Detection Of Target Object In A Static Image

Mr. Akhilesh Tayade
Me Ii Yr(Cse)
Hvpm's College Of Engg.& Tech,Amravati

Prof. Anjali B.Raut
Associate Professor
Hvpm's College Of Engg.& Tech,Amravati

Abstract: **In this paper, we deal with the salient object detection problem for images. We formulate salient object detection as a binary labeling task that separates a salient object from the background since; one pays more attention to salient object in an image as compared to the background of the image. Feature extraction methods are like edge detection, thresholding, multi scale contrast, Center surround histogram.**

While most previous approaches are either limited to special kinds of queries, or do not scale to large image sets, we propose a new method, efficient sub image retrieval (ESR), which is at the same time very flexible and very efficient. Relying on a two-layered branch-and-bound setup, ESR performs object-based image retrieval in sets of 100,000 or more images within seconds.. After normalization and linear/non-linear combination, a master map or a saliency map is computed to represent the saliency of each image pixel. Last, a few key locations on the saliency map are identified by winner-take-all, or inhibition-of-return, or other non-linear operations.

Keywords:- feature extraction, saliency computation, visual attention.

I . INTRODUCTION

The human brain and visual system pays more attention to some parts of an image which seems more important than others. The salient object can be defined as the object in an image which draws most visual attention. The object or an area of image which is of most interest or is more important is labeled as salient objects, or foreground objects that we are familiar with. The salient object detection problem for images is formulated as a binary labeling task, where the salient object is separated from the background.

There are many applications for visual attention like automatic image cropping, adaptive image display on small devices, image video compression. Visual attention helps object recognition, tracking, and detection as well.

II . FEATURE EXTRACTION:

A.SMOOTHENING AND EDGE DETECTION:

The Gaussian filter is used for smoothening of an image. The Gaussian blur is a type of image-blurring filter that uses a Gaussian function for calculating the transformation to apply to each pixel in the image. In two dimensions, it is the product of two such Gaussians, one in each dimension:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \quad (1)$$

where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution.

Edges characterize boundaries and are therefore a problem of fundamental importance in image processing. Edges in images are areas with strong intensity contrasts – a jump in intensity from one pixel to the next. Edge detecting an image significantly reduces the amount of data and filters out useless information, while preserving the important structural properties in an image. There are many ways to perform edge detection. However we will use the gradient method The gradient method detects the edges by looking for the maximum and minimum in the first derivative of the image. An edge has the one-dimensional shape of a ramp and calculating the derivative of the image can highlight its location. A pixel location is declared an edge location if the value of the gradient exceeds some threshold. As mentioned before, edges will have higher pixel intensity values than those surrounding it. So once a threshold is set, you can compare the gradient value to the threshold value and detect an edge whenever the threshold is exceeded.

Based on this one-dimensional analysis, the theory can be carried over to two-dimensions as long as there is an accurate approximation to calculate the derivative of a two-dimensional image. The Sobel operator performs a 2-D spatial gradient measurement on an image. Typically it is used to find the approximate absolute gradient magnitude at each point in an input grayscale image. The Sobel edge detector uses a pair of 3x3 convolution masks, one estimating the gradient in the x-direction (columns) and the other estimating the gradient in the y-

direction (rows). A convolution mask is usually much smaller than the actual image. As a result, the mask is slid over the image, manipulating a square of pixels at a time. The actual Sobel masks are shown be

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Fig1: Sobel masks for edge detection

The magnitude of the gradient is then calculated using the formula:

$$|G| = \sqrt{G_x^2 + G_y^2} \quad (2)$$

An approximate magnitude can be calculated using:

$$|G| = |G_x| + |G_y| \quad (3)$$

B. ADAPTIVE THRESHOLDING

Thresholding is used to segment an image by setting all pixels whose intensity values are above a threshold to a foreground value and all the remaining pixels to a background value. Whereas the conventional thresholding operator uses a global threshold for all pixels, adaptive thresholding changes the threshold dynamically over the image. This more sophisticated version of thresholding can accommodate changing lighting conditions in the image, e.g. those occurring as a result of a strong illumination gradient or shadows.

Adaptive thresholding typically takes a grayscale or color image as input and, in the simplest implementation, outputs a binary image representing the segmentation. For each pixel in the image, a threshold has to be calculated. If the pixel value is below the threshold it is set to the background value, otherwise it assumes the foreground value. The main approach to find the threshold is: local thresholding. The assumption behind method is that smaller image regions are more likely to have approximately uniform illumination, thus being more suitable for thresholding. To find the local threshold is to statistically examine the intensity values of the local neighborhood of each pixel. The statistic which is most appropriate depends largely on the input image. the local intensity distribution, The size of the neighborhood has to be large enough to cover sufficient foreground and background pixels, otherwise a poor threshold is chosen. On the other hand, choosing regions which are too large can violate the assumption of approximately uniform illumination. This method is less computationally intensive than the Chow and Kaneko approach and produces good results for some applications.

C. MULTI SCALE CONTRAST:

Contrast is the most commonly used local feature for attention detection because the contrast operator simulates the human visual receptive fields. Without knowing the size of the salient object, contrast is usually computed at multiple scales.

Visual data is decomposed into multi-scale sub-images which contain multi-scale details on the basis of Gaussian Pyramid, and contrast features are extracted from these multi-scale images for saliency map generation. The multi-scale contrast feature $fc(x, I)$ is defined as a linear combination of contrasts in the Gaussian image pyramid:

$$f_c(x, I) = \sum_{l=1}^{L} \sum_{x' \in N(x)} \| I^l(x) - I^l(x') \|^2 \quad (4)$$

where I^l is the lth-level image in the pyramid and the number of pyramid levels L is 6. $N(x)$ is a 9×9 window. The feature map $fc(\cdot, I)$ is normalized to a fixed range [0, 1].

A Gaussian pyramid is a technique used in image processing, especially in texture synthesis. The technique involves creating a series of images which are weighted down using a Gaussian average (Gaussian blur) and scaled down. When this technique is used multiple times, it creates a stack of successively smaller images, with each pixel containing a local average that corresponds to a pixel neighborhood on a lower level of the pyramid.

Multiscale contrast highlights the high contrast boundaries by giving low scores to the homogenous regions inside the salient object.

D.CENTER SURROUND HISTOGRAM:

Histogram-based methods are very efficient when compared to other image segmentation methods because they typically require only one pass through the pixels. In this technique, a histogram is computed from all of the pixels in the image, and the peaks and valleys in the histogram are used to locate the clusters in the image. Color or intensity can be used as the

measure. The salient object can always be distinguished by the difference of it and its context.

They are insensitive to small changes in size, shape, and viewpoint. Another reason is that the histogram of a rectangle with any location and size can be very quickly computed by means of an integral histogram .

Suppose the salient object is enclosed by a rectangle R. A surrounding contour R_S with the same area of R is constructed. Here the $\chi 2$ distance between histograms of RGB color is used. The most distinct rectangle R*(x) centered at each pixel x is found out by varying the size and aspect ratio and is given by:

$$R^*(x) = \arg\max_{R(x)} X^2(R(x), R_s(x)) \quad (5)$$

Then,the center-surround histogram feature $f_h(x, I)$ is defined as a sum of spatially weighted distances:

$$f_h(x, I) \propto \sum_{\{x' | x \in R^*(x')\}} w_{xx'} X^2(R^*(x'), R_s^*(x')) \quad (6)$$

where R*(x') is the rectangle centered at x' and containing the pixel x. The weight,

$$w_{xx'} = \exp(-0.5 \cdot \sigma_{x'}^{-2} \|x - x'\|^2) \quad (7)$$

is a Gaussian falloff weight. Finally, the feature map fh(·, I) is also normalized. Thus the salient object has a large center surround histogram distance.

III .*SALIENCY MAP COMPUTATION:*

The purpose of the saliency map is to represent the conspicuity— or "saliency"—at every location in the visual field by a scalar quantity and to guide the selection of attended locations, based on the spatial distribution of saliency. A combination of the feature maps provides bottom-up input to the saliency map, modeled as a dynamical neural network.

One difficulty in combining different feature maps is that they represent a priori not comparable modalities, with different dynamic ranges and extraction mechanisms. Also, because all feature maps are combined, salient objects appearing strongly in only a few maps may be masked by noise or by less-salient objects present in a larger number of maps. In the absence of top-down supervision, we propose a map normalization operator, N(.), which globally promotes maps in which a small number of strong peaks of activity (conspicuous locations) is present, while globally suppressing maps which contain numerous comparable peak responses. N(.) consists of:

1) Normalizing the values in the map to a fixed range [0...*M*], in order to eliminate modality-dependent amplitude differences;
2) Finding the location of the map's global maximum *M* and computing the average *m* of all its other local maxima; and
3) Globally multiplying the map by $(M - m)^2$.

Only local maxima of activity are considered, such that *N*(.) compares responses associated with meaningful "activation spots" in the map and ignores homogeneous areas. Comparing the maximum activity in the entire map to the average overall activation measures how different the most active location is from the average. When this difference is large, the most active location stands out, and the map is strongly promoted. When the difference is small, the map contains nothing unique and is suppressed.

Feature maps are combined into three "conspicuity maps", for intensity, & for colour, and for orientation, at the scale (*s* = 4) of the saliency map. They are obtained through across-scale addition, "Θ" which consists of reduction of each map to scale four and point-by-point addition:

$$\bar{I} = \bigcup_2^4 \bigcup_{s=c+3}^{c=4} N(I(c, s)) \quad (8)$$

$$\bar{C} = \bigcup_{c=2}^4 \bigcup_{s=c+3}^{c+4} [N(RG(c, s)) + N(BY(c, s))] \quad (9)$$

For orientation, four intermediary maps are first created by combination of the six feature maps for a given *q* and are then combined into a single orientation conspicuity map:

$$\bar{O} = \sum_{\theta=\{0,45,90,135\}} N(\bigcup_{c=2}^4 \bigcup_{s=c+3}^{c+4} N(O(c, s, \theta))) \quad (10)$$

**U corresponds to across-scale addition

The motivation for the creation of three separate channels I, C, O and their individual normalization is the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency map. The three conspicuity maps are normalized and summed into the final input

S= (1/3) (N(I)+N(C)+N(O)) (11)

At any given time, the maximum of the saliency map (SM) defines the most salient image location, to which the focus of attention (FOA) should be directed. We could now simply select the most active location as defining the point where the model should next attend. However, in a neuronally plausible implementation, we model the SM as a 2D layer of leaky *integrate-and-fire* neurons at scale four. These model neurons consist of a single capacitance which integrates the charge delivered by synaptic input, of a leakage conductance, and of a voltage threshold. When the threshold I reached, a prototypical spike is generated, and the capacitive charge is shunted to zero. The SM feeds into a biologically plausible 2D "winner-take-all" (WTA) neural network, at scale *s* = 4, in which synaptic interactions among units ensure that only the most active location remains, while all other locations are suppressed. The neurons in the SM receive excitatory inputs from *S* and are all independent. The potential of SM neurons at more salient locations hence increases faster (these neurons are used as pure integrators and do not fire). Each SM neuron excites its corresponding WTA neuron. All WTA neurons also evolve independently of each other, until one (the "winner") first reaches threshold and fires.

This triggers three simultaneous mechanisms:

1) The FOA is shifted to the location of the winner neuron;

2) The global inhibition of the WTA is triggered and completely inhibits (resets) all WTA neurons;

3) Local inhibition is transiently activated in the SM, in an area with the size and new location of the FOA; this not only yields dynamical shifts of the FOA, by allowing the next most salient location to subsequently become the winner, but it also prevents the FOA from immediately returning to a previously-attended location.

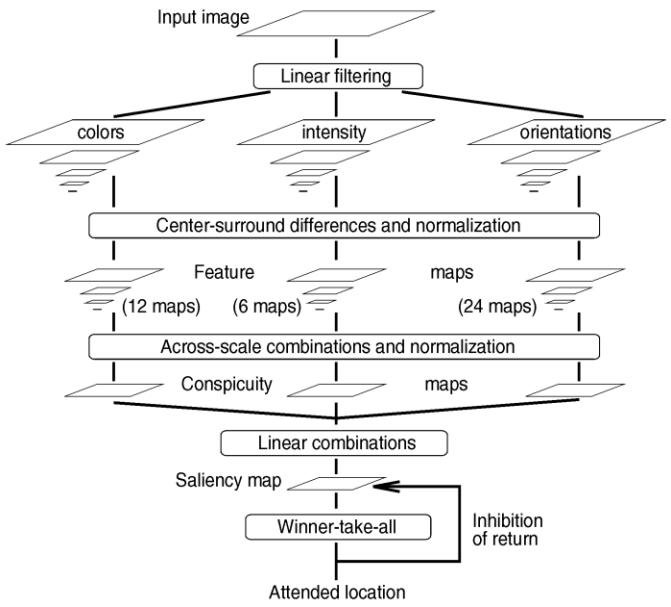

Fig 2: Itti's computational model for salient object detection

Such an "inhibition of return" has been demonstrated in human visual psychophysics. In order to slightly bias the model to subsequently jump to salient locations spatially close to the currently-attended location, a small excitation is transiently activated in the SM, in a near surround of the FOA. Since we do not model any top-down attentional component, the FOA is a simple disk whose radius is fixed to one sixth of the smaller of the input image width or height.

- This is Itti's computational model for salient object detection
- This is bottom up computational approach.

IV EVALUTION:

A.GROUND TRUTH CONSTRUCTION:

An image can have many salient objects and according to the user their idea of a salient object in an image may change. Hence for training of algorithm over a data set of number of images, the user is asked to specify the salient object in the image according to him. Henceforth, our algorithm puts a rectangle over the salient object. Then the ground truth of the salient image in the image is founded. The saliency probability map is,

$$g_x = \frac{1}{M}\sum_{m=1}^{M}(a)^m$$

(12)

Where,
M is the number of users.
$A^m = \{a_x^m\}$ the binary mask labeled by the m^{th} user.
With this saliency probability map the masked salient object is evaluated using three parameters viz. Precision, Recall and F-measure.
These are defined as,

B. PRECISION

Precision is defined as the ratio of correctly detected salient region to the detected salient region.

$$\text{Precision} = \frac{\text{correctly detected salient region}}{\text{detected salient region}}$$

$$\text{Precision} = \frac{\sum_x g_x a_x}{\sum_x a_x}$$

(13)

C .RECALL

Recall is defined as the ratio of correctly detected salient region to the ground truth of salient region.

$$\text{Recall} = \frac{\text{Correctly detected Salient Region}}{\text{Ground Truth salient region}}$$

$$\text{Recall} = \frac{\sum_x g_x a_x}{\sum_x g_x}$$

(14)

D. F-MEASURE

F-measure is the weighted harmonic mean of precision and recall with a non- negative α. F-measure is the overall performance measurement.

$$F_\alpha = \frac{(1+\alpha)\cdot\text{Precision}\cdot\text{Recall}}{\alpha\cdot\text{Precision}+\text{Recall}}$$

(15)

Where,

α =0.5

To obtain the salient object in an image we have to resolve the rectangles by maximizing the values for the expression:

$$f(x) = \sum_{x \in R}(1 - F(x)) + \sum_{x \text{ does not} \in R} F(x)$$

(16)

Where,
R is resolve rectangle;

F(x) $\in$ [0, 1] is the normalized saliency map.

V . CONCLUSION

Our project is based on bottom up computational approach only. We have completed basic image processing techniques related to image enhancement, feature extracion, segmentation and object reconisation. We have obtained feature maps by thresholding and boundary determination. We are looking forward to combine them to give saliecy map and determine salient object in an Image. We are currently working with grey scale static images. The algorithm developed in this paper can be extended to the colour (RGB) images and sequential images (video) also. We have defined the evalution criteria to calculate the efficiency of our algorithm. After completion of the implementation we can evalute our results through techniques mentioned.

Fig 3:Expected results after the application of the algorithm

REFERENCES

[1] Tie Lie, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, "Learning to detect Salient Object"

[2]L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[3] Visual attention home page "www. ilab.usc.edu/bu/"

[4] Saliency toolbox by Dirk Bernhardt-Walther
www.saliencytoolbox.net

[5] E. H. Adelson, C. H. Anderson, J. R. Bergen "Pyramid methods in image processing"-1984

[6[Y. Guillemaut, J. Kilner, J. Starck, A. Hilton, "Dynamic Feathering: Minimising Blending Artefacts 2009.

[7] View-Dependent Rendering" - IEEE transaction on Image processing 2008..

[8] Jiaya Jia, Chi-Keung Tang, "Eliminating Structure and Intensity Misalignment in Image Stitching" Research Grant Council of Hong Kong Special Administration Region, China (AOE/E-1999).

[9] H.-Y. Shum and R. Szeliski. "Construction of panoramic mosaics with global and local alignment". International Journal of Computer Vision, 36(2):101–130, February 2000. Erratum published July 2002,48(2):151-152.

[10] Xia Wan and C.-C. Jay Kuo" A New Approach to Image Retrieval With Hierarchical Color Clustering" ieee transactions on circuits and systems for video technology, september 1998