LOCAL AND GLOBAL SPECTRAL VISUAL SALIENCY ESTIMATION BASED ON HUMAN VISUAL BEHAVIOR

Oana Loredana BUZATU
Technical University of "Gh. Asachi", Iassy, Romania
Bd-l Carol I, nr. 11, corp A, lbuzatu@etti.tuiasi.ro

Abstract: In this paper a saliency detection algorithm based on local and global information is proposed. The proposed method extends a previous work by calculating the pixels saliency by means of local patches and the global dissimilarities between a patch and all other patches based on color contrast differences and spatial distances. The image patch scale is fixed and in accordance with the size of the regions of interest (ROI), while the spatial difference takes into account the fact that the focus of attention can subtend a certain region from the visual angle. Resulted algorithm is tested in an image retargeting application and proves its effectiveness.

Keywords: *local- global visual saliency, human visual sensitivity, image patch size, image retargeting.*

I. INTRODUCTION

It is well known that humans can identify all most instantly and with high accuracy salient regions from the visual field before performing an actual recognition task. Such a behavior is due to the intrinsic attentional mechanisms activated at the brain level in the presence of salient stimuli. The importance of studying and understanding these cognitive perceptual mechanisms is due to the fact that they can fill the gaps between the human visual system (HVS) as a biological organism and the same HVS as a complex system capable to process imagistic signals by applying known mathematical operations in a natural way. Imitating the perceptual behavior of the HVS as a saliency detection algorithm proved to be useful in many image analysis and synthesis applications such as automatic image cropping [1], image retargeting [2], image abstraction [3], image/video compression [4, 5], advertising design [6], object recognition, tracking and detection as well [2, 7, 8].

The selective behavior of the visual attention refers basically to the interaction between bottom-up and top-down mechanisms. The bottom-up mechanisms are referred as a perception process which automatically identifies salient regions in images, while the top-down ones are related to high-level prior information, such as knowledge, personal experience, emotions or current task [6, 9, 10]. Another important aspect when studying visual attention is to describe how it operates. In [11] visual attention is thought to operate as a two-stage process: in a first stage the attention is uniformly distributed over the entire visual field, operating parallel processes, while in the second stage only the most salient areas are attended by sequential operations. Similar approaches can be identified in many computation models which exploit local and global information for saliency detection [2, 7, 12-15].

There are also theories which state that attention acts as a zoom lens [16] by means of a trade-off between the size of the focused region and the process efficacy. It is thought that the focus of attention can subtend at least 1 degree from the visual angle, but for higher areas needs to engage more processing resources. Size of the attended areas is another perspective which will be further in this paper discussed.

The proposed algorithm is a bottom-up approach which extends the method from [17], exploiting local information in a global fashion. In [17], after a short overview concerning the quaternion based saliency detection methods, an algorithm which takes into account the amplitude information of the quaternion Fourier Transform (QFT) applied to a Lab image- filtered based on Contrast Sensitivity Functions (CSF), and the eigenangle and the eigenaxis of the QFT of the original image is proposed. The method performers comparable results with state-of-the-art methods in predicting human eye fixation locations, but the major draw-back is the fact the method is not able to detect entire salient objects, but rather their strong edges, when objects are large or their center when objects are smaller, and all this despite the multiscale adopted strategy. In order to overcome this inconvenience, in this framework the algorithm is locally applied and the pixels saliency is calculated in a global fashion based on the region rarity inside the image.

The remainder of the paper is organized as follows: in the second section some related works are overviewed; in section III and IV the proposed method is described and a brief experiment about how to choose the patches size is presented, followed by the evaluation sections and finally the concluding remarks section.

II. RELATED WORK

In literature there are several computational models which compute saliency maps for digital imagery. Most of the methods rely on Treisman Feature Integration Theory (FIT)

Manuscript received June 20, 2013; revised September 25, 2013

[9]. According to this theory salient locations in a natural scene automatically stand out due to their specific low-level features (such as color contrast, orientations, intensity) when a human looks at that scene. A representative work is the one proposed by Itti *et al.* [18], which uses Differences of Gaussian (DOG) as center-surround differences for color, intensity and orientation channels to estimate salient image locations. According to FIT theory these low-level features which describe a salient region are slightly different within a neighborhood. Due to this observation most of the existing computational methods refer to salient regions as regions with high contrast and adopt center-surround differences as in [18]. The major drawbacks of these algorithms is that their resulted saliency maps are most often blurry and emphasize small local features, making these approaches less useful in applications such as segmentation or object detection.

A different category of models are those applied in frequency domain [19, 20], which proved to achieve better performances than the FIT based methods, although they exhibit undesirable blurred areas and capture strong edges rather than entire salient areas. Both type of methods neglect an important characteristic of the HVS mentioned in the previous section and modeled by the zoom lens model [16] (the influence of an attended region decreases with the increase of spatial difference). The problems of these methods are generated by the fact that they consider either just a local contrast approach, either just a global approach. Local methods estimate saliency of a particular image region, namely patches, based on the surrounding regions as dissimilarities at the pixel level [2], as DOGs at different image scales [9] or by histogram analysis [7]. On the other hand global approaches treat the contrast relations over the entire image.

Unlike these methods, there are models which consider both local and global contrast, and proved to be more accurate in predicting salient regions and agree more with the two stage attention allocation model from [11]. In [2] a local-global saliency model which aims to detect regions which most consistently represent the image is proposed. The multi scale model computes saliency at pixel level considering image patches dissimilarities of color and spatial differences based on visual organization rules. In[7] salient objects are described locally, regionally and globally based on features such as multiscale contrast, center-surround color histograms and color spatial distribution, which are combined to form saliency maps by learning a Conditional Random Field.

In [12] saliency maps are computed based on a local-global approach. Borji and Itti [12] calculate saliency by implementing local center-surround operations and, then in a globally strategy, the rarity of each image patch is computed over the entire scene. Their method considers color features rarities by decomposing each color channel in non-overlapping patches, which are represented by vectors of coefficients. These coefficients are able to linearly reconstruct the patches from a learned dictionary of patches from natural scenes.

In [13], visual saliency is estimated based on color and orientations feature channels, which are commonly examined in parallel. For each feature channel, the local patch information is analyzed using covariance matrices, known as *region covariance*. Region covariances are able to better capture local image structures than standard linear filters, and they naturally provide nonlinear integration of features by modeling their correlations [13]. This method proved to be very efficient in several tasks as human fixations prediction, salient objects detection and image retargeting.

A similar with the proposed method approach is presented in [14], where saliency is estimated based on amplitude spectrum of the quaternion Fourier transform applied on local image patches. Saliency is calculated as dissimilarities between image patches and spatial distance. In this case spatial distance between patches takes into account the HVS behavior, which as shown in [16] is able to incorporate in its focused visual field only a region of a certain size and its visual sensitivity decreases with the increased eccentricity from the fixated region [21]. The method proved its efficiency in an image retargeting application. Another similar approach is proposed in [15]. This time, image patch saliency is computed as proposed in [2], considering a multi scale strategy. Different from [2], in this approach local patch meaningful structure is extracted using principal component analysis (PCA), which reduces noisy undesirable components. Based on color dissimilarities and spatial distances, the method performs good results for human fixations predictions and salient objects segmentation tasks.

In this framework, a method inspired by [2] and [14] is proposed. The method proposed in [17] is now applied locally on fixed size patches, and then both pixels and patches saliency are considered. Saliency at pixel level is computed based on the dissimilarities between amplitude values of the QFT spectrum calculated inside each patch. At the patch level, for each spatial distance between patches a weighting value is associated determined by the human visual sensitivity [21]. More implementations details are depict in the following sections.

The proposed paper aims also to solve the trade-off issue of the zoom-lens attention model [16]. It is well known that many state-of-the-art local methods deal with choosing the optimal patch size problem. Most of them apply their algorithm on small patches (regularly on 8x8 pixels) avoiding thus blocking artifacts in their final saliency maps. On the other hand, considering the fact that for a given *MxN* pixels image with patch size of *kxk* and patch overlapping percent λ, the computational complexity is computed as $(M*N)^2 \Big/ \left((1-\lambda)^2 k\right)^2$ and can increase as the patches size is smaller and they are more overlapped. On the other hand, in MIT dataset from [10], the authors concluded that the human eye is able to fixate and process at a time a region of interest of a certain size. Aiming to propose a biologically plausible algorithm, in this framework the size of the ROIs is also discussed. Using three different human eye fixations data sets, this topic is analyzed in the following section.

III. SIZE OF REGIONS OF INTEREST

Analyzing fixation data sets can provide significant insights into how human brain allocates visual attention. Qualitative studies proved that most likely eye gazes fall off on faces, persons, text, cars, human body parts, animals or any region which may reflect a coherent notion for the human brain. Semantically, these regions are grouped pixels with a unique appearance; this last observation is sustained by the bottom-up visual attentions mechanisms which interact further with

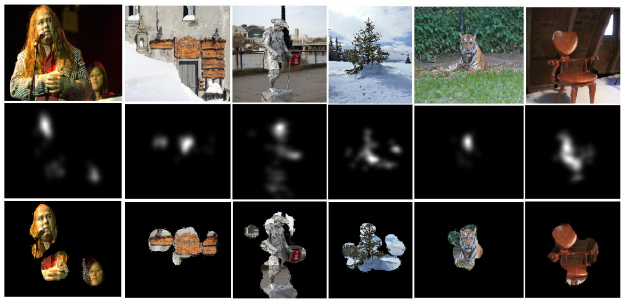

Figure 1. Regions of interests for sample images from MIT dataset: on rows from top to bottom, original images, corresponding fixation maps and bounding boxes which capture most potentially 20% salient fixations.

top-down mechanisms that help the observer to associate these grouped pixels with pre-defined known notions (e.g. persons, objects, parts of human body or objects). Although establishing the border between the influences of the two mechanisms is still an unknown path, it was proved in [22] that the most salient regions that humans fixate on are parts from proto-objects [23]. Furthermore, considering objects may be a good strategy in predicting eye fixations.

In their early eye tracking experiments, Judd *et al.* [10] noticed that human gazes fixate on different image locations, due to the detail level of an image. They analyzed the case of images containing faces and observed that when a face is shown far away, humans look at faces as a whole with one fixation, while when the face is shown close up, observers fixate parts of the faces (eyes, nose, mouth). From their experiments an interesting observation can be concluded: there is a certain size for the region of interest that a person fixates on. In order to analyze fixations on specific objects and image features incorporated in ROI's with specific size, bounding boxes were labeled for each image from three different data sets: *Toronto* [24], *MIT* [10] and *ImgSal* [25]. For each image a binary map was generated by thresholding the corresponding fixation map (see fig. 1). In fig. 1 the fixation maps were threshold in order to capture the majority of the fixations. It can be easily noticed that the majority of the proto-objects that humans fixate on can be incorporated in elliptical bounding boxes placed around connected areas of salient pixels on an image overlaid with its 20% salient mask. Finding the most suitable proprieties of these regions can lead to helpful tools in deciding the optimal ROI's size. In [26], the authors attempt to automatically segment images, proposing a novel framework for generating and ranking plausible objects hypothesis. They learn to rank an objects hypothesis by training a model on a set of low and mid-level features related to graph, region and Gestalt [27] proprieties and showed up that the minor axis length of the ellipse having the same normalized second central moments as the region is the most important propriety that can rank an object. For this framework, both the minor and the major axis length of the bounding regions were considered. In fig. 2 histograms of the minor and major axis length measurements are depicted. It can be easily noticed that for

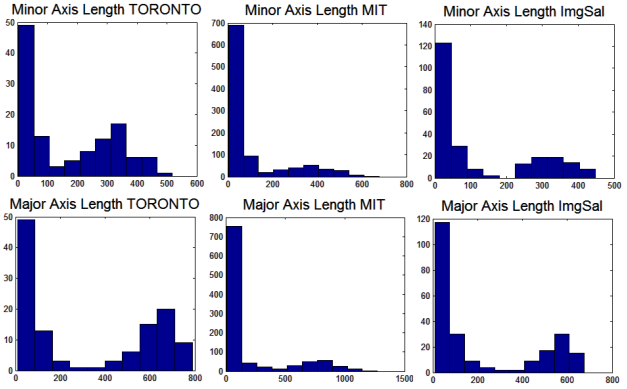

Figure 2. Histograms of Minor and Major axis length in pixels of the regions of interest on three different human fixations datasets: Toronto, MIT and ImgSal.

all three data sets, the peaks of the histograms are around a certain value which suggests that in generally the most axis have a length smaller than 100 pixels. Also the histograms depict a cluttered region with minor/major axis having larger sizes which peak around 300-400 pixels/600-700 pixels. These second axis length are due to the variety of the data sets, which may contain images with simple scene with a single or a few salient regions, and also images with higher complexity with more salient regions. An example are the first four images from fig. 1, where the presence of more than one region which human fixate on, regions that happens to be close to each other, produces a larger region of interest than in the case of last two images, which depict a single significant homogeneous region of interest. To have a better precision, the threshold was fixated to a larger value, and as can be noticed in fig. 3 only the most 3% most salient fixations were incorporated. For this situation similar values for the minor and major axis length were obtained (see fig. 4). The minor axis lengths range between values smaller than 100 pixels, with an average value around 50 pixels. Also, these values are, in both cases, regardless of the image size, which was kept as the original, and which varied from a data set to another, and inside the data set, as is the *MIT* case.

Due to the fact that was proved that minor axis length describes better than any region propriety the probability of a salient object region [26], further, only this measure will be considered in assessing the ROI's size. More precisely, for the local proposed approach, each image will be divided in overlapping patches with sizes of 50x50 pixels as suggested by the previous measurements.

III. PROPOSED ALGORITHM

The aim of the paper is to propose a biologically inspired strategy to estimate salient regions. Furthermore, due to the fact that human visual attention acts as a two stage process [11], the proposed local-global algorithm applies both locally and globally biologically inspired principles. The proposed bottom-up method considers color, intensity and orientation features to estimate visual saliency in *Lab* color space, which aspires to perceptual uniformity, its *L* component approximating human perception of intensity,

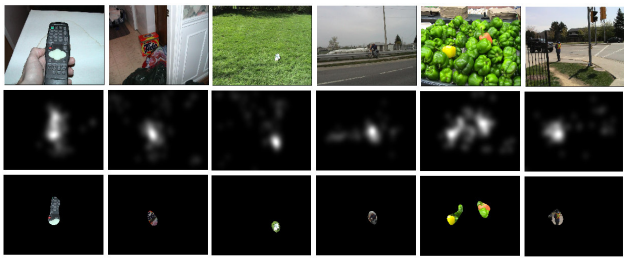

Figure 3. Regions of interests for sample images from Toronto dataset: on rows from top to bottom, original images, corresponding fixation maps and bounding boxes which capture 3% of the most salient fixations.

while the *a* and *b* channels matching the human chromatic opponent system. Unlike the previous work from [17], which adopts a global approach, in this context the same method is applied locally.

In [17] a spectral method for salient regions prediction is proposed. The model calculates saliency based on amplitude and phase spectrums of the QFT. It is commonly accepted that amplitude spectrum cares appearance and orientation information about a visual scene, while the phase spectrum specifies the location information. Thus, the amplitude spectrum of QFT can represent the color, intensity and orientation distributions within an image. The advantage of applying a quaternion transform, and, implicitly a vector representation of the image features, is given by the fact that in this way the difficulty of linear or *ad hoc* feature combination is overcome, since quaternion transform exploits and relies on the feature correlation [28].

The proposed model first transposes each input image to Intensity Color Opponent (ICOPP) space according to [25] which separates a image in a intensity I channel which corresponds to L component of the Lab color space, and two chromatic opponent color channels red-green RG and blue-yellow BY, corresponding to a and *b* channels. At this point the image can be represented in a matrix, vector form using quaternion representation [29]: $I_Q = I \cdot i + RG \cdot j + BY \cdot k$, where $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $jk = -kj = i$ and $ki = -ik = j$. In a second stage each input quaternion image I_Q is divided into partially overlapping patches of *50x50* quaternion pixels.

Unlike other methods which use linear filters as DOG, Gabor, PCA or region covariances, in this method the local information is extracted using filters based on human contrast sensitivity functions and phase spectrum of the QFT as proposed in [17].

Thus for a give quaternion image patch I_Q^x the local information is extracted by applying the polar form of the QFT of the patch as in relation (1):

$$I_{QFT}^x = \left|I_{QFT}^x\right| \cdot e^{\gamma\varphi} \tag{1}$$

In relation (1), the amplitude spectrum $\left|I_{QFT}^x\right|$ of the patch *x* is replaced with the quaternion matrix from (2), which considers human visual sensitivity to intensity and color-opponent contrasts:

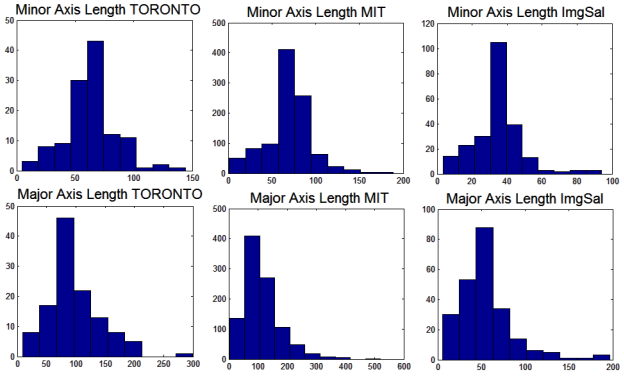

Figure 4. Histograms of Minor and Major axis length on Toronto, MIT and ImgSal datasets, when 3% of the most salient fixations are considered.

$$CSF_Q^x = \left|I_{FFT}^x\right| CSF_I i + \left|RG_{FFT}^x\right| CSF_{RG} j + \left|BY_{FFT}^x\right| CSF_{BY} k \tag{2}$$

where $\left|I_{FFT}^x\right|$, $\left|RG_{FFT}^x\right|$, $\left|BY_{FFT}^x\right|$ are the amplitude spectrums of the regular Fourier transform of the intensity and color opponent channels of the patch *x*, while CSF_I, CSF_{RG}, CSF_{BY} are the coefficient matrixes of the contrast sensitivity function based filters, one for each ICOPP channels. These *CSF* functions were experimentally developed based on the different responses of the HVS to visual stimuli [30, 31]. Biological evidences proved that human eye respond to visual stimuli above a certain contrast. This sensitivity threshold can be influenced by different parameters such as spatial frequency, viewing distance, orientation. CSF function, named also visual acuity, measures the sensitivity to various frequency and is different for intensity and the two chromatic pairs of color opponent stimuli RG and BY.

At this stage the algorithm extracts the local information from the amplitude spectrum. In order to reconstruct the patch, the phase spectrum of the patch *x* is used, by calculating the eigenaxis γ^x and the eigenangle φ^x as in [29]. Thus saliency of patch *x* is computed as:

$$S_{local}^x = \left|CSF_Q^x\right| \cdot e^{\gamma^x \varphi^x} \tag{3}$$

Furthermore, patch global saliency value is computed based on the differences between the patch and all other patches from the input image:

$$S_{global}^x = \sum_{y \neq x} \alpha_{xy} D_{(x,y)} \tag{4}$$

where α_{xy} is the weight for the patch difference between patches *x* and *y*, and $D_{(x,y)}$ is the difference matrix between the local saliency values of the pixels from patches *x* and *y*. The difference is computed as Euclidian distance between the pixels values located at the same spatial position inside the patches as in (5):

$$D_{(x,y)} = \sqrt{\left(S_{local}^x - S_{local}^y\right)^2} \tag{5}$$

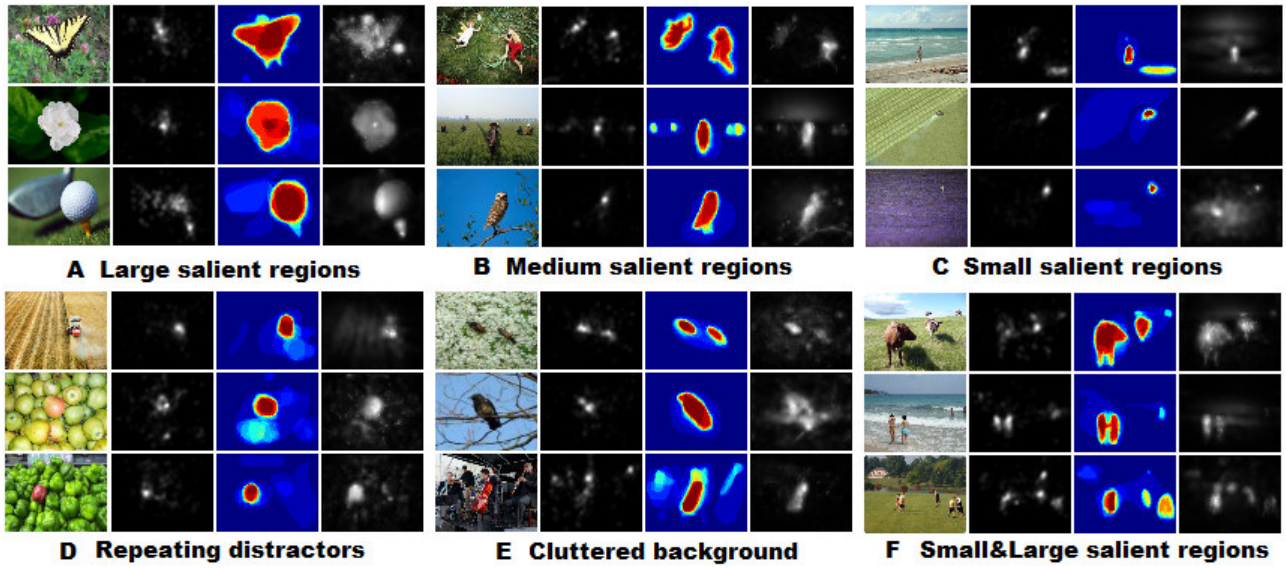

Figure 5. Qualitative results on several example images from each image category. For each type of images, on columns: original images, corresponding density maps of fixations, labeled salient regions and saliency maps with proposed algorithm.

Table I. Area under ROC curve (mean value) for each category of images

Image type	Large	Medium	Small	Repeating	Cluttered	Small&Large
Fixations	0.7906	0.7829	0.7627	0.8023	0.7958	0.8007
Regions	0.9566	0.8923	0.8680	0.9307	0.9318	0.9168

In relation (4) S^{x}_{global} and $D_{(x,y)}$ are matrixes of *50x50* pixels, while α_{xy} is a scalar value. These patch dissimilarities suggest the difference in appearance of each patch over the entire image. The patch global saliency increases as the differences are bigger.

As stated in the previous section, considering objects may be a good strategy for fixation estimation [22]. Semantically, objects are groups of pixels with homogeneous statistical proprieties, which may possess several centers of gravity about which the objects forms are organized [27]. Thus the positional distance between patches is another important factor and if the between pixels differences ensures the patch uniqueness, considering a spatial distance factor, resembling far-away patches will have a smaller contribution to the overall patch saliency. The observation that salient patches tend to group together is also sustained by the foveation theory from [21] and the zoom lens model [16], which suggest that HVS is less sensitive to image regions located far-away from the fixated area.

Unlike the methods proposed in [2, 15], which compute spatial distances between patches based on the spatial coordinates of the most similar patches, in this framework the strategy proposed in [14] is applied. It is generally accepted that the HVS is space-variant due to its biological construction which causes the decrease of its visual sensitivity with the increase of the eccentricity from the attended region [16, 21]. In [14], the human contrast sensitivity, expressed as a function of eccentricity (see relation 6), is used to establish the spatial distance based weights α_{xy} from (1).

$$C_S(f,e) = 1 \Big/ \left(C_0 \exp\left(\alpha f \frac{e+e_2}{e_2} \right) \right) \tag{6}$$

In relation (6), f is the spatial frequency (cycles/degree), e is the retinal eccentricity (degree), C_0 is the minimum contrast threshold, α is the spatial frequency decay constant, while e_2 is the half-resolution eccentricity. Experiments from [21] set the following values C_0=1/64, e_2=2.3, and α=0.106.

In [14], authors concluded that α_{xy} can be determined by normalizing $C_S(f,e)$ as in (7):

$$\alpha_{xy} = 1 \Big/ \left(C_0 \exp\left(\alpha f \frac{e+e_2}{e_2} \right) \right) \tag{7}$$

In relation (7) the eccentricity e remains the only one unknown parameter. This parameter can be calculated according to its relationship with the viewing distance v:

$$e = \tan^{-1}\left(\frac{d}{v} \right) \tag{8}$$

where d is the Euclidean distance between two patches. In [14] this relation is used to establish the image patch size (see relation 9). Since in this framework the patch size p was

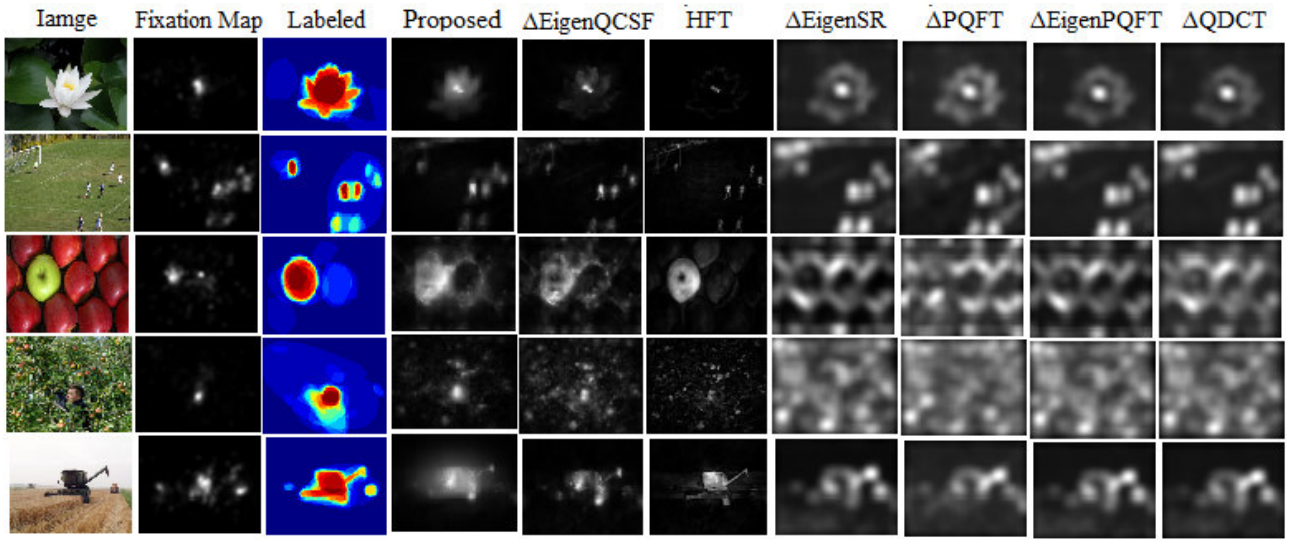

Figure 6. Qualitative comparison with several state-of-the-art spectral methods and the previous multi scale method from [17] between obtained saliency map and against fixations maps and labeled salient regions maps, respectively.

Table II. Comparison between AUC mean values of the proposed method and several spectral multi scale methods

ΔEigenQCSF	**HFT**	**ΔEigenSR**	**ΔPQFT**	**ΔEigenPQFT**	**ΔQDCT**	**Proposed**
0.7157	0.6724	0.6658	0.6610	0.6651	0.7034	0.7892
0.9251	0.8651	0.8695	0.8325	0.8709	0.8903	0.9160

already determined, and considering the eccentricity which ensures the best visual acuity [16], namely 1 degree, relation (9) is used to calculate the viewing distance v:

$$e = \tan^{-1}\left(\frac{p}{2v}\right) \tag{9}$$

From the procedures above, the saliency values of the pixels in patch x are computed as all the contributions from the patch pixels differences between this patch pixels and the pixels from all other patches in the image, as calculated in (4). Although in [15], where the patch size was empirically chosen, was shown that smaller patch size favors a more distinguishable saliency map, while larger patches produce blocking artifacts in the final saliency map, in the proposed method, this inconvenience is discharged by applying over the entire final saliency map a circular averaging filter with size of radius equal with half of the patch size.

IV. EVALUATION

In this section the proposed method is evaluated on natural images to prove its effectiveness. In the evaluation experiment the data set proposed in [25] is used as benchmark for comparison. The data set contains corresponding eye fixation maps and also manually labeled maps which incorporate most salient objects in the images. The data set *ImgSal* [25] consists of 235 natural images with fixed size 480x640 pixels which are divided into 6 categories: images with large salient images, images with medium salient regions, images with small salient regions, images with cluttered background, images with repeating distractors, and images with both small and large salient regions. Note that most of the images belong to more than one category since an image may contain different attributes. The data set was labeled with 19 naive subjects and includes two types of ground truth: fixation density maps and binary maps which label the most salient object in the image.

Unlike the previous method from [17], the proposed algorithm is applied at a single scale, which is the original image size. This is due to the fact that the size of patches was established when the images were analyzed at the original size. For qualitative evaluation, the receiver operator characteristic (ROC) curve is exploited by computing the average value of the areas under the ROC curve (AUC) of each image. In figure 5 qualitative results of the proposed algorithm for each image category are depicted, whereas in table I resulted AUCs for each image category are given.

Qualitative results from figure 5 show that the method produces fair saliency maps. In case of human fixations prediction, the algorithm performs better results in case of images with repeating distractors, as well as for images with salient regions of different sizes, whereas for images with small salient regions, the method shows worst results. The same situation happens in case of salient region detection. For this scenario, the method achieves best results when predicting large salient regions. This situation is quite challenging for many saliency detection methods, which are not able to uniformly capture the whole salient region . This advantage recommends the proposed algorithm as a proper pre-processing stage in an image retargeting application, as will further in this paper be shown. The most challenging situations are for sure the cases with cluttered background and repeating distractors. For both scenarios, the algorithm suppresses the repeating distractors, extracting the most salient fixations or regions (see fig. 5 D). Also in case of cluttered background fig. 5 E, the algorithm proves to be less sensitive to background noise.

Given the local-global strategy, the proposed algorithm is able to extract simultaneously salient regions of different

sizes from an image (fig. 5 F), while most of the existing state-of-the-art methods apply a multi stage strategy.

All these observations are both qualitatively and quantitatively sustained. To strength the method effectiveness, other 5 state-of-the-art spectral algorithms, as well as the previously proposed method [17] are employed in the comparison process: ΔEigenSR, ΔPQFT, ΔEigenPQFT, ΔQDCT proposed in [29] (Δ indicates a multi scale approach), HFT from [25] and ΔEigenQCFS [17]. All these methods propose a multi scale strategy, except from the HFT method which uses in the final saliency maps just one of the scales. Qualitative and quantitative results over the entire data set of the proposed and of the other 6 methods are depicted in figure 6 and table II. Resulted saliency maps from the methods proposed in [29] show the difficulty of these approaches to deal with background noise (see fig. 6 rows 3 and 4). In presence of repeating distractors the algorithm from [29] fails to suppress them; as for the cluttered background the saliency map is strongly affected by noise. In images with large salient regions, most of the methods fail in uniformly capturing whole salient regions, emphasizing strong edges or centers of the salient regions. The advantage of the multi scale methods from [29] is in the case of images with small salient regions as in first row example image from fig. 6. This is due to the fact that the algorithms are applied at reduced scale and, additionally, a multi stage strategy is employed. This last strategy makes the methods from [29] able to detect both large and small salient regions from an image (see last row in fig. 6). Although similar, after visual inspection of the saliency map computed with the approach from [25], the proposed method seems to achieve better qualitative and quantitative results. The method has the ability to better capture large salient region, to suppress repeating distractors and is less sensitive to background noise. These visual observations are also sustained by the AUCs mean value from table II. The proposed method achieves the best results in eye fixation prediction, but on salient region detection the method seems to be less accurate than the previously proposed version from [17]. Yet, the method proved to be more efficient in extracting large salient regions than any of the previously mentioned models. Furthermore, the method is tested in an image retargeting application.

V. APPLICATION IN IMAGE RETARGETING

Image retargeting, or content aware image resizing aims to resize images enlarging or shrinking the irrelevant information from the image, in order to make the image suitable for displays with different sizes or aspect ratios. Such an image application must be able to preserve the important content, without distortions on image content, at least on the salient regions. Hence, most of the image retargeting methods rely on an importance map, similar with a saliency map.

In this framework, the retargeting algorithm from [32] based on seams carving in a certain direction was used. The Matlab code (available at http://people.csail.mit.edu/mrub/) for retargeting method [32] was run using saliency maps computed according to the proposed method. For this purpose, *ReTargetMe* benchmark data set, also online available at *http://people.csail.mit.edu/mrub/retargetme/* was used. This data set contains 80 images with 92 different resizing situations such as increasing or reducing one of the image dimensions by 25% or 50%. In figure 7 a couple of

Figure 7. Different scenarios of image retargeting; on columns from left to right: input images, computed seams paths (red lines) with [32], respectively when using the saliency maps with proposed method, and corresponding retargeted images with seam carving, respectively proposed method.

retargeting results are presented. It can be easily notice that when using the proposed saliency maps, the retargeting method by seam carving can achieve more pleasant results. This is due to the fact that the proposed saliency maps remove the seams placed inside salient uniform regions such as those from *face* and *blueman* images from fig. 7, rows 2 and 4. The baseline method from [32] fails in these situations, producing significant distortions and artifacts inside the salient regions of the images. Quantitatively, due to the subjective nature of this application, there are just a few objective metrics which can assess these results. One of these metrics is the Earth Mover's Distance (EMD) defined as the minimal cost that must be paid to transform one histogram into the other, where there is a *ground distance* between the basic features that are aggregated into the histogram [33]. For this framework the global score of EMD for non-normalized histograms was computed. Average values of the metric on the *ReTargetMe* data set are given in table III. Smaller values of the metric indicate similar visual content. Although the results for the two scenarios are similar, the seam carving with the proposed saliency estimation achieves slightly smaller differences.

Table III. Average rankings of the EMD score of the proposed saliency method and the baseline Seam Carving method on the ReTargetMe data set.

Seam Carving [32]	**Proposed Seam Carving**
2,0425	2,0330

VI. CONCLUSIONS

In this study a bottom-up spectral method for visual saliency estimation was presented. Based on quaternion representation of the color images and hypercomplex transform spectrums, the proposed algorithm extends a previous work in a local global context. After applying the previous method locally, the proposed method computes saliency maps by considering global differences between image patches at pixels level and global spatial differences between patches based on the human visual sensitivity. Taking into account biologically inspired principles about ocular eccentricity and human's visual behavior, the previous work is applied on local patches with a specific size. The size of the patch is semi-empirically chosen, after analyzing the size of the ROIs on three different eye fixations data sets. Qualitative and quantitative results after applying the proposed algorithm on a benchmark data set proved the effectiveness of the method in predicting eye fixations and uniform salient regions as well. The method achieved comparable accuracy as several state-of-the-art methods, and proved its efficiency in an image retargeting application based on seam carving.

As future work, the method can be easily extended to video saliency detection, due to the fact that it is applied on quaternion color images. Furthermore, such an approach can be applied in a video retargeting application, also.

ACKNOWLEDGEMENT

This paper was realised with the support of POSDRU CUANTUMDOC "DOCTORAL STUDIES FOR EUROPEAN PERFORMANCES IN RESEARCH AND INNOVATION" ID79407 project funded by the European Social Fund and Romanian Government.

REFERENCES

[1] Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. F. Cohen, "Gaze-based interaction for semi-automatic photo cropping", *CHI, ACM,* pp. 771-780, 2006.
[2] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-Aware Saliency Detection", *IEEE Trans. Pattern Anal. Mach. Intell.* 34(10):1915-1926*,* 2012.
[3] M. Holtzman-Gazit, L. Zelnik-Manor, I. Yavneh, "Salient edges: A multi scale approach", *In: ECCV, Workshop on Vision for Cognitive Tasks*, 2010.
[4] C. Guo, L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression", *IEEE Transactions on Image Processing,* 19(1), pp. 185-198, 2010.
[5] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention", *IEEE Transactions on Image Processing,* 13(10), pp. 1304-1318, 2004.
[6] L. Itti, *Models of Bottom-Up and Top-Down Visual Attention,* PhD thesis, California Institut of Technology, Pasadena, 2000.
[7] T. Liu, Z. Yuan, J. S. 0001, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, "Learning to Detect a Salient Object", *IEEE Trans. Pattern Anal. Mach. Intell,.* 33(2), pp. 353-367, 2011.
[8] Z. Wang, B. Li, "A two-stage approach to saliency detection in images", *in ICASSP, IEEE*, pp. 965-968, 2008.
[9] A. Treisman, G. Gelade, "A feature-integration theory of attention", *Cognit. Psychol.*, 12(1), pp. 97-137, 1980.
[10] T. Judd, K. Ehinger, F. Durand, A. Torralba, "Learning to predict where people look", *ICCV,* 2009.
[11] J. Jonides*,* "Further towards a model of the mind's eye's movement"*, Bulletin of the Psychonomic Society,* 21(4), pp. 247-250, 1983.
[12] A. Borji, L. Itti, "Exploiting local and global patch rarities for saliency detection", *CVPR*, pp. 478-485, 2012.
[13] E. Erdem, A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariance", *Journal of Vision*, 13(4):11, 1-20, 2013.
[14] Y. Fang, W. Lin, B. S. Lee, C. T. Lau, Z. Chen, C. W. Lin, "Bottom-Up Saliency Detection Model Based on Human Visual Sensitivity and Amplitude Spectrum", *IEEE Transactions on Multimedia* 14(1): 187-198, 2012.
[15] J. Zhou, Z. Jin, "A new framework for multiscale saliency detection based on image patches", *Neural Processing Letters,* January 2013.
[16] C. Eriksen, J. St James, "Visual attention within and around the field of focal attention: A zoom lens model". *Perception & Psychophysics* 40(4): 225–240, 1986.
[17] O. L. Buzatu, A. Savin, "Saliency Based on Human Visual Sensitivity and Phase Spectrum of the Quaternion Fourier Transform", *ISSCS*, 2013.
[18] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", *IEEE Trans. Pattern Anal. Mach. Intell,*. 20(11), pp. 1254-1259, 1998.
[19] X. Hou, L. Zhang, Saliency Detection: A Spectral Residual Approach, *in CVPR , IEEE Computer Society,* pp. 1-8, 2007.
[20] C. Guo, Q. Ma, L. Zhang, Spatio-temporal Saliency detection using phase spectrum of quaternion Fourier transform, *in CVPR, IEEE Computer Society,* pp. 1-8, 2008.
[21] W. S. Geisler, J. S. Perry, "A real-time foveated multi-solution system for low-bandwidth video communication", *in Proc. SPIE*, 3299, pp. 294-305, Jul. 1998.
[22] W. Einhauser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency", *Journal of Vision*, 2008.
[23] D. Walther, C. Koch," Modeling attention to salient proto-objects", *Neural Networks,* 19(9):1395-1407*,* 2006.
[24] N.D.B. Bruce, J.K. Tsotsos, "Saliency based on information maximization", *NIPS*, 2005.
[25] J. Li, M. D. Levine, X. An, X. Xu, H. He, "Visual Saliency Based on Scale-Space Analysis in the Frequency Domain", *IEEE Trans. Pattern Anal. Mach. Intell.,* 35(4): 996-1010, 2013.
[26] J. Carreira, C. Sminchisescu,"Constrained parametric min-cuts for automatic object segmentation", *CVPR,* 2010.
[27] K. Koffka, *Principles of Gestalt Psychology*, Routledge & Kegan Paul, 1955.
[28] T. A. Ell, S. J. Sangwine, "Hypercomplex Fourier Transform of Color Images", *IEEE Trans. Im. Proc.,* 16(1), pp. 22-35, 2007.
[29] B. Schauerte, R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction", *in ECCV*, 2, pp. 116-129, 2012.
[30] J. M. Rovamo, M. I. Kankaanpaa, H. Kukkonen, "Modeling spatial contrast sensitivity functions for chromatic and luminance-modulated gratings", *Vision Research,* 39, pp. 2387-2398, 1999.
[31] J. L. Mannos, D. J. Sakrison, "The Effect of a Visual Fidelity Criterion on the Encoding of Images", *IEEE Trans. Infor. Theory*, IT-20(4), July 1974.
[32] M. Rubinstein, A. Shamir, S. Avidan, "Improved seam carving for video retargeting", *ACM Trans. Graph.,* 27(3), 2008.
[33] O. Pele, M.Werman, "Fast and Robust Earth Mover's Distances", *ICCV* 2009.