SIZE OF SALIENT REGION

The dynamic saliency map and static saliency map are described above. The final saliency map is their weighted sum. Both maps compete for saliency: the dynamic saliency map emphasizing its temporal saliency; the static salient map showing regions that are salient because of its spatial conspicuities. To make the maps comparable, M_d is normalized in advance to the same range as M_s . When fusing the maps, it is possible to determine the degree to which each map contributes to the sum. This is done by weighting the maps with a static salient map factor $t \in [0, \cdots, 1]$.

$$M = t \times M_d + (1 - t) \times M_s \tag{10}$$

The entropy maximum is considered to analyze the sizes of salient regions (Kadir and Brady, 2001).The most appropriate scale x_s for each salient region centered at location x in the final saliency map is obtained by Equation (11) which aims to consider spatial dynamics at this location:

$$x_s = arg\max_s\{H_D(s,x) \times W_D(s,x)\} \tag{11}$$

where D is the set of all descriptor values which consist of the intensity values corresponding the histogram distribution in a local region with size s around an attended location x in final salience map, $H_D(s,x)$ is the entropy defined by Equation (12) and $W_D(s,x)$ is the inter-scale measure defined by Equation (13).

$$H_D(s,x) \quad = \quad -\sum_{d\in D} p_{d,s,x} log_2 p_{d,s,x} \tag{12}$$

$$W_D(s,x) \quad = \quad \frac{s^2}{2s-1}\sum_{d\in D} |p_{d,s,x} - p_{d,s-1,x}| \tag{13}$$

where $p_{d,s,x}$ is the probability mass function, which is obtained by normalizing the histogram generated using all the pixel values in a local region with a scale s at position x in the final salience map, and the descriptor value d is an element in a set of all descriptor values D, which is the same set of all the pixel values in a local region.

5 EXPERIMENTAL RESULTS

Ban et al. (Ban et al., 2008) also proposed a dynamic visual selective attention model. The dynamic saliency map model part is described as follows. Firstly, a static saliency map is obtained by a frame in a video. Secondly, an optimal scale is calculated for each pixel location and for each static saliency map. Thirdly, those optimal scales and static saliency maps are used to calculate the entropy to form static entropy maps, then a static entropy map is obtained for every frame in the video. At last, all the entropy maps are used to calculate a new entropy map, which is called dynamic saliency map.

However, when the moving objects do not belong to the salient regions, Ban's approach is very hard to find the moving regions. Therefore, our proposed approach uses the information of every frame, instead of using the information of its saliency map to obtain a dynamic saliency map.

We choose five successive frames in a short video. The static saliency maps and scan paths for the first frame and the last frame are shown in Figure 5. Figure 6 (a) and (b) show Ban's dynamic saliency map obtained from the five successive static saliency maps and scan path generated by the dynamic saliency map, respectively. Figure 6 (c) and (d) show Ban's final saliency map and scan path, respectively. Figure 7 (a) and (b) show our proposed dynamic saliency map obtained from the successive frames and the scan path generated by the dynamic saliency map, respectively. Figure 7 (c) and (d) show our proposed final saliency map and scan path, respectively. In experiment, we take $t = 0.5$ which expresses that the dynamic map is as important as the static map for the final map. Figure 7 (d) includes scale information in salient region is represented by a scaled box on the corresponding salient region; other scan paths do not include any scale information and the boxs just express the location of salient region.

Comparing to the scan path of Figure 6 (b) and Figure 7 (b), we can find that our proposed approach is excellent to Ban's approach to detect the moving regions. The primary reason is that some moving objects do not belong to the static salient regions. Our proposed approach uses the information of every frame, instead of using the information of its saliency map. So our final scan path contains more dynamic regions. Our approach accords with the habit of human vision system.

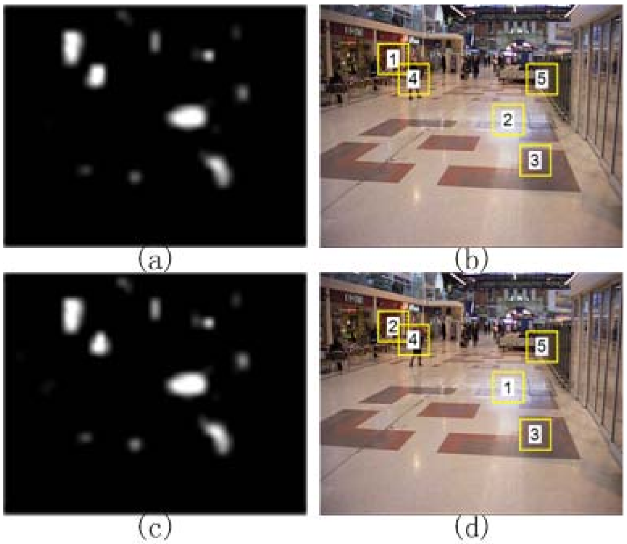

Figure 5: (a) and (b) are static saliency map and scan path for the 1st frame, respectively; (c) and (d) are static saliency map and scan path for the 5th frame, respectively.

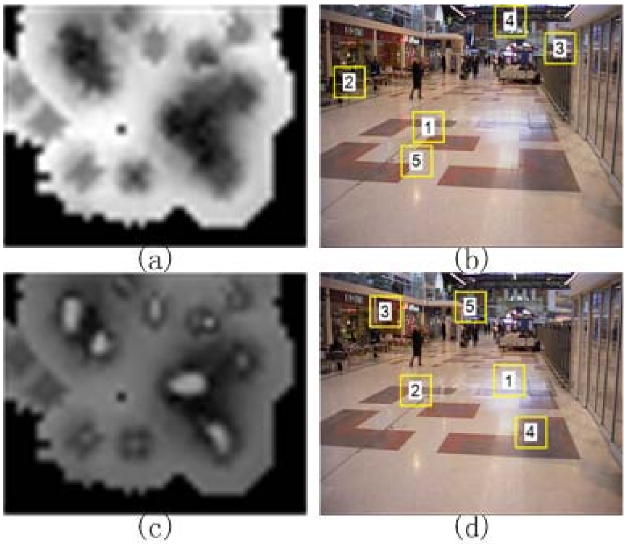

Figure 6: (a) and (b) are dynamic saliency map and scan path of Ban's model, respectively; (c) and (d) are final saliency map and scan path of Ban's model, respectively.

6 CONCLUSION

We have proposed a spatiotemporal saliency visual attention model based on maximum entropy in this paper. The main process is described as following. Firstly, a dynamic saliency map is based on the successive frames including current and some previous frames in a short video. In our work, the dynamic saliency map is created by calculating entropies in every corresponding local region between the current and previous frames in the video. Secondly, four low-level visual features including color contrast, intensity contrast, orientation and texture, are extracted and are combined into a static saliency map for the last frame in the

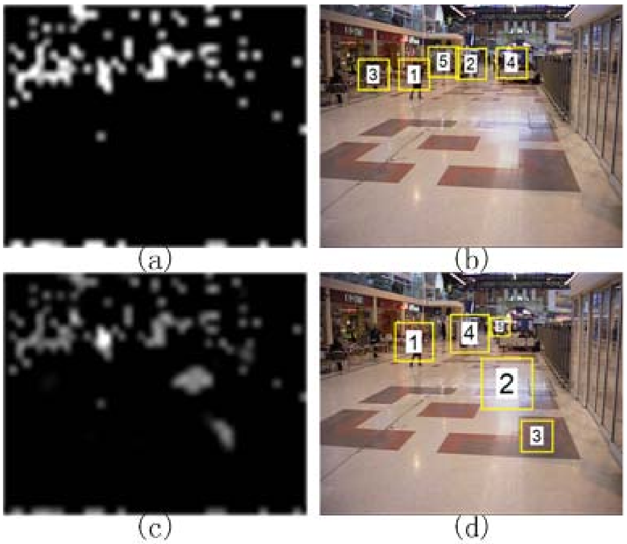

Figure 7: (a) and (b) are dynamic saliency map and scan path of our proposed model, respectively; (c) and (d) are final saliency map and scan path of our proposed model, respectively.

video. Our proposed model decides salient regions based on a final saliency map which is generated by integration of the dynamic saliency map and the static saliency map. Thirdly, the size of each salient region is obtained by maximizing entropy of the final saliency map. Experimental results indicate that: when the moving objects do not belong to the salient regions, our proposed model is excellent to Ban's model.

The key contribution of this paper is that the entropy value method is used to calculate dynamic saliency for some successive frames in a short video. The proposed selective attention model for a dynamic scene can play an important role as an initial vision process for a more human-like robot system. This method presents a new model for predicting the position of human gaze. In our future works, we will extend our spatiotemporal visual attention model to work in top-down environment by adding some prior knowledge.

ACKNOWLEDGEMENTS

This work was supported by the Chinese National 863 Grant No. 2009AA12Z10 and National Natural Science Foundation of China under contracts 60736010.

REFERENCES

Ban, S.-W., Lee, I. and Lee, M., 2008. Dynamic visual selective attention model. Neurocomputing 71, pp. 853–856.

Beaudot, W. H., 1994. The neural information in the vertebra retina: a melting pot of ideas for artificial vision. PHD thesis, Tirf laboratory, Grenoble, France.

Bogdanova, I., Bur, A., Hügli, H. and Farine, P., 2010. Dynamic visual attention on the sphere. Computer Vision and Image Understanding 114, pp. 100–110.

Bur, A., Wurtz, P., Müri, R. and Hügli, H., 2007. Motion integration in visual attention models for predicting simple dynamic scenes. Human Vision and Electronic Imaging XII.

Itti, L., Koch, C. and Niebur, E., 1998. model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, pp. 1254–1259.

Kadir, T. and Brady, M., 2001. Saliency, scale and image description. International Journal of Computer Vision 45(2), pp. 83–105.

Le Meur, O., Le Callet, P. and Barba, D., 2006a. Predicting visual fixations on video based on low-level visual features. Vision research 47, pp. 2483–2498.

Le Meur, O., Le Callet, P., Barba, D. and Thoreau, D., 2006b. A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, pp. 802–817.

López, M., Fernández-Caballero, A., Fernández, M., Mira, J. and Delgado, A., 2006. Motion features to enhance scene segmentation in active visual attention. Pattern Recognition Letters 27(5), pp. 469–478.

Lópeza, M. T., Fernández-Caballeroa, A., Fernándeza, M. A., Mirab, J. and Delgadob, A. E., 2006. Visual surveillance by dynamic visual attention method. Pattern Recognition 39, pp. 2194–2211.

Marat, S., Phuoc, T. H., Granjon, L., Guyader, N., Pellerin, D. and Gurin-Dugu, A., 2009. Modelling spatio-temporal saliency to predict gaze direction for short videos. International Journal of Computer Vision 82, pp. 231–243.

Massot, C. and Hérault, J., 2008. Model of frequency analysis in the visual cortex and the shape from texture problem. International Journal of Computer Vision 76, pp. 165–182.

Ojala, T., PietikÄainen, M. and Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition 29(1), pp. 51–59.

Ouerhani, N. and Hugli, H., 2003. A model of dynamic visual attention for object tracking in natural image sequences. Computational Methods in Neural Modeling.

Palmer and E., S., 1999. Vision science, photons to phenomenology. The MIT Press, Cambridge, MA.

Reinagel, P. and Zador, A., 1999. Natural scene statistics at the center of gaze. Network: Computation in Neural Systems 10, pp. 341–350.

Schwartz, S. H., 2004. Visual perception: a clinical orientation (3rd edn.). New-York: McGraw-Hill.

Shi, H. and Yang, Y., 2007. A computational model of visual attention based on saliency maps. Applied Mathematics and Computation 188, pp. 1671–1677.

Treisman, A. M. and Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychology 12, pp. 97–136.

Tsotsos, J. K., Culhane, S. M., Winky, Y. K., Lai, Y., Davis, N. and Nuflo, F., 1995. Modeling visual attention via selective tuning. Artificial Intelligence 78, pp. 507–545.

Veit, T., Cao, F. and Bouthemy, P., 2004. Probabilistic parameterfree motion detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 715–721.