Perceptual Grouping Using Global Saliency-Enhancing Operators†

Gideon Guy and Gérard Medioni

Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, CA 90089-0273

Abstract

We introduce saliency-enhancing operators capable of highlighting features which are considered perceptually relevant. We are able to extract salient ***curves*** *and* ***junctions*** *and generate a description ranking these features by their likelihood of coming from the original scene. We suggest the* ***global extension field*** *as means of describing the behavior of a curve segment, in terms of its continuation. We show that a directional convolution of an edge image with the above field can produce useful descriptions.*

Other fields are also used in the same manner to produce similar results for domain-specific applications. The scheme is particularly useful and robust as a gap filler and in the presence of noise.

It is interesting to note that all operations are parameter-free, non-iterative and the processing is linear in the number of edges in the input image.

1 Introduction

An area which is likely to improve results in computer vision is the one of perceptual grouping. Perceptual Grouping can be classified as a mid-level field directed toward closing the gap between what is produced by state-of-the-art low-level algorithms (such as edge detectors) and what is desired as input to high level algorithms (perfect contours, no noise, no fragmentation, etc.). Many researchers resort to using synthetic data as their input because of these weaknesses.

It is clear that humans are able to group objects into higher level entities, and that such skill is helpful (if not a must) in reaching intelligent decisions about a given scene, as well as reducing the complexity of later processing.

† This research was supported by the Advanced Research Projects Agency of the Department of Defence and was monitored by the Air Force Office of Scientific Research under Contract No. F49620-90-C-0078, and by NSF Grant No. 90-24369.

Figure 1(a) depicts an example of perceptual groupings easily experienced by the human visual system. The geometric shapes are easily distinguishable from the noisy background. Furthermore, we tend to fill the gaps and accept the fragmented curves as complete ones. A more striking example of *illusory* contours is found in the Kanizsa illusion[4] shown in figure 1(b). Here we perceive

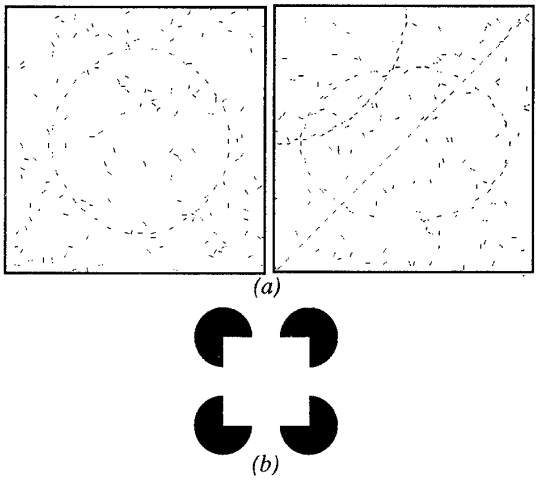

Figure 1 (a) Two instances of perceptual arrangements. (b) The Kanizsa square illusion.

edges which have no physical support whatsoever in the original signal. The process is known to be pre-attentive and hence domain-free.

It is obvious that extracting these structures is an exponentially problem, since in theory all possible subsets of the features need to be checked. Also, there is no easy way to define a consistent (and constant) metric with which to evaluate the possible groupings. Some methods to overcome these difficulties have been suggested over the years.

Lowe [5] discusses the Gestalt notions of co-linearity, co-curvilinearity and simplicity as important in perceptual grouping. Ahuja *et al.* [1] suggest methods for clustering

0-8186-2910-X/92 $3.00 © 1992 IEEE

and grouping sets of points having an underlying perceptual pattern.

Dolan and Weiss [2] demonstrate a hierarchical approach to grouping relying on compatibility measures such as proximity and good continuation.

Mohan and Nevatia [6] assume a-priori knowledge of the contents of the scene (i.e. aerial images). A model of the desired features is then defined, and groupings are performed according to that model. In a later work [7], groupings based explicitly on symmetries are suggested, but the first connectivity steps are performed locally.

Ullman *et al.* [10] suggest the use of a saliency measure to guide the grouping process, and to eliminate erroneous features in the image. The scheme prefers long curves with low total curvature, and does so by using an incremental optimization scheme (similar to dynamic programming). Others (like [11]) have also looked at similar problems. All of the above methods rely on some local operator to 'reveal' global structure and thus cannot perform well on noisy images.

2 Our approach

2.1 Overview

As was demonstrated before, the physical evidence extracted *locally* from images (through edge detectors) does not fully correspond to human perception of the image. It is thus desirable, we claim, to introduce *global* perceptual considerations at the low-level process.

In our method, each site (pixel or other cell) collects votes from *every* segment in the image. These votes contain orientation and strength information preferred by the voting segment. A measure of 'agreement' (in terms of orientation) is now computed, and sites which have high agreement values are considered salient.

Our voting scheme is somewhat related to the Hough transform approach [3], but can also detect shapes defined by their properties (smoothness,...) rather than by their exact shape (lines, circles,...).

The proposed approach is capable of 'highlighting' structures which are salient, as well as interpolating gaps in a smooth manner, while removing noisy edgels in a given image, all in a unified non-iterative and *parameter-free* scheme.

The process is likely to produce features more similar to what we perceive, both in terms of saliency and connectivity. Also, since noise is not likely to produce high agreement values by the above considerations, we expect to attenuate it and thus reduce the complexity of the image (e.g. in terms of the number of useful edges). The same process can also be used as a focus of attention mechanism to aid higher level processes in setting their priorities.

2.2 Model of the input

We would like to associate with each site of an image a direction, strength and a degree of uncertainty for that direction. So, in principle, one site could be classified as being a part of a curve with known orientation and no uncertainty, while another being a point with uncertain orientation. In practice, such input data is rarely available, and when an edge image is given, all segments are considered to have the same amount of uncertainty.

We can thus use as input either a thresholded output of any edge detector (with no linking) or even an un-thresholded version of the edge detector output. It can be shown that our system will yield almost the same results with different choices of this threshold, as long as a sufficient number of useful features are present.

2.3 Output description

Our model of the output is closely related to Ullman and Sha'ashua's[10] in the sense that a *saliency map* is first constructed from an edge image, and higher-level features are inferred later. The saliency map assigns a value and a direction to *every* position in the image.

Ideally, such saliency map should assign large values of likelihood along illusory lines (as well as along physical curves), and also specify a direction of most probable continuation of any given segment. This will enable us, at a later stage, to group features by following the salient connections between the primitives.

2.4 Perceptual constraints

Our underlying goal is to keep the interpretation as simple as possible in the 'Gestalt' sense. This translates into four major constraints:

1) *Co-curvilinearity* - In the lack of other cues, smooth continuation is the only interpretation, and so is co-curvilinearity.

2) *Constancy of curvature* - We tend to extend a curve of some constant curvature with the same curvature, keeping the interpretation as simple and regular as possible, yet consistent with our sensory information. This principle is called *Prägnanz* by Gestaltists (see figure 2).

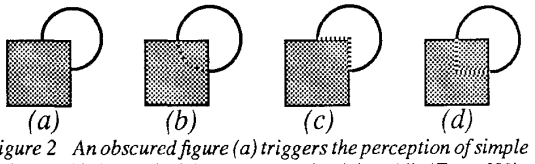

Figure 2 An obscured figure (a) triggers the perception of simple shapes (b), instead of the more complex (c) or (d). (From [9])

3) *Favoring low curvatures over large ones* - Humans seem to connect fragmented line segments in a way that the increase in total curvature is minimum (see Ullman *et al.* [10]).

4) *Proximity* - Influence decays over distance. Closer features will tend to influence each other more than distant ones.

With that in mind, we have devised a technique that implicitly imposes the above constrains in the form of an ***Extension Field*** emanating from each edge segment, as described next.

2.5 Extension Fields

An ***Extension Field*** is a non-normalized probability directional field describing the contribution of a single edge element to its environment in term of length and direction. In other words, it votes on the preferred direction and the likelihood of existence of every point in space to share a curve with the original segment. The field is of *infinite* extent, although in practice it disappears at a predefined distance from the edge. Figure 3 depicts such a field.

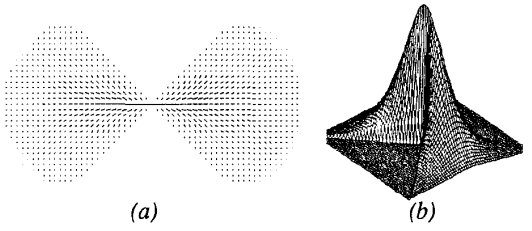

Figure 3 The basic Extension Field. (a) Direction, and (b) Strength.

2.5.1 Assignment of strengths and orientations

Since we favor small and constant curvature, field direction at a given point in space is chosen to be tangent to the osculating circle passing through the edge segment and that point, while its strength is proportional to the radius of that circle. Also, the strength decays exponentially with the distance from the origin (the edge segment).

The assignment of actual probabilities to the field is performed as follows. We consider two short edge segments, perpendicular to each other and apart.We assign probabilities to the field elements in such a way that all paths connecting these points are assigned roughly the *same* saliency, such that there *does not* exist any one best path between the two. Such a scenario removes most degrees of freedom as to the choice of values for the field. It is also in agreement with human perception.

2.6 Computation of the saliency map

The process of computing the saliency map can be thought of as a directional convolution with the above field (mask). The resulting map is then a function of a collection of fields, each oriented along a corresponding short segment. Each site accumulates the 'votes' for its own preferred direction and strength from every other site in the image. These values are combined at a site as described next.

Note that, although the process is local in essence, the fields impose some global order, and one line segment can implicitly 'vote' for a large curve without any explicit global reasoning involved.

2.6.1 Combining individual field elements

Ideally, we would want an averaged majority vote regarding the preferred orientation of a given position. We treat the contributions to a site as being vector weights, and compute moments of the resulting system. Such a physical model behaves in the desired way, giving both the preferred direction and some measure of the agreement. In practice, we use the direction of the principal axis ($EVmin$) of that physical model as the chosen orientation (See (1)).

$$\begin{bmatrix} m_{20} & m_{11} \\ m_{11} & m_{02} \end{bmatrix} = \begin{bmatrix} EVmin \\ EVmax \end{bmatrix} \begin{bmatrix} \lambda min & 0 \\ 0 & \lambda max \end{bmatrix} \begin{bmatrix} EVmin^T & EVmax^T \end{bmatrix} \quad (1)$$

This acts as an approximation to the desired majority vote, without the need to consider the individual votes, but rather the statistics of the ensemble.

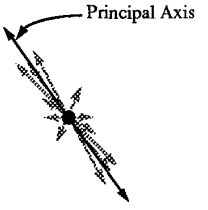

Figure 4 The principal axis of the votes collected at a site is taken as an approximation of the preferred direction.

The saliency map *strength* values are taken as the values of the corresponding λmax at each site. So, large values would indicate that a curve is likely to pass through this point. This map can be further enhanced by considering the eccentricity, or $1 - (\lambda min/\lambda max)$. When that value is multiplied by the previous saliency map we achieve better selectivity, and only curves are highlighted. This results in a map defined by $\lambda max - \lambda min$.

2.7 Detection of junctions

A junction is defined as a *salient* point having *low* eccentricity value. Regular (non-junction) points along a curve are expected to have high eccentricity values. On the other hand, junction points are expected to have low eccentricity, since votes were accumulated from many different directions. By combining the eccentricity and the eigenvalue at a point, we acquire a continuous measure of the likelihood of that site being a junction. This process creates a *Junction Saliency map*. Interestingly enough, this map evaluates to just λmin at every site, which simply means that the largest non-eccentric sites are good candidates for junctions. By finding all local maxima of the junction map we localize junctions.

2.8 The presence of noise

The addition of random noise to an image is expected to create a distributed map of votes (with low eccentricity values), and thus not to interfere with truly salient patterns. When an accidental formation of random segments does

give rise to high values in the map, that formation is perceived as significant by humans as well.

3 Other fields

So far, we have considered the input to be of maximum certainty in terms of the orientation of edges. Our input model allows for uncertainty in the input.

3.1 The Point field

Maximum uncertainty is modeled as a point without a direction. Thus, a suitable field will have circular symmetry, and in practice is constructed by convolving our original extension field with a multi-directional edge segment (Figure 5(a)). Other uncertainty patterns are also generated

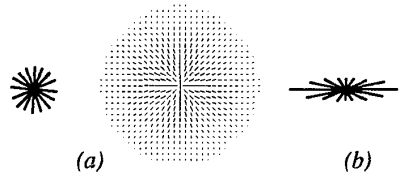

Figure 5 (a) A multi-directional edge and the resulting point field. (b) An uncertain edge element, with which to convolve the original field (Figure 3).

by the same method (Figure 5(b)). Atypical input is shown in figure 9, where a sine wave and a random set of points on a circle are embedded in noise. Obviously, perception is weakened by the loss of orientation data, and we are only able to handle (using the field in figure 5(a)) cases with a moderate amount of noise.

3.2 Special purpose fields

So far the fields described are meant to perform on any scene regardless of domain. When some knowledge is available about the domain of the input, special fields can be constructed to better enhance the specific features of that domain. A simple example would be a world of polygons, where only (approximate) straight lines are possible. The construction of the field is similar to what has been described for the point field (basically, a convolution with a long line).

4 High-level feature extraction

Once a saliency map (and a junction saliency map) is acquired, a process that actually groups salient shapes is started.

Grouping of features is done by a directional 'rooftop' following and linking algorithm on the saliency map. The linking process starts at the point of *largest saliency* and advances in the general direction dictated by a function of the orientation of the current position and the strength of neighboring points.

This process first removes the most salient curve, recalculates the saliency map, and then proceeds to remove the next most salient feature. The process complexity is thus proportional to the number of curves in the image. It is guaranteed to terminate since at each iteration the overall power of the field is strictly reduced by removing a feature. Each feature extracted is assigned a saliency value which can be used in later processing.

The output thus consists of a list of salient features, each with its own saliency measure. Since no thresholds are used throughout the process, *all* possible groupings are recorded. It is up to a specific application to prune that list according to its own constraints.

5 Complexity issues

A naive way to implement the algorithm requires $O(n^2k)$ operations, where n is the side size of the image, and k is the number of edge elements in the input image. In practice, the local density of edgels restricts the useful scope of the field. This means that a smaller *finite* field can be used. The complexity becomes now $O(k)$. This last modification has the disadvantage of not being able to bridge gaps larger then the size of the field. Alternatively, instead of computing a *dense* saliency map, we can compute the saliency of existing edgels only. This results in complexity of $O(k^2)$, and can be useful as a focus of attention map.

6 Results

We tested our approach with the synthetic data shown before. Figure 1(a) Shows a fragmented circle embedded

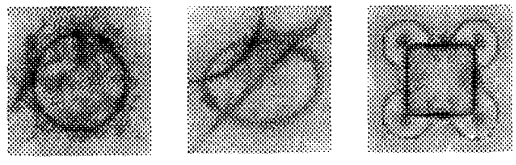

Figure 6 The Saliency maps of images in figure 1.

in a noisy environment. The saliency map produced is shown (strength only) as grey-level image in figure 6 and the result of following the path of highest saliency produces a "clean" circle. Figure 6 also shows the result of the same procedure for the other scenes. Figure 7 shows an example of the steps involved in producing a high-level description of a given image, using the junction map in conjunction with the saliency map

6.1 Real images

In figure 8 we show a real image example. The original image was processed with a simple edge detector (5x5 step masks), without any linking. Note that the edge image is fragmented and has a lot of noisy segments. Figure 8(c) shows the resulting Saliency map.

6.2 Using the Point Filed

We tested our system on the image in figure 6. Initially, the system was run using the Point field. This resulted in a

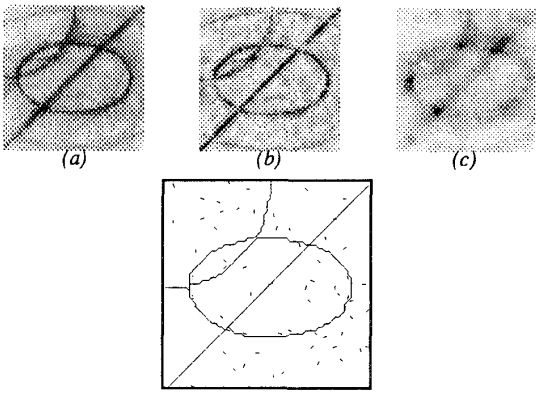

Figure 7 Extracting the most salient features. (a) Largest eigenvalue strength map. (b) Eccentricity enhanced map. (c) Junction saliency map, and (d) linking.

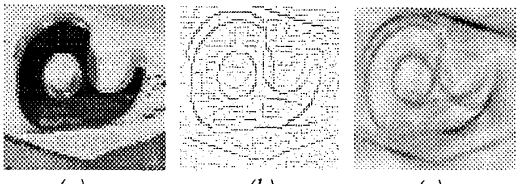

Figure 8 Example of a real image. (a) A tape dispenser image. (b) All edges. (c) Eccentricity enhanced saliency map.

saliency map *with* orientation data. A second phase of computation was performed now, using the directional Extension field (Figure 3). That stage produced the final saliency map as shown in figure 9(b). Note the quality of the

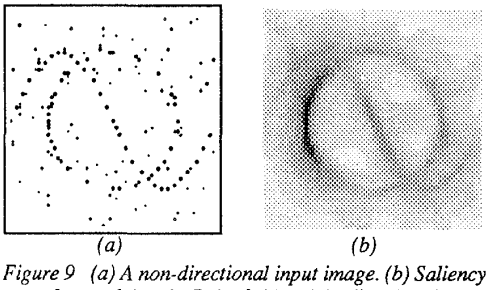

Figure 9 (a) A non-directional input image. (b) Saliency map, after applying the Point field and the directional extension field

saliency map is worse than one generated from a directional image with comparable noise level.

7 Summary and conclusion

We have introduced a unified way to extract perceptual features in edge images. By 'unified' we mean that all low-level features (edgels, points) are treated in a uniform way, and no special cases exist. The scheme is threshold-free and non-iterative. It is especially suitable for parallel implementation, since computations of the saliency maps are independent for each site, and parallel algorithms for line following are known and can easily be adapted. Also, calculations are simple and stable, as no curvatures or any other derivatives need to be computed on the digital curves.

The system can rank features based on their perceptual importance. This allows a real-time application to process as many features as time permits.

Some of the issues which have not been addressed are the resolution dependency of the description. At this time, only a one-level description is possible. Also, we have not tried to localize end-points of curves ending abruptly (e.g. in Figure 9).

Since all computations are performed on a discrete grid, quantization and rounding errors restrict the selectivity and amount of clutter the system can handle.

In the future we intend to incorporate additional types of fields capable of highlighting perpendicular end-point relations and symmetries (see [8] for a symmetry enhancing operator).

References

[1] N. Ahuja and M. Tuceryan, *Extraction of early perceptual structure in dot patterns: integrating region, boundary, and component Gestalt*, CVGIP 48, 1989, 304-356.

[2] J. Dolan and R. Weiss, *Perceptual grouping of curved lines*, Proceedings of the IUW 1989, 1135-1145.

[3] R. O. Duda and P. E. Hart, *Pattern Recognition and Scene Analysis*, John Wiley & Sons, New York, 1973.

[4] G. K. Kanizsa, *Subjective contours*, Scientific American, April 1976.

[5] D.G. Lowe, *Three-dimensional object recognition from single two-dimensional images*, Artificial Intelligence 31, 1987, 355-395.

[6] R. Mohan and R. Nevatia, *Perceptual Grouping for the detection and description of structures in aerial images*, IUW 1988.

[7] R. Mohan and R. Nevatia, *Segmentation and description based on perceptual organization*, Proceedings of CVPR 1989, 333-341

[8] D. Reisfeld, H. Wolfson, and Y. Yeshurun, *Detection of interest points using symmetry*, Proceedings of the ICCV 1990, 62-65.

[9] I. Rock and S. Palmer, *The Legacy of Gestalt Psychology*, Scientific American, December 1990, 84-90.

[10] A. Sha'ashua and S. Ullman, *Structural saliency: the detection of globally salient structures using a locally connected network*, Proceedings of the ICCV 1988, 321-327.

[11] S.W. Zucker, C. David, A. Dobbins, and L. Iverson, *The organization of curve detection: coarse tangent fields and fine spline coverings*, Proceedings of the ICCV 1988, 568-577.