Modelling Visual Saliency Using Degree Centrality

Rajarshi Pal, Animesh Mukherjee, Pabitra Mitra, and Jayanta Mukherjee
Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur,
India, Pin - 721302
Email: (rajarshi, animeshm, pabitra, jay)@cse.iitkgp.ernet.in

Abstract

Visual attention is an indispensable component of complex vision tasks. In this paper, a multi-scale, complex network-based approach for determining visual saliency is described. It uses degree centrality (conceptually and computationally the simplest among all the centrality measures) over a network of image regions to form a saliency map. The regions used in the network are multiscale in nature with scale selected automatically. Experimental evaluation establishes the superiority of the method over existing saliency methods, even in noisy environments.

1 Introduction

Attention is one of the most important component of primate vision. It is the mechanism to rapidly focus gaze to some selected portions of the visual input. Psychovisual experiments [1] suggest that, in the absence of any external guidance, attention is directed to visually salient locations in the image.

Modelling visually salient locations have been an important research direction over past decade. Feature integration theory [2] that explains various visual search strategies, is the foundation of many of the researches carried on in last couple of decades. Koch and Ullman [3] provides a basic framework to compute attended locations in an image. They suggest parallel extraction of various feature maps from the visual input and find out conspicuity maps according to each feature using center-surround differences. At the end, these maps are combined to get a single saliency map, encoding relative conspicuity of each location in the input scene. Models proposed in [4, 5] follow this framework. In [4], center-surround difference is implemented as the difference between feature map representation at finer and coarser scales. In [5], center-surround difference is implemented based on Difference-of-Oriented Gaussian (DOOrG) model. These models compute saliency based on low-level features such as colour, intensity and orientation. In [6], texture is effectively used to determine saliency where there is very little difference in terms of other features. Approaches described in [7]- [9] incorporate depth information in the computational model. Kadir and Brady [10] demonstrates the need for determining appropriate scale for computing

saliency and proposes a saliency computation method that automatically selects appropriate scale for analysis. There are several other approaches ([11]- [15]) that perform segmentation of input image and compute saliency at object level. Some models ([16, 17]) use Bayesian probability. Researchers have also tried to solve the problem from an information theoretic perspective ([18]- [21]) under the premise that salient regions are those that convey maximum information. Other approaches include selective tuning based [22], evolutionary programming based [23], contrast based [24], subspace estimation and analysis based [25, 26], learning from human eye movement data [27], and spectral residual [28] approaches.

An entirely different approach is stated in [29], where a set of complex networks is formed from the image. Nodes on such a network represent pixels of image and shift of attention is modelled using random walk over such network. Markov chain over such network is defined and equilibrium distribution on the Markov chain gives the saliency map.

In this direction a novel complex network based approach is presented in this paper to determine the visually salient locations of an image. Experiments reported in [2, 30] demonstrate that if a location/object significantly differs from all other locations and specifically from its surround, it is salient, i.e., it draws attention. This motivates to construct networks where nodes represent accumulation of similar pixels and the dissimilarity in terms of features between any pair of such accumulations is encoded as edge-weight between corresponding nodes. This dissimilarity measure is modulated by their positional proximity. Modulation by positional proximity ensures that difference with neighbouring locations gets more weightage. Such networks are constructed over multiple features (intensity and orientation) across multiple scale representation of the image. The network, termed as ViSaNet (Visual Saliency Network), constructed here is a suitable choice for determining salient locations as it combines both local and global conspicuity of a location. Incorporation of degree centrality analysis with this type of network suggests that a centrally situated node belongs to a salient location.

The proposed approach gives improved result than the approaches suggested in [4] and [29]. Moreover, performance of the proposed approach against noisy (zero-mean Gaussian) images is compared with those approaches. The result shows that the proposed approach is superior to [4] and [29] in presence of high noise.

The outline of rest of the paper is as follows: Section 2 depicts the proposed method in details. Section 3 shows the performance of our method compared to [4] and [29]. Robustness of the proposed method against noise is analyzed in section 4. Section 5 focuses on time complexity analysis and free parameter analysis of the proposed method. Finally, section 6 draws the conclusion.

2 Computing Saliency

2.1 Feature Extraction at Multiple Scales

A particular location/object becomes conspicuous when it is dissimilar from its surroundings in terms of one or more low-level features such as, intensity and orientation. Color is not considered here. In this paper, we focus only on monochrome images. Dissimilarity also depends on the scale at which regions are represented. At each scale, representation of input image is in terms of five set of feature maps - intensity, and orientation at four directions (0°, 45°, 90° and 135°). Gaussian Pyramid is a well known structure for representing intensity feature map at multiple scales. A Gaussian Pyramid, I^{ς} (ς varies from *0* to K), is constructed by repeated Gaussian filtering and subsampling of the intensity feature map I^0. Higher values of ς represent coarser scales. It may be noted that the value K is image-dependent. Procedure to fix the value of K for a particular image is stated in next subsection.

$$I^{\varsigma'}(x,y) = h * I^{\varsigma}(x,y)$$

$$= \sum_{\alpha=-\lfloor \delta/2 \rfloor}^{\lfloor \delta/2 \rfloor} \sum_{\beta=-\lfloor \delta/2 \rfloor}^{\lfloor \delta/2 \rfloor} h(\alpha,\beta) I^{\varsigma}(x-\alpha, y-\beta) \quad (1)$$

$$I^{\varsigma+1}(x,y) = I^{\varsigma'}(2x,2y) \quad (2)$$

where h is the 2D Gaussian filter of size δ-by-δ, $*$ denotes convolution operator and $\lfloor a \rfloor$ denotes the greatest integer lesser than or equal to a. If size of I^{ς} is $A \times B$, $I^{\varsigma+1}$ will be of size $\lfloor A/2 \rfloor \times \lfloor B/2 \rfloor$. Orientation at a particular direction θ is obtained at multiple scales through the creation of oriented Gabor pyramid [31] from the intensity feature map.[1]

2.2 ViSaNet: A Network Representation of Feature Map

Visual Saliency Network (ViSaNet) ($G = \langle V, E \rangle$) is constructed for each of the five features (intensity, and orientation at 0°, 45°, 90° and 135°) at each scale ς ($\varsigma \in [0 \ldots K]$). A block of connected and homogeneous pixels is represented by a single node V_i in the ViSaNet. This drastically reduces the number of nodes in the network as compared to number of pixels, and decreases the computation associated to find the edge-weights among these nodes. The homogeneity across a block of pixels in terms of a feature is determined by estimating the difference between maximum and minimum feature values within that block. If the difference between the maximum and the minimum feature values in a block is greater than a particular threshold (typically chosen, in our experiments, as 5.88% (*15/255*-th) of the dynamic range of the feature map), then the block is not homogeneous and is decomposed further using quadtree decomposition technique. Thus, sets of connected and homogeneous pixels are obtained that can be represented by individual nodes. If number of nodes in any of the five ViSaNet formed at a particular scale L is less than a predefined number η then computation at that scale is stopped and the value *(L - 1)* is assigned to K. η is chosen to be 100 in our experiments.

[1] The steerable pyramid in [32] also could have been used for this purpose.

The experiments in [2, 30] reports that a location/object is salient, i.e., draws our attention, if it differs significantly from its surroundings. The ViSaNet incorporates this by considering the edge-weight to be proportional to the difference between the features of those locations. Feature space distance $D_{f_{ij}}$ is estimated as the absolute difference of the mean feature values of these blocks and normalized with respect to the maximum value of feature distance among any pair of blocks.

$$D_{f_{ij}} = |\mu_{f_i} - \mu_{f_j}| / \overset{max}{\forall k, \forall l}(|\mu_{f_k} - \mu_{f_l}|) \tag{3}$$

where μ_{f_x} represents the mean feature value of the block corresponding to node x. The ViSaNet also gives more weightage to edges connecting nodes whose corresponding blocks are spatially close. In other word, the edge-weight is inversely proportional to spatial distance between those blocks. This ensures that the effect of feature space distance $D_{f_{ij}}$ decays with the increase of spatial distance and vice versa. A Gaussian function is used here to simulate the decay of feature influence with spatial distance, much like human foveal retinal vision [33]. The spatial distance $D_{c_{ij}}$ between two blocks corresponding to nodes V_i and V_j is computed as the Cartesian distance between the midpoints of these blocks and normalized with respect to the maximum possible distance among any pair of blocks. In summary, to estimate edge-weights a function $F(V_i, V_j)$ is used that represents distance between blocks corresponding to nodes V_i and V_j. Here, $F(V_i, V_j)$ is defined[2] as follows:

$$F(V_i, V_j) = D_{f_{ij}} \cdot e^{-D^2_{c_{ij}}/2\sigma^2} \tag{4}$$

σ is the standard deviation of the Gaussian function. Though the decay of feature influence with spatial distance is modelled using a Gaussian function, more generalization of this equation is possible. Theoretically, a function that is monotonically decreasing in the range $D_{c_{ij}} > 0$, can replace the Gaussian component of the above equation. The use of a function with higher kurtosis (super-Gaussian) gives relatively more weightage to the feature difference with a spatially close block compared to a function with lower kurtosis (sub-Gaussian). Experiments with super-Gaussian and sub-Gaussian functions in Eq. (4), instead of the Gaussian component, reveal no significant change in results.

2.3 Thresholding Edges in ViSaNet

Difference in feature values causes saliency. As edge weights are proportional to feature dissimilarity, edges with higher weight are of our interest in saliency computation. The subset of edges with weight greater than a particular threshold T is thus retained and other edges are discarded. In order to select the threshold T the distribution of edge-weights between all pairs of vertices is analysed. It is observed that the distribution takes the shape as shown in Figure 1. The aim is to detect a knee in the distribution. Since entropy measures have been used to detect threshold in distribution, we have adopted the approach stated below.

[2]Alternative to this measure, Eberly distance [34] could have been used to estimate the dissimilarity that, along with feature dissimilarity, accounts for both spatial distance and differential form of scale variations.

Let w_i be the weight of an edge E_i. (w_i is as same as $F(V_i, V_j)$). For, a particular threshold t, ratio of summation of weights for discarded set of edges to total set of edges is calculated as:

$$r = \sum_{w_i \leq t} w_i / \sum_i w_i \tag{5}$$

Therefore, the ratio of summation of weights for selected set of edges to total set of edges is (1 - r). Edge-weight entropy En of discarded and selected set of edges is defined as

$$En = -rlog(r) - (1-r)log(1-r) \tag{6}$$

Edge-weight entropy En varies with threshold t (Figure 2). Note that the maximum entropy value corresponds to the knee of the edge-weight distribution curve. Accordingly, the threshold for which edge-weight entropy is maximum is chosen as the edge-weight threshold T.

2.4 Analysing Degree of Nodes in Edge-Thresholded ViSaNet

Degree of a node, which is also a centrality [35] measure, is a structural attribute of the node in a network. It shows positional importance of the node - as it is defined by the number of other nodes to which it is directly connected. It can be easily estimated by row-wise/column-wise summation of the adjacency matrix $A_{i,j}$ of the network. Degree centrality DC_i of node V_i is given by

$$DC_i = \sum_j A_{i,j} \tag{7}$$

In the binarized ViSaNet, as an edge represents the dissimilarity between corresponding locations in the image and it is modulated by positional proximity, the degree of a node encodes how many other locations with which it has significant feature difference (modulated by positional proximity). As saliency of certain location is determined by how dissimilar it is form other locations, specially its surroundings, the degree of a node measures the conspicuity (saliency) of the location corresponding to that node. A conspicuity map is formed by mapping the degree centrality values of all the nodes to the pixels corresponding to their respective locations. Thus conspicuity map is computed for each feature at each scale. Peaks in the maps indicate the conspicuity of the concerned locations.

A typical plot of the degree distribution is given in Figure 3. From the distribution, it can be concluded that most of nodes have low degree and a few nodes have high degree. This nature of distribution is observed throughout experiments with all the images at various scales.

2.5 Combining the Conspicuity Maps for Different Scales and Features

In order to obtain a single saliency map, the conspicuity maps obtained for various features at multiple scales are to be combined. The difficulty in merging the maps is that these are of different dynamic ranges. Moreover, if there are features promoting less number of conspicuous locations than other features, those features (along with lesser conspicuous locations) should be highlighted [36]. This is due to the fact

that feature maps with relatively few conspicuous locations easily discriminate those few locations from other locations, i.e., relative saliency of those few locations are high. On the other hand, feature maps that can not isolate a few locations with high relative saliency, need to be suppressed. A comparative study of various map combination strategies is given in [36]. To inscribe this combination strategy, we, similar to the model proposed in [4], adopt the global nonlinear normalization followed by summation strategy described there. This nonlinear normalization strategy emulates a biological lateral inhibition mechanism, in which neighbouring similar features inhibit each-other [37]. According to this scheme, each conspicuity map C is processed as follows:

1. The values of a map are represented in a fixed range $[0, M]$.

2. Alongside M (which is the maximum of the map C) other local maxima are found and their average $\bar{m}$ is computed. To compute local maxima it is checked whether for some points p there exists some $\epsilon > 0$, such that $C(p) \geq C(q)$ when Euclidean distance between points p and q, $dist(p,q) < \epsilon$. Then values of C at points p, $C(p)$, are called local maxima.

3. The map C is multiplied by $(M - \bar{m})^2$. As a consequence, when in a map the global maximum differs largely from all other local maxima, the location corresponding to the global maxima is promoted. On the other hand, small difference indicates that the map contains nothing unique, and is suppressed.

Then for each feature, across-scale combination[3] of conspicuity maps is performed to determine salient locations in terms of that particular feature. At the end, across-feature fusion is performed to obtain a single saliency map that encodes relative saliency of locations considering all features (in this case, intensity and orientation) at multiple scales.

3 Evaluation

The saliency models are aimed at emulating human vision because of the later's efficiency to handle huge amount of information. To evaluate these models, human-specified salient locations are considered as ground truth data. The similarities/dissimilarities of results of the proposed method from the human-specified salient locations are compared with those of other schemes. This section, firstly, delineates the preparation of groundtruth data from human specified salient locations. Then two evaluation strategies along with the results are stated. A set of 50 images (Figure 4) are randomly selected for the experiment from a larger collection of images of which some are taken from iLab image database [4, 36], some from UCID [38] and some from the Internet.

[3]Alternative to this, usage of a scale-space distance measure such as Eberly distance [34], that account for differential nature of features in scale dimension, could have intrinsically accommodate across-scale merging.

3.1 Preparation of the groundtruth data

Ground truth data was prepared using the assistance of 62 volunteers. An image from our input set was shown to a volunteer for a very short time (100ms). A white canvas of the same size followed the image. She was asked to observe the image and to mark (on the white image) the centres of the locations which seem to be salient to her. Human vision has two components. When confronted with an unfamiliar scene, it is promptly directed to the visually salient locations. This is known as bottom-up component of our vision. With time human vision begins to be guided by recognition/interpretation of the objects/scene it observes. This is called top-down component. This justifies the very short display time of the input image as the aim here is to evaluate a bottom-up saliency model and avoidance of the influence of top-down component as much as possible. Again, human can not perceive a scene presented to her in less than *1/10*th of a second (100 msec). Therefore, presenting the image with less than 100 msec time is of no importance. Combination of these two logics explain the reason for 100 msec presentation time used in our experiments. Opinion of each volunteer was taken for 24 images randomly selected from the input set.

To prepare the ground truth data from the volunteers' markings, the following procedure is adopted for each images of the input set. Let P_x be the set of all points marked by user x for an image I and P be the set of points marked by all users for image I.

$$P = \bigcup_x P_x \tag{8}$$

Clustering is performed on set of points P and mean of spatial distribution μ_c and standard deviation of spatial distribution σ_c is estimated for each cluster c. Circular regions centring at each μ_c and a radius of corresponding σ_c is used to model the salient location. For each input image I, a binary image B (size as same as I) is constructed, where all the pixels encompassed by such circular regions are marked as 1, and all other pixels are marked as 0. Let q be a pixel in B.

$$q = \begin{cases} 1, & \text{if } \exists i, dist(q, \mu_{c_i}) \leq \sigma_{c_i} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$dist(a,b)$ is the Euclidean distance between two pixels a and b.

3.2 Evaluation Strategies and Results

3.2.1 Strategy 1

The basis of evaluation, here, is that vision processes only a few salient locations [39]. It is verified whether the high salient locations match with the circular regions indicated by groundtruth B. Let S be the saliency map obtained for image I and $S(Z)$ represents the collection of values in the saliency map S corresponding to pixel/group of pixels Z. For each circular region R_i in B, maximum of $S(R_i)$ is found and is represented by m_i.

$$m_i = max(S(R_i)) \tag{10}$$

It is checked whether there are other pixels that have saliency value greater than m_i and do not belong to any of the circular locations R_i's in B. Let Γ is the set of such pixels.

$$\forall y, y \in \Gamma \Leftrightarrow \forall j, S(y) > m_i \wedge y \notin R_j \tag{11}$$

If no such pixel exists, then Γ is a null set. If Γ is not the null set, then the error estimate ξ_i for R_i is the sum of normalised distances for all the pixels in Γ from R_i. Distances are normalised with respect to $\sqrt{X^2 + Y^2}$, where image I is of size $X \times Y$.

$$\xi_i = \sum_{y \in \Gamma} mindist(y, R_i)/\sqrt{X^2 + Y^2} \tag{12}$$

where, *mindist(a,A)* is the minimum Euclidean distance of a pixel a from a group of pixels A. The error estimate ξ for the saliency map S is the summation of such error estimates ξ_i for all values of i.

$$\xi = \sum_i \xi_i \tag{13}$$

Some of the saliency maps obtained using the proposed method, and the method in [4] and [29] are shown in Figure 5. The error estimates (Eq. 13) for all the input images are computed as stated above and displayed in Table 1. From the Table 1, one may easily observe that proposed method gives better result compared to the other two in most cases. It is to be noted that the evaluation strategy checks two things:

1. Whether the model predicts some other areas as more salient than the salient locations indicated by Groundtruth data. If, no such area exists, then the error estimate becomes zero.

2. If, such locations exist, then how far these locations are from the salient locations specified by Groundtruth data. Due to this reason the error estimate is pretty high in some cases.

3.2.2 Strategy 2: ROC Analysis

The comparison of performance of these methods is also done using a Receiver Operating Characteristic (ROC) curve which plots true positive rate against false positive rate. Figure 6 shows the mean ROC curves for the proposed method, the method in [4] and the one in [29]. These mean curves are obtained by averaging the loci of the ROC curves for 50 images obtained using each of the methods under comparison. The average area under curves suggests that the proposed method, the method in [29] and in [4] achieve 94%, 89% and 85% of the ROC-area of a human based control. Therefore, according to ROC analysis, the proposed method outperforms the other two.

4 Robustness Against Noise

Next we inspect how the proposed method performs under noisy conditions as compared to some well-established methods of saliency computation. As zero-mean Gaussian noise generates mathematically

tractable models and is used to simulate real-world situations, the robustness analysis is shown against this type of noise. Each of the input image is subjected to a zero-mean Gaussian noise of peak signal-to-noise ratio 10 dB.

The proposed method as well as the methods in [4] and [29] are applied to each of the noisy images. The obtained saliency maps are shown in Figure 7. To quantitatively compare the robustness of these techniques against noise, again errors are estimated (Eq. 13) on the results obtained from noisy images. From the Table 2 it can be concluded that the proposed method works better than those in [4] and [29] in the presence of white Gaussian noise. After carefully inspecting both the Tables 1 and 2, it can also be observed that increase in error due to presence of noise is higher in case of [4] and [29] compared to the proposed degree centrality based method. Therefore, the proposed method is more immune under noisy environments.

Figure 8 shows the mean ROC plots for noisy images for the three methods. The average area under curves suggests that the proposed method, the method in [29] and in [4] achieve 92.5%, 74.25% and 60.5% of the ROC-area of a human based control. Therefore, ROC analysis also suggests that the proposed method is less effected than the other two methods by the incorporated noise.

5 Analysing Time complexity and Free Parameters

5.1 Time Complexity Analysis

We compare execution time of the proposed method with the method in [29], which is also a complex network based approach. Both approaches have two basic steps - forming a complex network representation of an image and deriving saliency map using that representation.

In [29], each pixel is mapped to one node in the network. Complex network formation involves estimating edge and therefore, estimation has to be done for $\binom{N}{2}$ edges for an image with N pixels. In the proposed method, quadtree decomposition is applied to get blocks of pixels with similar feature value and one block is mapped to one node in the network. If it is assumed that N pixels are mapped into n blocks, then there are n nodes in the network and estimation has to be done for $\binom{n}{2}$. As N is much greater than n, the complex network formation process in the proposed method is faster than that of [29]. Table 3 demonstrates some sample execution time[4]of complex network formation from feature maps with N pixels and n blocks.

In [29], saliency is obtained using the equilibrium distribution of the Markov chain (which is formed by normalising weights of outbound edges for each node to 1) derived from the network. Computation of equilibrium distribution involves repeated multiplication of the Markov matrix with an initially uniform vector. For a network with κ nodes, time complexity of this process is in the order of $\mathrm{O}(\kappa^2 k)$, where k

[4]The measurements are taken using MATLAB 7.1 on an Intel Core(TM)2Duo 2.2GHz CPU. cputime command, which returns elapsed CPU time in seconds, is used for this purpose.

in number of iterations to meet the equilibrium. On the other hand, the proposed method only requires degree centrality computation which is in the order of $O(\kappa^2)$. This further speeds up the process. Table 4 gives some sample computational time of equilibrium distribution and degree centrality for a network with κ nodes.

Therefore, the proposed degree centrality based method is faster than the method in [29] in both steps.

5.2 Analysis of Free Parameters

Three free parameters have been used in the proposed method:

- threshold of dynamic range of a block above which quadtree decomposition is done on the block. This value is typically chosen, in our experiments, as 5.88% (*15/255*-th) of the dynamic range of the feature map. Therefore, this parameter is automatically tuned to individual feature maps.
- η to control number of spatial scales.
- σ, standard deviation of the Gaussian function in the formulation of edge-weight [Eq. (4)].

The model in [4] has two free parameters to determine the coarser and finer scales for center-surround differences. The graph based model described in [29] has three free parameters as follows:

- standard deviation of the Gaussian functions in the formulation of edge-weight.
- threshold to check whether equilibrium distribution is reached or not and subsequent controlling of the iteration.
- spatial scales to be used for computation.

Therefore, the proposed model is not using much extra free parameters than the models with which the proposed scheme has been compared. In this context, the notable attempt in [27] must be mentioned where a nonparametric approach to capture visual saliency is proposed.

6 Conclusion

In this paper, a complex-network based approach for determining visually salient locations in an image is depicted. It is a bottom-up approach using low-level attributes (intensity and orientation). It is also a multi-scaled approach as well where selected scales are image-dependent. The most important contribution of this paper is in demonstration of using degree centrality of a node to find visual saliency. Moreover, measuring degree centrality is computationally more efficient than the equilibrium distribution on the Markov chain (computationally equivalent to eigenvector centrality) based approach stated in [29]. The issue of high computational burden related to construction of complex network (as in [29]) is handled

here using a block-based approach. The proposed method gives superior results than methods stated in [4] and [29], even under noisy conditions.

A few things ought to be mentioned at the end. Firstly, as maximization of entropy leads to threshold selection, the selected threshold is image dependent. Noiseless and noisy versions of the experiment register different threshold value for the same image. Secondly, experiments with various threshold values reveal that the entropy maximization based scheme gives good threshold in an overall scenario. But there are cases where the selected threshold is not optimum. Therefore, optimum threshold selection for binarizing ViSaNet can motivate further research. Thirdly, though the proposed degree centrality based approach produces better results for most of the cases, Table I shows a few cases where other methods perform well. Therefore, a detailed interpretation of each of these cases can be a good research direction in future. Some categorization of images may come out to guide vision researchers on the selection of model based on image category.

Acknowledgment

This work was partially sponsored by the MCIT, Govt. of India through their grant no 1(23)/2006-ME&TMD, dated 07/03/2007.

Table 1: Comparison of error estimates [Eq. (13)] of the saliency maps obtained using the proposed method with maps obtained by [4] and [29].

Input image	Proposed method	Itti et al [4]	Harel et al [29]
1	**0**	46.1	**0**
2	2228.7	34003	**3.7**
3	**0**	833.4	**0**
4	**2.1**	965.3	2818.2
5	**346.6**	617.8	469.4
6	**2062.7**	5830.3	6416.2
7	**6885.6**	12172	7661.7
8	164.8	**145.3**	346.1
9	238.1	4727.6	**35.1**
10	**0.5**	2.4	76.4
11	**0**	269.3	66.5
12	**0**	212.5	**0**
13	**0**	**0**	**0**
14	**51.7**	316.2	831.2
15	**16.7**	21349	71.1
16	**8290.8**	14633	23723
17	**0**	**0**	**0**
18	2.8	44.3	**0.6**
19	**85.2**	383.3	327.7
20	**16726**	31614	24854
21	**0**	7637.2	15659
22	**0**	**0**	**0**
23	**226.5**	4673.7	409.3
24	**0**	**0**	2.4
25	410.8	6146.9	**150.4**
26	**68.3**	72.9	6269.9
27	**15.3**	381.8	145.5
28	**0**	**0**	408.7
29	**1.3**	124.2	3429.6
30	**0**	10138	**0**
31	**222.7**	382.1	2985.8
32	**0.1**	16096	11727
33	619.9	**536.7**	840.9
34	**0.9**	7625.8	13800
35	**8.9**	18.7	36.3
36	**5.7**	59.2	15.3
37	**0**	256.1	**0**
38	**0**	1681.9	**0**
39	**0**	**0**	**0**
40	**87**	305.6	152
41	**32.6**	846.2	875.7
42	**4.6**	12826	5306.6
43	**1.2**	6447.6	360.8
44	289.8	**231.5**	2604.5
45	**0**	62.7	58.5
46	460.1	**394.9**	2142.4
47	**0**	**0**	**0**
48	**0**	**0**	**0**
49	0.5	271.2	**0**
50	**5**	5.3	16.7
Average	**791.4**	4107.7	2702

Table 2: Comparison of error estimates [Eq. (12)] of the saliency maps obtained using the proposed method with maps obtained by [4] and [29] on noisy images(10 dB) (Gaussian).

Input image	Proposed method	Itti et al [4]	Harel et al [29]
1	**0**	323.6	1029.4
2	**2351.8**	52896	32769
3	**0**	1319.6	1083.8
4	**2.9**	5966.7	10546
5	**401.1**	1316.6	3637.9
6	**2130.1**	6896	6514.9
7	**7105.8**	12375	12471
8	**172.4**	187.8	910
9	**250**	9027.3	356.1
10	**1.2**	4429	88.32
11	**0**	2679	329.8
12	**0**	2101.3	**0**
13	**0**	**0**	**0**
14	**300.7**	1424.6	972.9
15	**21.1**	23280	85.24
16	**8440.8**	15661	24534
17	**0.3**	4.7	0.3
18	**3.8**	68	4
19	**207.8**	4017.4	394.7
20	**17596**	33112	59412
21	**0**	8792.6	32458
22	**0**	**0**	**0**
23	**387.3**	7791.1	2106.3
24	**0**	**0**	90.17
25	**433.5**	6309.8	451.4
26	**150.5**	1519.7	8103.6
27	**20.9**	1057.1	197.4
28	**0**	**0**	453.4
29	**2.1**	130.9	5587
30	**0**	11762	**0**
31	**309.8**	641.9	3662.7
32	**0.2**	24753	15593
33	**2193.8**	2620.1	5360.6
34	**1.3**	12179	41488
35	**74.2**	111.1	130.2
36	**22.3**	931.7	29.8
37	**0**	424.5	0.1
38	**0**	2553.7	**0**
39	**0**	130.4	**0**
40	**190.7**	423.5	233.5
41	**40.6**	3639.4	1018
42	**10**	53607	11516.6
43	**3.5**	9194.9	486.5
44	**306.4**	314.2	4697.7
45	**0.6**	181.6	157.6
46	**525.7**	621.3	3466
47	**0**	**0**	**0**
48	**0**	**0**	**0**
49	**1.5**	5656.8	97.5
50	**5.5**	5904.9	19.3
Average	**873.3**	6766.8	5850.9

Table 3: Time taken in network formation phase.

Size of feature map	Number of pixels	Time according to [29] (in sec.)	Number of blocks	Time according to proposed method (in sec.)
512x512	262144	1568.30	1210	38.41
256x256	65536	120.64	571	3.58
128x128	16384	4.16	205	0.08

Table 4: Comparison in computational time for equilibrium distribution and degree centrality computation process for a network with same number of nodes.

Number of nodes	Time for Equilibrium distribution (in sec.)	Time for degree centrality (in sec.)
1210	3.69	0.0313
571	2.52	0.0105
205	0.16	0.0009

References

[1] Constantinidis, C., and Steinmetz, M. A.: 'Posterior parietal cortex automatically encodes the location of salient stimuli', The Journal of Neuroscience, 2005, 25, (1), pp. 233-238.

[2] Treisman, A. M., and Gelade, G.: 'A feature-integration theory of attention', Cognitive Psychology, 1980, 12, (1), pp. 97-136.

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

[4] Itti, L., Koch, C., and Niebur, E.: 'A model of saliency-based visual attention for rapid scene analysis', IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20, (11), pp. 1254-1259.

[5] Cheoi, K., and Lee, Y.: 'Detecting perceptually important regions in an image based on human visual attention characteristic', Proc. Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, 2002, pp. 329-338.

[6] Hu, Y., Rajan, D., and Chia, L.-T.: 'Adaptive local context suppression of multiple cues for salient visual attention detection', Proc. IEEE International Conference on Multimedia and Expo, July 2005, pp. 346-349.

[7] Maki, A., Nordlund, P., and Eklundh, J.-O.: 'A computational model of depth-based attention', Proc. 13th International Conference on Pattern Recognition, August 1996, 4, pp. 734-739.

[8] Ouerhani, N., and Hugli, H.: 'Computing visual attention form scene depth', Proc. 15th International Conference on Pattern Recognition, September 2000, 1, pp. 375-378.

[9] Courty, N., Marchand, E., and Arnaldi, B.: 'A new application for saliency maps: synthetic vision of autonomous actors', Proc. of International Conference on Image Processing, September 2003, 3, pp. III-1065-1068.

[10] Kadir, T., and Brady, M.: 'Saliency, scale and image description', International Journal of Computer Vision, 2001, 45, (2), pp. 83-105.

[11] Osberger, W., and Maeder, A. J.: 'Automatic identification of perceptually important regions in an image', Proc. of 14th International Conference on Pattern Recognition, August 1998, 1, pp. 701-704.

[12] Sun, Y., and Fisher, R.: 'Object-based visual attention for computer vision', Artificial Intelligence, 2003, 146, pp. 77-123.

[13] Yu, Z., and Wong, H.-S.: 'A rule based technique for extraction of visual attention regions based on real-time clustering', IEEE Transactions on Multimedia, 2007, 9, (4), pp. 766-784.

[14] Liu, H., Jiang, S., Huang, Q., Xu, C., and Gao, W.: 'Region-based visual attention analysis with its application in image browsing on small displays', Proc. 15th International Conference on Multimedia, September 2007, pp. 305-308.

[15] Aziz, M. Z., and Mertsching, B.: 'Fast and robust generation of feature maps for region-based visual attention', IEEE Transactions on Image Processing, 2008, 17, (5), pp. 633-644.

[16] Itti, L., and Baldi, P.: 'Bayesian surprise attracts human attention', Advances in Neural Information Processing Systems, 2005, 18. [http://books.nips.cc/papers/files/nips18/NIPS2005_0199.pdf]

[17] Begum, M., Mann, G. K. I., and Gosine, R. G.: 'A biologically inspired bayesian model of visual attention for humanoid robots', Proc. 6th IEEE-RAS International Conference on Humanoid Robots, December 2006, pp. 587-592.

[18] Kohonen, T.: 'A computational model of visual attention', Proc. International Joint Conference on Neural Networks, July 2003, 4, pp. 3238-3243.

[19] Bruce, N. D. B.: 'Features that draw visual attention: an information theoretic perspective', Neurocomputing, 2005, 65-66, pp. 125-133.

[20] Bruce, N. D. B., Tsotsos, J. K.: 'Saliency based on information maximization', Advances in Neural Information Processing Systems, 2005, 18. [http://books.nips.cc/papers/files/nips18/NIPS2005_0081.pdf]

[21] Renninger, L. W., Coughlan, J., Verghese, P., and Malik, J.: 'An information maximization model of eye movements', Advances in Neural information Processing Systems, 2004, 17. [http://books.nips.cc/papers/files/nips17/NIPS2004_0869.pdf]

[22] Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., and Nuflo, F.: 'Modeling visual attention via selective tuning', Artificial Intelligence, 1995, 78, (1-2), pp. 507-547.

[23] Stentiford, F. W. M.: 'An evolutionary programming approach to the simulation of visual attention', Proc. 2001 Congress on Evolutionary Computation, May 2001, 2, pp. 851-858.

[24] Ma, Y.-F., and Zhang, H.-J.: 'Contrast-based image attention analysis by using fuzzy growing', Proc. 11th ACM International Conference on Multimedia, November 2003, pp. 374-381.

[25] Hu, Y., Rajan, D., and Chia, L.-T.: 'Robust subspace analysis for detecting visual attention regions in images', Proc. 13th Annual ACM International Conference on Multimedia, November 2005, pp. 716-724.

[26] Hu, Y., Rajan, D., and Chia, L. -T.: 'Scale adaptive visual attention detection by subspace analysis', Proc. 15th International Conference on Multimedia, September 2007, pp. 525-528.

[27] Kienzle, W., Wichmann, F. A., Scholkopf, B., and Franz, M. O.: 'A non-parametric approach to bottom-up visual saliency', Advances in Neural Information Processing Systems, 2006, 19. [http://books.nips.cc/papers/files/nips19/NIPS2006_0480.pdf]

[28] Hou, X., and Zhang, L.: 'Saliency detection: a spectral residual approach', Proc. IEEE Conference on Computer Vision and Pattern Recognition, June 2007, pp. 1-8.

[29] Harel, J., Koch, C., and Perona, P.: 'Graph-based visual saliency', Advances in Neural Information Processing Systems, 2006, 19. [http://books.nips.cc/papers/files/nips19/NIPS2006_0897.pdf]

[30] Sato, T., Murthy, A., Thompson, K. G., and Schall, J. D.: 'Search Efficiency but Not Response Interference Affects Visual Selection in Frontal Eye Field', Neuron, 2001, 30, pp. 583-591.

[31] Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., and Anderson, C. H.: 'Overcomplete steerable pyramid filters and rotation invariance', Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 1994, pp. 222-228.

[32] Simoncelli, E. P., and Freeman, W. T.: 'The steerable pyramid: a flexible architecture for multi-scale derivative computation', Proc. of IEEE International Conference on Image Processing, 1995, pp. III-444-447.

[33] Young, R. A.:, 'The gaussian derivative model for spatial vision: l. retinal mechanisms', Spatial Vision, 1987, 2(4), pp. 273-293.

[34] Pizer, S. M., Eberly, D., and Fritsch, D. S.: 'Zoom-invariant vision of figural shape: the mathematics of cores', Computer Vision and Image understanding, 1998, 69, (1), pp. 55-71.

[35] Sabidussi, G.: 'The centrality index of a graph', Psychometrica, 1966, 31, pp. 581-603.

[36] Itti, L., and Koch, C.: 'Feature combination strategies for saliency-based visual attention systems', Journal of Electronic Imaging, 2001, 10, (1), pp. 161-169.

[37] Cannon, M. W., and Fullenkamp, S. C.: 'A model for inhibitory lateral interaction effects in perceived contrast', Vision Research, 1996, 36, (8), pp. 1115-1125.

[38] Schaefer, G., and Stich, M., 'UCID - an uncompressed colour image database', Proc. SPIE Storage and Retrieval Methods and Applications for Multimedia, 2004, 5307, pp. 472-480.

[39] Tsotsos, J. K.:, 'Analyzing vision at the complexity level', Behavioral and Brain Sciences, 1990, 13, pp. 423-469.

Figures

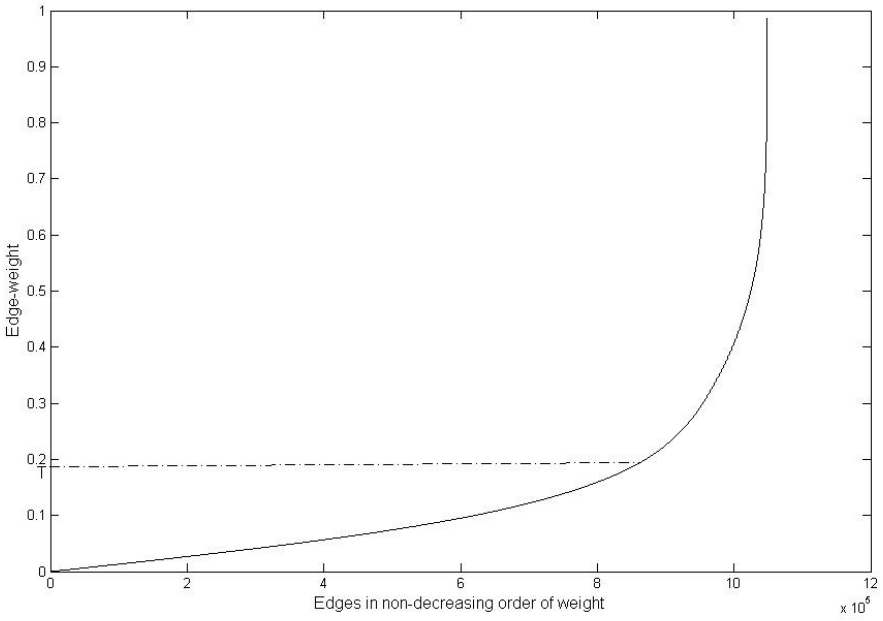

Figure 1: Edge weight distribution in ViSaNet.

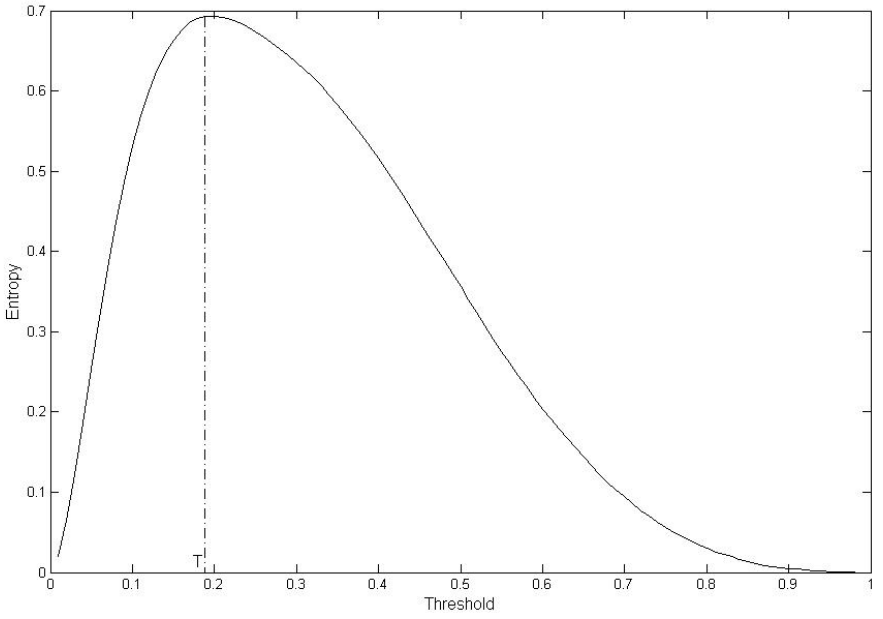

Figure 2: Entropy varies with threshold.

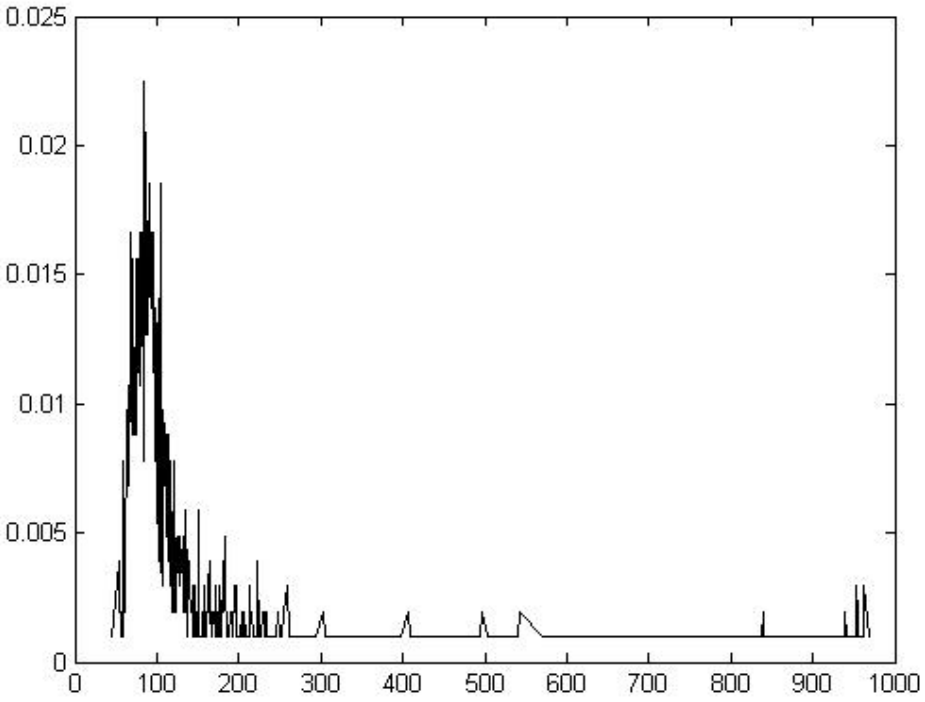

Figure 3: Degree distribution in ViSaNet.

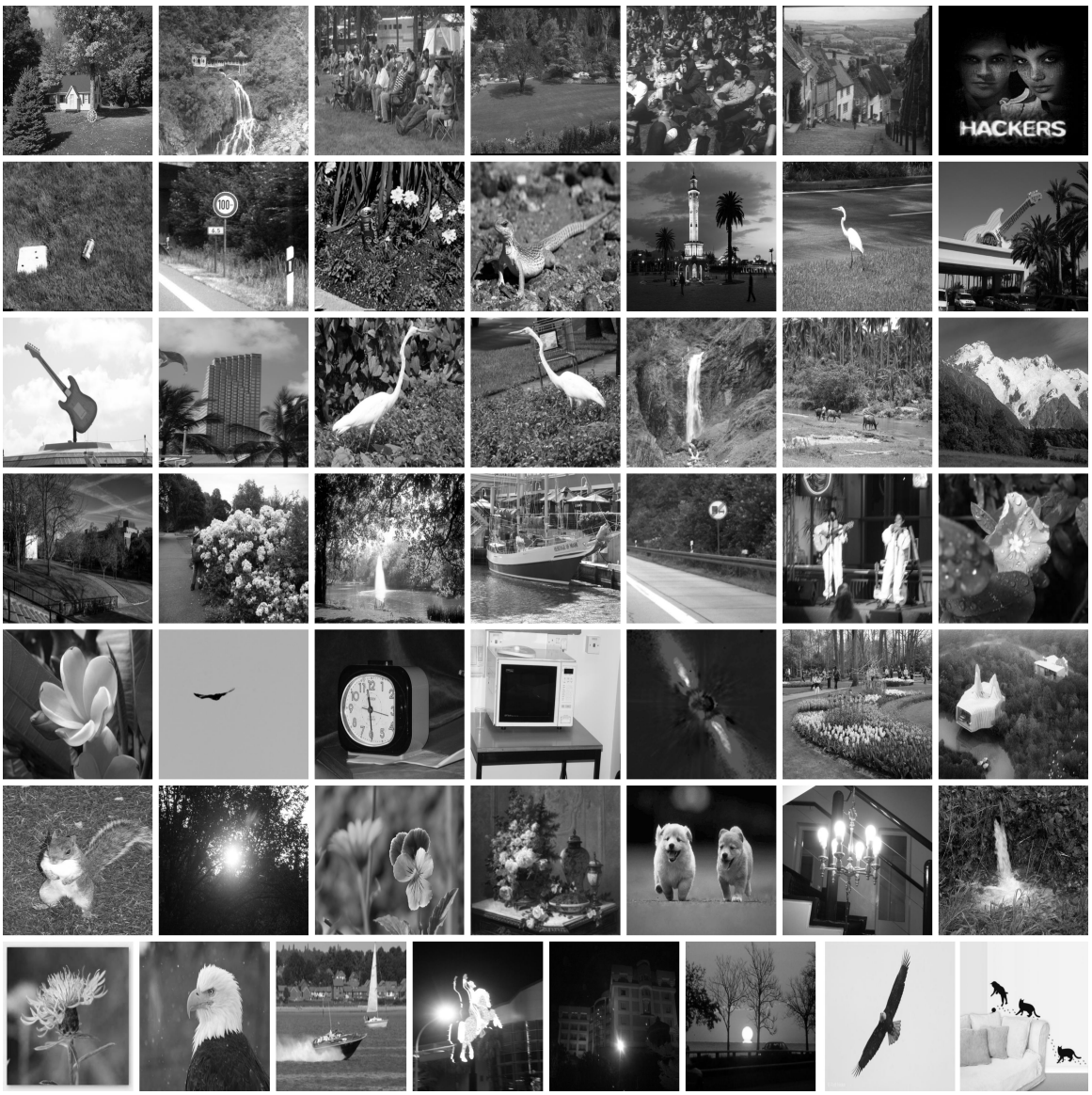

Figure 4: Image set used in experiments (aspect ratios are changed to fit all the images). Numbering is in row major order.

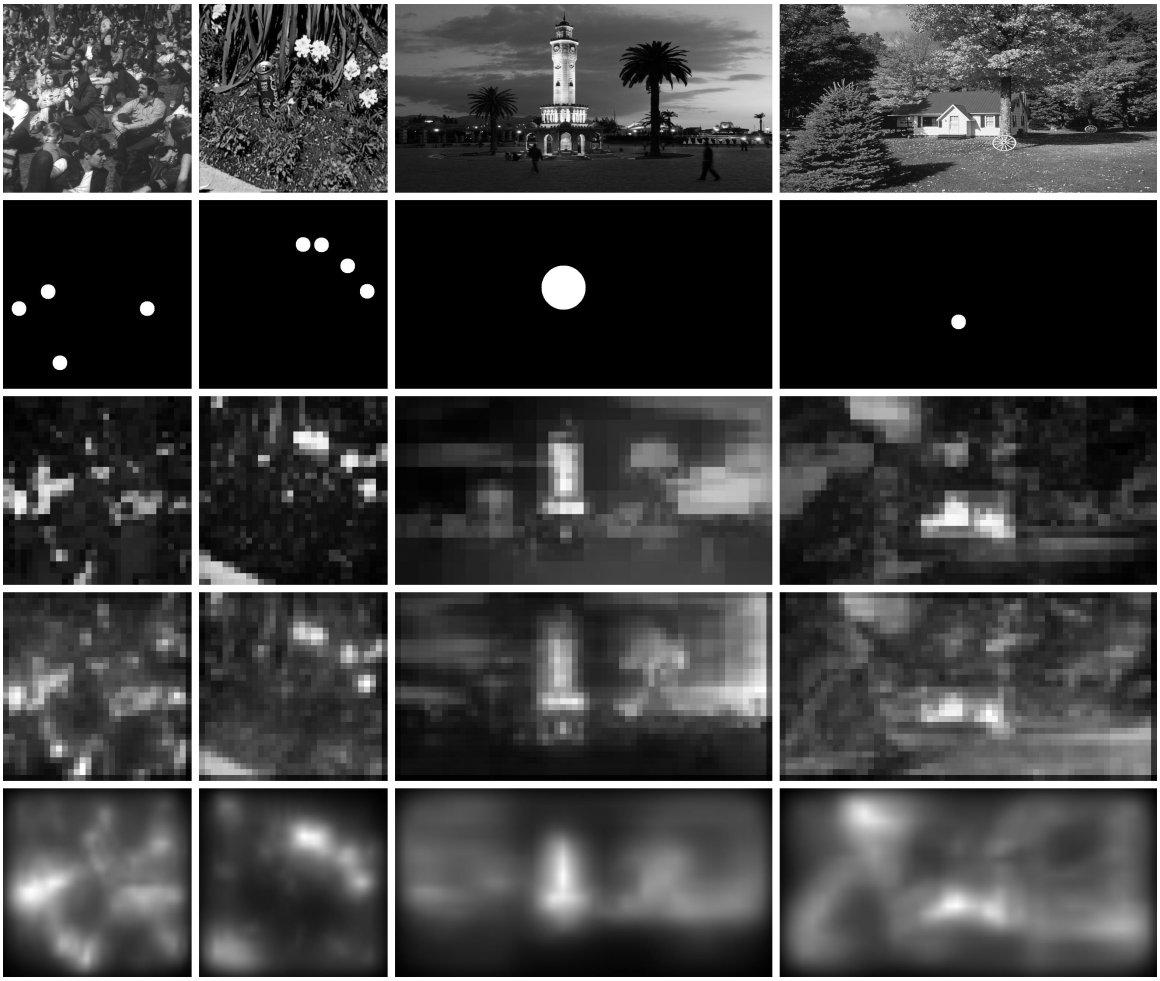

Figure 5: Rows from top to bottom: input image, groundtruth (binary image B stated above), saliency map according to proposed method, [4] and [29], respectively.

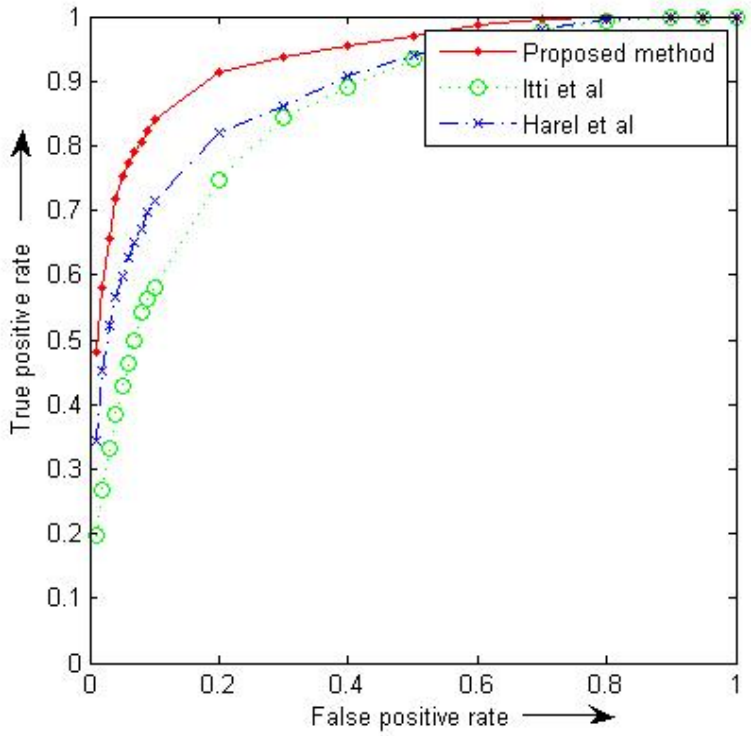

Figure 6: Mean ROC curves.

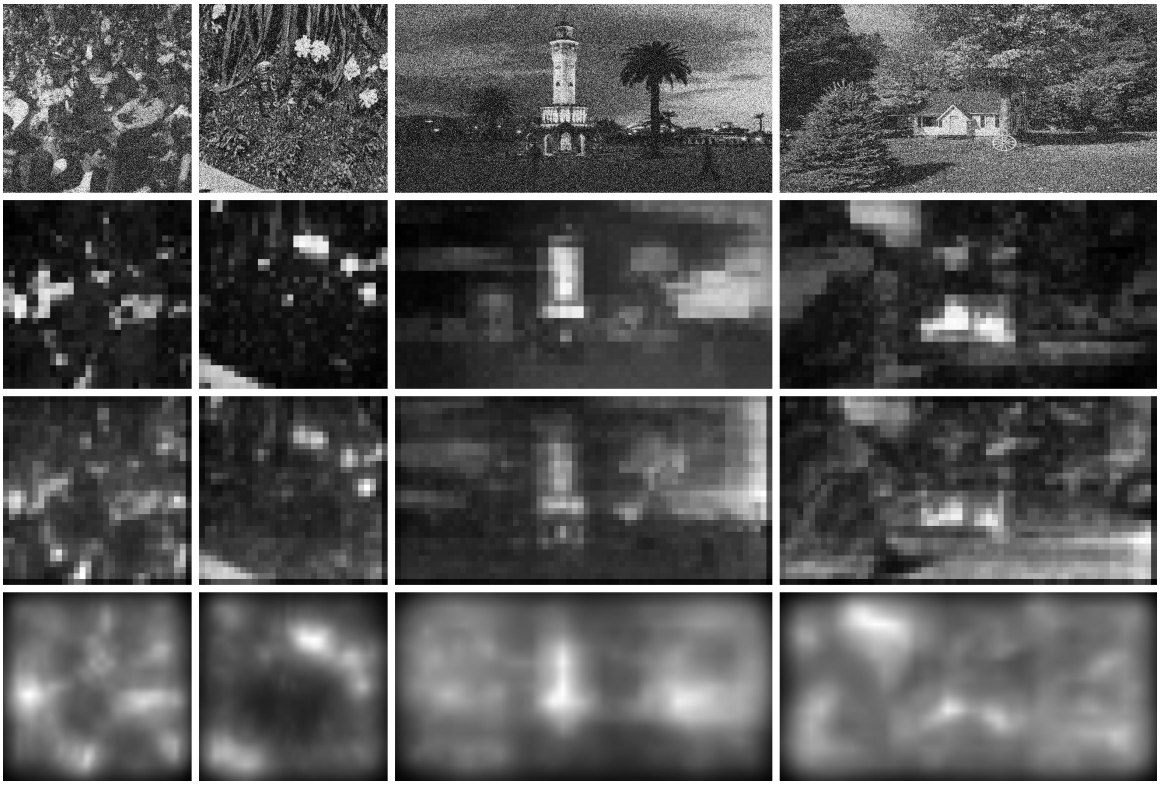

Figure 7: Rows from top to bottom: noisy input images (10dB), saliency maps of the noisy images according to proposed method, [4] and [29], respectively.

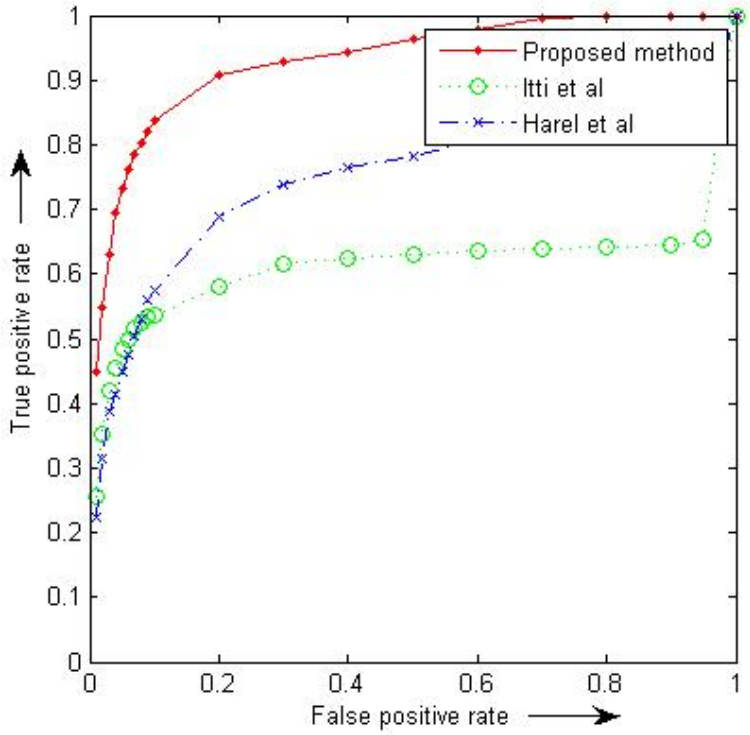

Figure 8: Mean ROC curves for noisy input images (10dB).

List of figures