I.J. Image, Graphics and Signal Processing, 2015, 8, 58-65
Published Online July 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2015.08.07

Scale Space Reduction with Interpolation to Speed up Visual Saliency Detection

Omprakash S. Rajankar
Ph.D. Research Scholar, MPSTME, NMIMS University, Mumbai, India
Email: online.omrajankar@gmail.com

Uttam D.Kolekar
Principal, A.P. Shah Institute of Technology, Thane, India
Email: uttamkolekar@gmail.com

***Abstract*—**The scale of salient object in an image is not a known priori, therefore to detect salient objects accurately multiple scale analysis is used by saliency detection models. However, multiple scale analysis makes the saliency detection slow. Fast and accurate saliency detection is essential to obtain Region of Interest in image processing applications. This paper proposes a scale space reduction with interpolation to speed up the saliency detection. To demonstrate the concept, this method is integrated with Hypercomplex Fourier Transform saliency detection which reduced the computational complexity from *O(N) to O(N/2)*.

***Index Terms*—**Hypercomplex Fourier Transform, Interpolation, Scale Space Analysis, Object scales, Visual saliency.

I. Introduction

Region of Interest (ROI) in the image are marked manually or it may be detected automatically from the image features. In general, fast and accurate ROI detection without human interaction is desirable in image processing applications such as object segmentation, object recognition and adaptive image compression. Visually interesting objects naturally draw our attention are called salient objects. The topographical representation of salient objects in an image is called as saliency map. Thus, saliency maps are gray scale images with larger pixel values for more salient points. ROI mask, which is useful in various applications, can then be derived by applying suitable threshold on the saliency map.

Visual salience is the state or distinct subjective perceptual quality which forces some items in the world that pop out from their adjacent items and quickly grabs visual attention [1]. Such as a colorful rainbow in the sky that makes it pop out from its surroundings and draw attention in an automatic and rapid manner. Similarly some parts of the image are more interesting or visually important, that catch visual attention. Thus, salient regions are described as uncommon or unique samples, of the image having distinct features. The regular samples of the image may then be termed as non-salient regions. Computing visual salience has been a topic of interest since last 25 years. Systematic scanning of images from left-to-right and top-to-bottom is the traditional method of attempting to locate objects of interest. Whereas, visual salience provides relatively rapid mechanism to select a few most likely objects and overlook the clutter present in an image [2]. The object scale and their spatial frequencies are associated [3]. Objects in an image become meaningful depending on its scale and distance of observation. For example, a green patch of land in the satellite image may turn out to be a forest in a closer look at the land from an airplane. An even closer look from the top of a building can show constituents of it such as trees, branches, and leaves. The green patch is observed at a coarse scale while the constituents are visible at the much finer scales. Analogous to objects in the real world, details in an image exist over a limited range of resolution. Thus, the concept of scale is very important while processing the image to highlight objects in it. However, scale of salient object in an image is not known in advance. Therefore to detect salient objects accurately multiple descriptions of the image becomes necessary. In the multiple scale analysis, a set of saliency maps is obtained and final saliency map is selected based on minimum entropy value. Scale space or the set of saliency maps is proportional to the size of the image, for example in the case of 128X128 and 256X256 size images the scale space used is 8 and 16 respectively. These multiple descriptions of the image are computationally complex and makes the saliency detection slow.

In the literature, saliency detection is done in different ways. Using simply the low-level image features such as intensity, contrast, color and orientation called bottom-up saliency detection or using some kind of task such as face recognition called as top-down saliency detection. Top-down saliency detection is much slower than bottom-up saliency detection [4]. Bottom-up saliency detection may be of spatial biological, computation or hybrid type. Frequency domain computational models of saliency detection are comparatively fast. Hence, frequency domain saliency detection models are becoming popular. Still they are not either not fast enough to match the requirement of real-time applications or are not sufficiently accurate.

Computationally efficient method called frequency-tuned saliency detection was proposed by Achanta et.al. in [5], which used a bandpass filter of appropriate bandwidth. A saliency map contains the wide range of frequencies, so the wide bandpass filter is formed by combining the outputs of several narrow band pass filters with contiguous pass bands. A wide band filter can include background and noise in the salient region. Thus, to avoid inclusion of background and noise, and to obtain only salient objects, narrow bandpass filters should be used. The accuracy of this method depends on proper tuning of the bandpass filters used, again the question arises how to tune the bandpass filters when the sizes or scales of salient objects are not known a priory.

The state of art visual saliency model of Itti and Koch [6] is a feature integration model. In this model, among all of the chosen features, saliency maps are generated by extracting the feature strength at several scales. The center-surround approach combines them to highlight the salient regions. Then, the individual feature saliency maps are summed to generate a master saliency map. Winner take all stage decides final saliency map. General architecture of this model is a massively parallel implementation to speed up saliency detection. This model forms the basis for other spatial saliency detection models.

SR[7], PFT and PQFT[8], HFT [9] and HSC[10] are the state of art frequency domain saliency detection models. The Computational models of saliency detection employ Fast Fourier Transform (FFT) or quaternion hypercomplex Fourier transforms (QHFT). FFT or QHFT of an image results in the amplitude and phase spectrum. The amplitude spectrum of image consists of all the frequency components present in the image. In the case of HFT model, the saliency map can be obtained by convolution of the amplitude spectrum with Gaussian kernel of a right scale, maintaining the other information unchanged. Fast and accurate saliency detection is the requirement for real-time applications in which ROI is used for further image processing. Scale space reduction is the solution to this problem, but it may affect the accuracy of the saliency map. Modules in [11] attempts to reduce the spectral scale space using bisection search, but occasionally the search trap in local minima. Heuristics search modules proposed in [12] is one of the solution to reduce the local minima trap, while it generate and search the saliency maps. This paper proposes a novel method of scale space reduction with interpolation to speed up the visual saliency detection while attempting to preserve the accuracy of saliency model. To demonstrate the concept, the proposed method is integrated with Hypercomplex Fourier Transform based saliency detection of Jian Li et. al. [9]. The reduction in the computational complexity achieved is *O(N/2).*

II. Background

A. Object Scale and spatial frequency

There is a close relation between the object scale and spatial frequency. In image low frequencies represent large objects and the background of the image; high frequencies define object boundaries; highest frequencies specify noise, coding artifacts and texture patterns; whereas medium frequencies generally represent the most important parts called as visual salient objects. In order to represent objects of different sizes, it is necessary to convolve a given image *I(x,y)* with 2D Gaussian kernel *g(x,y;σ).* Where, σ defines the width of the Gaussian kernel. In statistics, when we consider the Gaussian probability density function, we call it as standard deviation and σ^2 as variance. The series of Gaussian kernels for low pass filter is given by.

$$g(x,y;\sigma)=\frac{1}{\left(2\pi\sigma^2\right)}e^{-\left(x^2+y^2\right)/\left(2\sigma^2\right)} \tag{1}$$

Where, the term $1/\left(2\pi\sigma^2\right)$ is the normalization constant. With the normalization, the constant integral of the Gaussian kernel over its full domain is unity for every σ. Thus, the amplitude of the Gaussian kernel decreases rapidly with an increase in σ.

The Gaussian scale-space representation, $L(x,y;\sigma)$ is the convolution of the image *I(x, y)* with Gaussian kernel $g(x,y;\sigma)$ written as,

$$L(x,y;\sigma)=g(x,y;\sigma)*I(x,y) \tag{2}$$

Typically, only a finite discrete set of levels of *L* for $\sigma>0$ represents the scale-space. For, $\sigma=0$ $g(x,y;0)$ is an impulse function and $L(x,y;0)$ is simply the original image. Further, the Fourier transform of Gaussian *F(u,v)* is also Gaussian, therefore it can generate the linear scale-space representation that too, without introduction of new structures at the coarse scales [13][14][15].

The Fourier transform of Gaussian is given by,

$$F(u,v)=e^{\left(-2\pi^2\ \left(u^2+v^2\right)\sigma^2\ \right)} \tag{3}$$

As σ increases, smoothing of *I(x,y)* with a larger Gaussian filter results in removing high-frequency details. The object scale and cutoff frequency of low pass or band pass filters are very much related. The fineness in boundaries and the amount of high-frequency details required decides the higher cutoff frequency of the filters. The bandwidth of the filters should be narrow to highlight salient objects; at the same time, it should suppress background, noise, coding artifacts, texture, and repeated patterns. The use of the band-pass filter can also control the lower cutoff frequency to narrow the bandwidth. The larger is the scale of the object; the lesser is the lower cutoff frequency f_{LC} of the band pass filter. For salient object detection, it is necessary to highlight the entire object. For edge detection, f_{LC} is kept high which narrows down the bandwidth [16]. Similarly, a narrow bandwidth is used for corner and interest point detection [17], [5]. Nevertheless, it is difficult to detect the scale of large elongated object. If we know it in advance, segmenting the object can solve the problem to some extent else using all of the low-frequency contents is the only solution.

In fact, there is no way to know a priori what scales are appropriate for describing the objects in an image, hence, descriptions of the image at multiple scales is considered for salient object detection in hypercomplex Fourier transform saliency detection method.

B. Saliency detection using Quaternion Hypercomplex Fourier Transform

Hypercomplex Fourier Transform proposed in [18], [19], overcome the limitation of the traditional Fourier Transform; it can exploit the correlation between the three-color components and process color image in a holistic manner. For color images, to define the Hypercomplex Fourier Transform, the hypercomplex numbers specifically quaternion are used as.

$$f(n,m) = q_1 + q_2 i + q_3 j + q_4 k \tag{4}$$

Here q_1 is scalar part and $q_2 i + q_3 j + q_4 k$ is a vector part in which q_1, q_2, q_3 and q_4 are real numbers and i, j, k are complex operators that obeys,

$$i^2 = j^2 = k^2 = ijk = -1 \tag{5}$$

And $$ij = k, jk = i, ki = j, ji = -k, kj = -i, ik = -j \tag{6}$$

Given a hypercomplex matrix $f(m,n)$, (7) and (8) gives the left sided discrete version of the HFT and inverse HFT respectively. Both $F_H[u,v]$ and $f(m,n)$ are hypercomplex matrices.

$$F_H[u,v] = \frac{1}{\sqrt{MN}} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} e^{-\mu\pi\left(\left(\frac{nu}{N}\right)+\left(\frac{mv}{M}\right)\right)} f(m,n) \tag{7}$$

$$f(m,n) = \frac{1}{\sqrt{MN}} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} e^{+\mu\pi\left(\left(\frac{nu}{N}\right)+\left(\frac{mv}{M}\right)\right)} F_H[u,v] \tag{8}$$

Where, μ is unit pure quaternion and $\mu^2 = -1$ [9].

C. Representation of Feature Maps

The features considered for saliency detection in the HFT model of [9] are intensity and two opponent color space representations. A fourth feature vector of the quaternion can extend the saliency model to incorporate motion in video saliency detection. Feature maps f_2, f_3 and f_4 are computed from the three-color channels r, g, b of the image as,

$$f_2 = \frac{r+g+b}{3} \tag{9}$$

$$f_3 = \left(r - \frac{g+b}{2}\right)\left(g - \frac{r+b}{2}\right) \tag{10}$$

$$f_4 = \left(b - \frac{r+g}{2}\right) - \left(\frac{r+g}{2} - \frac{|r-g|}{2} - b\right) \tag{11}$$

The feature maps describe hypercomplex matrix $f(m,n)$ as,

$$f(m,n) = w_1 f_1 + w_2 f_2 i + w_3 f_3 j + w_4 f_4 k \tag{12}$$

Where w_{1-4} and f_{1-4} are the weights and the feature maps. In which, $f_{2,} f_{3,} f_4$ are features of the image and f_1 is a motion feature. For the static image, weight of motion feature is set as zero. Weights of image features like intensity f_2 and opponent colors f_3 and f_4 are set as $w_2 = 0.5$, $w_3 = 0.25$, $w_4 = 0.25$ respectively.

Polar form representation of $F_H[u,v]$ in (7) is as follows,

$$F_H[u,v] = \|F_H(u,v)\| e^{(\mu\phi(u,v))} \tag{13}$$

Where $\|F_H(u,v)\|, \phi(u,v)$ and $\mu(u,v)$ are called as amplitude spectrum $A(u,v)$, phase spectrum and a pure quaternion matrix or the eigenaxis spectrum respectively. For handling amplitude spectra, rewriting (1) as a series of Gaussian kernels for low pass filter at different scales *k*:

$$g(u,v;k) = \frac{1}{\sqrt{2\pi}\, 2^{k-1} t_0} e^{-(u^2+v^2)/(2^{2k-1} t_0^2)} \tag{14}$$

Convolution of the amplitude spectrum $A(u,v)$ with Gaussian kernels creates smooth Spectral scale space (SSS). $\Lambda = \{\Lambda_k\}$ represents SSS.

$$\Lambda(u,v;k) = \left(g(.,.;k) * A\right)(u,v) \tag{15}$$

Where, k is the scale parameter $k = 1,....K$ and it is determined by the image size. For given smoothed amplitude spectrum Λ_k and the original phase and Eigen-axis spectra, (17) perform the inverse transforms to give the saliency map at each scale: S_k

$$S_k = g * F_H^{-1} \left\| \left\{ \Lambda_k (u,v) e^{XP(u,v)} \right\} \right\|^2 \tag{16}$$

Here, g is Gaussian kernel at fixed scale. Fig. 1 shows a series of saliency maps S_k obtained, for Gaussian filtering of a 128x128 image at scales k=1 to 8. On a saliency map, salient regions are in the highlighted form and rest of the map is in suppressed form. It also reveals that when the scale is very small, the information contained in the amplitude spectrum plots retains quite well, but when it becomes very large, the pertinent information is lost [17]. With an expectation that the histogram of saliency map clusters around certain values, its entropy would be very small. With this assumption, in the HFT saliency model the final saliency map S is chosen from the set of saliency maps S_k for a specific scale based on minimum entropy.

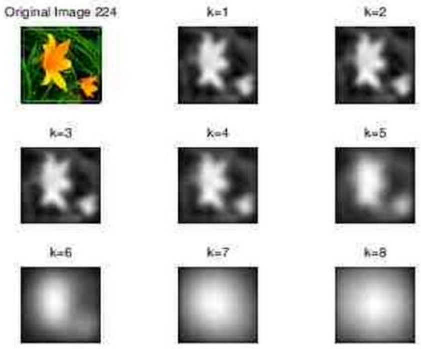

Fig. 1. Image filtered at eight different scales with Gaussian filter.

D. Model Parameters

The comparison of estimated saliency map S and the ground truth map G evaluates the saliency model. By inserting ones at eye fixation locations in a zero matrix, we built a fixation map. Smoothing of the fixation map with a Gaussian kernel produces the ground truth, map. A perfect ground truth map would come from an infinity observers. MIT Saliency benchmark is created from 39 viewers, it has a performance of 0.899 which captures about 95% of ground truth, and is assumed an accurate approximation [20]. To detect salient object regions in a scene, an algorithm should respond uniformly throughout the salient region. Ground truth map segment the entire portion of salient object, hence we use ground truth for comparison.

For analyzing saliency models, we considered more than one evaluation scores to make the conclusions independent of the choice of metric. Area Under the receiver's operating Characteristics (AUC) and Dice Similarity Coefficient (DSC) measures the performance of saliency detection algorithms and as a measure to evaluate the overlap between the thresholds applied saliency map and the ground truth respectively; The Peak value of the DSC (PoDSC) corresponds to the optimal threshold, and it is the best possible evaluation score [18].

Area under Curve (AUC): To obtain the ROC curve, G and predicted S are converted to binary using fixed and varying thresholds, respectively. ROC curve is then plotted as the true positive rate against the false positive rate averaged over the set of images and over distortion levels, if $G \in \{0.1\}$, $S \in \{0.1\}$ and the common points of saliency between the two are, $A = G \cap S$ Equations and (18) compute the True Positive Rate (TPR) and False Positive Rate (FPR) as:

$$TPR = \frac{N(A)}{N(G)} \tag{17}$$

$$FPR = \frac{(N(s) - N(A))}{(T(G) - N(G))} \tag{18}$$

Where the operator $N(X)$ gives the number of ones in X and $T(X)$ gives the total number of elements in X. AUC of ROC curve is then calculated [19]. The area under the curve indicates how well the saliency map predicts actual human eye fixations. AUC score of one indicate correct prediction whereas a score of 0.5 corresponds to chance level [21].

Linear Correlation Coefficient (CC): Pearson's linear coefficient CC measures the strength of linear correlation between two variables

$$CC(G,S) = \text{cov}(G,S) / \sigma_G \sigma_S \tag{19}$$

Where σ_G and σ_S are standard deviations of G and S maps, respectively. A CC score of ±1.0 indicate perfect linear relationship whereas a score of zero correspond to no linear correlation between two variables [22].

Normalized Scan path Saliency (NSS): NSS is the average value of the normalized saliency map at human eye fixation locations [23].

$$NSS = \frac{1}{N} \sum_{p-1}^{N} \frac{s_p - \mu_s}{\sigma_s} \tag{20}$$

Where the total number of eye fixations is N, the specific location of fixation *is* p and the saliency map variance is μ_s . A NSS score of one indicates that the subject look into the region whose predicted saliency is more than average by one standard deviation, whereas a score of zero corresponds to a chance in predicting human gaze.

Similarity measure (SIM): It measures the similarity between two different saliency maps when viewed as distributions [23]. It is the sum of the minimum values at each point in the distributions. The similarity between two maps S and G is obtained as

$$SIM = \sum_{x-1}^{X} \min(S(x), G(x)) \tag{21}$$

Where

$$\sum_{x-1}^{X}\{x-1\}S(x)=\sum_{x-1}^{X}G(x)=1 \tag{22}$$

A SIM score of one indicate that the distributions are identical whereas a score of 0.5 corresponds to completely different distributions.

The Earth Mover's Distance (EMD): EMD is a measure of the distance between two probability distributions over a region [20], [24]. EMD computes the minimal cost to transform the probability distribution of the saliency maps S into the one of the human eye fixations G [25] as:

$$EMD=\min_{f_{i,j}}\sum_{f_{i,j}} f_{i,j}d_{i,j}+\left|\sum_i G_i-\sum_j S_j\right|\alpha\max d \tag{23}$$

Where $f_{i,j}$ represents the amount transported from the i^{th} supply to the j^{th} demand. The ground distance between bin i and bin j in the distribution is d_{ij} An EMD score of zero indicates that the distributions are identical, whereas a score of one corresponds to no overlap and distributions are completely different [25].

III. Interpolation

Interpolation is the process to construct new data points within the range of a discrete set of known data points. Any two points in the plane, $(p_1,\ q_1)$ and, $(p_2,\ q_2)$ with $p_1 \neq p_2$ determine a unique first-degree polynomial in p whose graph passes through the two points. It may be generalized to more than two points. For given n points in the plane, $(p_k,\ q_k)$, $k=1,\ \ldots,\ n$, with distinct points, there exist unique polynomials in p of degree less than n that can connect all the points through a smooth curve [26]. This polynomial is called the interpolating function or interpolant. In general, there are three interpolation functions; linear, polynomial and cubic spline for each of the intervals. High-degree polynomials take more time to interpolate. Spline interpolation is used in this algorithm. Spline interpolation uses low-degree polynomials in each of the intervals and chooses the polynomial pieces such that they fit smoothly together. The resulting function is called as a spline. The natural cubic spline is piecewise cubic and twice continuously differentiable. Furthermore, its second derivative is zero at the end points. In this experiment only four saliency maps out of eight are obtained at scales 1,3,6 and 8. At the intermediate scales saliency maps are not obtained, instead they are predicted using the interpolation method. Fig. 2 compares true values of entropy with interpolated values for the two arbitrary sample images *imageA* and *imageB* from a MIT database at eight different scales. It is observed that interpolated values of entropy approach the true values of entropy at the intermediate scales 2,4,5 and 7.

Hence, the concept of interpolation is explored to reduce the number of saliency maps in the intermediate stage to make the saliency detection faster. Fig. 3 illustrates the procedure to obtain optimum saliency map with the proposed model. First of all color image is resized to 128X128 to make saliency detection time independent of image sizes. In the second stage, low-level feature matrices such as intensity and color difference are combined to form a quaternion hypercomplex matrix. In the third stage quaternion, Hypercomplex Fourier transform is applied to the quaternion hypercomplex matrix to obtain amplitude spectrum, phase spectrum and Eigen-axis spectrum. With this transform, whole color images can be transformed, rather than as color separated components. These three stages are same as Hypercomplex Fourier Transform (HFT) saliency detection model of [9]. In next stage, Gaussian kernels smooth the amplitude spectrum.

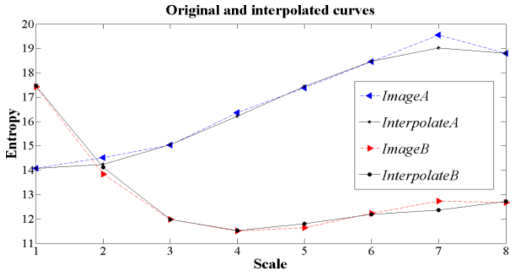

Fig. 2. Interpolation approach for a sample image from MIT Saliency Benchmark

Then, we obtained saliency maps for alternate scales. Entropy is calculated for the derived saliency maps. With the cubic spline interpolation, intermediate entropy values are obtained. Then minimum value of entropy predicts the scale of the optimum saliency map. To verify its correctness actual saliency map and its entropy is derived at the predicted scale. An optimum saliency map with minimum entropy is finally selected. Coding algorithm for interpolation approach is given below.

Algorithm:QHFT Visual saliency with Interpolation

Input: The resized color image C with a resolution $m \times n$

Output: Saliency map $S\ (m,\ n)$ of $C\ (m,\ n)$

Steps 1-3 are according to HFT algorithm of [9].

1) Calculate the feature maps $\{f_2, f_3, f_4\}$ form $C\ (m,\ n)$.
2) Combine these features to obtain the hypercomplex matrix $f(m,n)$.
3) Compute amplitude spectrum of $f(m,n)$ by taking its HFT. Preserve the exponential term which consists of the phase and Eigenaxis spectrum.
4) Smooth the amplitude spectrum with Gaussian kernels, according to (2) and obtain saliency maps according to (3) for a specific scale selected by the following interpolation approach. Let $k=1,\ 2,\ 3 \ldots N$ where N the maximum scale,

obtained from image size. For example for images of 128X128 sizes, *N* is considered as eight.

i. Derive saliency maps and calculate entropies for alternate scale for example $\{1,\ 3,\ 6\ and\ 8\}$. This set is called as true saliency set.
ii. Interpolate above set to predict entropies at intermediate scales $\{2,\ 4,\ 5\ and\ 7\}$ with the cubic spline method.
iii. Find the minimum value of entropy $E_{\min}$ and corresponding scale k .
iv. Derive actual Saliency map and its entropy for scale k only if it belongs to interpolated set. Include this saliency map into a set of true saliency maps.
v. Final optimum saliency map S is the map corresponding to minimum entropy map from a set of true saliency maps.

5) Return S.

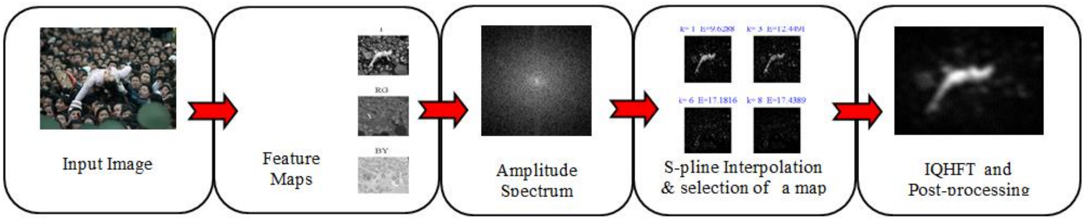

Fig. 3. Procedure to obtain an optimum saliency map

Sample image from MSRA Salient Object Database used in [27] are used to understand the algorithm. Fig. 4 illustrates the progress of the interpolation approach when minimum entropy is part of interpolated set. To verify its accuracy, the actual saliency map is derived at this scale. If the actual entropy found is not a minimum, then the final optimum saliency map is selected from the set of true saliency maps with a minimum entropy value. So in worst case $N/2+1$ is the scale space size. A red border shows the final saliency map, in the Fig. 4. In the second example, after interpolation the minimum entropy found belongs to the true set of entropy. Thus, the saliency map with minimum entropy belongs to the set of true entropy as shown in Fig. 5.

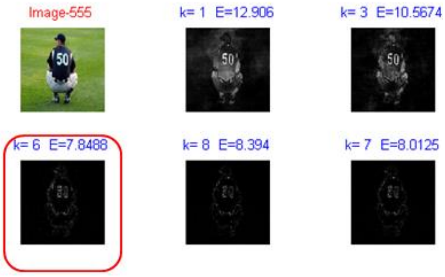

Fig. 4. Progress of interpolation approach when minimum entropy belongs to interpolated set

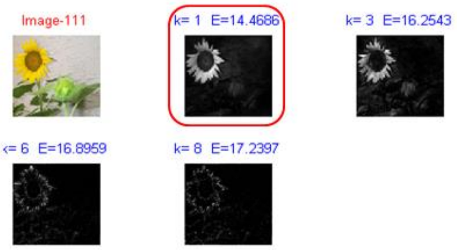

Fig. 5. Progress of interpolation approach when minimum entropy is in true set

IV. Experimentation And Results

In the experiments of saliency detection in natural images, to evaluate the performance; saliency map S generated by the three state of art saliency models Itti [6], HFT [9] AC [5] and the proposed model are directly compared with the human-labeled salient regions called as ground truth G . The test images and respective ground truth are from the ImgSal V1.0 database by Jian Li [28] which has 235 images of different category. Evaluation is also done with the Achanta's Salient Object database which has about 1000 images and accurately derived ground truth from the one presented in [29]. For analyzing saliency models, more than one evaluation scores are considered to make the conclusions independent of the choice of metric.

Fig. 6 and Fig. 7 shows the average Receivers Operating Characteristic (ROC) curves plotted as the true positive rate (TPR) against the false positive rate (FPR) averaged over the set of images and over distortion levels, for the proposed and the state of art models of saliency detection. From the figures we can see that, the proposed model HFT-S achived the highest AUC scores of 0.94 and 0.90, for the two databases *ImgSal* and *Achanta* respectively and are the same as the HFT model. The ROC curves and the AUC score of this model are better than the other state of art models. In general, the results of HFT and HFT-S models are consistent for the two databases, which proves the reliability of ROC curve and AUC score.

All the models were run on the same Pentium i-V 3.2GHz machine with 3GB RAM. The proposed algorithm of HFTs was implemented in MATLAB 13. The databases, including saliency maps and ground truth for Achanta's (AC) and Itti's models were downloaded from the site:

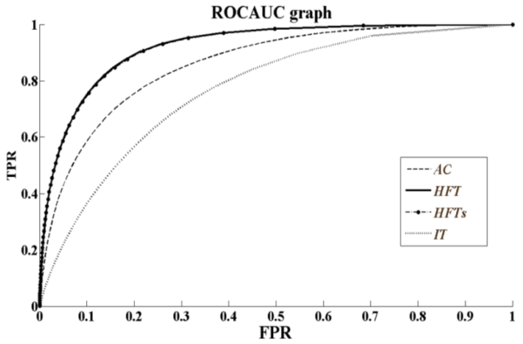

Fig. 6. Area under the receiver's Operating Curve for ImgSal V1.0 database

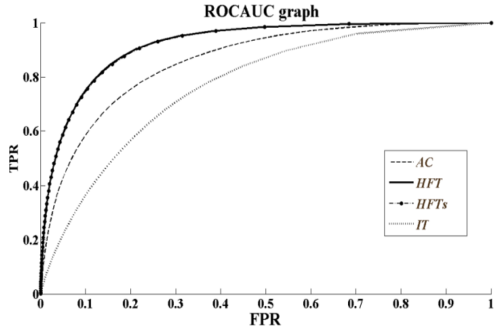

Fig. 7. Area under the receiver's Operating Curve for Achanta database

http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/GroundTruth/binarymasks.zip

http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/saliencymaps.zip

The database ImagSal was downloaded from http://www.escience.cn/people/jianli/DataBase.html .

Table 1 and Table 2 list the average evaluation scores AUC, PoDSC CC, NSS, SIM, EMD and time to obtain a saliency map of 128X128 resized color images for the two databases ImgSal and Achanta respectively. It is observed that evaluation scores of both HFT and HFT-S are better than AC and Itti. The evaluation scores of the proposed model are close to that of the HFT model. Whereas, the time required to detect saliency is reduced up to half as compared to HFT.

Table 1. Evaluation Scores for Database IMGSAL

Model	AUC	PoDSC	CC	NSS	SIM	Time to compute Saliency map (s)
Achanta(AC)	0.838	0.411	0.46	0.13	0.596	0.032
HFT	0.944	0.587	0.669	0.163	0.6305	0.234
HFT-S	**0.943**	**0.584**	**0.668**	**0.163**	**0.629**	**0.146**
Itti	0.899	0.462	0.518	0.126	0.579	1.1

Similar to HFT, AUC scores of HFT-S is close to one, which indicates the correct prediction of human eye fixation. A CC score of 0.668 indicates a reasonable linear relationship. An NSS score of the proposed method is greater than the other state of the art methods indicates that the subject look into the region. A SIM score of 0.629 indicates that the distributions are almost identical.

Table 2. Evaluation Scores for Database Achanta

Model	AUC	PoDSC	CC	NSS	SIM	Time to compute Saliency map (s)
Achanta(AC)	0.867	0.625	0.456	0.168	0.596	0.032
HFT	0.919	0.691	0.651	0.242	0.640	0.234
HFT-S	0.908	0.675	0.625	0.232	0.629	0.146
Itti	0.776	0.518	0.240	0.093	0.467	1.1

Interpolation approach is also studied for all the six category images of ImgSal database. The Table 3 lists the accuracy of each category. The average scale selection accuracy found is 82%. It indicates that the proposed method is fairly accurate in all the categories of images.

Table 3. Scale Selection Accuracy for Database IMGSAL

Category Salient Region size (Images)	Accuracy Against HFT (%)
Large (1-50)	92
Medium (51-130)	85
Small (131-190)	68.33
Cluttered backgrounds (191-205)	86.67
Large and Small (206-220)	93.33
Repeating Distracters (221-235)	86.67

V. Conclusion

The scale of salient object is an important factor, but it is not a known priori. In this context, HFT and other saliency detection models use multiple scale descriptions of the image. This paper proposed interpolation approaches to predict the appropriate scale and speed up saliency detection in the framework of HFT model. The interpolation approach only generated alternate saliency maps. Therefore, the proposed method reduced computational complexity from *O (N)* up to *O (N/2)* as compared to the HFT method. Scale selection accuracy of this method is about 82%. The performance scores for the proposed model and state of the art HFT model matched with each other. It concludes that the scale selection of the proposed method is not only accurate, but saliency detection is also faster than HFT model. It is worth noting that this method is free from local minima trap.

References

[1] Laurent Itti, "Visual salience - Scholarpedia," *Scholarpedia*, vol. 2, no. 9, p. 3327, 2007.

[2] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention.," *Vision Res.*, vol. 40, no. 10–12, pp. 1489–506, Jan. 2000.

[3] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *ICASSP '84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1984, vol. 9, pp. 150–153.

[4] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, “Top-down control of visual attention in object detection,” in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, 2003, vol. 1, pp. I–253–6.

[5] R. Achanta, S. Hemami, F. Estrada, S. Sabine, and D. L. Epfl, “Frequency-tuned Salient Region Detection,” *CVPR*, no. Ic, 2009.

[6] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[7] X. Hou and L. Zhang, “Saliency Detection: A Spectral Residual Approach,” in 2007 *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, no. 800, pp. 1–8.

[8] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression.,” *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–98, Jan. 2010.

[9] J. Li, M. D. Levine, X. An, X. Xu, and H. He, “Visual Saliency Based on Scale-Space Analysis in the Frequency Domain.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 1, Jul. 2012.

[10] C. Li, J. Xue, N. Zheng, X. Lan, and Z. Tian, “Spatio-temporal saliency perception via hypercomplex frequency spectral contrast.,” *Sensors (Basel).*, vol. 13, no. 3, pp. 3409–31, Jan. 2013.

[11] O. Rajankar and U. Kolekar, “Fast Visual Saliency Detection with Bisection search to Scale Selection,” in *IEEE Inter- national Conference on ICPC*, 2015, p. 6.

[12] O. Rajankar and U. Kolekar, “Heuristic Approach to Reduce Spectral Scale Space for Fast Visual Saliency Detection,” Image Vis. Comput., (in press),p. 13.

[13] L. M. Florack, B. M. ter Haar Romeny, J. J. Koenderink, and M. A. Viergever, “Scale and the differential structure of images,” *Image Vis. Comput.*, vol. 10, no. 6, pp. 376–388, Jul. 1992.

[14] J. J. Koenderink, “The structure of images.,” *Biol. Cybern.*, vol. 50, pp. 363–370, 1984.

[15] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at di erent scales,” J. Appl. Stat., 1994.

[16] P. P. Jonathan Harel, Christof Koch, *J. Harel, C. Koch*, and P. Perona, “Graph-Based Visual Saliency,” Nat. Rev. Immunol., Jun. 2006.

[17] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. J. Comput. Vis.*, vol. 60, pp. 91–118, 2004.

[18] T. Veit, J. Tarel, P. Nicolle, and P. Charbonnier, “Evaluation of Road Marking Feature Extraction,” in 2008 *11th International IEEE Conference on Intelligent Transportation Systems*, 2008, pp. 174–181.

[19] N. D. B. B. Bruce, J. K. Tsotsos, N. Bruce and J. Tsotsos, and J. K. Tsotosos, “Saliency Based on Information Maximization,” J. Vis., vol. 7, no. 9, pp. 155–162, Mar. 2006.

[20] T. Judd, F. Durand, and A. Torralba, “A Benchmark of Computational Models of Saliency to Predict Human Fixations A Benchmark of Computational Models of Saliency to Predict Human Fixations,” 2012.

[21] A. Borji and L. Itti, “State-of-the-art in visual attention modeling.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[22] U. Rajashekar, A. C. Bovik, and L. K. Cormack, “Visual search in noise: revealing the influence of structural cues by gaze-contingent classification image analysis.,” J. Vis., vol. 6, no. 4, pp. 379–86, Jan. 2006.

[23] R. J. Peters, A. Iyer, L. Itti, and C. Koch, “Components of bottom-up gaze allocation in natural images.,” Vision Res., vol. 45, no. 18, pp. 2397–416, Aug. 2005.

[24] O. Pele and M. Werman, “Fast and robust Earth Mover’s Distances,” in 2009 *IEEE 12th International Conference on Computer Vision*, 2009, pp. 460–467.

[25] O. Pele and M. Werman, *Computer Vision – ECCV* 2008, vol. 5304. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 495–508.

[26] C. Moler, “Interpolation,” in *Numerical Computing with MATLAB*, Electronic., Natick, MA: The MathWorks, Inc., 2004, p. 27.

[27] N. N. Zheng, X. Tang, Liu Tie;, Y. Zejian, S. Jian, W. Jingdong, Z. Nanning, A. Xiaoou, S. Heung-Yeung, T. Liu, Z. Yuan, J. Sun, J. Wang, and H.-Y. Shum, “Learning to detect a salient object.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–67, Feb. 2011.

[28] J. Li, M. Levine, X. An, and H. He, “ImgSal: A benchmark for saliency detection v1.0,” in *Proceedings of the British Machine Vision Conference*, 2011, pp. 86.1–86.11.

[29] Z. W. Z. Wang and B. L. B. Li, “A two-stage approach to saliency detection in images,” *2008 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 965–968, 2008.

Authors’ Profiles

Dr Uttam D. Kolekar received Ph.D. degree from the Bharathi Vidyapeeth University, Pune in Electronics Engineering in 2008. He has been working as Principal at A. P. Shah Institute of Technology, Thane. He has total 23 years teaching experience at various posts such as Lecturer, Assistant professor, Professor, Head of Department, Dean and Director at various Institutes of Mumbai University. He is a research guide at various universities in India and more than 5 candidates are working under him as research scholars from different institutes and Industries like TCS. He has more than 15 papers in national and international Journals and conferences against his name. He is also a reviewer for international journals like IEEE transactions. He is a fellow member of (IETE), Life member of ISTE & ISNT also he is a senior member of the international association of Computer science & information technology. He conducts seminars on mobile Ad-hock network, Mobile computing, GSM, GPRS, and UMTS. He is actively involved as a reviewer for many international conferences. He has addressed many workshops on wireless networks, Sensor networks, and Cyber security.

Mr. Omprakash S. Rajankar has received M.E. (Electronics Engineering) degree from SRTM University, Nanded in 2010. Currently pursuing Ph.D. from NMIMS University. He has 24 years of experience in teaching. Currently he is working as Principal at Zeal Polytechnic, Pune. His areas of interest are Image Processing and VLSI. He is a member of IEEE and life member of ISTE(India). He has 6 papers in International Conferences/journal to his credit.