Comparing Attention Operators for Learning Landmarks

Robert Sim, Sandra Polifroni and Gregory Dudek

TR-CIM-03-03 June 21, 2003

Centre for Intelligent Machines
McGill University
Montréal, Québec, Canada

Postal Address: 3480 University Street, Montréal, Québec, Canada H3A 2A7
Telephone: (514) 398-6319 Telex: 05 268510 FAX: (514) 398-7348
Email: cim@cim.mcgill.ca

Abstract

We consider the choice of which particular attention operator to use for visual recogniton and pose estimation applications. Visual attention plays an important role in reducing the amount of computation (biological or artificial) that is required for a particular task. However, it is not always clear which attention operator is best suited to any given task. This paper presents the results of an empirical study comparing the performance of a set of visual attention operators on the task of learning a set of landmarks for pose estimation. These landmarks are defined visually in the local environment of a moving robot. The performance of the operators is quantitatively evaluated based on the utility of the learned landmarks for the task of robot pose estimation.

Contents

List of Figures

1. **Introduction**

This paper evaluates several alternative attention operators in the context of mobile robot localization (i.e. pose estimation). Attention operators (also known as *interest operators*) have become increasing important in computational models of both biological vision. Attention also seem to be of growing importance for an growing range of purely computational tasks. In particular, the successful application of attention can significantly reduce the computational cost of performing generic visual tasks [**3**]; in fact it may be a necessity in many contexts. In view of this principle, it is not surprising that there is a large body of prior work on visual attention and computational methods for computing points of fixation. It is sometimes the case that these attention operators are based on biologically motivated criteria, but some are based also on analytic or even *ad hoc* criteria. Given the range of possible attention operators and their differing computational characteristics, it is important to ask how the choice of operator relates to performance.

In this paper, we consider visual attention as it is applied to the task of robot localization. We build on prior work that addresses the localization problem in the context of learning a set of visual landmarks in the environment [**4**]. Our approach depends on the reliable output of an attention operator in order to identify initial candidates for landmarks. The open question remains as to which attentional criteria is best suited to landmark learning and is likely to provide the best locatization system. To this end, we compare the empirical results of employing four different attention operators to learning landmarks in a variety of environments. We present quantitative results to support the assertion that stability is the dominant requirement for the extraction of useful landmarks. As a byproduct, we provide results that allow us to claim improved pose estimation performance from visual cues.

While our experimental data is collected and quantified using one particular localization framework as a testbed, there are several related schemes for both robot positioning and object recognition that employ (or could employ) an attention-like process in a manner consistent with the experiments described here. As such, we believe these results are directly applicable to these other methodologies as well [**5, 6, 7, 8, 9**].

The remainder of this paper is divided as follows. We review the field of visual attention in Section 2, and discuss the specific operators used in our experiments in Section 3. We then briefly describe our landmark learning mechanism in Section 4. Our experimental approach and results are presented in Section 5, and finally we discuss the results and our conclusions in Section 6.

2. **Visual Attention**

Visual attention has been demonstrated to be crucial to human vision and is becoming a critical approach for machine vision. Several models of how to allocate attention have been proposed in the computational and biological literature. The majority of these models use low-level image features to select locations of interest in an image. These operators are often based on features such as edge density, edge orientation and contour closure (Figure 1) and there is evidence that strongly suggests the such features are exploited by the primate visual system at a level prior to object

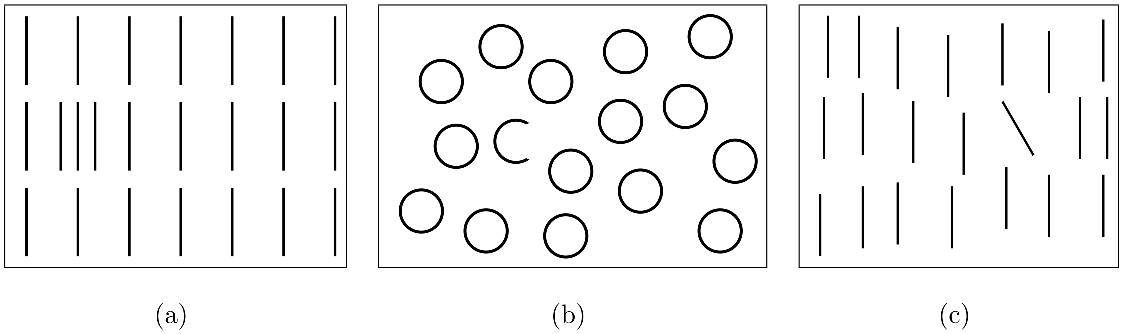


FIGURE 1. Edge density (1(a)), contour closure (1(b)) and edge orientation (1(c)) are presumed to be selected by human visual system at a level prior to image understanding.

recognition and image understanding. These features act as cues to the visual system to attend locations for further analysis [**10, 11, 12**], reducing the amount of the image to be analysed [**3**].

It has been observed that less effort and time are needed to identify changes if they occur in regions which would be used to describe an image (regions of central interest) [**13**]. It has also been demonstrated that attention plays an important role in recognizing previously observed scenes [**14**]. Our prior work indicates that attention is also useful in the task of estimating position, where changes in a region of interest can be mapped to a corresponding change in pose [**4**].

Computational attention operators can use different models to select locations of interest in an image. These models can be based on psychophysical studies, on neurological frameworks of human vision or on the physics of image formation. Psychophysically motivated models generally rely on one feature which has been shown to be preattentively selected by the human visual system and act as spatial filters for these features. Locations in the image are assigned interest values based on the degree to which they exhibit the feature used for selection (for example, degree of convexity [**15**], degree of symmetry [**16, 17, 18, 19**]) or are based on the local variations in image statistics (differences in edge orientation [**1**], differences in edge density [**1**], curvature variation [**20**]). Methods relying on image structures examine the features present in the image and assign values of interest based on certain properties of these features. These methods can be based on signal processing [**21**] and scale-space models [**22, 23**], among others. Finally, neurophysiologically-based methods [**24, 2, 25, 26**] attempt to model the the neural mechanisms controlling human visual attention.

We are interested in determining which attentional operators are best suited to robot position estimation. In the following section, we describe the set of operators that were employed in our experiments.

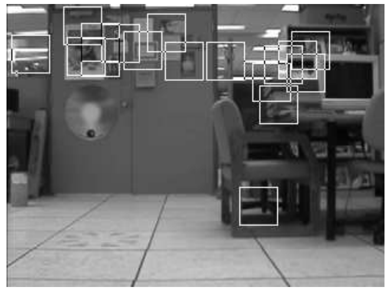


FIGURE 2. Landmarks selected by the edge density operator.

3. **The Operators**

Several interest operators exist which attempt to model aspects of human preattentive vision. Preattentive vision is the ensemble of parallel processesing stages that precede cognitive processing and determine what immediately attracts our attention. Most of these operators concentrate on analysing one or two features which have been shown in the psychophysical literature to be preattentively selected by the human visual system. In this section, the operators used for this research will be described. The operators are an edge density operator [**1, 4**], a radial symmetry operator [**16**], a corner detector [**27, 28**], and a combination luminance/edge orientation/colour operator [**2**] (for the sake of brevity, this last operator will be refered to as the Caltech operator). For the purposes of comparison, we will also employ a "Random" operator which simply selects image points at random.

3.1. **Edge Density Operator** [**1**]**.** The edge density operator was developed for a robotic mapping task involving the assembly of images from an environment to be used in a virtual tour. It is also the default operator used in our prior work on the localization problem [**4, 29**]. This operator is motivated by work by Treisman which showed that edge density is one of the feature primitives preattentively selected by human vision [**12**]. The operator works by selecting regions in the image which deviate the most from the mean density of the whole image. An edge map is created where each element is assigned an intensity corresponding to the strength of its associated edge. This edge map is convolved with a Gaussian windowing operator to give an edge density map. Interest is then defined as the location where the local density varies maximally from the mean density over the whole image (see Figure 2).

3.2. **Radial Symmetry Operator.** This operator was developed as a low-level mechanism for guiding gaze control in an active vision system [**16**]. Whereas the density operator is based on edges, the radial symmetry operator is based on differential properties of image intensity function. The radial symmetry operator does not attempt to select regions which are perfectly symmetric, but attempts to quantify the extent to which any pixel exhibits *local symmetries.*

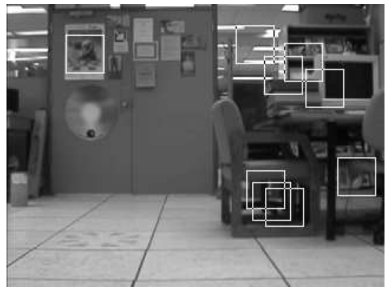


FIGURE 3. Landmarks selected by the radial symmetry operator.

For a point p in the image, the gradient of the intensity at that point is denoted by

$$\nabla p = (\frac{\partial p}{\partial x}, \frac{\partial p}{\partial y}). \tag{1}$$

The magnitude, r, and phase, θ, at point p are given by

$$r = \log(1 + \|\nabla p\|) \tag{2}$$

$$\theta = \arctan(\frac{\partial p}{\partial y} / \frac{\partial p}{\partial x}). \tag{3}$$

The angle α_{ij} denotes the angle that the line through points p_i and p_j makes with the horizontal. For any point p in the image, radial symmetry contributions come from pairs of points which have p as a midpoint. The symmetry contribution made by each pair is the product of the magnitudes of the intensities at each point weighted by a distance term, D_σ, and a phase term, P:

$$D_\sigma(i,j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|p_i - p_j\|}{2\sigma}} \tag{4}$$

$$P(i,j) = [1 - \cos(\theta_i + \theta_j - 2\alpha_{ij})][1 - \cos(\theta_i - \theta_j)]. \tag{5}$$

The direction of the radial symmetry contribution is defined as the average of the phases of the two contributing points. The radial symmetry for any point is the sum of all contributions weighted by the difference between the phase of the point and the phase where contribution is the greatest. Interest is defined as points with the greatest radial symmetry (see Figure 3).

3.3. **The Smallest Eigenvalue Criterion.** The "smallest eigenvalue" operator was developed for detecting and tracking points of interest in an image [**28, 27**]. This operator can be classified as a generic *corner detector* (Figure 4). The operator defines interest in terms of the local variation of the intensity function, in particular, the condition number of the covariance matrix of image intensity in a local neighborhood. The criterion for an interest point requires that the smallest eigenvalue of the 2×2

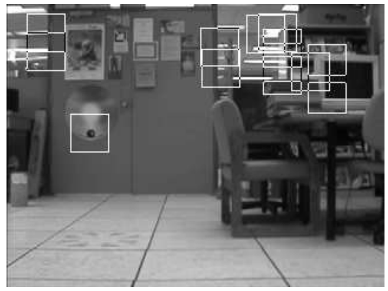


Figure 4. Landmarks selected by the smallest eigenvalue operator.

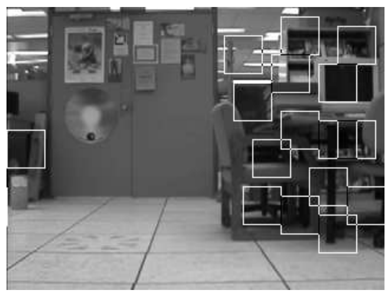


Figure 5. Landmarks selected by the Caltech operator.

intensity gradient covariance matrix Z at any particular point exceed a threshold. That is, for any point p_i in the image, we compute the covariance matrix

$$Z_i = \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix} \tag{6}$$

where g_x and g_y are the directional derivatives of the intensity image at the point p_i. p_i is a point of interest if λ^i_{min}, the smallest eigenvalue of Z_i, exceeds a particular threshold and is maximal over all λ^j_{min} in the neighbourhood of p_i.

This requirement is equivalent to requiring that the image gradient is strong in two orthogonal directions (for example, corners or checkerboard textures). In this sense, it is related to the edge density operator; however it avoids the need for an explicit edge computation and requires variation in two orthogonal directions. For example, whereas the edge density operator is equally likely to select a point anywhere along the edge of a doorframe, the Kanade, Lucas and Tomasi operator will prefer the corners of the doorframe. On the other hand, by avoiding the explicit edge detection stage it is sensitive to local variations in illumination and shading.

3.4. **Caltech Operator**[**2**]. This operator is by far the most complex we employed and can only be described here in summary form. This operator is based on the winner-take-all model of attention by Koch and Ullman[**24**]. This operator uses multi-scale saliency maps and three interest features: colour, luminance contrast and orientation. Features are extracted using a center-surround method implemented as the difference between fine and coarse scale responses. The feature maps created

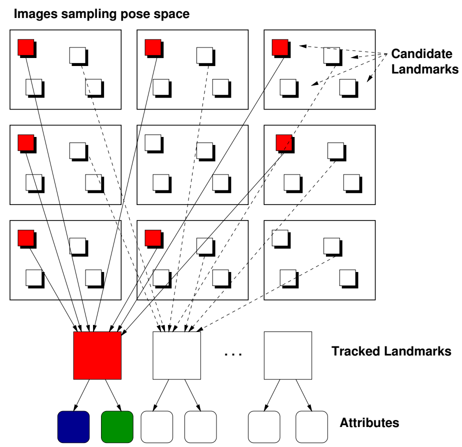


FIGURE 6. The offline training method.

using the center-surround method are fed into conspicuity maps. The colour features are extracted using red-green and blue-yellow center surrounds based on colour pop-out. Four Gaussian pyramids are created: one each for red surround and green center, green surround and red center, blue surround and yellow center and yellow surround and blue center. The luminance contrast map is made using dark center and light surround as well as light center and dark surround. Gabor pyramids[1] are used to find locations of orientation contrast between the center and the surround.

The three conspicuity maps are then normalized and summed to give the saliency map, which is then fed through a winner-take-all network of inhibition and return so that the global salient points can be located without selecting the same points twice (see Figure 5).

4. Landmark Learning

In this section, we briefly describe the landmark learning framework that we employ for our experiments. These descriptions are necessarily terse and further details have appeared elsewhere [**30, 4**].

Our localization framework consists of two distinct phases; an initial, off-line *learning* or *exploration* phase, and an on-line pose estimation phase. In the initial off-line phase a set of landmarks is extracted from image data and grouped for future recognition. A set of attributes of the learned groups, otherwise known as *tracked landmarks*, are encoded using a generic parameterization method, which is later exploited for characterizing the landmark as a function of camera position. The on-line phase, which is employed whenever the pose of the camera is required, consists of detecting and classifying landmarks from the current view, and thereby computing a pose estimate from the attributes of the observed landmarks. The framework is depicted in Figures 6 and 7 and described below.

- Off-line learning phase. (Figure 6):

[1] A Gabor filter consists of a sinusoidal grating with a Gaussian envelope.

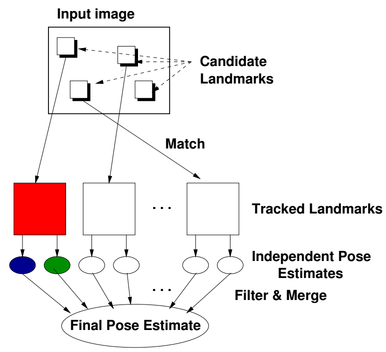


FIGURE 7. The online pose estimation phase.

(1) **Exploration:** Images are collected sampling a range of poses in the environment.
(2) **Detection:** *Landmark candidates* are extracted from each image using a model of visual attention.
(3) **Matching:** *Tracked landmarks* are extracted by tracking visually similar candidate landmarks over the configuration space.
(4) **Parameterization:** The tracked landmarks are parameterized on the basis of a set of computed landmark attributes (for example, position in the image, intensity distribution, edge distribution, etc), and then measured in terms of their *a priori utility* for pose estimation.
(5) The set of sufficiently useful tracked landmarks is stored for future retrieval.

- On-line pose estimation (Figure 7):
  (1) When a position estimate is required, a single image is acquired from the camera.
  (2) Candidate landmarks are extracted from the input image using the same model of visual attention used in the off-line phase.
  (3) The candidate landmarks are matched to the tracked landmarks learned in the off-line phase.
  (4) A position estimate and associated uncertainty is obtained using each computed attribute for *each* matched candidate landmark.
  (5) A final position estimate and uncertainty is computed by merging the individual estimates of the observed candidates.

5. Experimental Results

Our experimental methodology is as follows. We ran experiments for each attention operator in the domains depicted in Figure 8. The first domain consisted of a gantry-mounted robot arm with a camera affixed to the end effector. The camera was directed towards a simple scene constructed from a variety of objects. This domain was selected for the ease of measuring ground truth (to less than a hundredth of a millimetre). The pose space was restricted to a ten centimetre square.

Operator	Mean Error (cm)	Std. Deviation (cm)	Percent Error
Edge Density	0.088	0.00643	8.8%
Radial Symmetry	0.13	0.01	13%
Smallest Eigenvalue	0.092	0.005	9.2%
Caltech	0.13	0.018	13%
Random	2.9	5.6	290%

TABLE 1. Experimental Results for Simple Scene

Operator	Mean Error (cm)	Std. Deviation (cm)	Percent Error
Edge Density	12.5	19.6	62%
Radial Symmetry	22.7	45.4	113%
Smallest Eigenvalue	12.2	19.0	61%
Caltech	11.4	14.5	57%
Random	63.1	3265	315%

TABLE 2. Experimental Results for Laboratory Scene

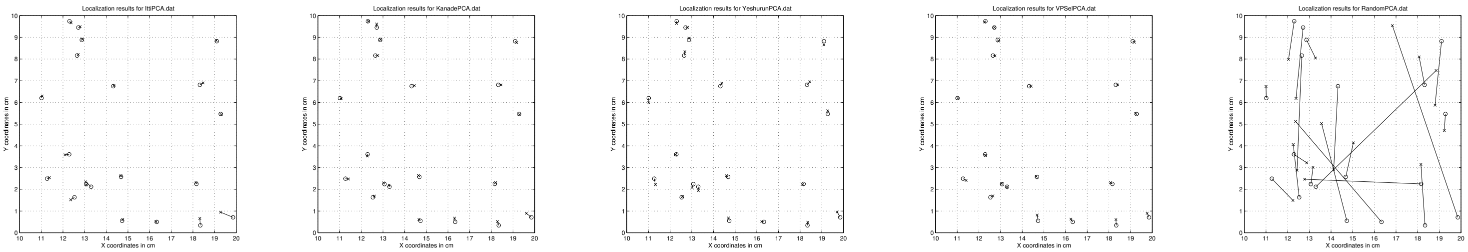
The second domain consisted of a Nomad 200 mobile robot operating in a two metre by two metre pose space in our laboratory. This domain was selected for its generic nature. As such, ground truth could be estimated only to within 0.3cm.

FIGURE 8. Simple Scene and Laboratory Scene.

For each domain under consideration, a set of training images was collected over a uniform sampling of the pose space; at one centimetre intervals for the simple scene and twenty centimetre intervals for the lab scene. For each attention operator a set of landmarks was learned from the training images. Finally, a second set of test images (twenty for the simple scene, thirty for the lab scene) were taken from random points in the pose space and applied to the online phase of each trained set of landmarks.

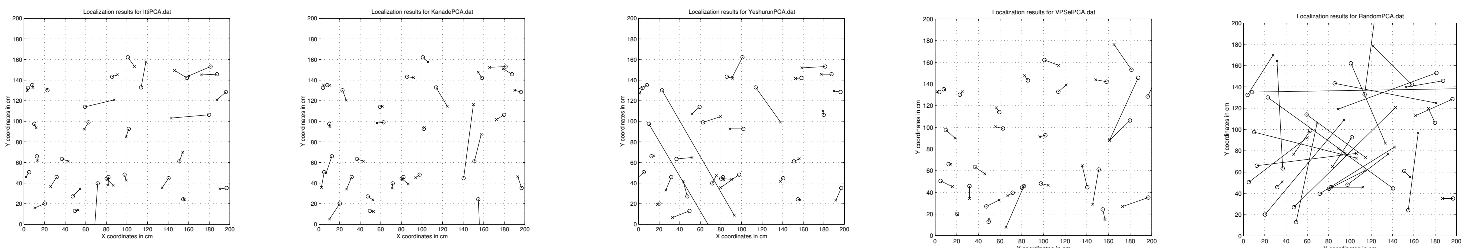
The mean pose estimate error and mean uncertainty was computed for each attention operator and are reported for the Simple Scene in Table 1 and for the Laboratory Scene in Table 2. In addition, we define the percent error as the mean error divided by the sample interval for the particular scene. Note that a localization scheme that simply chooses the "nearest" image would have an expected percent error of 50%.

Figures 9 and 10 depict the set of pose estimates for each attention operator in each domain, as plotted against their ground truth.



(a) (b) (c) (d) (e)

FIGURE 9. Simple Scene Results: a) Caltech, b) Smallest Eigenvalue, c) Symmetry, d) Edge Density, e) Random



(a) (b) (c) (d) (e)

FIGURE 10. Laboratory Results: a) Caltech, b) Smallest Eigenvalue, c) Symmetry, d) Edge Density, e) Random

The results for these experiments indicate that the pose estimation procedure was consistently more effective in the simple scene that in the much more challenging laboratoty scene across all attention oerators, as we would expect. In every case, the errors in the laboratory scene most of the errors was due to a small number of outlier measurements which could have been removed by subsequent post-processing (as discussed in our prior work on pose estimation). In the current experiments, we have left these outliers in place as their presence is part of what is being evaluated.

In all experiments, the radial symmetry operator produced worse performance than the other operators except the random operator. While the Caltech operator produced results commensurate with the edge and eigenvalue operators, the complexity of the operator and the substantial computing time involved suggest that it would not a a preferred candidate in practice (for example, it is the only operator that explicitly computes a multi-scale result).

6. **Conclusions**

In this paper we have evaluated the use of alternative attention operators for the task of robot pose estimation. This task also includes an implicit image recognition subtask and hence these results appear to be relevant for that related problem as well. Our pose estimation scheme is based on the selection of subwindows from an image under the assumption that they are locally stable in the face of (small) changes in illumination, translation, scaling, rotation and other factors.

Unlike a related operator evaluation by Schmid[**31**] the present study includes several complex "high-level" operators and considers them in the context of a pose estimation task.

Our results indicate that, in general, all four operators perform very well for the task of pose estimation, particularly in comparison with the performance of the random operator. The observant reader will note that the slightly poorer results of the Symmetry Operator in the Laboratory Scene are due to a small number of outlier estimates. Overall, we can conclude that stability is the dominant factor that leads to the good results demonstrated here. That is to say that each operator selects different kinds of points, but each does so with substantial stability. That stability is an essential property of a good attention operator is evidenced by the abysmal performance of the random operator.

Open questions include how to combined the results of more than one operator in the context of pose estimation and related tasks. The manner in which multi-scale data from various operators can be combined or exploited also remains to be considered more fully.

Bibliography

[1] Eric Bourque and Gregory Dudek, "Viewpoint selection – an autonomous robotic system for virtual environment creation", in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Victoria, Canada, October 1998.

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

[3] John K. Tsotsos, "The complexity of perceptual search tasks", in *International Joint Conference on Artificial Intelligence*, 1989, pp. 1571–1577.

[4] R. Sim and G. Dudek, "Learning visual landmarks for pose estimation", in *Proceedings of the IEEE International Conference on Robotics and Automation(ICRA)*, Detroit, MI, May 1999, IEEE Press.

[5] F. Pourraz and J. L. Crowley, "Continuity properties of the appearance manifold for mobile robot position estimation", in *Proceedings of the 2nd IEEE Workshop on Perception for Mobile Agents*, Ft. Collins, CO, June 1999, IEEE Press.

[6] G. Dudek and D. Jugessur, "Robust place recognition using local appearance based methods", in *International Conference on Robotics and Automation*, San Francisco, April 2000, IEEE Press.

[7] David G. Lowe, "Object recognition from local scale-invariant features", in *International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 1150–1157, IEEE Press.

[8] Matjaž Jogan and Aleš Leonardis, "Robust localization using panoramic view-based recognition", in *15th International Conference on Pattern Recognition.* September 2000, vol. IV, pp. 136–139, IEEE Computer Society.

[9] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition", in *Proc. 6th European Conference on Computer Vision*, Dublin, June 2000.

[10] Anne Treisman, "Perceptual grouping and attention in visual search for features and for objects", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, pp. 194–214, 1982.

[11] Anne Treisman, "Preattentive processing in vision", *Computer Vision, Graphics, and Image Processing*, vol. 31, no. 2, pp. 156–177, August 1985.

[12] Anne Treisman, "Features and objects in visual processing", *Scientific American*, vol. 255, no. 5, pp. 114B–125, Nov 1986.

[13] R.A. Rensink, J.K. O'Regan, and J.J. Clark, "To see or not to see: The need for attention to perceive changes in scenes", *Psychological Science*, vol. 8, pp. 368–373, 1997.

[14] Sandra Polifroni, "Evaluation of computational attention operators using human image recognition", Master's thesis, McGill University, 2000.

[15] A. Tankus and Y. Yeshurun, "Detection of regions of interest and camouflage breaking by direct convexity estimation", in *IEEE Workshop on Visual Surveillance*, Los Alamitos, CA, USA, 1998, pp. 42–8.

[16] D. Reisfeld, H. Walfson, and Y. Yeshurun, "Context free attentional operators: The generalized symmetry transform", *International Journal of Computer Vision*, vol. 14, pp. 119–130, 1995, Special Edition on Purposive Vision.

[17] M.F. Kelly and M.D Levine, "Annular symmetry operators: A method for locating and describing objects", in *ICCV '95*, 1995, pp. 1016–1021.

[18] Gal Sela and Martin D. Levine, "Real-time attention for robot vision", *Real-Time Imaging*, vol. 3, pp. 173–194, 1997.

[19] Yoram Bonneh, Daniel Reisfeld, and Yehezkel Yeshurun, "Quantification of local symmetry: Application for texture discrimination", *Spatial Vision*, vol. 8, no. 4, pp. 515–530, 1995, Feature Issue on Symmetry.

[20] Amnon Sha'ashua and Shimon Ullman, "Structural saliency: The detection of globally salient structures using a locally connected network", in *ICCV '88*, 1988, pp. 321–327.

[21] Carl-Fredrik Westin, Carl-Joahn Westelius, Hans Knuttson, and Gosta Granlund, "Attention control for robot vision", in *Computer Vision and Pattern Recognition*, 1996, pp. 726–733.

[22] Tony Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch", *International Journal of Computer Vision*, vol. 11, no. 3, pp. 238–318, 1993.

[23] Martin Jägersand, "Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach", in *Proceedings of the 5th International Conference on Computer Vision*, 1995, pp. 195–202.

[24] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry", *Human Neurobiology*, vol. 4, pp. 2428–2434, 1985.

[25] Sean M. Culhane and John K. Tsotsos, "An attention prototype for early vision", in *Computer Vision – European Conference on Computer Vision '92*, G. Sandini, Ed., Santa Margherita Ligure, Italy, 1992, Second European Conference on Computer Vision, pp. 551–560, Springer-Verlag.

[26] John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando J. Nuflo, "Modeling visual attention via selective tuning.", *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

[27] J. Shi and C. Tomasi, "Good features to track", in *Proc. IEEE International Conf. Computer Vision and Pattern Recognition (CVPR)*. 1994, IEEE Press.

[28] Carlo Tomasi and Takeo Kanade, "Detection and tracking of point features", Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, April 1991.

[29] R. Sim and G. Dudek, "Mobile robot localization from learned landmarks", in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems(IROS)*, Victoria, Canada, October 1998, IEEE Press.

[30] R. Sim and G. Dudek, "Learning and evaluating visual features for pose estimation", in *Proceedings of the Seventh IEEE International Conference on Computer Vision(ICCV)*, Kerkyra, Greece, sept 1999, IEEE Press.

[31] Cordelia Schmid, Roger Mohr, and Christian Bauckhage, "Comparing and evaluating interest points", in *International Conference on Computer Vision*, 1998, pp. 230–235.