AN ATTENTION MODEL FOR EXTRACTING COMPONENTS THAT MERIT IDENTIFICATION

Mohammad Jahangiri and Maria Petrou

Imperial College London
Department of Electrical and Electronic Engineering

ABSTRACT

Cognitive systems are trained to recognise perceptually meaningful parts of an image. These regions contain some variation, i.e. local texture, and are roughly convex. We call such regions "blobs". *We define blobs to be components that merit further analysis by a higher level interpretation module* as they very likely constitute semantically meaningful units, rather than characteristic features or salient spots. A scheme, independent of scale and colour, is proposed, based on the use of Gaussian kernels and mathematical morphology for the extraction of blobs. For understanding how well the extracted blobs match the meaningful regions, we present an eye-tracking experiment using 20 subjects and 20 different colour images using the hypothesis that the gaze of the viewers are more attracted to the meaningful regions/objects of a scene. We show that the gaze of the subjects is attracted more to the regions which were extracted by our model in comparison with the regions which were extracted by the saliency map model, proposed by Itti and Koch.

***Index Terms*—** Blob Detection, Attention Model

1. INTRODUCTION

For reducing the complexity of the scene, primates analyse a subset of the available sensory information. In vision these subsets correspond to the interesting parts of a scene which contain some variation i.e. some local texture. For example, eye tracking experiments have shown that the gaze of the viewer when looking at a face is attracted by the eyes and the mouth and far less by the cheeks which are in comparison flat areas [1]. In general, parts of an image that may contain interesting structures and may require interpretation are those where the image gradient shows significant spatial variation. Flat regions are hardly informative on their own. In an automatic system of interpretation, therefore, a mechanism is required to draw attention of the system to those parts of the image that are most likely to contain useful or meaningful information. This mechanism should not be scale or colour sensitive and it should be such that regions of interest may be easily identified in its output in the form of "blobs", by a simple thresholding. Based on these ideas, we propose in this paper a methodology for blob extraction based on the calculation of local gradient magnitude at various scales.

"Blob" in the computer vision literature refers to points and/or regions in the image that are either brighter or darker than their surroundings [2]. There is an extensive work in the literature for extracting blobs in images some of which may be found in [2, 3, 4, 5, 6, 7, 8, 9]. In this paper we propose a blob detector which is inspired by the human vision system and in particular by the complex cells of the V1 region that are composite and act as line and edge detectors. In contrast to [7], our model does not include colour as one of the pre-attentive features: we are interested in identifying a door for further interpretation either it is red or blue, and we are not interested in interpreting first a red door that sticks out more prominently in a yellow wall than a white door does. So we are not ranking the extracted regions in any sense of saliency. Further, unlike [9], we are not interested in the recognition of specific objects. *Our strategy is to extract regions which merit further analysis by higher levels of interpretation*[1], as they very likely constitute semantically meaningful units, rather than characteristic features or salient spots.

2. METHODOLOGY

Steps of the proposed methodology for extracting blobs in colour images are shown in figure 1. In the first step an edge

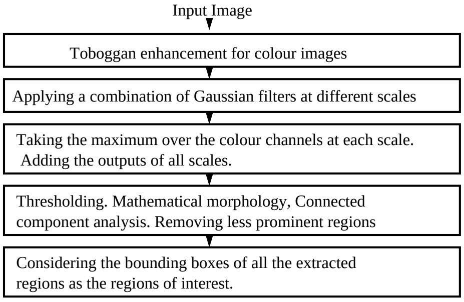
Input Image
Toboggan enhancement for colour images
Applying a combination of Gaussian filters at different scales
Taking the maximum over the colour channels at each scale.
Adding the outputs of all scales.
Thresholding. Mathematical morphology, Connected component analysis. Removing less prominent regions
Considering the bounding boxes of all the extracted regions as the regions of interest.

Fig. 1. The block diagram of the proposed algorithm

[1] By "higher levels of interpretation" we mean applying a pattern recognition algorithm for object recognition.

preserving smoothing algorithm based on the work of [10] is applied to the image (see figure 2-b). In the second step, since we wish to identify regions of interest on the basis of the variation they show, irrespective of colour and scale, we apply a combination of Gaussian based filters to the enhanced image. As we are not interested in the positive or negative edges, dark or bright lines and blobs, we use the first and second derivatives of the Gaussian as filters and take the absolute value of all outputs. Furthermore, as we wish to measure local structure independent of colour, we apply these operators in each band separately and take the maximum response of all three bands for each combination of filters. Finally, as we are not interested in any particular scale, we apply filters of various sizes, just like the human vision system does, having cells that work at a variety of scales. Once all these filter outputs have been computed, their values are added, to produce the final output which is expected to highlight the regions of interest where an interpretation module should be directed. The chosen scales for each Gaussian are $\left(\sqrt{2}, \sqrt{2}^2, ..., \sqrt{2}^s\right)$. This is because these scales have linear characteristics in the scale space [2]. In our experiments we set $s = 3$. These steps yield an interest map an example of which it is shown in figure 2-c.

For extracting blobs we first binarise the interest map by selecting a global threshold using an automatic thresholding algorithm [11]. We implement the following steps subsequently on the binarised map to extract the regions which contain blob structures.

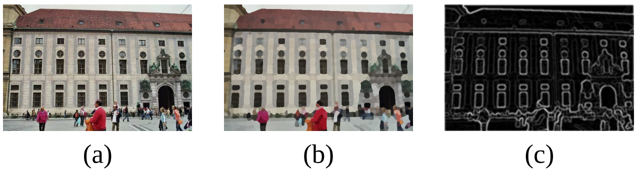

Fig. 2. a) Original Image b) Edge preserving smoothing c) Interest map of the combined approach of Toboggan enhancement and Gaussian filtering.

1) Fill in the closed contours.
2) Use morphological opening to remove thin extrusions. The structural element for opening is selected automatically from the mode of the histogram of the thickness of regions (vertical and horizontal directions separately) in the binary map obtained in the previous step.
3) Compute the following regularity criterion for each of the connected components in the binarised map

$$Regularity = \frac{\#\{pixels \in (R \cap C\})}{\#\{pixels \in C\}} \tag{1}$$

where C is the set of pixels that make up the boundary of the convex hull of the region, and R is the set of pixels that make up the boundary of the region. Regions with regularity more than a threshold are considered as regions of interest. The threshold in our experiments was chosen to be equal to 0.4.
4) Delete the identified regions in the previous step from the binary map. For extracting more regions from the binary map, we apply connected component analysis to the black pixels of the binary image. We identify any connected component (in the black regions) which does not touch the border of the image and is also regular. The identified regions are dilated and considered as another set of extracted blobs. The pixels of this new set of regions are turned to black in the input binary image. From the remaining regions, we select connected regions which fulfil the regularity criterion.
Results of applying these steps to the interest map shown in figure 2-c, are shown in figure 3.

For improving the extraction performance when we have overlapping blobs, we delete the regions which are less prominent in comparison with the other ones, using the constructed interest map. In the final stage the bounding boxes of the extracted binary regions, which fulfil some regularity criterion, are considered as extracted regions. Consider a connected component, Ω, which consists of more than two overlapping regions. In other words $\Omega = \omega_1 \cup \omega_2 \cup \cdots \omega_n$ where $\omega_1, \omega_2, \cdots, \omega_n$ are some extracted blobs and $n \geq 2$. We use the following steps to remove the regions which may be of least interest.

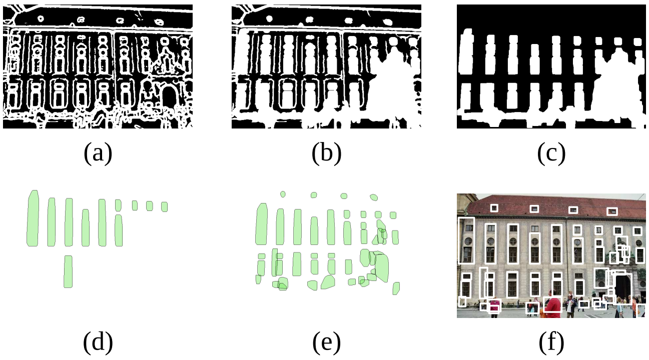

Fig. 3. a) Binarisation b) Filled contours c) Opening d) Initial blobs extracted e) Further blobs extracted f) Bounding boxes of all the blobs extracted

1) Rank each region according to the number of regions with which it has an overlap.
2) Select the region with the highest rank, say ω_m, which, say, is overlapped with regions ω_{m1}, ω_{m2} and ω_{m3}.
3) Compute the prominence of each of the regions ω_m ω_{m1}, ω_{m2} and ω_{m3}. The prominence is defined as follows:

$$Prom(\omega_i) \equiv \frac{\sum_k R\left(k\right)}{Area\left(\omega_i\right)} \tag{2}$$

where the summation in the numerator is over all pixels that belong to the part of the region that does not overlap with any other region, matrix R is the interest map and the $Area$

function computes the total area of the region.
4) Omit the region which has the minimum value of $Prom$ among all regions in Ω.
5) Repeat steps 1 to 4 until no regions overlap.
The result of applying the proposed technique on the extracted regions shown in figure 4-a is shown in figure 4-b.

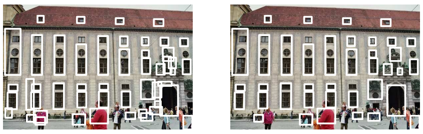
(a) Before removal (b) After removal

Fig. 4. Result of removing overlapping regions

3. EXPERIMENTAL RESULTS

The main purpose of the proposed methodology is extracting regions which *merit* further analysis by a higher level interpretation module. These are the regions which attract the gaze of most of the viewers when looking at a scene investigatively. Therefore, we designed an experiment, using an eye-tracker, to understand which regions are mostly seen by the viewers while looking at an image investigatively. From this experiment we extracted a map. This was used to compare the results of our algorithm with those obtained by the algorithm in [7]. The details of our experiments are as follows.

Eye-tracking experiment design The Tobii T60 infrared eye tracking system was used to record the eye position every 16 ms. Twenty different images were presented in a random order to each of the subjects on a $17''$ TFT $(1280 \times 1024$ Pixels) monitor. The participants were seated comfortably in an ordinary office facing the screen from a distance of about one meter. Twenty persons contributed to this experiment, so 400 gaze paths were collected.

The image ensemble included 20 images of faces (Caltech faces), football match images, buildings and some other outdoor and indoor scenes. Before starting each experiment, the subject was informed that a question succeeds each presented image. The questions were chosen to be some general simple questions like “Was the image of a woman or man?”, “What game was the match?”,“How old was the building?” etc. This was for encouraging the subjects to look at the image investigatively and at the same time do not bias their way of viewing. A brief central fixation cue preceded each 10 s image presentation. The tracking error was less than 0.5° of visual angle. The analysis was based on eye positions from 0.4-10 s after presentation of the image. The first 0.4 s were omitted to avoid any bias due to the central fixation cue preceding each image.

Building the gaze Map Each fixation point was computed by taking the average of the right eye fixation and the left eye fixation point. The smallest distance (in pixels) that separates the fixations was set to 25. Therefore, for constructing the gaze map of each image we took the fixation point and considered a circle with radius equal to 25 around it. The pixel values inside the circle were set to be equal to the amount of time (in ms) that a viewer had gazed at that fixation point. We then accumulated the maps of the participants to construct the gaze map which we used for our further analysis. Some sample images and the accumulated gaze maps are shown in figure 5.

Comparison with saliency map We used the Saliency toolbox developed by Itti and colleagues to extract regions from the saliency map model. This toolbox can be downloaded freely from http:/www.saliencytoolbox.net. In order to have a fair comparison, the number of successive salient regions is chosen to be equal to the number of extracted regions in our approach. Next, for each extracted region in each image the following criterion was computed:

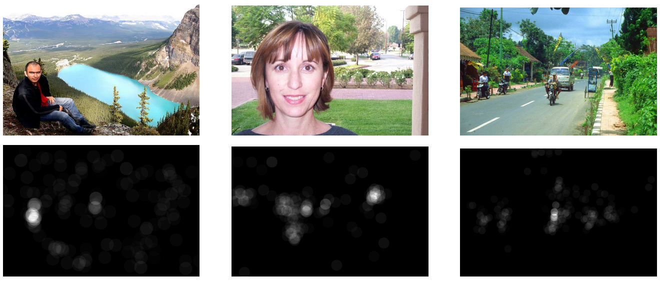

Fig. 5. Some of the images in our database are shown. Below each image we show its gaze map. The whiter a pixel, the more the gaze of the viewers was attracted to that pixel.

$$ME = \sum_{k=1}^{M} \frac{Gazemap(Pixels \in R_k)}{Area\,(R_k)} \quad (3)$$

where, $Gazemap$ is the accumulated map constructed by the eye-tracking experiments, M is the number of extracted regions in our approach and R_k is the k_{th} identified region. We computed ME both for the extracted regions in our model and the extracted regions of the saliency map model. These values are plotted in figure 6 against the indices of images in our database. From this figure it can be seen that ME in our approach is more than the saliency map model in almost all the cases. Therefore it can be inferred that in this ensemble of images the extracted regions of our model attract the gaze of the viewers more than the regions which were extracted by the saliency map model.

In figure 7 the bounding boxes of the extracted regions are superimposed on some sample images. It can be seen that the extracted regions usually contain a region which can be further classified to be a window, door, chimney, player, face,

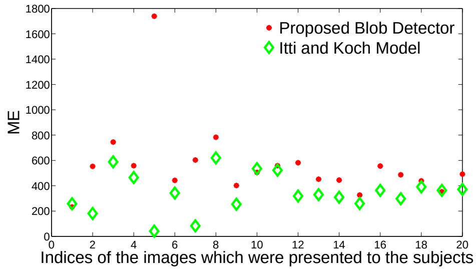

Fig. 6. ME for different images are plotted against the indices of the images in the database. In 85% of the images, the value of ME in our approach is more than the saliency map model.

eyes, flag, lip, building etc. Finally, we applied our algorithm, which is fully automatic, and without changing any of its parameters, to a database of 280 images of building facades from https://www.ipb.uni-bonn.de/svn/etrims-img-dbs/ with manually outlined and perceptually meaningful components. For testing whether an extracted region corresponds to a manually segmented region or not, we define:

$$O \equiv \max_{S} \left\{ \frac{\#\left\{pixels \in (R \cap S)\right\}}{max\left(\#\left(pixels \in R\right), \#\left(pixels \in S\right)\right)} \right\} \quad (4)$$

where R is the set of pixels which belong to a blob, and S is the set of pixels which make up a hand segmented region of interest. If O is more than 0.6 then we infer that the blob correspond to one of the manually segmented regions. Our algorithm detected 5118 blobs from these images from which 2864 regions correspond to one of the manually segmented regions (out of the total number of 3589 segmented regions)

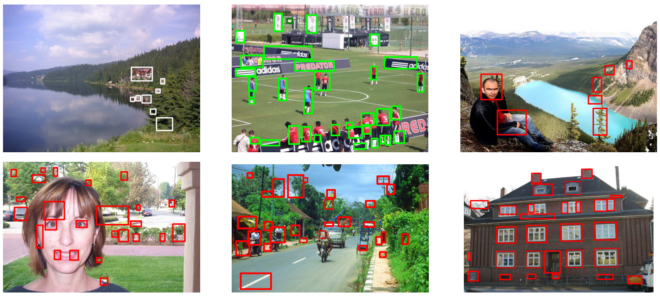

Fig. 7. Bounding boxes of identified blobs are superimposed on the images.

4. CONCLUSIONS

In this paper we define a blob to be a region that merits further analysis by a scene interpretation module, possibly depicting a perceptually meaningful entity. We proposed a fully bottom up approach for extracting such blobs in images. For this we proposed a fully automatic, bottom up scheme independent of scale and colour, based on the use of Toboggan enhancement, Gaussian kernels, mathematical morphology and connected component analysis. We designed an eye-tracker experiment which aimed at making the viewer look at an image in an investigative way. Our experiments showed that the performance of the proposed algorithm is far better than the state of the art saliency model [7], in extracting the regions which attract the gaze of the viewer when looking at a scene investigatively. This is because the saliency map models pre-attentive saliency, while our approach is designed to model *investigative* driven attention which is also different from goal driven attention where the person looks for a specific object. Further experiments with 280 images of building facades which contain thousands of manually segmented perceptually meaningful regions showed that our fully automatic, fully bottom up approach, could extract 79.8% of them.

5. REFERENCES

[1] A. L. Yarbus, *Eye Movements and Vision*, Plenum Press New York, 1967.

[2] T. Lindeberg, “Scale space theory in computer vision,” in *Kluwer*, 1994.

[3] C. Steger, “Subpixel-precise extraction of watersheds,” in *ICCV*, 884–890,1999.

[4] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *IJCV 60 (2)*, 91–110, 2004.

[5] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *IJCV 60(1)*, 63-86, 2004.

[6] J. Matas, O. Chum, M.Urban, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” in *British Machine Vision Converence*, 2002.

[7] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” in *PAMI 20(11)*, 1254–1259, 1998.

[8] R. P. Roa, G. Zelinsky, M. Hayoe, and D. H. Ballard, “Eye movements in iconic visual search,” in *Vision Search 42(11)*, 1447–1463, 2002.

[9] T. Serre, L. Wolf, S. Bileschi, M. Riesnhuber, and T. Poggio, “Robust object recognition with cortex-like mechanisms,” *PAMI 29(3)*, 411–425, 2007.

[10] J. Fairfield, “Toboggan contrast enhancement,” in *SPIE 1708*, 282–292, 1992.

[11] M. Petrou and P. Bosdogianni, *Image Processing: The Fundamentals*, Wiley, 1999.