Perceptual Grouping from Gabor Filter Responses*

María J. Carreira[1], James Orwell[2], Ramón Turnes[1], James F. Boyce[2],
Diego Cabello[1] and John F. Haddon[3]
[1]Department of Electronics and Computer Science,
University of Santiago de Compostela,
15706 Santiago de Compostela, SPAIN
mjose@dec.usc.es
[2] Physics Department, King's College London,
Strand, WC2R 2LS London, UK
[3]D.E.R.A.,
Farnborough, GU14 6TD Hants., UK

Abstract

Perceptual organisation can be defined as the ability to impose structural organisation on sensory data, so as to group sensory primitives arising from a common underlying cause. Our organisational philosophy is hierarchical, with complex organisations being formed from simpler ones. In this paper, directional features extracted from Gabor responses are used as the primitives for perceptual grouping. In previous work, we extracted Gabor features in 8 directions and then applied two SOMs, thus classifying each pixel in the image within a 8x10 neuron-map, each corner of which represents one of four main directions, (horizontal, vertical, left diagonal and right diagonal). In the present work we group pixels with similar directional features, thereby detecting salient structures within an image. These detected-structures will be used as tokens from which to create the next level of abstraction in the hierarchy of the system. This approach is an alternative to the use of sets of edges as primary features: the directional features that Gabor filters provide are a potentially richer source of information. Preliminary results obtained from application to forward-looking infrared (FLIR) images are very promising. At present only four main directions are utilised, *i.e.* vertical, horizontal, right diagonal and left diagonal: the technique may be readily extended to the eight utilised in previous work. The next stage will be to group the tokens by the application of additional Gestalt-laws in order to detect objects.

1 Introduction

The phenomenon of perceptual organisation enables humans to detect such relationships among image elements as collinearity, parallelism, connectivity, and repetitivity. In

 This work was part funded by the DRA (Farnborough), under Research Contract SF/U740.

BMVC 1998 doi:10.5244/C.12.34

computer vision, perceptual grouping is the study of how features are clustered for object recognition. Inspired by biological studies, especially the Gestalt school, its purpose is to group feature elements prior to recognition. Perceptual organisation has been studied by investigators in psychology [1, 2] in an attempt to classify the behaviour of grouping phenomena in the human visual system, with *laws* of symmetry, proximity, simplicity, closure *etc*. proposed as the mechanism for grouping features such as edges, corners, regions.

In the application of perceptual organisation to computer vision [3-7], Marr [8] was first to suggest incorporating groupings based on curvilinearity into larger structures in his *primal sketch*. Witkin and Tenenbaum [4], in their structure-based vision paradigm, recognised the broad implications of perceptual organisation for computational theories in machine vision, whilst Lowe [3, 9] used grouping based on simple organisation, such as parallelism and collinearity, to demonstrate the consequent reduction of computational complexity. Perceptual organisation has been used to detect straight lines [10] and curves [11], Reynolds and Beveridge [12] detect groupings of parallel lines and proximate orthogonal lines (among other relationships) in aerial images, while Quan *et al.* [13] have successfully used scene-level groupings to detect and estimate motion, and Mohan and Nevatia [14] have implemented a vision system that uses perceptual organisation to detect objects with particular shapes but without explicit models.

The hierarchical nature of organisation has been emphasised by many researchers [6, 10, 11]. Complex, high-level, organisations are built incrementally from simpler lower-level, features, permitting better computational control of the process.

Perceptual organisation has also been based on network formalism. Sha'ashua and Ullman [15] use a locally connected network to derive salient structures from local characteristics: the output is a saliency map, a representation which emphasises salient locations. Mohan and Nevatia [14] use concepts of perceptual organisation in detecting and describing buildings in aerial images. They demonstrated the usefulness in complex image understanding of the structural relationships which are made explicit by perceptual organisation. All reasonable feature groupings are first detected, and the promising ones then selected by a constraint satisfaction network, the results being extended to curved segments in [7]. Sarkar and Boyer [6] developed a model for perceptual organisation with explicit knowledge about various Euclidean geometric structures; the main contribution being a formalism based on Bayesian networks for geometric knowledge-base representation.

The goal of our work is to develop an artificial system for grouping and recognition, which learns the Gestalt principles of perception (proximity, similarity, smooth continuity, and closure). It will utilise both supervised and unsupervised networks to build up clusters of features useful for recognition in typical images, and so will comply with the laws of grouping. In a previous paper [16] we have discussed the formation of perceptual features. In this work, we describe how pixels with similar features may be grouped into tokens.

2 Previous work: Extraction of Perceptual Features

In the first stage of the system perceptual primitives are extracted in successive stages [16]. The application of several orientations and scales of Gabor filters maps the original

image intensity to a high dimensional space. A Self Organising Map (SOM) is then used to find a sub-space representation and sampling which provides useful feature maps. Following the work of Lampinen and Oja [17], a hierarchy of two layers groups data firstly from different orientations and then different scales. The objective of the two SOMs is to reduce the 54 dimensions of the input space (3scales x (8 orientations + 1) x 2) to a two-dimensional feature space which summarises all of the useful information provided by the input space. Images classified using these first and second layer maps are the input to the perceptual grouping stage we are described in this paper.

Figure 1 illustrates the selection of the winning neurons. For each pixel *(x,y)* in the original image, a feature vector, a *gaborjet*, is formed for 3 different scales and 8 different orientations:

$$H^k(x,y): H_0^k(x,y), H_1^k(x,y), \ldots, H_7^k(x,y); \quad k = 0, 1, 2 \ . \tag{1}$$

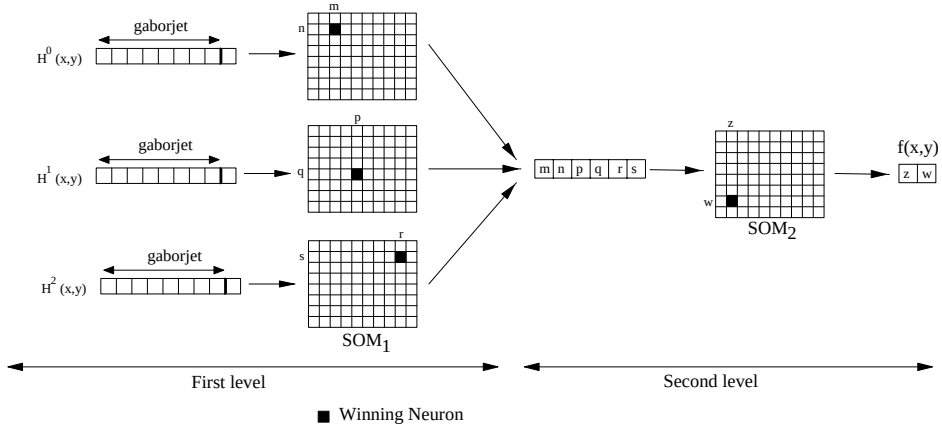


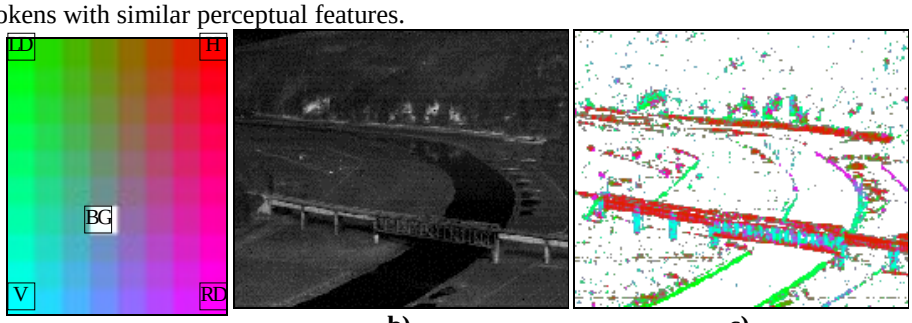
Figure 1. Level 1 and Level 2 maps and the feature resulting from a pixel (x,y) of the original image.

In order to reduce inter-image variability, the *response* of a gaborjet is normalised to zero mean and unit variance. Additionally, to correct for leptokurtic tendencies in the input data, the response is bounded to the range *{0,1}* by passing it through an *arctangent* sigmoid function. The response is added, as an extra (ninth) component to the gaborjet, to form the data input to the map (Figure 1). The first eight components of a gaborjet are scaled such that their modulus is equal to the response. The consequent redundancy of coding is offset by the utility of expressing the response explicitly in the input of the SOM. A consequence of the scaling is that the separation of two gaborjets, which differ only in response, is much larger than previously, since the difference is coded in all dimensions.

The main feature for each neuron in the second map, SOM_2, may be represented using a graduated colour map, as shown in Figure 2a. Background neuron has been transformed to white for the purpose of illustration, though the colour representation of the remainder becomes distorted.

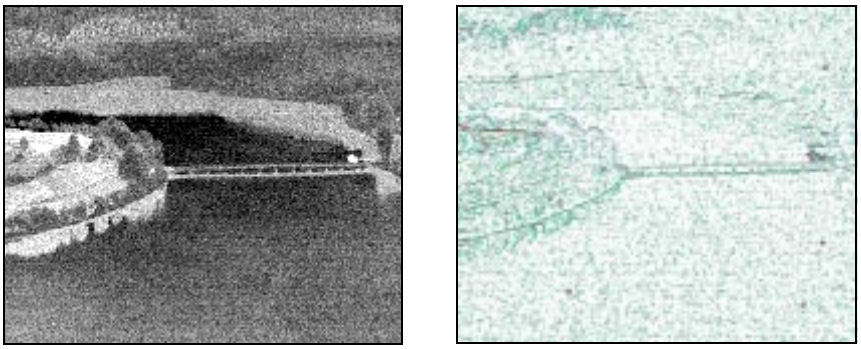
If the main feature of pixel *(x,y)* lies in neuron *(z,w)* (see Figure 1), then the colour of pixel *(x,y)* in the output image will be the colour of position *(z,w)* in the *8×10* colormap. Mapping the results for each pixel for the FLIR image of Figure 2b yields a perceptual feature image (Figure 2c). Figure 3 shows the results from another bridge image with

results inferior to those in Figure 2 due to the poorer quality of the original image. In the next section pixels with vertical properties will be grouped to build the pillars of this bridge. The task of this paper is to show how to group these individual pixels to form tokens with similar perceptual features.



a) **b)** **c)**

Figure 2. a) 8×10 colormap, with main features labelled (BG=background, LD=left diagonal, H=horizontal, V=vertical, RD=right diagonal), b) oldbridge2 input image and c) labelled image after first stage processing.



a) **b)**

Figure 3. a) Bridge image and b) labelled image after first stage processing.

3 Perceptual Grouping

As shown in Figure 2, the first stage of processing yields reasonable results, especially if compared with those of a classical segmentation of the images. It appears that pixels having similar perceptual features have formed groups in the feature space. However they do not map into single neurons, for example, the pillars of the bridge have colours similar, but not identical, to that in position (*0,9*) representing a neuron with a vertical feature. Although they have similar colour neurons (*1,9*), (*0,8*), *etc*., formally, they have *different* perceptual features.

Pixels with *similar* features will be grouped in order to build tokens useful for the recognition of larger, distributed, structures. An exhaustive examination of several kinds of aerial infra-red images (10) yielded the conclusion that their characteristics are almost the same, and all of them have the following properties:

- The four main directions lie in the four corners of the second layer map (*8×10*), with neuron (*0,0*) usually corresponding to the right diagonal, neuron (*7,0*) to the horizontal, neuron (*0,9*) to the vertical and neuron (*7,9*) to the right diagonal (see Figure 4).
- The neuron corresponding to the background of the image was always the same (*3,6*), independently of which kind of images were examined.

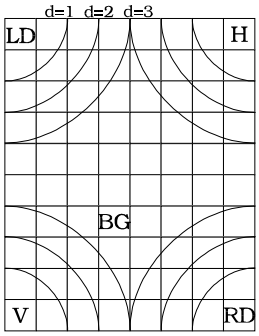


Figure 4. Representation of 8x10 SOM_2 map with neurons representing each main perceptual feature (H: horizontal, V: vertical, LD: left diagonal and RD: right diagonal. BG is the neuron representing the background).

In order to extend the analysis from the four directions considered at present to the eight of the previous system [16] we may utilise the continuity of the SOM map representation of direction. For example, a neuron mid-way between the corner (*7,0*), horizontal direction, and corner (*0,0*), right diagonal direction, will represent the mid-way direction between them, $\pi/8$. Similar observations apply to the remaining three directions, so obtaining a more precise grouping. In the present work, as previously mentioned, only the four principal directions have been utilised.

The important decision when considering how to group the neurons on the SOM_2 map (*8×10*) with the classes of the 4 corners is not to regard it as a clustering problem. The main property to be expressed is that neurons near the bottom left corner of the map must have a high vertical component. The chosen solution is to form groups of neurons for each corner by defining corresponding neighbourhoods of radius *d*, as shown in Figure 4. Any pixel whose winning neuron lies less than a distance *d* to the feature corner is assigned the direction corresponding to that corner. The maximum distance considered is $d_{max}=4$, half the size of the horizontal dimension of the SOM_2 map.

To each pixel of the image is associated the co-ordinates of the winning neuron of the SOM_2 and hence the distance from the feature being considered. If this distance is less than a threshold, then the pixel can initiate the process of grouping by means of a flood-fill filter specific to the direction of orientation. Figure 5 illustrates the filters for grouping pixels corresponding to direction features *vertical*, *horizontal*, *left* and *right diagonal* pixels, Figure 5a, b, c and d respectively.

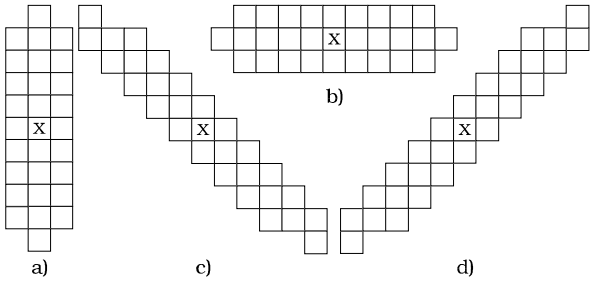


Figure 5. Kernels of floodfill filters applied to group pixels with a)vertical features, b)horizontal features, c)right diagonal features and d)left diagonal features.

The effect of the floodfill filters is to implement the concept of collinearity from Gestalt laws. This is because they interpolate small gaps between the pixels forming a line with the effect that, if the gap is not too big, all the pixels together form a unique group. In addition, the filter is very narrow, so that pixels in a parallel direction are not joined to the group unless they were neighbours of the starting pixel. The floodfill filter is applied iteratively until no further pixels can be added to the group. The main program then tries to find another pixel with $d<d_{max}$ to initialise another group.

The end result of the process is a new, labelled, image, each group of pixels bearing a different label, containing pixels with similar directional features.

The above grouping of perceptual features which have been detected previously is performed in several steps, and depends on several parameters whose values require definition:

1. The maximum distance to each feature *d*.
2. The shape of the kernel for the floodfill filter.
3. The feature upon which grouping depends.

Tests have been performed with different distances, from 1 to 4 in steps of 0.5 in order to observe the behaviour of the grouping. The method of growing each region has yet to be investigated: we consider it to be an important determinant of the final confidence factor of a region with a given feature.

The final result of the complete process is a reasonable number of groups with a significant perceptual feature, which can help us to better interpret the image.

4 Results

As stated above, tests have been performed to determine the dependence of the grouping of directional features upon the distance of the winning neuron in the SOM map from the map location characteristic of the feature. Results are showed for vertical features in Table 1 and for horizontal ones in Table 2. The number of regions when we are able to detect all of the *important* features in each image is shown in bold. The set of test images comprises; bridge, an image of a bridge (Figure 3); powerstation, two images from a sequence approaching a powerstation; oldbridge, a sequence of three images approaching a bridge (the second is in Figure 2b); and an image of a runway.

	DISTANCES						
IMAGES	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**
Bridge (Figure 3**)**	7	14	14	18	16	60	**71**
Powerstation1	5	6	6	8	9	43	**59**
Powerstation2 (Figure 8**)**	5	8	8	10	10	51	**62**
Oldbridge1	20	21	21	**27**	25	34	33
Oldbridge2 (Figure 2**)**	2	11	11	35	34	**49**	50
Oldbridge3	4	14	14	24	24	**48**	43
Runway	0	2	2	4	4	42	**59**

Table 1. Number of groups obtained grouping pixels with vertical features

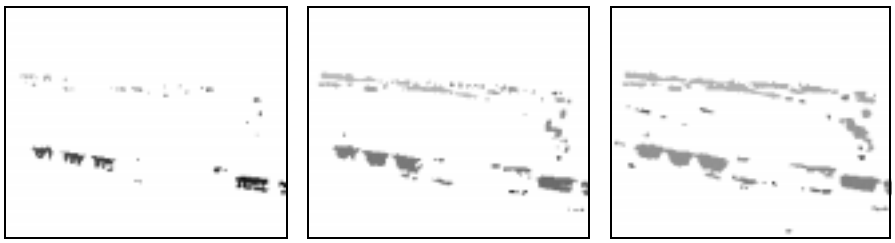
	DISTANCES						
IMAGES	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**
Bridge (Figure 3**)**	2	4	4	16	16	25	**31**
Powerstation1	0	2	2	6	6	17	**18**
Powerstation2 (Figure 8**)**	1	1	1	14	12	16	**19**
Oldbridge1	31	41	**41**	37	34	37	35
Oldbridge2 (Figure 2**)**	1	5	5	12	15	38	**38**
Oldbridge3	8	40	40	**26**	26	27	31
Runway	0	1	1	5	5	5	**4**

Table 2. Number of groups obtained grouping pixels with horizontal features

The results show that the number of regions finally obtained is not excessive. In addition, if account is taken of their confidence factors, and we select only those regions having significant confidence, we can *see* the structure of the objects to be detected in the images. More results are shown in Figures 6 and 7, with grey levels representing each different group. Figure 6 shows the vertical groups obtained from the image of Figure 2, while Figure 7 shows the horizontal groups for the same image. Figure 8 shows the results for the powerstation image, where the chimneys appear as independent groups. Another case is that of Figure 3. As noted previously, Figure 3b is unclear and it might be inferred that the results are inadequate. However, if vertical features are grouped, as in Table 1, few regions are obtained for distance 1 but many for distance 3.5 with significant verticality. These results, (Figure 9), show that the regions that appear first are the pillars of the bridge, although it appeared from Figure 3b that they had not been detected.

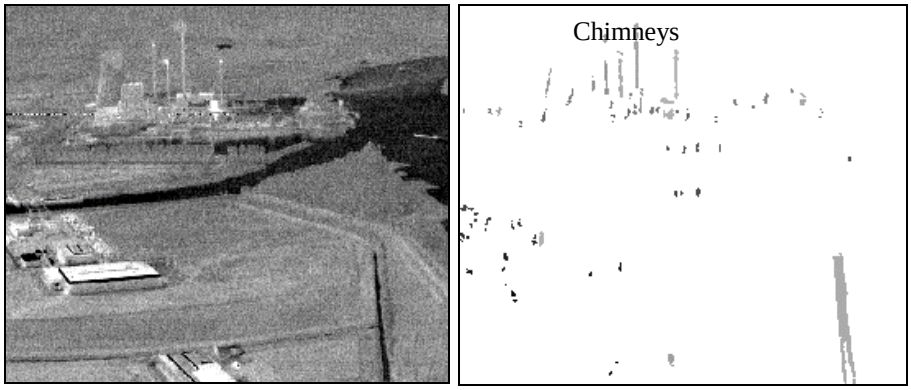
a) **b)** **c)**

Figure 6. Vertical groups for Figure 2b with distances 1, 2 and 3 respectively.



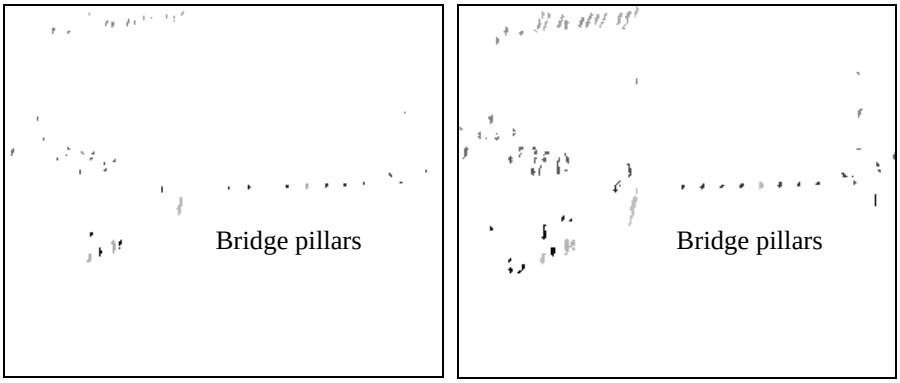
a) **b)** **c)**

Figure 7. Horizontal groups for Figure 2b with distances 1, 2 and 3



a) **b)**

Figure 8. Vertical groups for powerstation image with distance 4



a) **b)**

Figure 9. Vertical groups for bridge of Figure 3a with distances 3.5 and 4

5 Discussion and Future Work

The results presented above have shown that the output of the first stage of the system is very effective in obtaining significant directional features of images. Nevertheless, we

need to make some grouping of pixels with similar perceptual features, and so must define what will be similar. In this work we have defined similarity in terms of the distance in the SOM from the winning neuron corresponding to a pixel to the winning neuron characteristic of a feature. As a consequence, depending on the quality of the image and the distance from the camera to the objective (bridge, chimneys of powerstation, *etc*.), the process of grouping pixels should be halted. Future work will consider the way of growing regions with similar features, and the decision of when to stop the process. For example, for the nearest image of the oldbridge, from Table 1 and Table 2, we observe that, with a SOM distance of 2.5, the significant regions representing the bridge are extracted. In addition, by optimising the method of growing, it may be possible to distinguish the vertical bars of the bridge. In contrast, the image of the powerstation is of very poor quality: Figure 8a has been contrast enhanced for the purpose of viewing. From the results of Table 1, it is clear that we need to increase the SOM distance compared to that used in the previous image in order to detect the chimneys of the powerstation.

A confidence factor will be computed for each group, based on the property that neurons close to the characteristic neuron of a directional feature should contribute more to the confidence factor of the group than those that are distant. A possible expression of confidence that embodies this property is as follows:

$$cf_i^k = \frac{\sum_{p/d_p^k < d_{max}} e^{-0.2 d_p^k}}{n_i} \tag{2}$$

where cf_i^k is the degree of confidence that region i belongs to feature k, n_i the number of pixels in group i, p the point being analysed, d_p^k the distance from the neuron corresponding to p to the neuron corresponding to feature k, and d_{max} the maximum distance a neuron can be from a feature in order to be joined to a group. The exponential is attenuated by a factor of 0.2, in order to obtain a smooth slope, since neurons at unit distance must be very close to confidence factor 1. A possible alternative to the above implementation of confidence factors would be to use fuzzy logic. The perceptual grouping of pixels with similar features would be based on a fuzzy measure of belonging to a feature, and leading to an overall fuzzy measure of each group in being vertical or horizontal, *etc*.

6 Conclusions

In previous work [16], we showed that Gabor features can be grouped hierarchically to create a feature map that is perceptually useful, smoothly varying with the input data, intrinsically parallel, and able to achieve massive reduction of dimensionality. We have continued this development and have shown how to easily reduce the large quantity of information that the first stage provides. We have formed groups of pixels with significant perceptual features in images of poor quality and preliminary results are very promising, as it is easy to find the structures (bridge, chimney, *etc*.) we are looking for. In addition to this conclusions, we consider that it would be easy to apply this process to other kinds of images, as shown in [16], and that it is applicable to a sequence of images, in order to follow detected structure, so that it would be applicable to the tracking of objects.

References

[1] D. Katz, *Gestalt Psychology: Its Nature and Significance*. New York: Ronald, 1950.

[2] S.E. Palmer, "The psychology of perceptual organization: A transformational approach", in *Human and Machine Vision* (J. Beck, B. Hope and A. Rosenfeld, Eds.). New York: Academic, 1983, pp. 269-339.

[3] D.G. Lowe, *Perceptual Organization and Visual Recognition.* Hingham, MA: Kluwer Academic, 1985.

[4] A.P. Witkin and J.M. Tenenbaum, "On the role of structure in vision", in *Human and Machine Vision* (J. Beck, B. Hope and A. Rosenfeld, Eds.). New York: Academic, 1983, pp. 545-567.

[5] S.W. Zucker, "The diversity of perceptual grouping", in *Vision, Brain, and Cooperative Computation* (M.A. Arbib and A.R. Hanson, Eds.). Cambridge, MA: MIT Press, 1987, pp. 231-262.

[6] S. Sarkar and K.L. Boyer, "Integration, inference, and management of spatial information using Bayesian networks: perceptual organization", *IEEE Trans. Patt. Anal. Machine Intell..*, 15(3): 256-274, 1993.

[7] R. Mohan and R. Nevatia, "Perceptual organization for scene segmentation and description", *IEEE Trans. Patt. Anal. Machine Intell.*, 14(6): 616-635, 1992.

[8] D. Marr, *VISION*. San Francisco, CA: W.H. Freeman, 1982.

[9] D.G. Lowe, "Three-dimensional object recognition from single two-dimensional images", *Artificial Intell.*, 31: 355-395, 1987.

[10] M. Boldt, R. Weiss and E. Riseman, "Token based extraction of straight lines", *IEEE Syst. Man Cybern*, 19(6): 1581-1594, 1989

[11] J. Dolan and R. Weiss, "Perceptual grouping of curved lines", in *Proc. DARPA Image Understanding Workshop*, 1135-1145, 1989

[12] G. Reynolds and J.R. Beveridge, "Searching for geometric structure in images of natural scenes", in *Proc. DARPA Image Understanding Workshop*, 1987

[13] L. Quan, R. Mohr and E. Thirion, "Generating the initial hypothesis using perspective invariants for a 2D image and 3D model matching", *Proc. Int. Conf. Comp. Vis.*, 1988

[14] R. Mohan and R. Nevatia, "Using perceptual organization to extract 3-D structures", *IEEE Trans. Patt. Anal. Machine Intell.*, 11(11): 1121-1139, 1989.

[15] A. Sha'ashua and S. Ullman, "Structural saliency: The detection of globally salient structures using a locally connected network",*Proc. Int. Conf. Comp. Vis.*, 321-327, 1988.

[16] J. Orwell, R. Turnes, M.J. Carreira, D. Cabello, J. Boyce and J.F. Haddon, "Towards self-organized feature maps from Gabor filter responses", in *Workshop on Self-Organizing Maps* (*WSOM'97*), 220-226, 1997.

[17] J. Lampinen and E. Oja, "Distortion tolerant pattern recognition based on self-organizing feature extraction", *IEEE Trans. On Neural Networks*, 6(3), 1995.