SALIENCY BASED ADAPTIVE IMAGE SIGNATURE USING BINARY HOLISTIC IMAGE DESCRIPTOR

RAJESH MATE[1], PRAVIN DERE[2] & SANJAY M. HUNDIWALE[3]

[1]M.E Student, Department of EXTC, ARMIET College of Engineering, Sapgaon, Mumbai, Maharashtra, India

[2]Associate Professor, Department of Electronics, Terna College of Engineering, Nerul, Mumbai, Maharashtra India

[3]Associate Professor, Department of EXTC, ARMIET College of Engineering, Sapgaon, Mumbai Maharashtra, India

Abstract

We introduce a simple image descriptor referred to as the image signature. We show, within the theoretical framework of sparse signal mixing, that this quantity spatially approximates the foreground of an image. We experimentally investigate whether this approximate foreground overlaps with visually conspicuous image locations by developing a saliency algorithm based on the image signature. This saliency algorithm predicts human fixation points best among GBVS (Graph based visual saliency), Itti/Koch saliency map and sun saliency does so in much shorter running time. And we experimentally investigate that the distance between images induced by the image signature is closer to human perceptual distance than can be achieved using other saliency algorithms, pixel-wise, or GIST descriptor methods.

Problem definition

One of the most severe problems of perception is information overload. Peripheral sensors generate afferent signals more or less continuously and it would be computationally costly to process all this incoming information all the time. Thus, it is important for the nervous system to make decisions on which part of the available information is to be selected for further, more detailed processing, and which parts are to be discarded. Furthermore, the selected stimuli need to be prioritized, with the most relevant being processed first and the less important ones later, thus leading to a sequential treatment of different parts of the visual scene. This selection and ordering process is called selective attention. Among many other functions, attention to a stimulus has been considered necessary for it to be perceived consciously.

What determines which stimuli are selected by the attention process and which will be discarded? Many interacting factors contribute to this decision. It has proven useful to distinguish between bottoms-up and top-down factors. The former are all those that depend only on the instantaneous sensory input, without taking into account the internal state of the organism. Top-down control, on the other hand, does take into account the internal state, such as goals the organisms has at this time, personal history and experiences, etc. A dramatic example of a stimulus that attracts attention using bottom-up mechanisms is a fire-cracker going off suddenly while an example of top-down attention is the focusing onto difficult-to-find food items by an animal that is hungry, ignoring more "salient" stimuli.

Introduction

In this paper, we provide an approach to the figure-ground separation problem using a binary, holistic image descriptor called the "image signature." It is defined as the sign function of the Discrete Cosine Transform (DCT) of an image. As we shall Demonstrate, this simple descriptor preferentially contains information about the foreground of an image—a property which we believe underlies the usefulness of this descriptor for detecting salient image regions. In Section 2, we formulate the figure-ground separation problem in the

framework of sparse signal analysis. We prove that the Inverse Discrete Cosine Transform (IDCT) of the image signature concentrates the image energy at the locations of a spatially sparse foreground, relative to a spectrally sparse background.
Then, in Section 3.1, we demonstrate that a saliency map derived from the image signature outperforms many leading saliency algorithms on a benchmark data set of eye-movement fixation points. In Section 3.2, we show that the distance between images induced by the image signature most closely matches the perceptual distance between images inferred from these data among competing measures derived from other saliency algorithms, the GIST descriptor, and simpler pixel measures.

Related Work

Holistic image processing short-circuits the need for segmentation, key-point matching, and other local operations. Bolstered by growing general interest in large-scale image retrieval systems, holistic image descriptors have become a topic of intense study in the computer vision literature. GIST [2] is an excellent example of such an algorithm in this field. Other holistic scene models focus on the separation of foreground and background. For example, Candes et al. [4] introduced a sparse matrix factorization model. A more relevant study comes from Hou and Zhang [5], motivated by Oppenheim et al.'s early discovery [6], [7]. They found that the residual Fourier amplitude spectrum, the difference between the original Fourier amplitude spectrum and its smoothed copy, could be used to form a saliency map. The residual retains more high-frequency information than low, where the smoothed copy is similar to the original. The image signature, in comparison, discards amplitude information across the entire frequency spectrum, storing only the sign of each DCT component, equivalent to phase for Fourier decomposition. The image signature is thus very compact, with a single bit per component, and as we shall show in the remainder of this paper, possesses important properties related to the foreground of an image.

IMAGE SIGNATURE

Preliminaries

We begin by considering gray-scale images which exhibit the following structure:

$$\boldsymbol{x = f + b}, \qquad \boldsymbol{x, f, b \in IR^N} \qquad \textbf{(1)}$$

where f represents the foreground or figure signal and is assumed to be sparsely supported in the standard spatial basis. b represents the background and is assumed to be sparsely supported in the basis of the Discrete Cosine Transform. In other words, both **f** and $\hat{\mathbf{b}}$ have only a small number of nonzero components. Please refer to Table 1 for important definitions used throughout the rest of this section.

Performing the exact separation between b and f given only x and the fact of their sparseness is, in general, very difficult. For the problem of figure-ground separation, we are only interested in the spatial support of f (the set of pixels for which f is nonzero). In this paper, we show, first analytically, then empirically, that given an image which can be decomposed as 1, we can approximately isolate the support of f by taking the sign of the mixture signal x in the transformed domain and then inversely transform it back into the spatial domain, i.e., by computing the reconstructed image

$$\bar{\boldsymbol{x}} = \boldsymbol{IDCT}[\boldsymbol{sign}(\hat{\boldsymbol{x}})].$$

Formally, the image signature is defined as

Image Signature(x)=sign(DCT(x)). **(2)**

If we assume that an image foreground is visually conspicuous relative to its

background, then we can form a saliency map m (see [8] for classic use) by smoothing the squared reconstructed image defined above

$$\boldsymbol{m = g * (\bar{x}\, 0\, \bar{x})} \qquad \textbf{(3)}$$

Introduction to Visual Saliency

Our attention is attracted to visually salient stimuli. It is important for complex biological systems to rapidly detect potential prey, predators, or mates in a cluttered visual world. However, simultaneously identifying any and all interesting targets in one's visual field has prohibitive
computational complexity making it a daunting task even for the most sophisticated biological brains , let alone for any existing computer. One solution, adopted by primates and many other animals, is to restrict complex object recognition process to a small area or a few objects at any one time. The many objects or areas in the visual scene can then be processed one after the other. This serialization of visual scene analysis is operationalized through mechanisms of visual attention: A common (although somewhat inaccurate) metaphor for attention is that of a virtual *spotlight,* shifting to and highlighting different sub-regions of the visual world, so that one region at a time can be subjected to more detailed visual analysis .

Visual attention may be a solution to the inability to fully process all locations in parallel. However, this solution produces a problem. If you are only going to process one region or object at a time, how do you select that target of attention? Visual saliency helps your brain achieve reasonably efficient selection. Early stages of visual processing give rise to a distinct subjective perceptual quality which makes some stimuli stand out from among other items or locations. Our brain has evolved to rapidly compute saliency in an automatic manner and in real-time over the entire visual field. Visual attention is then attracted towards salient visual locations.

Visual saliency is sometimes carelessly described as a physical property of a visual stimulus. It is important to remember that saliency is the consequence of an interaction of a stimulus with other stimuli, as well as with a visual system (biological or artificial). As a straight-forward example, consider that a color-blind person will have a dramatically different experience of visual salieny than a person with normal color vision, even when both look at exactly the same physical scene). As a more controversial example, it may be that expertise changes the saliency of some stimuli for some observers. Nevertheless, because visual saliency arises from fairly low-level and stereotypical computations in the early stages of visual processing, the factors contributing to saliency are generally quite comparable from one observer to the next, leading to similar experiences across a range of observers and of behavioral conditions.

Introduction to Saliency Map

Definition :Saliency map has its root in Feature Integration Theory and appears first in the class of algorithmic models above . It includes the following elements (see Figure 1):

1. An early representation composed of a set of feature maps, computed in parallel,

permitting separate representations of several stimulus characteristics.

2. A topographic saliency map where each location encodes the combination of properties across all feature maps as a conspicuity measure.

3. A selective mapping into a central non-topographic representation, through the topographic saliency map, of the properties of a single visual location.

4. A winner-take-all (WTA) network implementing the selection process based on one major rule: conspicuity of location (minor rules of proximity or similarity preference are also suggested).

5. Inhibition of this selected location that causes an automatic shift to the next most conspicuous location. Feature maps code conspicuity within a particular feature dimension.

The saliency map combines information from each of the feature maps into a global measure where points corresponding to one location in a feature map project to single units in the saliency map. Saliency at a given location is determined by the degree of difference between that location and its surround. The drive to discover the best representation of saliency or conspicuity is a major current activity; whether or not a single such representation exists in the brain remains an open question with evidence supporting many potential loci.

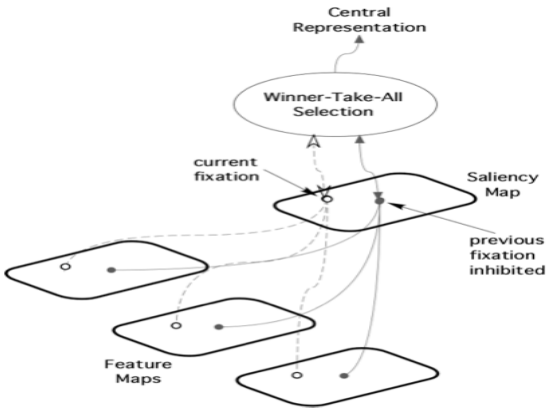


Fig.1 The Saliency Map Model as originally conceived by Koch & Ullman 1985. (figure adapted from Koch & Ullman 1985)

Proposed Algorithm

Quad Tree decomposition segmentation Method:

A **quadtree** is a tree data structure in which each internal node has exactly four children. Quadtrees are most often used to partition a two dimensional space by recursively subdividing it into four quadrants or regions. The regions may be square or rectangular, or may have arbitrary shapes. This data structure was named a quadtree by Raphael Finkel and J.L. Bentley in 1974. A similar partitioning is also known as a *Q-tree*. All forms of Quadtrees share some common features:

- They decompose space into adaptable cells
- Each cell (or bucket) has a maximum capacity. When maximum capacity is reached, the bucket splits
- The tree directory follows the spatial decomposition of the Quadtree.

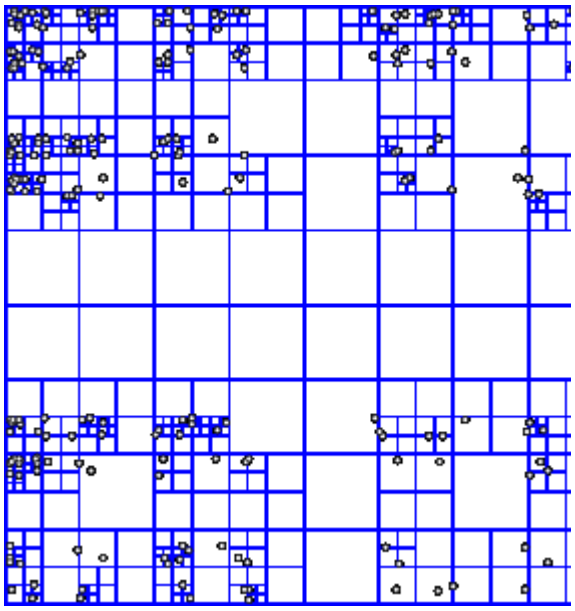


Fig: A region quadtree with point data

3 THE EXPERIMENTS

3.1 Image Signature on Synthetic Images

In this section, we use synthetic images to demonstrate its behavior in carefully constructed cases. In later sections, we will demonstrate the utility of the image signature for practical applications.

Let **f,b,x** $\in$ **IR**^{64X64}. The support of the foreground is a 5 X 5 block $(|T_f| = 25)$ that appears at a random location. The support for $\hat{b}$ is randomly selected in the DCT domain, with $|\Omega_b| = 500$. For $i \in T_f$, the amplitude of each pixel fi is drawn from a normal distribution. Similarly, for $j \in \Omega_b$, each $\hat{b}_j$ is drawn from normal distribution. Fig. 1 shows f , b, and x in both the spatial and the DCT domains.

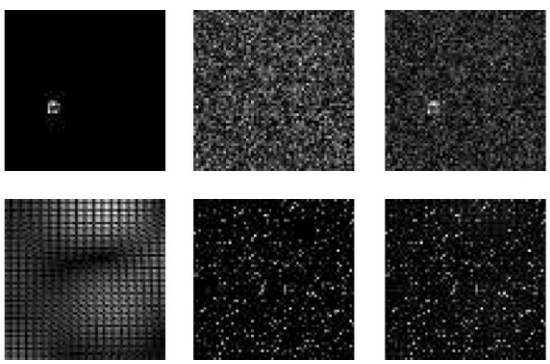


Fig.1. An illustration of the randomly generated images. The first row: f , b, and x in the spatial domain. The second row: The same signals represented in the DCT domain: $\hat{f}, \hat{b}$ *and* $\hat{x}$.

The image signature reconstruction is illustrated in Fig. 2. Note that a Gaussian blurring is used to suppress the noise introduced by the sign quantization. Ideally, the standard deviation of the Gaussian kernel should be proportional to the size of the object of interest. We here choose σ = 0.05 of the image width (in other words, we implicitly assume that the width of the object is about 10 percent of the image width).

Fig.2. An example of the input image x, the reconstructed image $\bar{x}$, and the saliency map m.

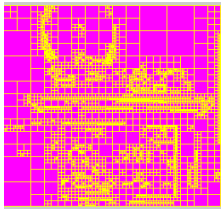
3.2 Generating the Saliency Map of an Image

Here, we report our experimental findings in saliency detection using the image signature. As we demonstrated earlier, the reconstructed image detects spatially sparse signals embedded in spectrally sparse backgrounds. We will show that the saliency map (3) formed from the reconstruction greatly overlaps with regions of human overt attentional interest, measured as fixation points on an input image.

The exact details of the saliency algorithm are as follows: First, a color image is resized to a coarse 64X48 pixel representation. Then, for each color channel x^i, the saliency map is formed from the image reconstructed from the image signature

$$m = g * \sum_i (\bar{x}^i \; o \; \bar{x}^i) \; . \qquad (19)$$


Original image


Saliency Map

Comparatively Algorithm

1)Itti/Koch algorithm

Based on the earlier publication of one of the authors [1], it was assumed that it is possible to design an algorithm reflecting the behaviour of the human eye and nervous system while looking at the image. The list of distinguishing features from the previous paper was refined and in its final form it includes colour, intensity and orientation. Itti and Koch's saliency model is one the earliest and the most widely used in later works for comparison purposes [4]. The authors present a draft of an algorithm examining the differences of colours, saturation and orientation on a given image. The saliency value is dependent on detected local spatial discontinuities of those features. Moreover, the algorithm measures relative
positions of isolated fragments and attaches less saliency to elements lying close to each other and more to those far apart. In the first phase the image is converted to nine different scales (from one to eight reduction factor). An operator denoting "acrossscale difference" is introduced for the "centre-surround" operation to detect locations which stand out from their surroundings. In the second step, the image is decomposed into a series of 42 feature maps.

The first set of feature maps consists of intensity contrast and colours double opponency maps, generated from colour channels of the image. First, the intensity *I* is calculated as an average of colour values:

$$I=\frac{r+g+b}{3}$$

Then, the four colour channels for red, green, blue and yellow are calculated, normalised by intensity and additionally reduced to zero in places where the intensity doesn't reach the threshold of 1/10 of the maximum value for the image. Also, the colours are modified in relation to each other according to pattern:

$$R=\frac{r(g+b)}{2} \quad \text{and} \quad Y=\frac{r+g}{2}-\frac{|r-g|}{2}$$

with negative values set to zero. With five Gaussian pyramids for those five channels for each scale of the image, six feature maps are generated using the introduced centre-surround difference operator on scaled intensity maps, denoted **I**(s,c), where $c \in \{2,3,4\}$ and $s = c + \delta$, $\delta \in \{3,4\}$
are scale reduction factors. This set of maps is made to represent the intensity contrast, which, in mammals, is detected by neurons sensitive either to dark centres on bright surrounds or to bright centres on dark surrounds [5].

Next, there are twelve chromatic double opponency maps calculated for the centre-surround differences between red-green (RG(c,s)) and blue-yellow (BY(c,s)) pairs of opposing channels. They are paired this way because in the centre of their receptive fields, neurons are excited by one colour (e.g., red) and inhibited by another (e.g., green), while the opposite is true in the surround. Such spatial and chromatic antinomy exists for the red/green, green/red, blue/yellow, and yellow/blue colour pairs in human primary visual cortex [6]. One double opponency map is responsible for two such pairs: (RG(c,s)) for red/green and green/red and (BY(c,s))for blue/ yellow and yellow/blue. The final map set consists of the local orientation maps generated via oriented Gabor pyramids from intensity maps. There are twenty-four such maps, where is the preferred orientation encoding orientation differences. All of those forty-two feature maps of various scales are collapsed into three conspicuity maps for intensity ,, colour and orientation . In the next step they are normalised using a normalising operator designed to globally promote maps where a small number of strong peaks of activity is present, while globally suppressing maps

which contain numerous comparable peak responses. The output saliency map is then in consequence

$$S=\frac{N(\bar{\imath})+N(\bar{c})+N(\overline{o})}{3}$$

The algorithm identifies the maximum value of the saliency map (SM) as a centre of attention and then uses the winnertakes-all neural network to identify secondary maximums to generate an approximation of switching the visual attention of the human eye. The so called Itti/Koch algorithm is a very popular method of looking for salient elements on the image, mostly because of the iLab Neuromorphic Vision C++ Toolkit, the freely available implementation in C++ licensed under GPL and developed by Itti and his team.

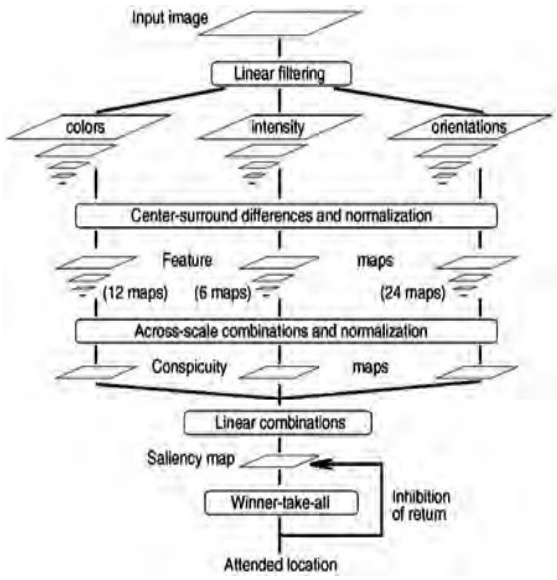
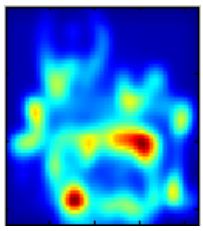


Fig. General architecture of Itti/Koch model. Source: [2] Rys. Ogólna architektur a model Ittiego/Kocha

Original image



Saliency Map

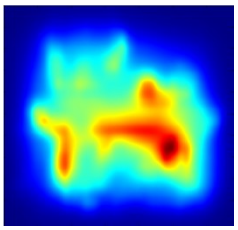
GBVS(Graph based visual saliency)

We present a method of finding saliency map radically different form the classic Itti/Koch algorithm. The process is divided into three stages: extraction of features to feature vectors, generating activation maps from feature vectors and normalisation and combination of activation maps into a single saliency map. The extraction part is omitted in this work and the algorithm itself assumes there are pre-existing feature maps. Both activation and normalisation phases use Markov chain interpretation of the image. The first step of the algorithm is to decompose the image into series of feature maps and perform the preliminary analysis by other methods. The selection of those features is also delegated to outside solutions. The purpose of this step is to reduce the image
resolution by switching from "pixels" to "regions" to simplify the calculations. The remainder of the algorithm deals with a single feature map M. The dissimilarity function is introduced to measure differences between regions (pixels) of the feature map.

The map is now converted to a fully-connected directed graph **GA** by treating each region as a node and drawing a fully-connected graph on those nodes. The weight of the directed edge is defined as directly proportional to dissimilarity of its two ends and their relative position on the map: where and σ is a free parameter. The next step is to define a Markov chain from the **GA** graph by treating the nodes as states and the edge weights as transition
probabilities. The weights of outbound edges of each node are normalised to 1 beforehand. In the equilibrium distribution of this chain the mass accumulates in the states with high dissimilarity with their surrounding nodes. The chain itself is ergodic because the graph that it is based on is strongly connected, so the equilibrium distribution exists and is unique. This equilibrium distribution is the basis to create an activation map A. The authors refer to this solution as "organic", because as neurons work independently but are influencing each other, here each state/node works independently and the result is the

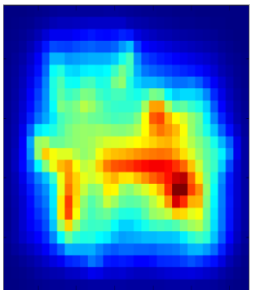
equilibrium state achieved with the input from all of states/nodes. The activation map has to be normalised in the process similar to the former. The map is once again converted to a graph **GN** with nodes representing regions and edges with weight proportional to the activation map value and relative distance between nodes:
with F defined as before. With the outbound edges weights normalised to 1, the Markov chain as states and weights as transition probabilities. In the equilibrium distribution the mass flows to the nodes with high activation value.
This equilibrium distribution is the basis to create the output saliency map. We experimental research shows that this algorithm predicts human fixations on the salient regions more reliably than other tested algorithms (including Itti/Koch algorithm). Also, the salient regions found using this method are more cohesive than with other methods while maintaining high accuracy.


Original image


Saliency Map

Sun Saliency map
We propose a probabilistic formula for saliency. The algorithm can be based on both difference of Gaussians (DoG) feature selection known from Itti/Koch and similar methods (called by authors "biologically plausible linear filters"), or the IC A algorithm known from AIM method. The DoG version of this method begins with the separation of three colour components of an image (red, green and blue). Next, the structures of colour channels and intensity are calculated. Next, the DoG filter is convolved with three colour channels (*I*, *RG* and *BY*) at four σ- scales (4, 8, 16 and 32 pixels) to create 12 feature response maps. In this algorithm it is necessary to measure the probability distribution over the features. The authors decided to use natural image statistics from a collection of natural images, so the next step is to use the twelve feature response maps in conjunction with 138 images of natural scenes to generate the estimation of the probability distribution over the observed values of each of the 12 features. Using the exponential power distribution, the estimated distribution is then parameterised where Γ is the gamma function, θ is the shape parameter, σ is the
scale parameter and f is the response filter output. The feature maps can also be obtained via IC A algorithm applied
to a training set of natural images. In such a case 362 feature maps were generated (from 11×11 pixel patches in 3 colour
channels). The output saliency is calculated from probability logarithm generated from either *DoG* filters or IC A-acquired patches, with the only difference in the span of the summation,where i spans from 1 to 12 in the *DoG* generated filters or from 1 to 362 in the IC A features case.


Original image


Saliency Map

The results are summarized in all Among 3 comparative algorithms, the Hamming distance between Lab-signature descriptors correlates best with reaction times. That is, among the methods tried here, the perceptual distance between change blindness pairs is best explained by the image signature descriptor. Given our understanding of the connection between foreground information and the signature, a difficult change blindness trial is likely one in which the removed object is perceived as

part of the background, for in such a trial, we expect a small signature distance.

Conclusion

We introduced the image signature as a simple yet powerful descriptor of natural scenes. We proved on the basis of theoretical and experimental arguments that this descriptor can be used to approximate the spatial location of a sparse foreground hidden in a spectrally sparse background. predicting them better than leading saliency algorithms at a fraction of the computational cost. We also provided experiment in which the perceptual distance between slightly different images was predicted most accurately by the image signature descriptor. In future we are investigate experimentally the other saliency algorithm which will be giving better result with less time

REFERENCES

[1] C . Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry.," Human Neurobiology 4, pp. 219–227, 1985.

[2] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis.," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, p. 1254–1259, 11 1998.

[3] X. Hou and L. Zhang, "Dynamic Visual Attention: Searching for coding length increments," in Advances in Neural Information Processing Systems, D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, Eds., 2009, pp. 681–688.

[4] L. Zhang, T. K. Marks and M. H. Tong, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," Journal of Vision, vol. 8, no. 7, pp. 1–20, December 2008.

[5] A. G. Leventhal, Ed., The Neural Basis of Visual Function: Vision and Visual Dysfunction, vol. 4., Boca Raton, Fl: CRC Press, 1991..

[6] S. Engel, X. Zhang and B. Wandell, "Colour Tuning in Human Visual Cortex Measured With Functional Magnetic Resonance Imaging," Nature, vol. 388, no. 6,637, p. 68–71, July 1997.

[7] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in IC ME '06: Proceedings of IEEE international conference on multimedia and expo, 2006.

[8] N . D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," Journal of Vision, vol. 9, no. 3, pp. 1–24, March 2009.

[9] J. Harel, C. Koch and P. Perona, "Graph-Based Visual Saliency," Proceedings of Neural Information Processing Systems (NIP S)., 2006.

[10] X. Hou, J. Harel and C. Koch, "Image Signature: Highlighting Sparse Salient Regions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pp. 194–201, January 2012.

[11] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," International Journal of Computer Vision, vol. 42, no. 3, pp. 145–175, 2001.

[12] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[13] R . Achanta, F. Estrada, P. Wils and S. Susstrunk, "Salient region detection and segmentation.," in International Conference on Computer Vision Systems, 2008.

[14] R . Achanta, S. Hemami, F. Estrada and S. Susstrunk, "Frequencytuned Salient Region Detection," IEEE International Conference

on Computer Vision and Pattern Recognition (CVPR 2009),
pp. 1597–1604, 2009.
[15] Y. Fang, Z. Chen, W. Lin and C.-W. Lin, “Saliency-based image retargeting
in the compressed domain,” in Proceedings of the 19th ACM
international conference on Multimedia, New York, 2011.
[16] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang and S.-M. Hu, “Global
contrast based salient region detection,” in Proceedings of the
2011 IEEE Conference on Computer Vision and Pattern Recognition,
Washington, DC, 2011..