Saliency Detection Making Use of Human Visual Perception Modelling

Cristina Oprea, Constantin Paleologu, Ionut Pirnog, and Mihnea Udrea
Dept. of Telecommunications
Politehnica University of Bucharest
Bucharest, Romania
cristina@comm.pub.ro, pale@comm.pub.ro, ionut@comm.pub.ro, mihnea@comm.pub.ro

***Abstract*—This paper proposes an algorithm for accurate detection of salient areas from a given scene. We used a complex model for the human visual system, in order to simulate the visual perception mechanisms. Human visual system modelling requires accurate knowledge about the entire visual pathways. This work focuses on the following features of the human vision: the color perception mechanism, the perceptual decomposition of visual information in multiple processing channels, contrast sensitivity, pattern masking, and detection/pooling mechanism present in the primary visual cortex. Pattern masking is considered within a complex approach, combining data from distinct dimensions. The results are shown to correlate well with the subjective results obtained from an eye-tracking experiment.**

Keywords – human visual system; saliency map; visual perception; masking; perceptual decomposition.

I. INTRODUCTION

This paper is focused on a region identification question and the regions that we are looking for are the ones having the best saliency from the perceptual point of view, as presented in [1]. The main idea is to be able to decide wich are the most important areas in a given scene, image or video frame. Such an algorithm has several applications, some of the most important beeing in the video preprocessing stage (coding), in order to optimize the compression scheme and in watermarking schemes that should hide information more effectively in images. In such applications, the characteristics and especially the limitations of the human visual system can be exploited to obtain the best performance with respect to visual quality of the output. Physiologists and psychologists have performed psycho-visual experiments aiming to understand how the human visual system (HVS) works. Engineers apply the results of those psychovisual experiments in their applications, but to do so, they use the simplified human vision models. This paper presents an attempt to integrate a such computational model of the human visual system into a tool for perceptual important areas detection. However, the experimental conditions used in the psycho-visual experiments are not representative for all types of image processing applications. Using the simplified human vision models with little knowledge regarding the applicability of these models under the new conditions limits the precision of the results.

The computational model of the human visual system that we have used is a model derived from the one introduced by [2]. This model is based on the multi-channel architecture, as first proposed by Watson in [3] who assumed that each band of spatial frequencies is dealt with by a separate channel. The contrast sensitivity function (CSF) is the envelope of the sentivities of those channels. The detection process occurs independently in any channel when the signal in that band reaches a threshold. In addition, several models proposed later, including [4] and the present work, take into consideration temporal channels as well as chromatic sensivities and orientation selectivity. The perceptual decomposition in multiple channels is then performed in both domains, spatial and temporal. The temporal channels will deal with the dynamic stimuli from the visual scene.

The paper is structured in four sections. Section II introduces the latest achievements in the area of perceptual region detection and human visual system modelling. Section III contains a detailed presentation of the proposed method, while in the final section of this paper we show that the results obtained with this algorithm are a good approximation for the perceptual regions detected with subjective experimental testing.

II. PREVIOUS WORK

Previous work related to this approach was mainly conducted in the field of visual attention modelling. Although visual assessment task in humans seems simple, it actually involves a collection of very complex mechanisms that are not completely understood. The visual attention process can be reduced at two physiological mechanisms that combined together result in the usual selection of perceptual significant areas from a natural or artificial scene. Those mechanisms are bottom-up attentional selection and top-down attentional selection. The first mechanism is an automated selection performed very fast, being driven by the visual stimulus itself. The second one is started in the higher cognitive areas of the brain and it is driven by the individual preferences and interests. A complete simulation of both mechanisms can result in a tremendously complex and time-consuming algorithm.

The process of finding the focus of attention in a scene is usually done by building feature maps for that scene, following the feature integration theory developed by Treisman [5]. This theory states that distinct features in a scene are automatically registered by the visual system and

coded in parallel channels, before the items in the image are actually identified by the observer. Independent features like orientation, color, spatial frequency, brightness, and motion direction are put together in order to construct a single object being in the focus of attention. Pixel-based, spatial frequency and region-based models of visual attention are different methods of building feature maps and extracting saliency.

The pixel-based category is represented by Laurent Itti's work concerning the emulation of bottom-up and top-down attentional mechanisms [6]. Another possibility of building feature maps is by applying different filtering operations in the frequency domain. Most common type of such filtering is done using Gabor filters and Difference of Gaussians filters. The work in [7] applies the opponent color theory and uses contrast sensitivity functions for high contrast detection. Last category of visual attention models are the region-based algorithms. In this case it is usually performed a clustering operation like region segmentation on the original image and then feature maps are computed using these clusters [8].

Regarding the HVS modelling, there have been studied and evaluated by the Video Quality Experts Group (VQEG) several video quality metrics that are using such models for the visual system. Based on a benchmark by the VQEG in the course of the Multimedia Test Phase 2007-2008, some metrics were recently standardized as ITU-T Rec. J.246 [9] and J.247 [10]. The first recommandation, J.246 presents several methods for perceptual visual quality assessment for cable television. Such networks have the advantage of permitting the transmission of some information about the reference or even a reduced bandwidth reference.

The second recommandation J.247 states a new set of methods dedicated to perceptual video quality measurement when the entire reference is available. One of those methods, PEVQ or Perceptual Evaluation of Video Quality performs a pre-processing step that extracts a region of interest from the reference and the distorted signals. All the following calculations are then performed only on that region of interest. This step is based on the observation that distortions nearest to the border are not really noticed by viewers and often get ignored. This idea can be developped into a more precise analysis and one can identify a region of interest that best fits the perceptual saliency. All further calculations can be performed for that specific area found to be closest to the human focus of attention, consuming less time and resources in the application under consideration.

III. PROPOSED ALGORITHM

Human visual system modelling requires accurate knowledge about the entire visual pathways. In present, only certain aspects of vision are well understood and so, human visual system models have been developed in order to simplify the bahaviours of what is a very complex system. As the knowledge about the real visual system improves, the model can be upgraded. Such models are used by experts and researchers in image processing, video processing and computer vision, dealing with applications related to biological and pshychological processes.

Many HVS features have their origins in evolution, since people needed to hunt for food and defend from other predators. For exemple, motion sensitivity is higher in peripheral vision with the purpose of early detection of any danger coming from wild animals. Also, motion sensivity is stronger then texture sensivity since it was crucial to scan the landscape and detect any camouflaged animals.

The model used in this work focuses on the following features of the human vision: the color perception mechanism, the perceptual decomposition of visual information in multiple processing channels, contrast sensitivity, pattern masking, and detection/pooling mechanism. In the following presentation, each feature is integrated into an algorithm processing step. The main target is to obtain at the end of the algorithm a map indicating the most salient areas from a scene. Perceptual saliency detection stands for the identification of objects, persons, visual stimuli in general that have the quality of standing out relative to neighboring items or simply beeing eye-catching. This task is similar to finding the focus of attention, that means recreating the mind's perceptual function to direct its inner awareness upon a specific target.

A. Color processing

The chromatic information from the visual scene is processed in the retinal stage according to the trichromatic theory. In the following stages of the visual pathway, specifically in the lateral geniculate nucleus, the color data is encoded according to the opponent colors theory, a technique that removes redundancy from the data stream.

At the first stage of color perception in the retina, photoreceptor cells convert the light energy into neural signals. The basic process performed by photoreceptors is absorption of photons from the field of view and signalling this information through a change in the membrane potential. This mechanism provides the subsequent cortical areas with the necessary information about the scene comprised in the field of view. There are two types of photoreceptor cells: rods and cones, and they have different functions. Rods are found primarily in the periphery of the retina and are used to see at low levels of light. Rods are not sensitive to color, only to light/dark or to black/white. Rods can function in less intense light than can the other type of photoreceptors, cone cells, and they are concentrated at the outer edges of the retina beeing used in peripheral vision. Cones are located especially in the center of the retina. There are three types of cones that differ in the wavelengths of light they absorb; they are usually called short or blue (S), middle or green (M), and long or red (L). Cones are used to distinguish color at normal levels of light.

In later stages of visual information processing, the color is to be coded differently. From the three primaries given by cones and the intensity given by rods, the color is eventually encoded as one luminance channel (magnocellular cells from the lateral geniculate nucleus - LGN) and two crominance

channels: one for red-green cones (parvocellular cells in LGN) and another one for blue-yellow connes (koniocellular cells in LGN).

The color processing block in our algorithm is conducting a conversion from the usual YCbCr color-space to an opponent color space, similar to the one discovered at the LGN level. The resulting color components are: W-B for white-black, R-G for red-green, and B-Y for blue-yellow. These opponent colors can be associated to a luminance signal and two color difference signals. The colors selected are not random, they are considered opponent because under normal circumstances there is no hue one could describe as a mixture of opponent hues [11].

In order to obtain those components, the trichromatic values (RGB computed from YCbCr) undergo a power-law nonlinearity to counter the gamma correction used to compensate for the behaviour of conventional CRT displays. In LCD displays, the relation between the signal voltage and the intensity is very nonlinear and cannot be described with gamma value. However, such displays apply a correction onto the signal voltage in order to approximately get a standard γ=2.5 behavior.

The linear RGB values produced are then converted to responses of the L, M and S cones on the human retina, based on the spectral absorbtion measured for these cells. This conversion is performed in two steps: first, RGB color space is converted to CIE XYZ color space; second, from XYZ components will be computed the LMS values. For the first transformation we have used a matrix defined in ITU-R Rec. BT.709-5 [15]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412 & 0.358 & 0.180 \\ 0.213 & 0.715 & 0.072 \\ 0.019 & 0.119 & 0.950 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

The responses of the L, M, and S cones from the human retina are computed according to CIECAM02, the most recent color appearance model ratified by CIE Technical Committee (International Commission on Illumination) [16]:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.7328 & 0.4296 & -0.1624 \\ -0.7036 & 1.6975 & 0.0061 \\ 0.0030 & 0.0136 & 0.9834 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (2)$$

There is no unanimity of opinion regarding the particular values of the coefficients used in those transformations and several papers still use the classical von Kries transformation. We prefered the transformation matrix proposed in the latest standard published by ITU-R since it comes from more recent studies and measurements.

Knowing the L, M, S cones absorbtions rates, the convertion to an opponent color space becomes possible due to the transformation matrix proposed by Poirson and Wandell [16]. The same transformation matrix has also been used by Winkler in his Perceptual Distortions Metric [2]:

$$\begin{bmatrix} W-B \\ R-G \\ B-Y \end{bmatrix} = \begin{bmatrix} 0.990 & -0.106 & -0.094 \\ -0.669 & 0.742 & -0.027 \\ -0.212 & -0.354 & 0.911 \end{bmatrix} \cdot \begin{bmatrix} L \\ M \\ S \end{bmatrix} \quad (3)$$

The color space proposed by Poirson and Wandell was developed aiming to completely separate the color processing from the pattern perceptual processing. Keeping apart the color from the pattern makes easier to simulate the mechanisms in the human vision.

The opponent color space agrees with the color processing at higher levels in the human brain, especially in the cortical area called V1. This type of color encoding de-correlates the signals coming from the retina and removes redundancy. In fact, in area V1 it has been proven to exist two types of double-opponent cells: red-green and blue-yellow. Red-green cells confront the relative amounts of red-green in one part of a scene, with the amount of red-green in a neighboring part of the scene; such cells respond best to local color contrast (red next to green).

B. Multi-channel decomposition

The multi-channel decomposition is performed according to a theory that explains the visual perception of a scene including multiple visual stimuli: each feature from the input scene is processed separately. Many cells in the human visual system and mainly in the visual cortex have been proven to be selectively sensitive to certain types of signals such as patterns of a particular frequency or orientation.

The visual cortex is made from the combination of several areas: V1 (or primary visual cortex), V2, V3, V4, and V5. Neurons in the visual cortex respond to visual stimuli that appear within their receptive field by sending action potentials. The receptive field of one neuron is the region within the entire visual field which causes a response from that neuron. Each neuron responds best only to a subset of stimuli within its receptive field. This mechanism is neuronal tuning. First visual areas (for exemple V1 area) have neurons with simpler tuning that will respond to stimuli falling in their receptive fields such as vertical lines or textures with particular spatial frequencies. In later visual areas, neuronal cells have complex tuning that is much more complicated cu simulate. For instance, a neuron in the inferior temporal cortex may only react when a certain face appears in its receptive field.

During the experiments regarding the primary visual cortex, it has been noticed that the tuning properties of V1 neurons differ greatly over time. Evidence shows that there are at least two temporal mechanisms that affect neuronal responses in V1. The overall functioning of V1 can be thought of tiled sets of selective spatiotemporal filters. This is why the multi-channel decomposition splits the input into a number of channels, based on the spatio-temporal mechanisms present in area V1 from the visual cortex. In theory, these filters together can carry out neuronal processing of spatial frequency, orientation, motion, direction, speed (thus temporal frequency), and other spatiotemporal features.

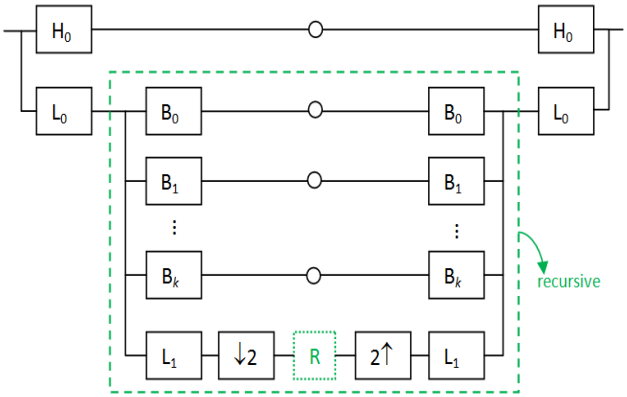


Figure 1. Simoncelli's steerable pyramid. Downsampling by a factor of 2 and upsampling by 2 are used. The recursive construction of the pyramid is achieved by inserting a copy of the diagram contents enclosed by the dashed rectangle at the location of the block "R".

Temporal mechanisms are modeled with a perceptual decomposition in the temporal domain. We used two filters for two temporal mechanisms: the sustained and transient mechanisms, the same filters used in [2] and proposed by Fredericksen and Hess [17]. Finite impulse response (FIR) filters with linear phase are computed by means of a least-squares fit to the normalized frequency magnitude response of the corresponding mechanism as given by the Fourier transforms of $h(t)$ and $h''(t)$, the second derivative of $h(t)$, from the following equation:

$$h(t) = exp[-(5\ln(t/\theta))^2] \quad (4)$$

The sustained mechanism is implemented by a low-pass filter, while the transient mechanism – by a band-pass filter. Both FIR filters are applied only to the luminance channel in order to reduce computing time. This simplification is based on the fact that our sensitivity to color contrast is reduced for high frequencies.

Spatial mechanisms are modeled by means of a steerable pyramid decomposition [12], first proposed by Simoncelli. In this linear decomposition, an image is subdivided into a collection of subbands localized in both scale and orientation. Similar multiscale transforms have often been used in image processing and image representation. For exemple, the wavelet transform was proven usefull in applications were scalable video coding was needed.

The scale tuning of the filters is constrained by a recursive system diagram, as illustrated in Fig. 1. The left part of the diagram is called the analysis filter bank, while at the right side, the synthesis filter bank performs the reconstruction of the original image. The orientation tuning is constrained by the property of steerability, which means that the transform is shiftable in orientation. A set of filters form a steerable basis if :

(i) they are rotated copies of each other and

(ii) a copy of the filter at any orientation may be computed as a linear combination of the basis filters.

The pyramid's algorithm itself is based on recursive application of two types of operations: filtering and subsampling. First, the input signal or the original image/frame is divided into a low-pass and a high-pass portions. The low part will be further subdivided into bandpass portions and another low-pass one; each of the bandpass filters select features having distinct orientations. The last low-pass portion obtained is subsampled by a factor of 2 and the algorithm will be repeated in recursive cascades. The bandpass divisions are not subsampled in order to avoid aliasing, while for the subsampled low-pass subimage, the aliasing issue is prevented by using low-pass radial filters especially designed.

In addition to having steerable orientation subbands, this transform can be designed to produce any number of orientation bands, k. The resulting transform will be overcomplete by a factor of $4k/3$, meaning that the coefficient output rate is greater then the input signal sample rate. Note that the steerable pyramid retains some of the advantages of orthonormal wavelet transforms, but improves on some of their disadvantages (e.g., aliasing is eliminated; steerable orientation decomposition). One obvious disadvantage is in computational efficiency: the steerable pyramid is substantially overcomplete.

Six sub-band levels with four orientation bands each plus one low-pass band are computed; the bands at each level are tuned to orientations of 0, 45, 90, and 135 degrees, as illustrated in Fig. 2. The same decomposition is used for the W-B, R-G and B-Y channels, meaning that all color channels go through the same steerable pyramid transform. This approach agrees with the primary visual cortex architecture regarding color, spatial frequency, and orientation processing.

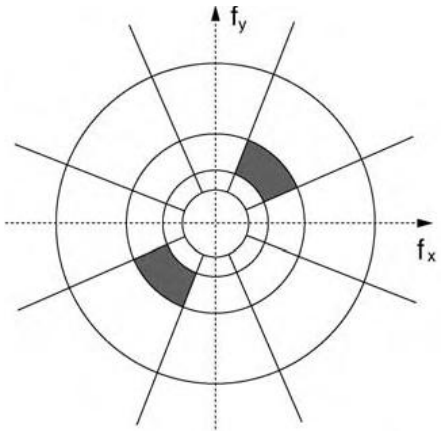
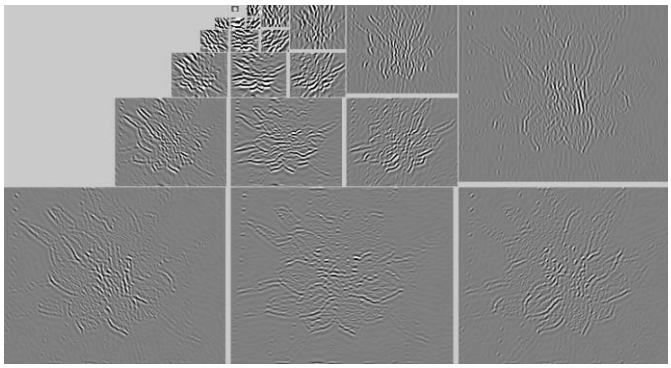


Figure 2. Spatial frequency plane partitioning in the steerable pyramid transform. The gray region indicates the spectral support of a single sub-band oriented at 45 degrees [2].

At this point of the algorithm, the input image is subjected to the steerable pyramid transform and the result is illustrated in Fig. 3 for a set of two images. The first image illustrates a flower who's petals have radial disposition, thus containing all orientations. In the output of

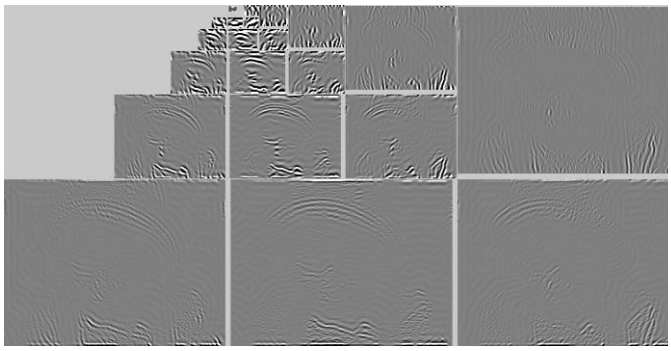
the steerable pyramid transform, at the first decomposition level it is easy to recognize the extracted feature's orientations.

a)



b)

c)



d)

Figure 3. a) and c) are original test images; b) and d) are the outputs from the steerable pyramid transform (four orientations and five levels decomposition).

C. Contrast sensitivity

Contrast is a visual property that makes an object distinguishable from neighboring elements or background. The human visual system is more sensitive to contrast than to absolute luminance and the human eye itself is designed to react only to luminance variations.

Researchers built contrast sensitivity functions from experimental measurements and beside the classic luminance contrast given by white/dark association, we now have color contrast sensitivity curves for the two chromatic channels: red-green and blue-yelow [18], [19] as it can be seen in Figure 4. The contrast sensitivity function shows a typical band-pass shape peaking at around 4 cycles per degree with sensitivity dropping off either side of the peak, meaning that human vision is most sensitive in detecting contrast differences occurring at 4 cycles per degree. The high-frequency cut-off represents the optical limitations of the visual system's ability to resolve detail and is typically about 60 cycles per degree. Tipically, our sensitivity to color contrast is reduced for high frequencies.

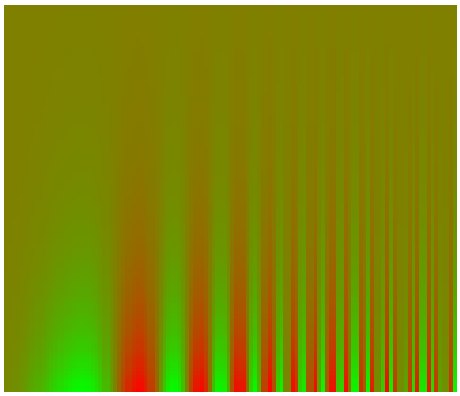


Figure 4. The contrast sensitivity function for the red-green channel is the envelope of the visible gratings.

Next step after the temporal and spatial decomposition is a shortcut in computation efficiency. Instead of pre-filtering the W-B, R-G and B-Y channels with their respective contrast sensitivity functions, which is the accurate approach, we searched for a set of weighting factors for each channel. The weights were determined intending to obtain a filter set that approximates the spatio-temporal contrast sensitivity of the human visual system. It was preferable since it conducts to a more simple implementation and the simulation time improves.

D. Masking

Visual masking is a perceptual phenomena that stands for the reduction or the elimination of the visibility of one brief stimulus called the "target" due to the presentation of a second brief stimulus, called the "mask". Within the framework of quality assessment it is helpful to think of the distortion or the coding noise as being masked by the original image or sequence acting as background. Masking

explains why similar coding artifacts are disturbing in certain regions of an image while they are hardly noticeable in others. In order to be possible for visual masking to appear, both the target and the mask must be briefly presented, less then 50ms.

Our human visual system model implements both intra-channel and inter-channel masking. Masking is known to be stronger between visual stimuli of the same type (located in the same decomposition channel), so called intra-channel masking. This type of visual masking appears for a pair target-mask that have the same characteristics: belong to the same frequency band, same orientation, and even identical chromaticity. But masking also happens, at a lesser extent, between stimuli coming from different channels, beeing called inter-channel masking. We approached the masking perceptual process as a question of multiple excitations and inhibitions flows in the cortical pathways. For a neuron's excitation stronger than the associated inhibition from other neurons, we obtain the evidentiation. The opposite phenomenon, an excitation weaker than corresponding inhibitions will emulate the perceptual masking. An accurate modelling of evidentiation and masking operations will bring foreward salient features and objects from the input image.

In neural networks, the neuron's excitation or inhibition can be simulated with the following linear equation:

$$y_j^N = \sum_i g\left(w_{ji} x_i + c_j\right) \tag{5}$$

where the output or the response of the *j*-th neuron is given by all inputs to that neuron, indexed by *i* and denoted as *x*, weighted by the coefficients *w* according to that neuron's specialization. Excitation appears for a positive weight, while inhibition follows a negative weight.

Our model takes into consideration the excitatory behaviour of specialised neurons inhibited by a pool of responses from other nervous cells in the visual cortex. Instead of a linear model, we adopted a nonlinear model where the weights were removed, thus eliminating the problem of choosing their values. The excitation is modeled by a power-law nonlinearity, where the input *x* is raised at power *p*. The inhibition follows the same modelling rule having an exponent *q*.

$$y = \frac{x^p}{c + h_1 * x^q + h_2 * x^q} \tag{6}$$

Equation 6 illustrates that the excitatory behaviour can be modelled by means of a power-law nonlinearity with exponent *p* greater then the inhibitory exponent *q*. The numerator models the excitation and *x* is a coefficient from the perceptual multi-channel decomposition. Such a coefficient comes as output from one of the filters in the filter bank that comprises the steerable pyramid. Therefore, *x* is a coefficient that carries information about a feature in the input image having precise characteristics: spatial frequency, color, and orientation. The denominator contains a constant *c* that prevents division by zero and two convolutions: h_1 represents a gaussian pooling kernel for coefficients from the same decomposition channel, while h_2 is another gaussian pooling kernel for different channels interactions. This approach has proven to be more accurate then using a single pooling kernel for all coefficients. In the inhibitory path, filter responses are pooled over different channels by means of two convolutions, combining coefficients from the dimensions of space and orientation.

E. Detection

The information coded in multiple channels within the VI area of the visual cortex is integrated in the subsequent cortical areas. This process can be simulated by gathering the data from these channels according to rules of probability or vector summation, also known as pooling.

Then, the steerable pyramid is reconstructed only for the luminance channel.

a) b) c)

Figure 4. a) Original image "Lena"; b) Saliency map obtained with our algorithm; c) Importance map obtained with TBQM metric [13]. The brighter pixels have higher saliency/ perceptual importance.

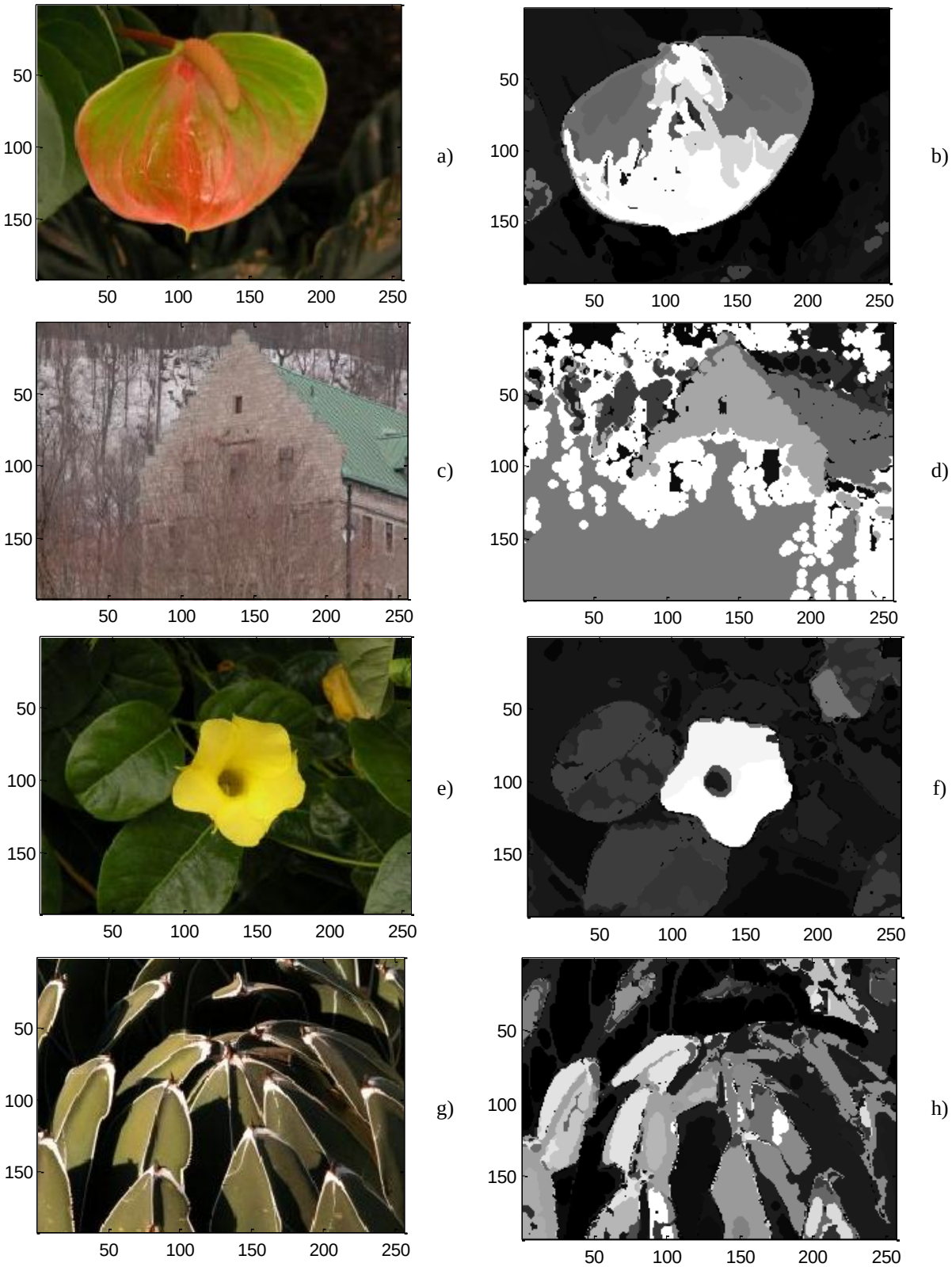


Figure 5. a), c), e) and g) are original test images from the eye-tracking experiment database [14]; b), d), f) and h) are saliency maps obtained with the algorithm described previously. The brighter areas have a stronger perceptual importance, while the dark zones designate features without saliency.

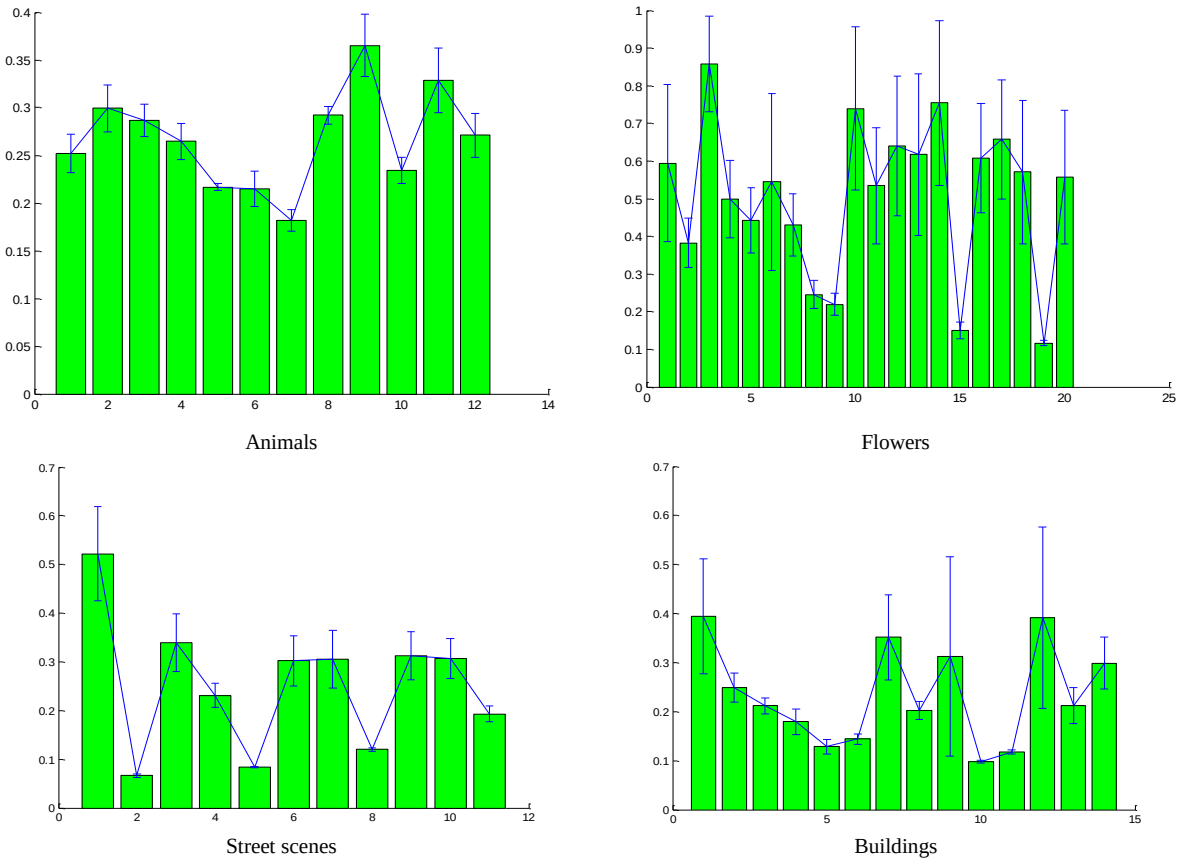


Figure 6. Correlations between human fixation maps from the eye-tracking experiment and objective saliency maps.

Our model of the human visual system uses at this stage a thresholding operation for the set of coefficients *y* resulted in the previous processing stage. Threshold values are model constants determined by experimental simulations, taking into consideration the saliency related to multiscale representation. The chromatic information was already included by the inter-channel masking processing. The resulting image pixel values are brought to a subunitary domain and they represent the saliency contained in the original image.

IV. RESULTS

For comparison of our results with one of the existing models, we provided the output of saliency maps by [13] in Figure 4. The model proposed in this paper is inclined towards totally eliminating the noninteresting areas because of being strict in selection. This tendency confines the exploration to a limited number of spots and the probability to skip a moderately prominent object in a visual search turns high. The method for importance map construction by [13] shown to be unable to discriminate the saliency of naturally prominent colors and also it does not consider the global context in the given scene. The proposed method has eliminated such weaknesses by incorporating theory of colors into the model and by including the influence of local and global neighborhood on the saliency of objects. In Figure 5 there are presented four test images and their objective saliency maps determined with our algorithm.

The saliency maps are also correlated with the subjective results obtained for a 29 test images database, containing eye-tracking data [14]. Such data are highly accurate due to the experimental setting and the testing subjects carrefully selected. Eye-tracking data result in the only subjective saliency maps that can be used for comparison with objective methods. In order to compare the saliency maps with the human data, we used a correlation method proposed in [14]. The value of comparison is given by the correlation coefficient ρ:

$$\rho = \frac{\sum_{x,y}(OM(x,y) - m_{OM})(SM(x,y) - m_{SM})}{\sqrt{\sigma_{OM}^2 \sigma_{SM}^2}} \quad (7)$$

where $OM(x,y)$ is the objective map, $SM(x,y)$ is the subjective map, and m, σ^2 are the mean and the variance of the values from these maps. A positive correlation coefficient indicates similar structure in both maps. Our objective maps result in correlation coefficients greater then those obtained with TBQM metric in 73% cases. In Figure 6 are illustrated the correlation coefficients for four

different types of images: natural scenes with animals, natural scenes with flowers, street scenes with peoples and cars and finally, building scenes. All the images come from the database provided by [14]. Each bar represent the mean correlation coefficient for the computed correlations between the 31 human fixation maps and our saliency map. The error bars give the confidence intervals.

V. CONCLUSIONS

The model proposed exploits spatiotemporal information and provides an efficient preprocessing step (salient spatiotemporal event detection) that will limit the application of high-level processing tasks to the most salient parts of the input. Our model simulates only the behaviour of the primary visual cortex (V1), which is necessary for conscious vision. As future work, the algorithm will be upgraded with emulations of the superior extrastriate visual cortex areas that will replace the final detection operation performed in the current work.

ACKNOWLEDGMENT

This work was supported by the UEFISCDI grant PN-II-RU-TE no. 7/5.08.2010.

REFERENCES

[1] C. Oprea, C. Paleologu, I. Pirnog, M. Udrea, "Saliency Detection Based on Human Perception of Visual Information" pp.96-99, 2010 Sixth Advanced International Conference on Telecommunications, 2010.

[2] S. Winkler, “Digital Video Quality Vision Models and Metrics”, John Wiley & Sons, 2005, ISBN 0-470-02404-6.

[3] A.B. Watson, “The cortex transform: Rapid computation of simulated neural images”, Computer Vision, Graphics, and Image Processing, 1987, 39,3, pp. 311-327.

[4] M. A. Masry and S. S. Hemami, “A metric for continuous quality evaluation of compressed video with severe distortions”, Signal Proc.: Image Communication, 2004, vol. 19, (2), pp. 133-146.

[5] A. M. Treisman and G. Gelade, “A feature-integration theory of attention”, Cogn. Psychol., vol. 12, pp. 97-136, 1980.

[6] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention”, IEEE Trans. Image Process., 2004, vol.13, (10), pp. 1304 – 1318.

[7] O.L. Meur, P.L. Callet, D. Barba, D. Thoreau, “A coherent computational approach to model bottom-up visual attention”, IEEE Trans.Pattern Anal. Mach. Intell., 2006, vol. 28, pp. 802-817.

[8] C. Oprea, I. Pirnog, C. Paleologu, M. Udrea, “Perceptual Video Quality Assessment Based on Salient Region Detection”, Proceedings of AICT 2009, May 2009, pp. 232-236.

[9] ITU-T Rec. J.246, “Perceptual audiovisual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference”, 2008.

[10] ITU-T Rec. J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference”, 2008.

[11] M. Nida-Rumelin, J. Suarez, "Reddish Green: A Challenge for Modal Claims About Phenomenal Structure". Philosophy and Phenomenological Research 78: 346. doi:10.1111/j.1933-1592.2009.00247.x, 2009.

[12] Q. Wu, M. A. Schulze, K. R. Castleman, “Steerable Pyramid Filters for Selective Image Enhancement Applications”, Proceedings of ISCAS ‘98, 1998.

[13] G. Zhai, W. Zhang, X. Yang, Y. Xu,” Image quality metric with an integrated bottom-up and top-down HVS approach”, IEE Proc.-Vis. Image Signal Process., Vol. 153, No. 4, August 2006.

[14] G. Kootstra, A. Nederveen, B. De Boer, “Paying attention to symmetry”, Proc. British Machine Vision Conference, UK, 2008.

[15] ITU-R Rec. BT.709-5, “Parameter values for the HDTV standards for production and international programme exchange”, ITU, Geneva, Swiss, 2002.

[16] A. B. Poirson, B. A. Wandell, “Pattern-color separable pathways predict sensitivity to simple colored patterns”, Vision Research vol. 36 (4), pp. 515-526, 1996.

[17] R. E. Fredericksen, R. F. Hess, “Estimating multiple temporal mechanisms in human vision”, Vision Research, vol. 38(7), pp.: 1023-1040, 1998.

[18] F. W. Campbell, J. G. Robson, “Application of Fourier analysis to the visibility of gratings”, Journal of Physiology, 1968.

[19] E. Peli, “Contrast in complex images”, Journal of the Optical Society of America, vol. 7 (10), pp.2032-2040, 1990.