THE SALIENCY GROUPING FIELD

Maurizio Pilu
HP Labs Europe
Bristol, UK

ABSTRACT

A saliency map specifies the relevance of locations in an image and good methods exist today for recovering it automatically in a bottom-up fashion. This paper addresses the problem of finding organization in an actual saliency map of everyday images through a grouping field in order to elicit structure that can be used for a variety of purposes. After calculating the saliency using a version of the Itti *et al.* model, we retain only the maxima of the map as features. Then we calculate an initial grouping field by convolving the salient locations with orientation-selective grouping kernels. Finally we iteratively reinforce the field to enhance perceptually salient components.

1. INTRODUCTION

A saliency map is a map associated to an image that indicates the relevance of a particular image location based upon various criteria. Many of the usable saliency detection methods, e.g. [1], take an input image such as that in Figure 1-left and produce a map similar to that of Figure 1-center. These maps have started to be used in some applications [2], but a lot is still to be done. One of the reasons is that although they indicate what could be salient in an image, they are still relatively low-level (each pixel has a saliency value) and in fact their main use so far has been in directing attention for higher-level processing stages. In a recent work [2] the saliency map is binarized to give a score to regions of interest that are pre-specified.

In this work we use realistic saliency maps, with the limitation they entail, to elicit image structure. We do this by recovering what we call the *saliency grouping field*, a vector field akin to [3] indicating at each salient location and in a set of directions the probability of having other salient locations associated to it.

Over the years many perceptual organization works have tried to perform grouping of low-level, often conceptual image features. The present works build upon them but uses realistic saliency features, the maxima of the saliency map as suggested by Itti *et al.* [1]. An example is given in Figure 1-right.

After calculating the saliency using a version of [1], we retain only the maxima of the map as features. Then we calculate an initial grouping field by convolving the salient location with the orientation-selective grouping kernels of [4]. Finally we iteratively reinforce the field at locations where a particular grouping direction stands out from others while also being consistent with neighbors in that direction.

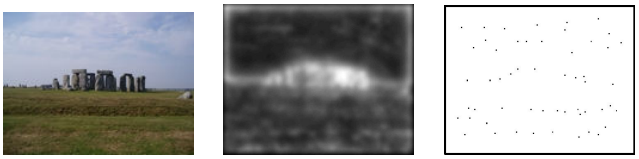

Fig. 1. An image (left), its saliency map computed as in Section 2 (center) and the maxima (right) which we consider, idem-potentially, as salient features.

2. SALIENCY MAP USING THE ITTI *ET AL.* MODEL

A number of works have attempted to define a model for bottom-up saliency, or more precisely the generation of a context-free interest map from an image. Privitera and Stark [5] provide a taxonomy of the various methods and in a later work even compare the results with actual eye movements.

The method we employed is that based on biologically plausible principles due to Itti *et al.* [1], which is thought to be one of the best bottom-up saliency models. We are not going to report here the details of the approach, which is only briefly overviewed below.

The method works by first determining three image-topic maps, or channels, the intensity channel, the color channel and the orientation channel. Each of the three channels are computed by combining center-surround responses across different spatial scales. The intensity map represents intensity contrast, scoring highly when something is brighter or darker than its surround. The color map is based on the color opponent theory and uses the contrast between red/green in the surround and green/red in the center. The orientation maps are computed combining the center-surround responses of a bank of Gabor filters. Finally, a normalization operator that emphasizes maps with few, strong conspicuous locations is then applied to each channel and the saliency map is simply computed as the average of the three normalized channels.

Figure 1-center shows an example of saliency map S

0-7803-8554-3/04/$20.00 ©2004 IEEE.

obtained from the image of Figure 1-left. More results are visible in Figure 5.

3. THE GROUPING KERNELS

Following [4] we use a parametric polar lemniscate kernel to weight saliency responses that are at distance ρ and angle θ from the current location:

$$k(\rho,\theta) = \begin{cases} e^{\frac{1}{2}\frac{\rho^2}{\sigma^2}\cos^{2n}(\theta)} & \text{if } \frac{\pi}{2} \leq \theta \leq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases}$$

The parameter σ defines the spatial support of the kernel whereas n is associated to the orientation selectivity.

An actual $(2h+1) \times (2h+1)$ kernel K of bandwidth h is built by applying Eqn. 1 for the polar coordinates ρ and θ of the kernel elements with respect to the center of the kernel. Figure 2-left shows a kernel of bandwidth $h = 15$ for $\sigma = 2$ and $n = 5$.

We use a bank of four rotated grouping kernels $K_n, n = 1 \ldots 4$ covering a finite number for orientation angles $\alpha_i = \{0, \frac{\pi}{2}, \pi, \frac{3}{4}\pi\}$. Each rotated kernel is simply created by applying $k(\rho, \theta - \alpha)$.

These kernels have been used in some other perceptual organization works, such as [6].

4. INITIALIZING THE GROUPING FIELD

Salient locations in an image are determined by non-maxima suppression of the saliency map, as suggested by [1]. Moreover we assume that the salient locations are idem-potent; the reason for this choice is explained in the discussion in Section 7.

Let us define for convenience an operator $\mathcal{M}(A)$ that turns a discrete scalar field A into a binary one that is 1 at the location of all the local maxima of A and 0 elsewhere.

Using the saliency map S, we then form a new binary map $S_{max} = \mathcal{M}(S)$ that is non-zero only at the location of all the m maxima $M_i = \{x_i, y_i\}$ of S.

We then convolve this map with the kernels K_n (see Figure 2-right) to create four scalar fields $G'_n = K_n * S_{max}$ which are normalized at each location to obtain four normalized scalar fields:

$$G_n(x,y) = \frac{G'_n(x,y)}{\sum_{n=1}^{4} G'_n(x,y)}$$

with $\sum_{n=1}^{4} G_n(x,y) = 1$, indicating the strength of each response along the corresponding orientation.

Figure 4-left shows the grouping field overlaid on the saliency map. At every maxima location $M_i = \{x_i, y_i\}$ there associated four scalars $G_n(x_i, y_i)$.

Note that given that S_{max} is a linear combination of m Dirac pulses at all the M_i (Figure 2-right), the grouping field is also the superposition of the m grouping kernels placed at all the M_i. Moreover, since we are using a sparse matrix, more precisely non-zero only at the location of the maxima, the convolution could be speeded up considerably by taking into account only the non-zero element of S_{max}, effectively implementing a weighted sum only at the locations of the maxima. The convolution notation has been used only for clarity.

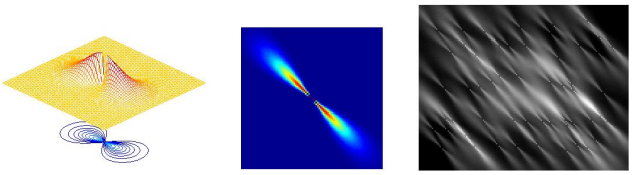

Fig. 2. Left: An example of grouping kernel of Section 3. Center: A rotated kernel and (right) the results of the convolution of the saliency maxima map S_{max} with this kernel.

5. REINFORCING THE FIELD

Although the grouping kernels manage to capture the relationship between saliency maxima, in particular when two or more fall within the support of the kernel, in most cases there is considerable ambiguity in the responses. As we said earlier this is due to the fact that the saliency model of [1] bears no direct orientation information, unlike, for example, what was assumed in [4].

In order to elicit the perceptual structure of the saliency maxima we use an iterative reinforcement process that progressively reinforces perceptually aligned components of the field and inhibits components that exhibit no perceptual structure.

We have experimented with a number of reinforcement criteria and here we present an approach that produces good results and manifests good convergence properties.

The reinforcement iterations consists of the following updates of the fields:

$$G_n(x_i, y_i) \longleftarrow G_n(x_i, y_i) \cdot [1 + \mu \mathcal{R}_n(x_i, y_i)]$$

where $\mathcal{R}_n(x,y)$ is a scalar reinforcement function which is positive when the n^{th} component of the field is to be reinforced at (x,y) and negative when is to be inhibited. $\mu \ll 1$ is a gain factor.

The reinforcement function implements two simple principles: a) reinforcing the strongest field component at each location and b) reinforcing a component if it has other neighbors whose strongest field components have the same orientation.

Computationally we express each element of $\mathcal{R}_n$ as

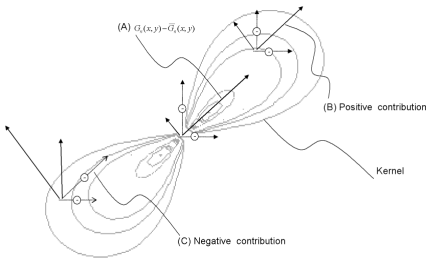

Fig. 3. Illustration of the principle behind the reinforcement updates. See text.

$\mathcal{R}_n(x,y) = \frac{\mathcal{R}'_n(x,y)}{\sum_{i=1}^{4} \mathcal{R}'^{i}(x,y)}$, from the un-normalized map

$$\mathcal{R}_n^{'} = K_n * (G_n - \overline{G}). \quad (1)$$

The scalar field $\overline{G} = \frac{\sum_{i=1}^{4} G_i}{4}$ represents the average orientation response at each location. The second factor of Eqn. 1 is positive when response G_n is above the average response at that location, thereby ensuring that the strongest ones are favoured. The convolution with the (same) grouping kernels K_n used in Section 4 ensures that neighboring components along the same directions that are dominant too have a positive contribution on $\mathcal{R}'_n$ whereas those that are not so dominant have a negative contribution.

Figure 3 illustrates what is going on for a simple case. In (A) there is shown the four components of the vector field at a location (x, y) after taking off the their average, that is $G_n(x,y) - \overline{G}(x,y)$. The $\ominus$ symbol indicates a negative value. The strongest response in at $\pi/4$. At two other locations (B) and (C) the respective components are shown. In (B) a strong positive component is found in the same $\pi/4$ direction as in (A), whereas in (C) the strongest response is at $3\pi/4$ while at $\pi/4$ we have a negative contribution, due to the fact that in that direction the response was below average. By convolving with the kernel K_n we weigh these contributions according to the grouping field described in Section 3; in the specific case in the figure, the overall effect is that $\mathcal{R}_n(x,y)$ is slightly positive, since the kernel will give a bigger weight to (B) rather than (C), (B) being more aligned with the reference direction and closer to (x, y). Figure 4 shows how the initial field has been "sharepend" by this reinforcement process, now showing dominant grouping directions.

6. RESULTS

This section shows some results that illustrate what grouping field is to be expected when using this method to a range of images; we shall also explain, case by case, what kind of structure is elicited by the grouping field.

The examples in Figure 5 are ordinary photos down sampled to about $20,000$ pixels, shown in column (a). The saliency maps calculated using the method of Section 2 are shown underlaid to the fields in column (c). The binary feature maps corresponding to these images and for which we recover the grouping field are shown in column (b).

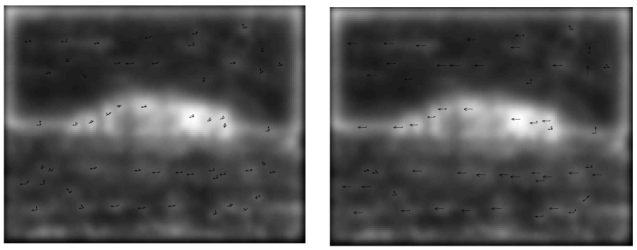

Fig. 4. Initial grouping field (left) and final field after the reinforcement iterations (right).

The vector fields in column (c) are the final grouping fields after the reinforcement process of Section 5. With similar results as in Figure 4, the final fields are much "sharper" than the initial ones (not shown here for lack of space) and the strongest response usually indicates a dominant direction along which a number of salient locations are present.

Note that at several locations two strong responses are still present; this is an extremely good property of the process (a consequence of its convergence properties) that preserves as much as possible strong alternatives. Another manifestation of the efficacy of the approach is that often a dominant direction after the initialization is superseded by another direction after the reinforcement, since the process converges to a solution that takes into account global, rather than local, structure.

In the first example - first row - from the saliency maxima in (b) a couple of strictures pop out to our eyes, notably two leaving from the middle on the left border and leading one to the chin the other to the hand, respectively. In the resulting grouping field in (c) these two structures are made clear. In the second example - second row - the mainly horizontal structure of the image is well captured by the final field (c). Note the two other minor structures elicited, the one of the high tree on the left margin and the bent foliage of the tree three-quarters to the right. Smaller structures that were present in the saliency map were suppressed by the reinforcement, which has privileged the mainly horizontal course of the scene. In the third example - third row - two main structures of the saliency maxima in (b) have been elicited by the field shown in (c), the outline of the massif, and the horizontal arrangement at the road level. In the final example - fourth row - there are a number of structures elicited by the field in (c) which also pop out in the maxima of (b), in particular one going along the waterfall, one along the ridge of the cliff , one along the tree on the right and two more subtle ones, the first crossing horizontally from the boulder in the lower-right across the subject legs and the

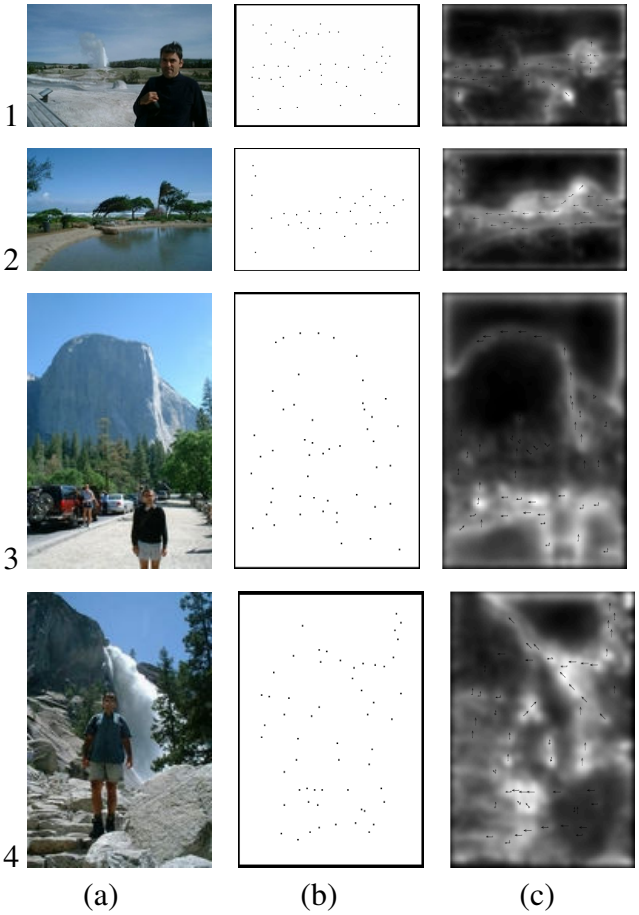

Fig. 5. Four example showing the input images (a), the salient location map (b) and the recovered grouping fields overlaid on the top of the resepctive saliency maps (c).

other along the body.

7. DISCUSSION

In this work we have addressed the problem of finding structure in an image by applying one of the strongest principles of perceptual organization, good continuation, to the output of an actual saliency computation algorithm, which models what is important in an image.

Along with many other perceptual organization works, we express perceptual structure in the form of a grouping vector field which specifies features association probabilities [3]. We believe this is an essential pre-processing stage for further higher level processing but we make no claim about modelling the purposive visual scanning path in human vision.

Looking at the results of Figure 5 it is important to notice that what we have presented is not an "expensive" edge detector. In fact the only information that we are using to produce the field is the saliency maxima of column (b), which sometimes just happen to occur along the length of edges. However, the maxima could be originating from any feature, e.g. some cars and a face, as it happens in the third example in Figure 5.

To recover the field we have not used the the value of saliency. In fact, as explained in [1], saliency models often look globally for salient regions while repressing others, which could also be salient had we had contextual knowledge. A next step would be to incorporate this and other high-level contextual or semantic information such as region homogeneity, or mid-level information, such are grouped edges. Here we have relied only on point-like features, which however originated from an actual computation of a saliency map using [1].

An application of the method proposed is the automatic browsing of images on low-resolution screens (e.,g. for mobile phones) where we would like to zoom-in and pan across the images showing to the user interesting locations. Grouping fields such as the ones in Figure 5 will drive the automatic generation of such viewing paths without relying on semi-manual mark up like in [2].

8. REFERENCES

[1] Laurent Itti, Christof Koch, and Ernst 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.

[2] Hao Liu, Xing Xie, Wei-Ying Ma, and Hong-Jiang Zhang. Automatic browsing of large pictures on mobile devices. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 148–155. ACM Press, 2003.

[3] Lance R. Williams and David W. Jacobs. Stochastic completion fields: A neural model of illusory contour shape and salience. *Neural Computation*, 9(4):837–858, 1997.

[4] F. Heitger and R. von der Heydt. A computational model of neural contour processing: Figure-ground segregation and illusory contours. *International Conference on Computer Vision*, 93:32–40, 1993.

[5] C.M. Privitera and L.W. Stark. Evaluating image processing algorithms that predict regions of interest. *Pattern Recognition Letters*, 19(11):1037–1043, September 1998.

[6] A.A. Robles-Kelly and E.R. Hancock. Grouping line-segments using eigenclustering. In *Proceedings of the British Machine Vision Conference*, 2000.