A Video Image Compression Method based on Visually Salient Features

Hongchang Ke[1], Hongbin Sun[1], Lei Gao[2], Hui Wang[3]
[1]School of Computer Technology and Engineering
Changchun Institute of Technology, Changchun 130012, China
[2]CSIRO Land and Water, Glen Osmond, SA 5064, Australia
[3]College of Computer Science and Engineering
Changchun University of Technology, Changchun 130012, China
khch_2000@163.com

Journal of Digital Information Management

ABSTRACT: *This study presents a visual attention model for determining image areas to receive different extents of video compression in order to minimize perceived program artefacts whilst maximizing the compression possible. The model integrates features related to motion with existing video image compression algorithms. The proposed visual attention model extracts the color, intensity, textural, and motion features of a video to determine the predicted region of interest (ROI). First, color, intensity, and texture saliency maps are generated by applying the "center-surround" method and a motion saliency map is produced using a difference operator. Then, a multi-channel weighting method is used to generate a global saliency map and to determine the ROI according to a winner-takes-all network (WTA). The proposed video image compression algorithm performs either low or no compression on the ROI while a high degree of compression is applied to the other regions. Tests indicate that the proposed visual attention model is able swiftly to identify the ROI, allowing the proposed compression algorithm to exert a high compression efficiency yet with minimally noticeable visual degradation.*

Subject Categories and Descriptors
I.2.10 [Vision and Scene Understanding]: Video Analysis; **I.4.10 [Image Representation]**

General Terms: Video Image Compression, Video Analysis

Keywords: Visual Attention, Region of Interest, Motion features, Digital Information Management, Simmilarity

Received: 4 July 2014, Revised 18 August 2014, Accepted 21 August 2014

1. Introduction

Video image compression is important in the fields of multimedia, digital information management and Internet communication. From the perspective of information theory, an image can be considered as a source, which contains redundant amounts of data [1–3]. The redundancy of image data represents spatial redundancy as a result of the correlation between the pixels of adjacent parts of the image; timing redundancy as a result of the correlations between frames in the image sequence; and spectrum redundancy caused by the correlations between different color planes or frequencies [4–7]. Image compression aims to reduce the bit requirement by eliminating data redundancy. However, it is feasible to allow certain regions of an image to be affected by compression artifacts if this is conducive to gaining an increase in compression ratio. For example, as the human eye is often the ultimate recipient of image information in most applications, it is possible to optimize the image compression in line with human visual characteristics, allowing us to discard some information to which the eye is insensitive. Thus both the compression ratio and visual quality can be enhanced on image recovery [8–11]. In short, video image compression aims to eliminate the data redundancies of sequential images by reducing the number of bits required to represent the data and to reconstruct it, while the resulting

video minimizes storage and transmission issues compared with the original [12]. Hence, a process which mimics salient features of the human visual system can be used effectively to determine the various data redundancies in sequential images [13].

Users are normally only particularly aware of the part of the image known as the region of interest (ROI) [14]. Therefore, the ROI compression method focuses on limiting the compression of this key area, compared with other parts of the image. However, ROI is very difficult to determine automatically, especially where the region of interest is not static. At present, therefore, ROI is typically processed manually [15–16].

The proposed visual attention mechanism can quickly locate the salient regions of an image. Thus, our computer model of visual attention can rapidly identify the ROI of an image without human intervention [17]. Itti and Koch proposed a significant spatial visual attention computer model [18] that extracts the feature vectors and filters the input image through multi-channel and multi-scale filtering. After feature extraction, the model simulates the receptive field properties of brain cells using a center-surround operator to obtain a saliency map. A neural network known as 'winner-take-all' (WTA) is then utilized to determine the ROI. This model continues to be the basis of the work of many researchers, but it is limited with respect to dynamic information as it is only based on spatial information.

Therefore, this study is the first to propose a visual attention model that also integrates motion features and then establishes a video image compression algorithm based on the aforementioned visual attention model. This visual attention model can detect the ROI without human intervention and then the proposed video image compression algorithm performs either low or no compression on the ROI, and high compression on the remaining regions. The proposed algorithm results in a higher compression ratio than the JPEG algorithm, while still maintaining important image quality and visual effects.

2. Visual Attention Model Combining Motion Features

In this model, the color, intensity, and textural features of the current frame are extracted. The static saliency maps (color saliency, intensity saliency, and texture saliency maps) are generated using a Gaussian pyramid, wavelet decomposition (Mallat), and center-surround operator. These multi-scale features are then normalized to generate the saliency map for each channel. The motion saliency map is generated by extracting the dynamic differences between the current and the previous frame using a multi-scale differential filter. Finally, a global saliency map is produced using a multi-channel weighting method so that the ROI can then be determined with the WTA mechanism. Figure 1 shows our visual attention model that includes integration of the motion features.

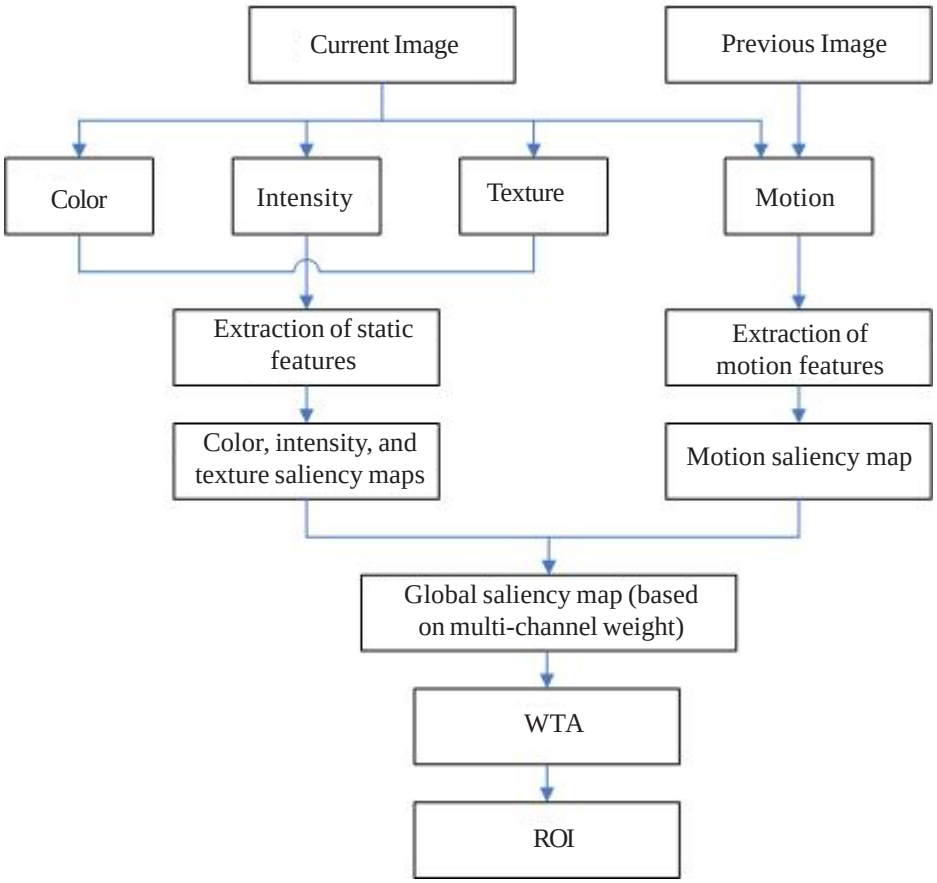

Figure 1. Visual attention model that includes motion features

2.1 Generation of the Static Saliency Map

First, we extract the static salient features, namely, the color, intensity, and textural features, of the input image. The image features are filtered using a nine-scale Gaussian pyramid. Scales 0 denotes the original image. The height and width of each overlying scale are sequentially halved. The model takes scales $\{2, 3, 4\}$ of the Gaussian pyramid as the central scale c. The surround scale is $s = c + \delta$, where $\delta \in \{3, 4\}$. Corresponding filters are applied to each scale of the pyramid, and we can generate six different combinations by subtracting the pixel values between the center and the surrounding scales: $\{2-5, 2-6, 3-6, 3-7, 4-7, 4-8\}$.

Assume that r, g, b represents the red, green, and blue features of the input image, respectively R, G, B and Y stand for red, green, blue, and yellow, respectively, in terms of the color features. These features can be calculated using the following equations:

$R = r - (g + b) / 2$, $G = g - (r + b) / 2$, $R = b - (r + g) / 2$ and $Y = (r + g) / 2 - | r - g | / 2 - b$. If the result is negative, the value is taken as 0. A Gaussian pyramid is then used in filtering. The equations for the color features are:

$$RG(c, s) = | R(c) - G(c) | \ominus | G(s) - R(s) |$$

$$BY(c, s) = | B(c) - Y(c) | \ominus | Y(s) - B(s) | \quad (1)$$

where c represents the center scale; $c \in \{2, 3, 4\}$, namely, scales 2, 3 and 4 from among the 9 scales; s represents the surround scale; $c \in \{3, 4\}$, namely, scales 3 and 4 out of the 9 scales; and $\ominus$ represents the point-by-point subtraction between the two feature maps based on the Difference of Gaussians (DoG) algorithm:

$$G(x, y) = \frac{1}{2\pi\sigma_1^2} e^{-\left(\frac{x^2+y^2}{2\sigma_1^2}\right)} - \frac{1}{2\pi\sigma_2^2} e^{-\left(\frac{x^2+y^2}{2\sigma_2^2}\right)} \quad (2)$$

where $\sigma_1 < \sigma_2$, $G(x, y)$ is the 2D ON DoG operator; $\sigma_1 > \sigma_2$; $G(x, y)$ is the 2D OFF DoG operator; σ_1 controls the center (fovea) of the sensitive area; and σ_2 manages the surround (i.e. periphery) of the sensitive area.

Twelve feature maps can be obtained for the color channel. Intensity information I can be derived from the following: $I = (r + g + b) / 3$. Thus, the intensity features $I(c, s)$ of the input image can be computed using

$$I(c, s) = | I(c) \ominus I(s) | \quad (3)$$

Six feature maps can be generated for the intensity channel.

The extraction of textural features differs from those of the color and intensity features. The model involves nine-scale wavelet decomposition using the Mallat wavelet transform algorithm. The three high-frequency components of wavelet decomposition comprise the texture pyramid. Therefore, the textural features $T(c, s, \theta)$ can then be obtained according to multi-scale difference:

$$T(c, s, \theta) = | T(c, \theta) | \ominus | T(s, \theta) | \quad (4)$$

where $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ represents four different orientations.

Twenty-four feature maps can be obtained for the texture channel.

Then three-channel saliency maps can be generated by normalization.

$$\tilde{C} = \bigoplus_{C=2}^{4} \bigoplus_{s=3}^{4} N(C(c, s)) \quad (5)$$

$$\tilde{I} = \bigoplus_{C=2}^{4} \bigoplus_{s=3}^{4} N(I(c, s)) \quad (6)$$

$$\tilde{T} = \bigoplus_{C=2}^{4} \bigoplus_{s=3}^{4} N(T(c, s, \theta)) \quad (7)$$

where $N(.)$ is a normalized function; "$\oplus$" represents the point-by-point addition between two feature maps; and $\tilde{C}$, $\tilde{I}$ and $\tilde{T}$ denote the color, intensity, and texture saliency maps, respectively.

2.2 Generation of the Motion Saliency Map

The motion features between the current and previous frame images are extracted [19]. The different scales images are filtered between the sequentially adjacent frame images. I^σ denotes the scale σ of the image, where $\sigma \in \{0, 1, 2, 3, 4\}$. If the current image is $I_t(x, y)$, then the previous image is $I_{t-1}(x, y)$. Each image scale can be generated by difference iteration from the previous image. Equation (8) is expressed as follows:

$$I^\sigma(x, y) = \frac{1}{4} I^{\sigma-1}(2x, 2y) + \frac{1}{8} \Big(\sum_{\gamma \in \{1, -1\}} I^{\sigma-1}(2x + \gamma, 2y) + \sum_{\gamma \in \{1, -1\}} I^{\sigma-1}(2x, 2y + \gamma) \Big) + \frac{1}{16} \sum_{\gamma \in \{1, -1\}} I^{\sigma-1}(2x + \gamma, 2y + \gamma) \quad (8)$$

where σ is the pyramid scale and the value of γ is either 1 or -1.

Equation (9) shows the range of x and y, where $0 \le 2x \le w_k^{\sigma-1}$ and $0 \le 2y \le h_k^{\sigma-1}$. Therefore, the width W_k^σ and height h_k^σ of I^σ are the maximum integers and meet the two conditions given by

$$W_k^\sigma \le \frac{W_k^{\sigma-1} + 1}{2} \quad (9)$$

$$h_k^\sigma \le \frac{h_k^{\sigma-1} + 1}{2} \quad (10)$$

The template of filter M is computed using (10):

$$M = m \times m^T$$

where $m = \{m_0, m_1, ..., m_{n-1}\}$ is an n-dimension vector and

$m_k = m_{n-k-1}$ meets

$$\sum_{k=0}^{n-1} m_k = 1 \quad (11)$$

The motion feature map $M_t(c, s)$ is written as (12):

$$M_t(c, s) = | I_t^c \ominus I_{t-1}^s | \quad (12)$$

where $c, s \in \{0, 1, 2, 3, 4\}$. "$\ominus$" represents the point-by-point subtraction of the different scales of the sequence of images.

The motion saliency map $\tilde{M}$ is expressed as (13):

$$\tilde{M} = \bigoplus_{c=0}^{4} \bigoplus_{s=0}^{4} N(M_t(c, s)) \quad (13)$$

where "$\oplus$" represents the point-by-point addition between two feature images. Finally, the static and the motion saliency maps are merged into the global saliency map. Equation (14) is written as follows:

$$\tilde{S} = \alpha\tilde{C} + \beta\tilde{I} + \gamma\tilde{T} + \lambda\tilde{M} \quad (14)$$

where α, β, γ and λ are the weight coefficients that meet $\alpha + \beta + \gamma + \lambda = 1$.

3. Video Image Compression Algorithm Based on Visually Salient Features

The ROI of a given video image can be obtained, based on the proposed visual attention model incorporating motion features. In general, the ROI of the image should have only a low compression ratio or not be compressed at all while the background region of the image will be subjected to the highest compression ratio [20]. Thus, the ROI influences the extent of image compression [21], and so the method used to determine the ROI is critical. In the proposed model, the ROI can be located from the global saliency map. The focus of attention is the ROI position that displays maximum saliency [22].

Image compression is conducted in blocks; so, the salient value of each image block is calculated using a process known as block saliency. First, the normalized saliency map is interpolated to make it the same size as the reconstructed image. Assume that the saliency map is divided into several similarly-sized blocks, B_i stands for the current block i. Now, the block saliency value of B_i can be defined as (15):

$$\varphi = \sum_{j \in B_i} S_j / N \quad (15)$$

where φ is the block saliency value of B_i. N denotes the total number of pixels in the block and S_j represents the saliency at position j of the saliency map.

Once the saliency map has been normalized, the gray values of all its pixels fall within the range of $0 - 255$ ($0 \le \varphi \le 255$).

The video image compression algorithm is then applied as follows:

(1) S is calculated given the motion features based on our proposed visual attention model with integrated motion features. S is the global saliency map of the current image in the input video;

(2) φ is computed. φ is the salient value of each image block B_i;

(3) The image block that contains the ROI of the global saliency map S is determined. The ROI is presumably located in block i, that is, B_i. If the block's saliency value φ meets:

$\varphi / \sum_{(x,y)} S_I(x, y) \ge \lambda$, (5) is applied,

where $\sum S_I(x, y)$ is the sum of the salient values of the saliency map and λ is the given threshold.

(4) The ROI is transferred using the inhibition of return and the WTA mechanism. The ROI is then detected for the next block B_i. If this block saliency φ meets:

$\varphi / \sum_{(x,y)} S_I(x, y) \ge \lambda$, (5) is utilized; otherwise, (4) is applied.

(5) The search for the ROI is concluded;

(6) The JPEG2000 algorithm is used to encode the ROI at either no compression or at only low compression ratio, and to compress the area surrounding the ROI with a high compression ratio. The areas with low image saliency can be compressed at higher compression ratios depending on their different saliencies.

4. Experimental results

The proposed algorithm was simulated using MATLAB to verify its accuracy and validity. Ten groups of video image sequences (30 images per group) were selected for the experiment. All test video image sequences were obtained from website http://media.xiph.org/video/derf/. Three groups of bitmap image sequences were selected at random and all achieved satisfactory results. Figures. 2 and 3 and Tables 1 and 2 depict the experimental results.

Figure 2 shows the multi-channel saliency map based on our proposed visual attention model incorporating motion features. Each channel's saliency map was obtained and the four features weighted and normalized, allowing the ROI of the original image to be determined. Figure 2(a) illustrates the original experimental image, along with the three selected groups of image sequences. Figures 2 (b)—2 (e) display the color, intensity, texture, and motion saliency maps respectively. The static features of the scene are effectively suppressed, as shown in Figure 2 (e). The part of the image that displays motion features is effectively highlighted to make the moving target that exhibits these features stand out.

Figure 2. Partial results of the multi-channel saliency map

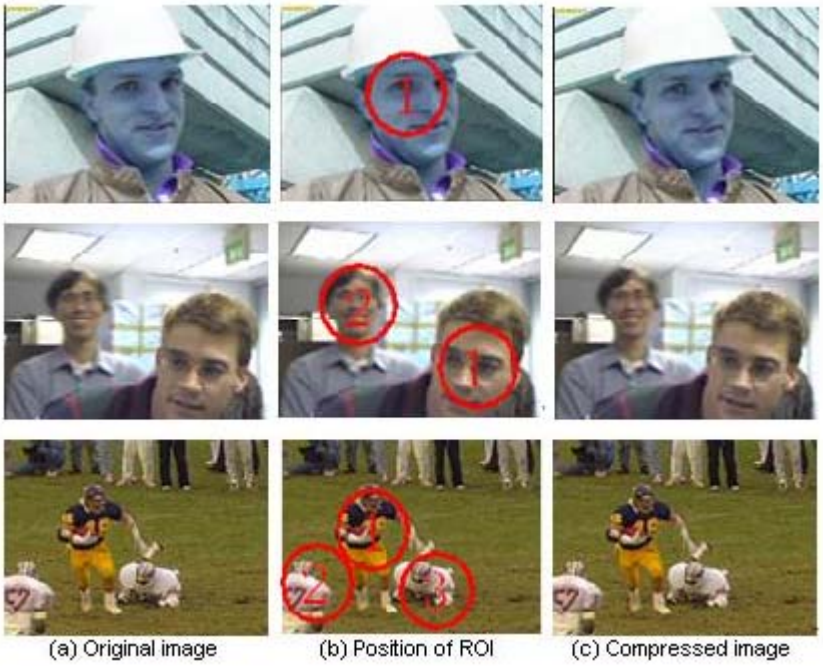

Figure 3. Component results of video image compression

Table 1 presents some of the experimental results from the averaged multi-channel weighting. According to the literature [23], each ROI has greater saliency in its central region. Once the features of each image have been extracted, the saliency map of eye movement can be generated. The similarity between the ROI of eye movement and the ROI calculated by the proposed algorithm can be obtained. This simulation thus aims to evaluate the accuracy and validity of the proposed algorithm. The four feature maps (color map, intensity map, textural map, and motion map) are assigned different weights whose sum is 1. The saliency maps of these four features are then superimposed to generate the global saliency map S. The global saliency map is normalized from 0 to 255. The ROI is then extracted. In addition, the four weightings of the three groups of image sequences (30 images per group) is used to denote the average weighting of each group of images. Table 1 indicates the average weighting of each channel and shows that this corresponds to the greatest similarity between the ROI of eye movement and the ROI obtained using the proposed algorithm.

Table 1 also demonstrates the differing influence of original features on the ROI as well as showing that, although the same original features can influence different kinds of image, their degrees of influence can vary. Figures that display clear motion features have the maximum weighting in the three groups of image sequences. In the third group

Video category	Color weighting	Intensity weighting	Texture weighting	Motion weighting
Foreman	0.20	0.13	0.09	0.58
Laboratory	0.11	0.21	0.03	0.65
Football	0	0	0	1.0

Table 1. Averages of the multi-channel weightings

Result / Algorithm	Test video		
	Foreman	Laboratory	Football
Image size	192 × 144 bitmap82944 bytes	192 × 144 bitmap82944 bytes	192 × 144 bitmap82944 bytes
JPEG algorithm	0.739 bpp7657 bytes	0.824 bpp8544 bytes	0.944 bpp9792 bytes
Proposed algorithm	0.581 bpp6024 bytes	0.635 bpp6578 bytes	0.719 bpp7456 bytes

Table 2. Comparison of the proposed algorithm and the JPEG algorithm

of image sequences, the three football players are running, and so their motion features are very obvious. Thus, the weighting of this motion feature channel is 1, whereas the weights of the other three feature channels are 0. Thus, here, the ROI of the image can be described by the motion feature channel alone, and hence, the proposed algorithm reduces the amount of calculation required.

Figure 3 shows the partial results of video image compression. Figure 3 (a) depicts the original experimental image, along with the three groups of selected video image sequences. Figure 3 (b) indicates the ROI image, with the position of the ROI being marked with a red circle and where the numerical values indicate the saliency of the ROIs, and accordingly, where different compression ratios should be used. In general, a low compression ratio, or no compression at all, should be applied to the ROI. The region of lower-interest (ROLI, which surrounds the ROI) exhibits the highest compression ratio; in the experiment, while the proposed video image compression algorithm performs no compression on the ROI, the other regions used a compression ratio of 15:1. Figure 3 (c) presents the results of this image compression.

Table 2 compares the two algorithms. The compression rate of the proposed algorithm is higher than that of the JPEG algorithm. Furthermore, the bits per pixel (bpp) and bytes of the compressed image are better in the proposed algorithm than in the JPEG algorithm. Using the proposed algorithm, the ROI also displays high fidelity, whereas the background has low fidelity. The peak signal-to-noise ratio (PSNR) of the compressed video is higher when obtained with the proposed algorithm than that calculated using the JPEG algorithm. The proposed algorithm presents a compromise with respect to ROLI because of its high compression rates; nonetheless, the visual effect of the image is improved as a whole, relative to that processed with the JPEG algorithm.

5. Conclusion

In this study, we have proposed a visual attention model saliency-based visual attention for rapid scene analysis. that combines motion features along with a video image compression algorithm. The proposed video image compression algorithm is advantageous with respect to the video's motion features. The proposed visual attention model can swiftly determine the ROI, and the proposed video image compression algorithm has a high compression efficiency while still optimizing the important visual effects. Future research should determine the optimum relationship between video quality and the strength of compression to be used in order to improve the final PSNR of the compressed video.

6. Acknowledgement

This work is supported in part by: (1) A Project Supported by the Scientific Research Fund of Jilin Provincial Education (20120268). (2) A Project Supported by the Scientific and Technological Planning Project of Jilin Province (20120332). (3) A Project Supported by the Scientific and Technological Planning Project of Jilin Province (20100565). (4) A Project Supported by the Scientific Research Fund of Jilin Provincial Education (2013434).

References

[1] Lee, S. H., Moon, J., Lee, M. (2006, July). A region of interest based image segmentation method using a biologically motivated selective attention model. In Neural Networks. IJCNN'06. *International Joint Conference on* (p. 1413-1420). IEEE.

[2] Huang, C., Liu, Q., Yu, S. (2011). Regions of interest extraction from color image based on visual saliency. *The Journal of Supercomputing*, 58 (1) 20-33.

[3] Yan, L., Yu, Z., Han, N., Liu, J. (2013). Improved Image Fusion Algorithm for Detecting Obstacles in Forests. *Journal of Digital Information Management*, 11 (5).

[4] Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. Image Processing, IEEE Transactions on, 13 (10) 1304-1318.

[5] Liu, L., Fan, G. (2003). A new JPEG2000 region-of-interest image coding method: partial significant bitplanes shift. *Signal Processing Letters, IEEE,* 10 (2) 35-38.

[6] Schmid, C., Mohr, R., Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, 37 (2) 151-172.

[7] Siagian, C., Itti, L. (2007, October). Biologically-inspired robotics vision monte-carlo localization in the outdoor environment. In Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on (p. 1723-1730). IEEE.

[8] Hou, X., Zhang, L. (2007, June). Saliency detection: A spectral residual approach. In Computer Vision and Pattern Recognition. CVPR'07. IEEE Conference on (p. 1-8). IEEE.

[9] Andreopoulos, Y., van der Schaar, M., Munteanu, A., Barbarien, J., Schelkens, P., Cornelis, J. (2003, April). Fully-scalable wavelet video coding using in-band motion compensated temporal filtering. In Acoustics, Speech, and Signal Processing. *In*: Proceedings. (ICASSP'03). 2003 IEEE International Conference on (3, p. III-417). IEEE.

[10] Le Meur, O., Le Callet, P., Barba, D., Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28 (5) 802-817.

[11] Hu, Y., Rajan, D., Chia, L. T. (2005, July). Adaptive local context suppression of multiple cues for salient visual attention detection. InMultimedia and Expo. ICME. IEEE International Conference on (p. 4-pp). IEEE.

[12] Cotronei, M., Lazzaro, D., Montefusco, L. B., Puccio, L. (1999). Image compression through embedded multiwavelet transform coding. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 9 (2) 184-189.

[13] Ho-Phuoc, T., Guyader, N., Guérin-Dugué, A. (2010). A functional and statistical bottom-up saliency model to reveal the relative contributions of low-level visual guiding factors. *Cognitive Computation*, 2 (4) 344-359.

[14] Amerijckx, C., Verleysen, M., Thissen, P., Legat, J. D. (1998). Image compression by self-organized Kohonen map. *IEEE Transactions on neural networks*, 9 (3) 503-507.

[15] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints.*International journal of computer vision*, 60 (2) 91-110.

[16] Baluch, F., Itti, L. (2010). Training top-down attention improves performance on a triple-conjunction search task. *PloS one*, 5 (2) e9127.

[17] Wang, H., Dang, Y., Ke, H., Liu, X. (2013). A Method of Target Region Detection Based on Multi-channel Weighted Visual Attention. *Journal of Computers*, 8 (10) 2478-2482.

[18] Itti, L., Koch, C., Niebur, E. (1998). A model of IEEE Transactions on pattern analysis and machine intelligence, 20 (11) 1254-1259.

[19] Ke, H. C., Wang, H., Li, H. Y. (2011). An Attention Target Detection Method Based on Dynamic Saliency Map. *Advanced Materials Research*, 308, 574-578.

[20] Parkhurst, D., Law, K., Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research,* 42 (1) 107-123.

[21] Ryu, G. G., Suh, I. H., Lee, S. (2009). Covert Visual Attention by Object-based Selective Visual Features and Their Saliency Map. In IPCV (p. 170-173).

[22] Ke, H., Wang, H., Zhao, H., Liang, K. (2010). Visual mental imagery memory model based on weighted directed graph. *Journal of Computers*, 5 (8) 1256-1263.

[23] Jost, T., Ouerhani, N., Wartburg, R. V., Müri, R., Hügli, H. (2005). Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100 (1) 107-123