A Simple Method for Detecting Salient Regions

Paul L. Rosin
School of Computer Science
Cardiff University
Cardiff CF24 3AA, Wales, UK
email: Paul.Rosin@cs.cf.ac.uk

Abstract

A simple method for detecting salient regions in images is proposed. It requires only edge detection, threshold decomposition, the distance transform, and thresholding. Moreover, it avoids the need for setting any parameter values. Experiments show that the resulting regions are relatively coarse, but overall the method is surprisingly effective, and has the benefit of easy implementation. Quantitative tests were carried out on Liu *et al.*'s dataset of 5000 images. Although the ratings of our simple method were not as good as their approach which involved an extensive training stage, they were comparable to several other popular methods from the literature. Further tests on Kootstra and Schomaker's dataset of 99 images also showed promising results.

Keywords: salience map, importance map, focus of attention, distance transform.

1 Introduction

The salience map is a basic part of many theories of visual attention. Each pixel in the map represents the importance of the object in the scene that is projected into the visual field at that position. In the human visual system there is considerable neurophysiological evidence for the salience map [1], and moreover that it is computed in a bottom-up manner [2]. Its purpose is to direct eye movements to the most relevant parts of the visual field, allowing high resolution analysis of the fixated region by the fovea. In a similar manner, salience maps are useful in computer vision as they provide an efficient means of processing complex scenes by locating likely areas of interest for directing analysis. Thus it has wide applicability, and salience maps are used in object recognition, tracking, compression (e.g. JPEG2000), determination of suitable locations for embedding watermarks, etc.

The most influential model for generating salience maps was proposed by Koch and Ullman [3], and has been further developed over the years [4, 5]. Essentially, saliency at each location is determined by its dissimilarity from its neighbourhood in terms of intensity, colour, orientation, etc. Each basic property is computed and summed over several scales to form "conspicuity" maps. After normalisation these are further combined to form the salience map. Positions of the maxima, ordered by their strength, can then be sequentially processed. The extents of the maxima are determined by thresholding the salience map or the relevant conspicuity map. However, it has been noted that this approach requires many design parameters such as the number, types and sizes of the filters, as well as weights, normalisation schemes, etc. Finding and justifying appropriate parameter values can be problematic, but their values have significant impact on the behaviour of the system [6].

A related recent approach by Aziz and Mertsching [7] has similar problems. It first extracts regions, and then computes five properties for each region: colour, size, symmetry, orientation and eccentricity. The rarity or contrast of these values determines the feature saliency values, which are then combined into a single measure. Again a problem with this approach is the number of parameters. It requires sixteen values defining scaling factors, thresholds, etc., which potentially compromises the robustness of the system. Although this is an extreme example of "parameter proliferation", the majority of methods in the literature for measuring salience require at least a few parameters. In contrast, we will present a method that is entirely parameter free.

There are many other approaches developed in recent years. For instance, Kadir and Brady [8] define saliency as local complexity, which is measured as scale localised features with high entropy. Kootstra and Schomaker [9] used several variations of Reisfeld *et al.*'s local symmetry operator [10] to measure salience, which they found to be a better match to experimental human data than Itti and Koch's [4] contrast based saliency measure. Whereas normally computation of salience maps is considered to be a bottom-up process,

Moosmann *et al.* [11] use prior high level knowledge: an ensemble of decision trees is trained to classify variable sized windows into various object categories. Salience maps of new images are then determined by running the classifier, and the degree to which a pixel is classified as non-background indicates its salience.

Ma and Zhang proposed an approach based on contrast and fuzzy growing [12], where contrast was defined as the Gaussian weighted sum of the differences between a pixel and its neighbouring pixels. This was adapted by Liu and Gleicher [13] to operate at multiple scales using a pyramid. In addition, they noted that the interior of regions had low saliency, which they attempted to overcome using regional information. The image was segmented (using the mean shift image algorithm) and the saliency value of each region was calculated as the average saliency value of the pixels within it. The drawback of such an approach is that it depends on good quality segmentation, which is unrealistic given the current state of the art.

2 Salient Region Detection Algorithm

We present in this paper a scheme to compute salience maps based on just edge density. Our goal is not to design an algorithm to provide more accurate salience maps than currently available in the research literature, but rather to offer the following advantages. Namely, it is extremely simple to implement, it requires no parameters, it dispenses with colour, but nevertheless works well in many instances. Basically, edge magnitudes are generated, e.g. using the Sobel operator, and dense regions of the edge map correspond to salient locations.

Edge density could be determined using a Parzen estimator, i.e. by blurring the edge map. However, this requires choosing an appropriate kernel size for each image. To avoid requiring any parameters to be set we use instead the distance transform (DT) [14] to propagate the edge information. Since the standard DT is defined for binary images rather than grey level images we apply *threshold decomposition*. That is, all the possible thresholded images are generated from the grey level image (or in practise a subset is often sufficient – we have used 66), each is processed separately, and then the set of DT maps is summed. The result is the salience map, in which lower values indicate greater salience. Pixels close to edges have low values, and the higher the edge magnitude then the lower the DT value.

We note that our previous work on the salience distance transform (SDT) [15] would be an alternative approach to computing the salience map. The SDT took as input a sparse edge strength map (e.g. the edge detection should incorporate non-maximal suppression) and computed the distances from each non-edge to the closest edge. Unlike a standard DT the distances were inversely weighted by the edge strengths (or other edge properties), so that strong edges had greater effect than weak noisy edges. However, for the purposes of determining salient regions, the threshold decomposition scheme used in this paper is more effective. Since the SDT is defined such that even weak edges have zero distance values on the edges themselves then this will

create many spurious local low (i.e. nominally salient) patches unless they are eliminated by thresholding. As opposed to applying the DT with threshold decomposition, the SDT has the advantage that it only needs to be applied to a single image. However, the difference in computation time is not so great since the SDT requires the image to be scanned by the chamfer masks multiple times (the number depends on the complexity of the image, but around 10-20 iterations is typical) whereas the DT only requires a single scan of each mask.

The final step is to detect salient regions, which are of unknown number, size and shape. Rather than search for local minima (which would still not directly indicate size and shape) the DT maps are thresholded – we have used Tsai's moment preserving algorithm [16] – and the resulting regions tend to capture the important objects in the scene quite well. The overall algorithm is very straightforward, and is summarised in figure 1.

The results of the proposed method are shown in figure 2 along with results from Moosmann *et al.* [11] and Walther and Koch [5][1] for comparison (in all cases lower values in the salience maps indicate higher significance). Although Moosmann *et al.* do not extract regions from the salience map we have added this step using Tsai's thresholding for easier comparison of the three methods, and the salient regions are overlaid on the images. It can be seen that the proposed method is not as specific as the other methods, including more responses to high contrast areas such as the fence in the first image and the shrubs in the second image. The thresholded regions also tend to extend beyond the object boundaries and to merge together several objects. Nevertheless, in all images the proposed method has successfully located all the main objects, and rejected the majority of the background.

Further comparison is provided in figure 3 on the image from the DT sequence used for testing by Kadir and Brady [8]. The results of Walther and Koch's algorithm are shown in figure 3b. It can be seen that many of the vehicles are missed, and localisation is poor. Kadir and Brady's algorithm more successfully locates the vehicles (position and scale only are determined) and also some of the other objects on the street. However, it also produces many multiple responses on the vehicles, as well as responding to road markings. As seen in figure 3d the proposed method also successfully locates the vehicles and street signs.

3 Possible Extensions to the Algorithm

3.1 Edge Detection

Many variations of the basic scheme can be found and some are demonstrated here. For instance, a possible avenue for improvement is to replace the simple Sobel edge magnitudes by more sophisticated estimates of edge significance [17]. We give two examples.

The first is based on Bischof and Caelli's [18] measure of edge "stability". Edges are detected using the Canny operator at multiple scales, and are then tracked from fine through to coarse scales ($\sigma \in [1, 16]$) with a fixed window centred at the finest scale. In our modified version the gradient magnitudes at each scale are summed during the tracking process, which effectively combines edge lifetime and gradient magnitude into the edge map.

The second computes a Gaussian pyramid, where each level is reduced in size to a quarter area, and the number of levels in the pyramid is set to $N = \log_2 \frac{S}{10}$ where S is the minimum of the image width and height [13]. For simplicity and efficiency the Sobel is applied at each level (rather than the Gaussian weighting function used by Liu and Gleicher [13]), and then after upsampling the levels are averaged.

Some results of edge detection are shown in figure 4; in comparison to the Sobel the Canny edges have single pixel width, and the scale-space tracking has suppressed the noise and clutter to some extent. On the other hand, using the pyramid has had the opposite effect as the coarser scales have caused the edge response for dominant large scale edges to spread out.

3.2 Blob Detection

A second variation is to look for blobs rather than edges; this is done using Difference of Gaussians (DoG) which have been shown to be effective for interest point detection [19]. Whereas the edge maps are effective even for a single scale the DoG operator is strongly dependent on scale, and so a multi-scale approach is necessary. The Gaussian kernel parameters for the DoG are set to $\{\sigma, 2\sigma\}$ where $\sigma = \{2^1, 2^2, \ldots, 2^N\}$, and $N = \log_2 \frac{S}{10}$ as above. The absolute values of the DoG responses at each scale are scaled to the range $[0, 255]$ and the results summed to form a single feature map. While the summed DoG responses could be directly thresholded to find the salient regions we have found the results to be fragmented. Substantially better results are obtained by applying the DT (via threshold decomposition) to the summed DoG and then thresholding the output, as demonstrated in figure 5.

[1]For displaying Walther and Koch's results their salience map was thresholded. The number of components was used to determine the number of salient regions to be located by their algorithm in the (raw) salience map and extracted from the conspicuity maps.

3.3 Incorporating Colour

Colour is obviously a potentially very useful property, and following Walther and Koch [5] two opponent colour maps for red-green and blue-yellow were generated. The above edge and blob schemes were then applied to the two colour maps as well as the intensity image to produce three DT maps which were finally scaled to $[0, 255]$, summed and thresholded. The process is summarised in figure 6. However, the results were disappointing, as demonstrated in figure 7 which shows the results of processing the lorry image from figure 2. Since the lorry is bright red it shows up strongly on the red-green colour map (figure 7a). The summed DoG responses on this colour map *do* highlight the lorry (figure 7b), but the peak response is from the smaller, more circular, wheel hubs, while the response at the top edge of the lorry is weak. The missing lorry top can be seen in the overlay of the thresholded DoG (figure 7c). The DT map of the summed DoG of the red-green colour map exhibits the same weakness at the top of the lorry (figure 7e) and does not improve on the result of processing just the intensity image. Naturally, combining the other two channels does not improve the delineation of the lorry. Other examples revealed further problems with incorporating colour. In same cases good responses in one of the three colour/intensity channels would be diluted to such an extent when the three salience maps were combined that the overall detection of salient regions was degraded. Also, creating the opponent colour maps potentially magnifies noise, although Walther and Koch's [5] scheme of zeroing the opponent colours if all the red, green and blue components are less than a tenth of the dynamic range mitigated this problem.

3.4 Comparing variations of the algorithm

The basic algorithm and some of its variations are demonstrated in figure 8. The images come from Walther and Koch [5][2] and the Berkeley Segmentation Dataset [20]). In addition to Walther and Koch's results (row 2), results are shown for the basic DT method (row 3) and the following variations: the Sobel edge detector was replaced by multi-scale "stability" edges (row 4); the single scale Sobel edge detector was replaced by a multi-scale pyramid version (row 5); the Sobel edge detector was applied to the opponent colour maps and the intensity image and then combined (row 6), the Sobel edge detector was replaced by summed DoG responses (row 7); the DoG was applied to the opponent colour maps and the intensity image and then combined (row 8). It can be seen that none of the variations are uniformly successful. The results are generally similar, and it is hard from observation to decide which (if any) is best.

However, some general comments can be made. First, the DoG, with or without colour information, does not appear to improve on the basic DT method. Second, relying on edges alone does impose limitations on

[2]All the images from Walther and Koch were originally framed by a wide dark border which we have removed. Otherwise, the border causes a large strong edge that detracts from the edges within the frame.

the scope of the method. This is illustrated by the results on the last image which is less successful that the preceding results. For the horses the top edge of the horse's back is weak, and moreover there is a lot of fine detail in the grass. In other words, the semantically significant boundary of the horse is assigned low magnitude edges while the grass is insignificant, but has relatively high edge magnitudes, and moreover covers a large area. This is directly reflected in the selected salient region which incorrectly includes the grass and excludes the top of the horse (row 3). The multi-scale edge detection manages to reduce the effect of the grass, but does not improve the horse's back (rows 4 and 5).

4 Experiments

Given the difficulty in the previous section in comparing the results from the various methods, it is obvious that a more quantitative evaluation is necessary. Various attempts have been made in the literature at quantitatively testing the performance of algorithms for computing salience (although many papers do not in fact go beyond presenting example salience maps). Some focus on testing the robustness and repeatability of the positions of detected regions in the image when it is modified, e.g. after scaling or cropping [21]. This gives some indication of the effectiveness of the salient regions as features for tracking applications [8], but does not indicate if the detected salient regions are really the most significant salient regions to be found.

In order to evaluate the perceptual quality of the salient regions, ground truth provided by people is required. Of course, for a large scale dataset this is a time consuming task, requiring several man-months. Fortunately, this has already been carried out by Liu *et al.* [22] in order to evaluate their Conditional Random Field (CRF) method. We use their dataset $\mathcal{B}$ which contains 5000 images that were annotated by nine different users who each marked up an axis aligned rectangle contained the most salient region. The final ground truth mask for each image was taken as the majority vote of the nine users' masks, which generally resulted in an almost rectangular mask. Liu *et al.*'s method uses image features such as contrast and colour spatial distribution which are combined using the CRF framework to estimate their relative weights. Since their method required a learning phase they used 2000 images from a separate dataset ($\mathcal{A}$) plus 1000 images randomly selected from $\mathcal{B}$ to construct a training set. Testing was then carried out on the remaining 4000 images from $\mathcal{B}$.

We note that most of Liu *et al.*'s images were collected from the web, and perhaps as a consequence of this many of them (we estimate about $5\% - 10\%$) contain some type of added border effect. Such images cause problems for our proposed algorithm, since the border is often the highest contrast feature in the image, and is also large if it extends around all four sides. It therefore tends to dominate the saliency map. An example is shown in figure 9; it can be seen that the identified salient region is essentially highlighting the border at the expense of the remainder of the image content. Nevertheless, since such images were

presumably included by Liu *et al.* for testing purposes, we have retained them in our experiments without special treatment.

Liu *et al.*'s method of comparing a binary salience map against the corresponding ground truth mask was to find the smallest axis aligned rectangle in the salience map that contains 95% of the detected salient region. Given corresponding masks the detection rate for each image was now quantified by the following terms, where the F-value is an overall performance measure, and α was set to 0.5.

$$\begin{aligned} \text{precision} &= \frac{area(\text{detected salient region}) \cap area(\text{ground truth region})}{area(\text{ground truth region})} \\ \text{recall} &= \frac{area(\text{detected salient region}) \cap area(\text{ground truth region})}{area(\text{detected salient region})} \\ F_\alpha &= \frac{(1+\alpha) \times \text{precision} \times \text{recall}}{\alpha \times \text{precision} + \text{recall}} \end{aligned}$$

Since the ground truth used a single rectangle to identify a single object (perhaps with multiple components) we have modified our method to only output the single largest region out of the full set of detected salient regions. We also note that since our method tends to systematically overextend the salient regions, then a better match with the ground truth will be obtained by systematically shrinking the output regions. This was carried out using morphological erosion with a disk structuring element, and as figure 10a shows, produced a considerable improvement of the F scores.

Liu *et al.* [22] provided the mean performance values for the 4000 images for their method ($F = 0.80$) as well as Walther & Koch's method ($F = 0.68$) and Ma and Zhang's method ($F = 0.61$) [12]. These can be compared to the results of testing some of the versions of the proposed method on the full set of 5000 images of dataset $\mathcal{B}$, as shown in figure 10a. Note that a baseline level of performance is $F = 0.43$, which can be obtained by considering the entire region as the detected salient region (which would result in high recall, but poor precision). Unsurprisingly, our proposed method is outperformed by Liu *et al.*'s more sophisticated method. Nevertheless, our method still has the advantages of great simplicity of construction, and also not requiring a (computationally expensive) training phase. It is also instructive to compare our results against the other two methods. Our approaches' scores are comparable with Walther & Koch's, and is significantly better than the fuzzy growing method. Regarding Walther & Koch's method, our method again has the advantage of simplicity and avoids the problem of parameter selection [6].

Comparing figure 10a to the conclusions from our informal testing in the previous section, we find that DoG with colour *does* in fact improve upon the performance of the basic method. However, the best performance is from the multi-scale "stability" edges – which do not use colour.

Not all the saliency map algorithms proposed in this paper are scale invariant. In particular, for the basic method, edge detection is performed at a single scale. Nevertheless, this is not necessarily substantially detrimental to its performance. Figure 10b demonstrates the effect of rescaling all the images in dataset $\mathcal{B}$

to reduce their area by a factor of four. It is expected that this loss of information will degrade the quality of the salience map, but in most cases the drop in performance is modest. Given the smaller size of image less erosion is required; if disk diameters 54 and 45 are applied to the original and quarter area images, then the reduced image size leads to a reduction of F to about $94\% - 95\%$ of full scale performance.

Of course, useful as the ground truth is, it is crude since it is only made up of single bounding rectangles. This can lead to problems such as the one demonstrated in figure 11. The input image contains two significant objects, but just one is selected as ground truth: figure 11b. Our saliency detector finds both objects, but combines the cat with the fence, and misses the top half of the boy (figure 11c). Since the larger region is the cat the estimated rectangle does not overlap with the ground truth (figure 11d), and consequently the result receives an F score of zero, effectively underestimating the quality of the saliency detection. There are further limitations due to the approximate nature of delineation of the ground truth. However, there is no reason to assume that these limitations should bias the F measure for or against any of the saliency measures.

As a second test we have used the smaller data set produced by Kootstra and Schomaker [9]. This consists of 99 images which were selected from five different categories. Ground-truth salience was based on human fixation data which was recorded from an eye-tracking experiment in which subjects were asked to freely view the images (i.e. they were not given a task). From each trial of each subject the distance transform was applied to the fixation data points and inverted, to produce the *fixation distance map*. Evaluation can then be carried out by computing the correlation between a salience map and the corresponding fixation-distance map. To reduce the effects of variability across subjects we have used the combined fixation-distance maps, in which the maps for each image have been averaged across all subjects.

Although the source images were 1024×768, the fixation distance maps were only available at a sixteenth area resolution, and therefore we computed the salience maps on the input images reduced to this scale. In addition, for comparison, the salience maps were also computed at a further sixteenth reduction in area, and then rescaled to 256×192 before computing correlation. Correlation values are given in table 1, and should be compared against the values obtained by Kootstra and Schomaker [9] for their symmetry operator based methods: 0.68 (isotropic), 0.66 (radial), 0.68 (colour) and Itti and Koch's [4]: 0.26. Note that the isotropic and radial methods operate on intensity rather than colour images. According to the correlation criterion it can be seen that Itti and Koch's method performs very poorly. In comparison, our results are considerably better although not generally as good as Kootstra and Schomaker's. The exception is the stability method, whose correlation value increases substantially to match their performance when the input image is shrunk to $\frac{1}{256}$ the original area.

The explanation for the large difference in performance of the stability method for the two image sizes

is found in examples such as those shown in figure 12. These three examples show images containing (a) little dominant structure, (e) a large over-sized foreground object filling the image, and (i) highly patterned and structured background competing with the foreground. In such cases reducing the image size to $\frac{1}{256}$ the original area tends to focus salience on the centre of the image, which generally yields appropriate salience maps for these types of images. Although the above examples are extreme cases, as figure 13 shows, most (25%) of the salience maps are improved by using the smaller image size.

5 Conclusions

A simple and non-parametric method for detecting salient regions in images has been described. The simplest version is based on applying the Sobel edge detector to gray-level images, followed by threshold decomposition, the distance transform and thresholding. Improvements in performance were gained by using multi-scale edge detection or by using multi-scale DoGs applied to opponent colour maps and intensity. Surprisingly, the inclusion of colour did not produce the large benefits expected – however this is in agreement with Kootstra and Schomaker's [9] recent findings. We obtained our best results using the stability (i.e. intensity-only multi-scale Canny edge detector) method. This may be attributed to its simplicity. Both multi-scale blob detection and colour analysis introduce the need for fusing multiple maps, and this stage can introduce errors which offsets their benefits.

Results of large scale testing on Liu *et al.*'s dataset of 5000 images, and comparison with other more complex methods, show that the proposed methods are generally as effective as some of them, but without their disadvantages. Further testing on a Kootstra and Schomaker's small dataset also showed promising results. Even though threshold decomposition effectively increases the number of images to be processed, computation time is still reasonably efficient, e.g. it takes about a second for the basic method to process a 512×512 image on a standard PC.[3]

While the proposed method can be seen to work surprisingly well, considering its simplicity, it has limitations. In particular since it is edge based then it will fail if many strong edges exist in the image that do not belong to salient objects, e.g. from strongly textured backgrounds.

Acknowledgements

I would like to thank Dirk Walther and Timor Kadir for making their software available, and Gert Kootstra for providing his images and fixation distance maps. The image from the DT sequence was made available by the KOGS/IAKS Universität Karlsruhe,

[3]Source code for the basic method is available at http://users.cs.cf.ac.uk/Paul.Rosin.

References

[1] J.H. Fecteau and D.P. Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences*, 10(8):382–390, 2006.

[2] L. Elazary and L. Itti. Interesting objects are visually salient. *J. Vision*, 8(3):1–15, 3 2008.

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

[4] 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*, 20(11):1254–1259, 1998.

[5] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[6] W. Kienzle, F.A. Wichmann, B. Schölkopf, and M.O. Franz. A nonparametric approach to bottom-up visual saliency. In *Proc. NIPS*, pages 689–696, 2006.

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

[8] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[9] G. Kootstra and L.R.B. Schomaker. Prediction of human eye fixations using symmetry. 2009.

[10] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: The generalized symmetry transform. *International Journal of Computer Vision*, 14(2):119–130, 1995.

[11] F. Moosmann, D. Larlus, and F. Jurie. Learning saliency maps for object categorization. In *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*, 2006.

[12] Y.F. Ma and H.J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Int. Conf. on Multimedia*, pages 374–381, 2003.

[13] F. Liu and M. Gleicher. Region enhanced scale-invariant saliency detection. In *Proc. ICME*, pages 1477–1480, 2006.

[14] G. Borgefors. Distance transforms in arbitrary dimensions. *Computer Vision, Graphics and Image Processing*, 27:321–345, 1984.

[15] P.L. Rosin and G.A.W. West. Salience distance transforms. *CVGIP: Graphical Models and Image Processing*, 57:483–521, 1995.

[16] W.H. Tsai. Moment-preserving thresholding. *Computer Vision, Graphics and Image Processing*, 29:377–393, 1985.

[17] P.L. Rosin. Edges: saliency measures and automatic thresholding. *Machine Vision and Applications*, 9:139–159, 1997.

[18] W.F. Bischof and T. Caelli. Parsing scale-space and spatial stability analysis. *Computer Vision, Graphics and Image Processing*, 42:192–205, 1988.

[19] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[20] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Int. Conf. Computer Vision*, volume 2, pages 416–423, 2001.

[21] D.W. Lin and S.H. Yang. Wavelet-based salient region extraction. In *Pacific Rim Conf. Multimedia*, volume 4810 of *Lecture Notes in Computer Science*, pages 389–392. Springer, 2007.

[22] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum. Learning to detect a salient object. In *Proc. Conf. Computer Vision Pattern Recognition*, pages 1–8, 2007.

List of Figures

Input: Grey level image G

Output: Binary salience mask S

Calculate gradient magnitude $E = \text{edgeDetect}(G)$;

Initialise image running total $T = 0$;

foreach *intensity* i **do**

  Extract E_i = threshold decomposition for slice i ;

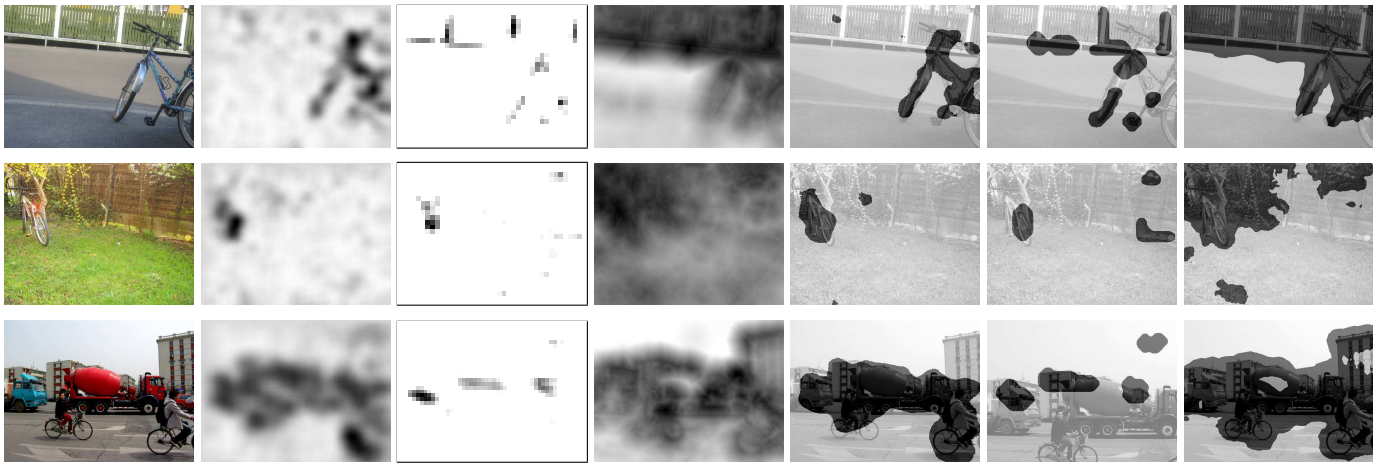
  Calculate distance transform: $DT(E_i)$;

  Update image running total T += $DT(E_i)$;

end

Threshold summed distance transforms $S = \text{Tsai}(T)$;

Figure 1: The basic algorithm for detecting salient regions

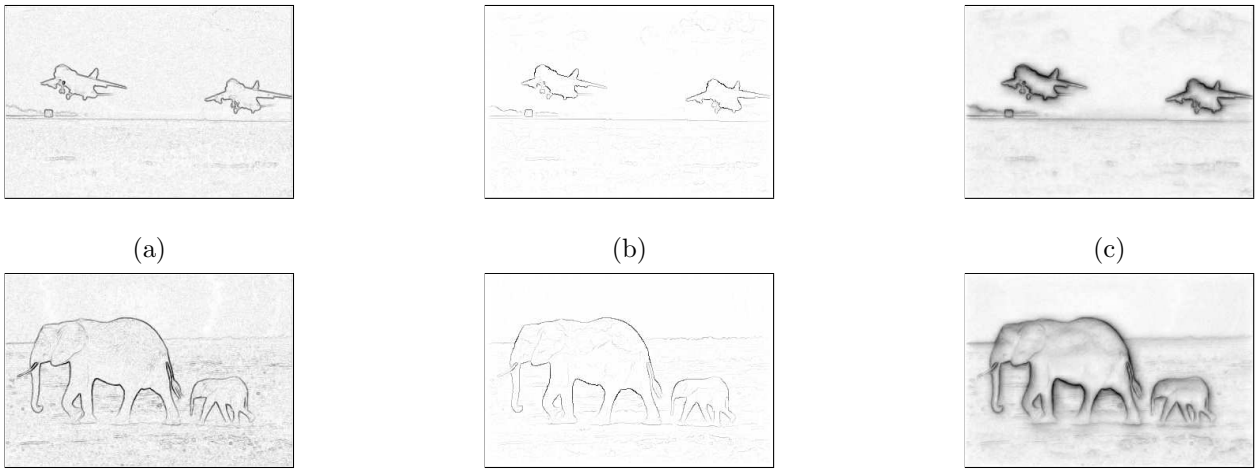


(a) (b) (c) (d) (e) (f) (g)

Figure 2: (a) Images from Moosmann *et al.* [11]. (b) Moosmann *et al.*'s salience map, (c) Walther & Koch's salience map, (d) salience map of proposed method, (e) salient regions from Moosmann *et al.*'s salience map, (f) salient regions from Walther & Koch's salience map, (g) salient regions from the proposed method.

(a) (b) (c) (d)

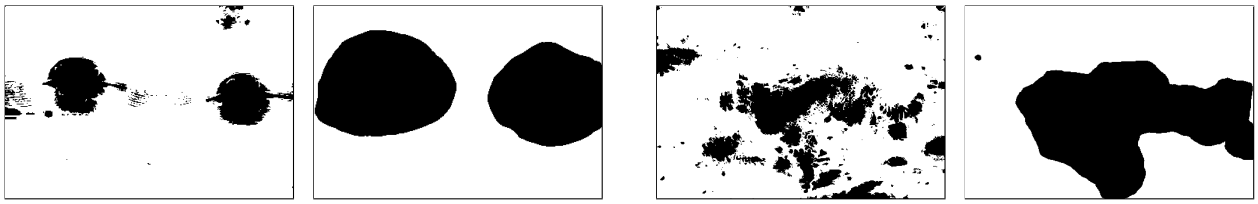
Figure 3: a) Image from DT sequence; b) Walther & Koch's salient regions; c) Kadir & Brady's salient regions; d) overlay of regions from the proposed method.



(a) (b) (c)

(d) (e) (f)

Figure 4: Extracted edges from the aircraft and elephant images in figure 8; a) & d) Sobel operator; b) & e) Canny operator tracked and summed over multiple scales ("stability"); c) & f) Sobel operator summed in pyramid.



(a) (b) (c) (d)

Figure 5: Extracted salient region masks from the aircraft and cheetah images in figure 8; a) & c) directly thresholding the summed DoG responses b) & d) thresholding the DT of the summed DoG responses.

Input: Colour image C

Output: Binary salience mask S

Calculate opponent colour maps $\{RE, BY\}$;

Generate G = grey level version of C ;

foreach *channel* $I \in \{RG,BY,G\}$ **do**

  foreach *scale* σ **do**

    Calculate difference of Gaussian: DoG(G, σ) ;

  end

  I_{DoG} = (normalise($\sum_{\sigma}$ DoG(I, σ)) ;

  Initialise image running total $T(I) = 0$;

  foreach *intensity* i **do**

    Extract I_i = threshold decomposition for slice i of I_{DoG};

    Calculate distance transform: $DT(I_i)$;

    Update image running total $T(I)$ += $DT(I_i)$;

  end

  normalise($T(I)$) ;

end

$T = \sum_I(T(I))$;

Threshold summed distance transforms S = Tsai(T) ;

Figure 6: The modified DoG and opponent colour algorithm for detecting salient regions

(a) (b) (c) (d) (e)

Figure 7: Incorporating colour information. a) the red-green opponent colour map of the lorry image in figure 2; b) summed DoG responses; c) thresholded and overlaid summed DoG responses; d) DT of the summed DoG responses; e) thresholded and overlaid DT of summed DoG responses.

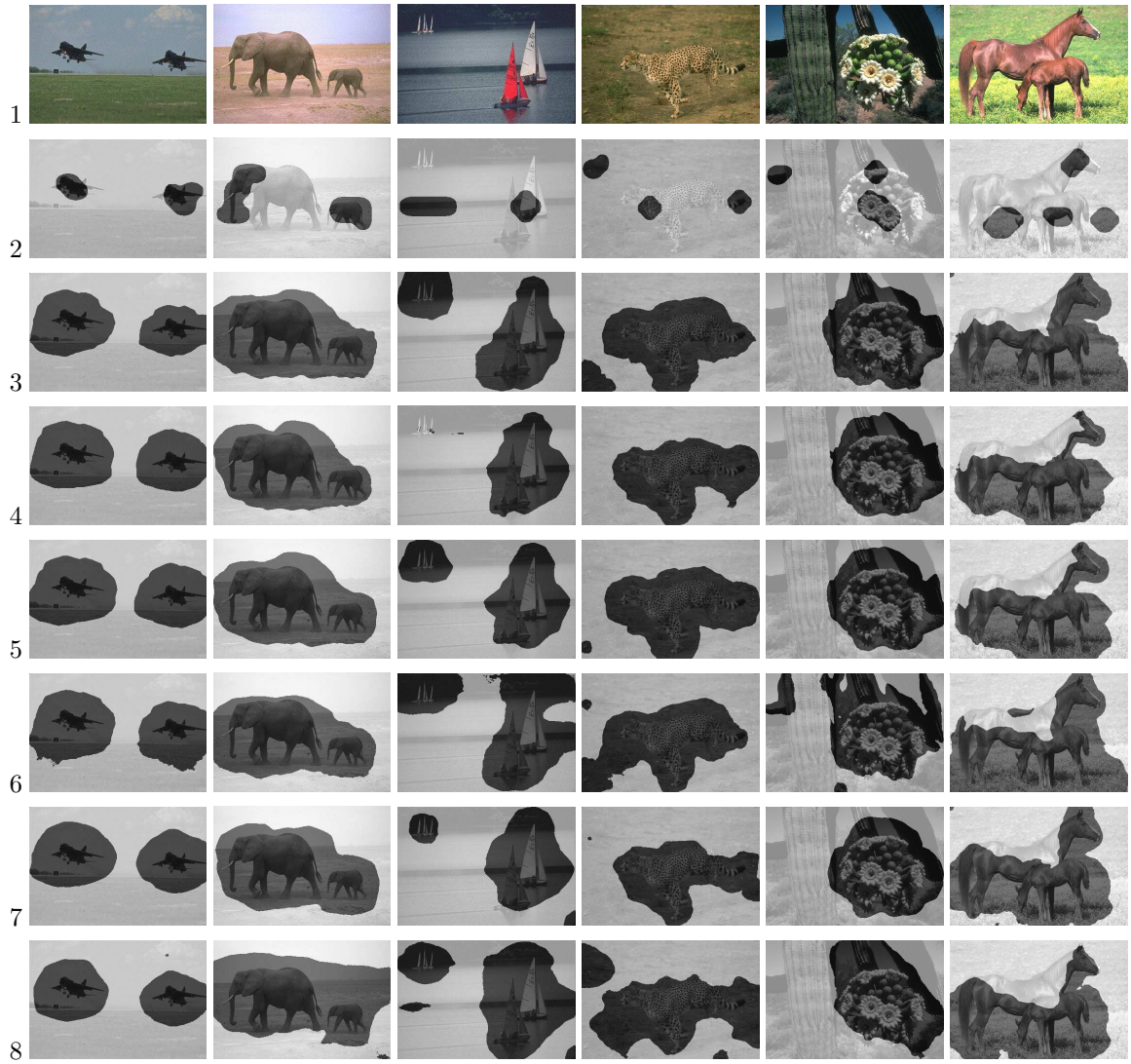
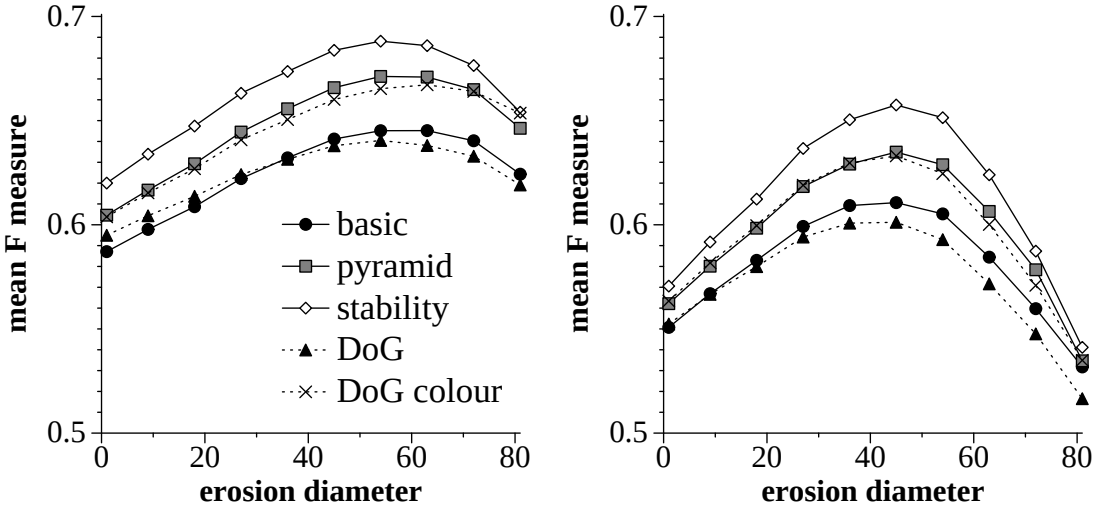


Figure 8: Each row shows: 1/ original images; 2/ thresholded and overlaid Walther & Koch's salience maps; 3/ thresholded and overlaid DT maps; 4/ thresholded and overlaid DT maps using multi-scale edges; 5/ thresholded and overlaid DT maps using pyramid edges; 6/ thresholded and overlaid summed DT maps of opponent colour maps and intensity; 7/ thresholded and overlaid DT maps of summed DoG responses. 8/ thresholded and overlaid summed DT maps of summed DoG responses from opponent colour maps and intensity.

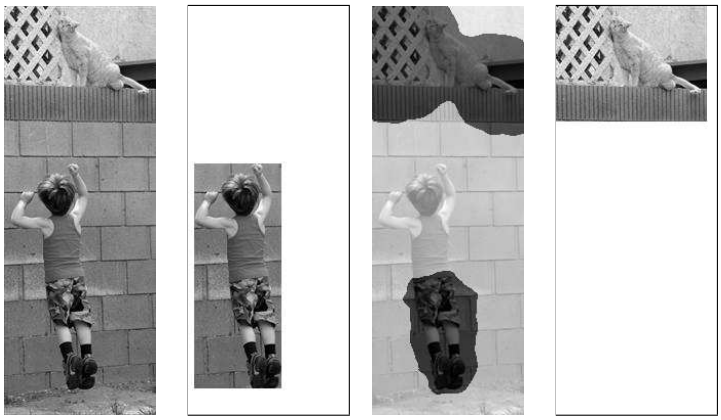
(a) (b)

Figure 9: a) Image containing a double border pattern (grey level version); b) detected salient region shown in black using the proposed basic method.



(a) (b)

Figure 10: Mean F values ($\alpha = 0.5$) for proposed methods applied to Liu *et al.*'s dataset $\mathcal{B}$ calculated for different erosions of the salient regions, and showing the effects of rescaling the images in the dataset. a) full size images; b) quarter area images.



(a) (b) (c) (d)

Figure 11: a) Image (grey level version); b) ground truth rectangle; c) detected salient regions; d) minimum rectangle containing 95% of largest region.

method	$\frac{1}{16}$ area	$\frac{1}{256}$ area
basic	0.46	0.48
pyramid	0.46	0.47
stability	0.52	0.67
DoG	0.43	0.43
DoG colour	0.46	0.53

Table 1: Mean values of the salience maps of the proposed methods correlated against fixation-distance maps for Kootstra and Schomaker's dataset.

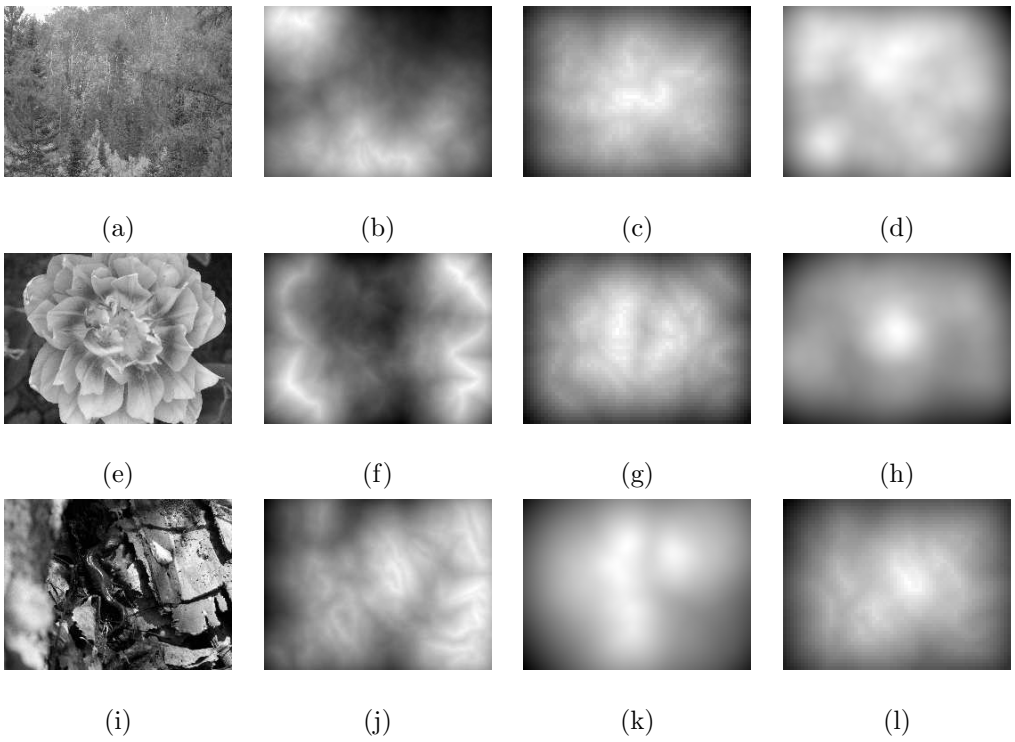


Figure 12: first column: input image; second column: stability based salience at $\frac{1}{16}$ original image area; third column: stability based salience at $\frac{1}{256}$ original image area; fourth column: fixation-distance maps (ground truth).

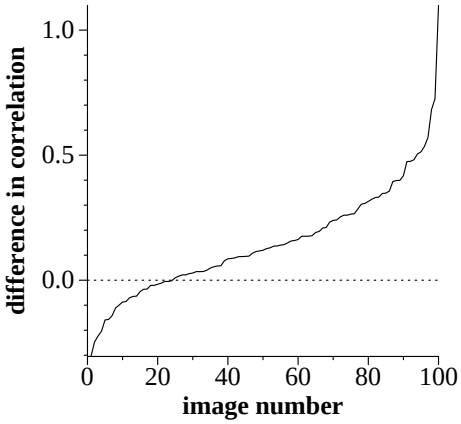


Figure 13: Difference in correlation values for the stability based salience method applied to Kootstra and Schomaker's dataset using $\frac{1}{256}$ compared to $\frac{1}{16}$ reduced area versions of the original images.