Attention in Iconic Object Matching

T. D. Grove and R. B. Fisher
Department of Artficial Intelligence, Edinburgh University
5 Forrest Hill, Edinburgh EH1 2QL
PHONE: 0131-650-3098 FAX: 0131-650-6899
rbf@aifh.ed.ac.uk

Abstract

In this paper we present the attention sub-system of an iconic, picture based, vision system. Our system is based on an interest map that records the saliency of potential foveation points. These saliency scores are computed on the basis of both photometric features and knowledge of the likely relationships between object sub-components. We demonstrate our attention algorithm on artificial and natural images.

1 Introduction

This paper describes the attention mechanism developed during an investigation [5] into an iconic vision system, that is, a system that performs visual tasks (e.g. object recognition) using pictorial data obtained directly from images. In its use of 2D models, it bears some similarity to the eigenshape approaches of [1] and [13] and the raster-based Principle Component Analysis approaches of [15] and [11]. This contrasts with the alternative approach, which is to derive symbolic or geometric descriptions, which are attractive from a computational point of view, because they are compact and easy to manipulate. Unfortunately, they are also extremely hard to derive with any degree of reliability. On the other hand, although iconic data is bulky and hard to manipulate, it is readily accessible.

The view taken in this paper is that the accessibility of iconic representations outweighs their disadvantages, and that many of the apparent disadvantages of iconic representations can, to some extent, be overcome. This is especially true now that computer memory and processors have become faster and cheaper. We have been investigating a problem decomposition involving three parts:

1. A feature extraction mechanism that provides the input to the system, capturing important detail and suppressing noise and artifacts due to the imaging system.
2. A model matching mechanism that copes with variability in the relative rotation, scale, illumination, etc. between the models and data.
3. A visual attention mechanism responsible for locating items of interest. This is important in an iconic vision system, since we cannot afford to search the entire visible world for models as this would be computationally too expensive for large images. The use of visual attention also represents an approach

to the 'figure-ground' problem: When a region is attended, it is foveated. Foveation isolates the attended region from the background, effectively segmenting the image into figure and ground.

An additional goal was to develop this system in a manner which was broadly connectionist and biologically plausible, following the belief that the human (and animal) vision system depends to a large extent on an iconic subsystem, as well as a geometric subsystem [3].

In this paper we concentrate on the attention mechanism and the features that it uses as bottom-up cues to guide visual search. The matching mechanism is described in detail in [5].

2 Feature Extraction

This project adopts a foveated (r, θ) polar coordinate system for retino-centric coordinates. We use 20 bands, each of which contains 48 sectors. The receptive fields (i.e. the area of the (i, j) image from which they take input) of each pixel in the (r, θ) representation increases in size as r grows larger, in order to cover the entire foveated area. This increase in size is logarithmic, with the receptive fields of pixels being 1.2 times larger than those in the preceding band. The pixels in the innermost bands take their input from only one or a few pixels, averaging the value. This gives us high resolution around the origin. Pixels in the outermost bands average over large numbers of pixels, giving lower resolution. Pixels overlap by about 33% to avoid gaps which leads to a certain amount of blurring. A polar representation is attractive because it maps rotation and scaling into translation, and this feature is used in the matching algorithm to deliver scale and rotation invariance.

A foveated region is stored as a stack of (r, θ) images. This stack includes the raw (r, g, b) image data and this data at 1/2 and 1/4 scale (obtained by subsampling). This forms a small image-pyramid (or, strictly speaking, an image-cone) from which further features are extracted using low-level feature operators. The image stack contains a total of 42 (r, θ) feature images.

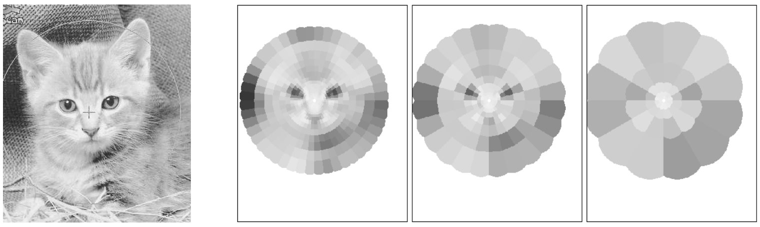
We want to detect features that represent interesting properties of the world, rather than artifacts of the imaging system and we would also like to be able to compute these features locally in agreement with the neurophysiology of early vision. The features used here are those proposed by Marr [9] and are based on the standard interpretation of neurophysiological investigations of the retina and primary visual cortex. These features are extracted from the outer half (i.e. the (r, θ) pixels with large r) of the planes in the image stack. They are:

- Positive and negative intensity blobs at three scales
- Positive and negative bars at three scales
- Corners at three scales
- Positive and negative R-G opponent colour blobs at three scales

P0	P1	P2
P3	X	P4
P5	P6	P7

The 3x3 Neighbourhood

Figure 1: The 3x3 Neighbourhood and Edge Templates



a. Input image   b. Foveated red channel at 1, 1/2 and 1/4 scale

Figure 2: (r, θ) data at three scales

All feature detection operators are based on a 3x3 operator (see Figure 1), and are applied to each point in the (r, θ) image to yield a new image. Rather than use larger operators, we reapply the 3x3 operator to a scaled intensity image. Halving the scale of the source image is equivalent to doubling the size of the operator. In this project, we use three scales : 1 , 1/2 and 1/4. This is illustrated in Figure 2.We apply most of our feature detectors to a gray-scale (r, θ) intensity image obtained from the (r, g, b) data.

Many authors emphasise the importance of corners in vision (e.g. [2]). Figure 3 shows the appearance of a typical corner in (r, θ) space. Equation (1) will give a strong response to a corner such as the one presented in Figure 3. The first term in this function will return a large value if there is a large difference between the centre pixel (x) and pixels p_1, p_3 and p_4. This defines the "tip" of the corner. The second term will suppress the first if this "tip" is not joined to a base (i.e, if there is a large absolute difference between p_x and p_6.) This term is scaled, since there is likely be some intensity difference between the "tip" and "base", due to the tip occupying less area of the pixel. The output of the corner detector is thresholded at zero. Functions similar to (1) are defined for detecting positive (white-on-black) and inverted (the 'corner' pixels are p_1 and x) corners.

$$negativecorner = \min(p_1 - x, p_3 - x, p_4 - x) - \frac{|x - p_6|}{2} \tag{1}$$

A blob is a local maximum (a positive blob) or minimum (a negative blob), i.e. a pixel which is either lighter or darker than all of its neighbours. A positive blob detector can therefore be defined as:

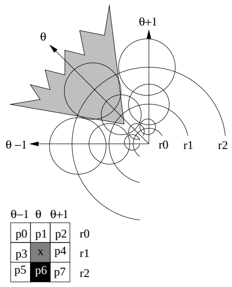


Figure 3: A corner in (r, θ) space

$$positiveblob = \min\left(x - p0, x - p1, \ldots x - p7\right) \tag{2}$$

The result of this operator is thresholded at zero. The negative blob has an analogous definition, with the x and p_n terms reversed.

Four bar detectors are used; a positive and negative vertical bar detector, and a positive and negative horizontal bar detector. The positive vertical bar detector is defined as:

$$\begin{aligned} +verticalbar = \min\left(p1 - p0, x - p3, p6 - p5, p1 - p2, x - p4, p6 - p7\right) \\ - \max\left(|x - p1|, |x - p6|\right) \end{aligned} \tag{3}$$

The first term finds the minimum difference along the sides of the bar. This is suppressed by the second term, which finds the absolute difference along the bar.

$R - G$ colour blobs are found by applying the blob detector to an image derived from the difference of the red channel from the green channel. This will give a strong response where there is a large local change in colour.

3 Attention

The purpose of the attention subsystem (see Figure 4) is to locate items of interest (that is, items that exist in the system's model base). A representation called an interest map controls this task by providing a record of the saliency of candidate foveation points. The interest map obtains values from low-level features extracted from the currently foveated region, and higher level context information derived from the matcher.

Desirable properties of an attention mechanism are that it 1) has a high likelihood of foveating regions containing matchable models, 2) explores the visual space, rather than continually refoveate the same region, 3) operates on both natural scenes and artificial scenes, 4) detects potential threats and changes in the environment and 5) is biologically and psychologically plausible (i.e. is broadly connectionist and exhibits the property of graceful degradation).

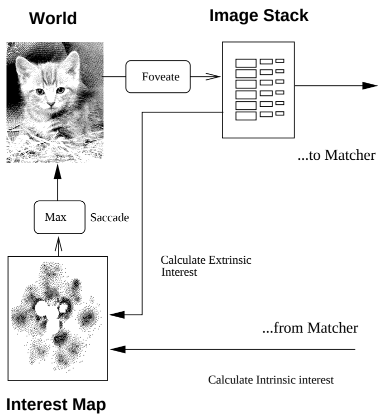


Figure 4: System Overview

Approaches to modelling visual attention include using 3D information derived by stereopis [14] and colour [10] scale-space blobs [8] and motion (e.g. [12] and [1]). Our approach uses 2d photometric features of the image combined with an 'interest map'. Each point in this map refers to a point in space to which attention could be directed via a saccade and has a saliency score. This map is the same size, and uses the same coordinate system as the world-image. After each match has taken place, the interest map is updated. The point with the maximum value is chosen as the new foveation point. The interest map can be regarded as analogous to the accumulator array used in the Hough transform. As in the Hough transform, bins in the interest map accumulate evidence from different features and different foveation points. This integration of evidence from a number of different sources can be expected to lead to robustness with respect to noise. For this reason, an interest map approach should be more reliable than a foveation stack (used, for example, by [7]) based technique. A less important feature that makes the interest map attractive in this project, is its potentially simple neural implementation as a sheet of mutually ihibitory cells. However, we are unaware of any evidence that supports the existence of such a structure in the human or animal vision system. Three sources modify the interest map: intrinsic interest, extrinsic interest, and suppression.

3.1 Extrinsic Interest

An extrinsic interest image is constructed by summing the various feature planes described in Section 2, evenly weighting each feature response. This image is then defoveated (mapped back into the (i, j) coordinate system) into the interest map. Before use, the interest map is initialised with a gaussian function centered on the image center, so that in the event of there being no stimuli, the system will perform a spiral search outward from the center of the image.

3.2 Intrinsic Interest

The intrinsic interest mechanism increases the interest of any areas likely to contain objects, based on what the system knows as a result of having already observed part of the image.

In addition to image and feature data, the image stack (and models) hold information about the relative position of models that commonly co-occur. This is used both to assist matching (allowing a weak image match to be sufficient to generate a model match, providing that it occurs in the right context) and attention. When a model is matched, the locations of models that occur with it are found and added to the interest map once the scale and rotation of the model for which the match occurred has been taken into account.

The effect of this is to increase the strengths of those parts of the interest map that correspond to the locations of unseen objects associated with the recently matched model. For example, matching an eye (without having previously matched any other features) will result in an increase in interest at locations corresponding to the locations of other features associated with this eye. (e.g. the mouth, the other eye, etc). The intrinsic interest component is weighted more heavily than the extrinsic component, on the grounds that it can be expected to be more reliable.

3.3 Suppression

In order to avoid constantly refoveating the same region, it is necessary to suppress the interest map. Whenever a model is matched, a large negative constant is added into the region of the interest map corresponding to the model's location. The size of the suppressed region is the same as that of the model. A smaller region is also suppressed whenever a foveation occurs, regardless of whether a model is matched. The interest of every point increases by a small constant on each iteration to ensure that points outside saccadic loops will eventually become sufficiently interesting to attract attention.

4 Attention Experiments

The extrinsic interest function consists of three components. Each component was tested independently on a simple image consisting of circles and squares. Figure 5 shows the use of blobs and bars in guiding ttention. For the first 20 saccades, the system spends a large amount of its time examining the middle circle. This is unsurprising, since its edges are close to being ideal (r, θ) space bars. After a number of foveations, all the interest points in the vicinity of this circle have been suppressed and the blob detector sends the system to foveate the filled circle, and then on to the box, after which the experiment was stopped.

The corner driven attention in Figure 6 displays a kind of edge tracking behaviour. This stems from the operator definition which defines a corner as an end point. If a bar passes through the fovea, the system will detect its termination or

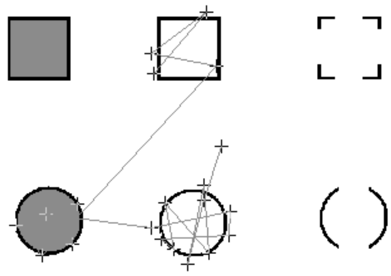


Figure 5: Foveations resulting from blob and bar detectors

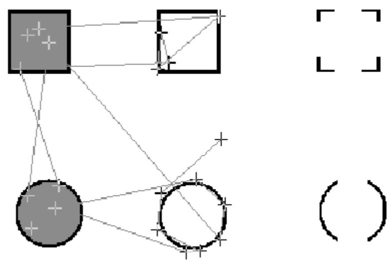


Figure 6: Foveations resulting from corner detection

a change in curvature as a corner, providing that this change is great enough. If there is nothing more interesting to look at, the system will track the boundary, which is what is happening in the case of the circle in Figure 6.

Figure 7 demonstrates the use of the opponent colour blob detectors. These can be used to detect the two coloured objects in the scene - being the lefthand filled square (green) and circle (red). The system cannot initially see either of the coloured objects, and so begins searching outward from the centre. After a number of unsuccessful foveations, the system foveates the circle. It does not, however, foveate the square as this appears to be just outside the periphery and is therefore not seen. The remaining colourless (i.e. black and white) items are not attended.

Two tests were carried out on real images. Figure 8.a image shows a cat, containing large quantities of texture. Figure 8.b shows a Ugaritic cuneiform script clay tablet, which is a much more artificial object, having sharp edges and well defined features. Both extrinsic and intrinsic interest are used.

In Figure 8.a three models are used: an eye, a face and a nose model. These have all been derived from the image. These models contain information related to the relative positions of the other models, so that once one model has been matched, intrinsic interest will assist in locating the other models.

Figure 8.a shows the saccadic path generated in the course of 30 saccades. The figure-ground separation is distinct in this image, since there are fewer strong features and they are further apart. The system seems to be strongly attracted to the eyes (which stand out as negative blobs) and the mouth (positive blobs). This

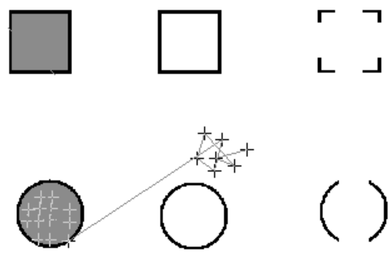


Figure 7: Foveations resulting from opponent colour blob detectors

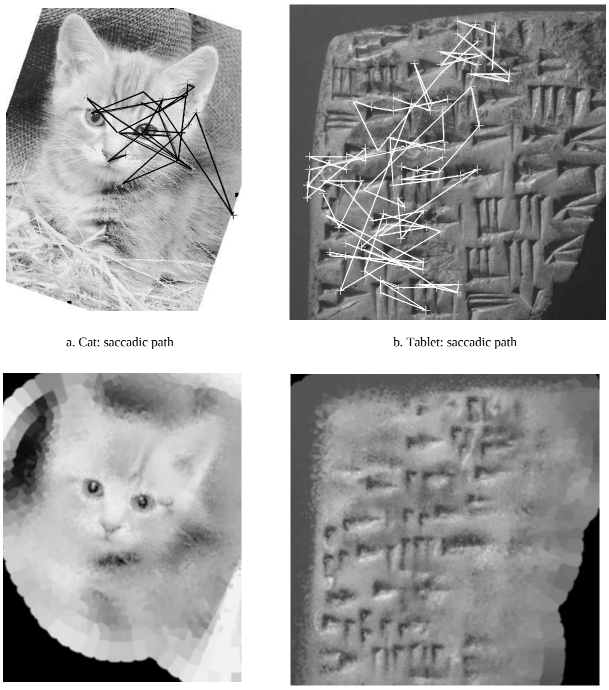
is slightly suggestive of human experiments [17]. There is also some attraction to the dark black marks on the cat's coat (negative bars). A reconstructed image (Figure 8.c) was created by defoveating the attended points back into an (i, j) image. The face stands out clearly in this image, and it is reasonable to say that the system as implicitly identified the features of the face as 'figure'. The system makes 5 successful matches in the course of 30 saccades. The left eye is matched twice.

In Figure 8.b, the system has two models: one for a horizontal wedge and one for a vertical one. The vertical wedge model states that there should be further vertical wedge on either side. While this is not always the case, there is a common character made of three vertical wedges. By defining a vertical wedge in this way, intrinsic interest enables the system to locate further wedges.

Figure 8.b shows the saccadic path generated after the system has explored about 50% of visual space during the first 90 saccades. Figure 8.d shows the reconstructed image. The system has made a reasonable attempt at separating the figure from the ground. Areas that do not contain any wedges, such as the top-left hand corner of the image, are not foveated and consequently appear as uniform in the reconstructed image. Many of the points visited also lie on or near wedges, causing the wedges to appear relatively distinctly in the reconstructed image. 16 models are correctly matched in the course of the search.

5 Conclusions

Tests on real images indicate that the attentional mechanism does separate figure from ground in the manner desired. It also explores a reasonable amount of the visual world and locates objects with sufficient precision to make it possible to match them using the algorithm described in [5]. The interest map approach implemented also has properties that the human visual system is also likely to possess: 1) it is attracted by features, 2) it is calculated locally from the registered feature map and 3) it is robust since local damage is isolated from other attention foci. The interest map could be neurally implemented as a sheet of mutually inhibitory cells. The purpose of the mutual inhibition would be to compute the max operation necessary to locate the peak corresponding to the most interesting point in the visual field. The types of features we are using to drive attention



a. Cat: saccadic path

b. Tablet: saccadic path

c. Cat: reconstructed image

d. Tablet: reconstructed image

Figure 8: Saccadic paths and reconstructed images

also seem plausible, although here only the low-level features such as intensity differences and corners are used, rather than higher level features such as faces. Treating blobs as figure has some similarity with the Gestalt idea of "enclosedness" [6]. There is also support from eye-tracking experiments [4] that observers are statistically more likely to foveate enclosed regions, than ones which exhibit poor "enclosedness". They conclude that this is due to the low-frequency channels dominating in peripheral vision. This agrees with the approach taken in this current project.

The attention mechanism could also be augmented by including higher level features such a symmetry and high level task direction. The features could be extended by the addition of 3D data (e.g. from stereopis) to give labelled surfaces discontinuities and including some treatment of motion.

References

[1] A Baumberg and Hogg DC. An adaptive eigenshape model. In David Pycock, editor, *British Machine Vision Conference*, volume 1, pages 87–97. BMVA, September 1995.

[2] M Brady. Computational vision. In D BroadBent, editor, *The Simulation of Human Intelligence*, chapter 5. Blackwell, 1993.

[3] MJ Farah. *Visual agnosia: disorders of object recognition and what they tell us about normal vision.* MIT Press, 1990.

[4] JM Findly. Local and global influences on saccadic eye movements. In DF Fisher, RA Monty, and JW Senders, editors, *Eye Movements, Cognition and Visual Perception.* Lawrence-ErlBaum Associates, 1981.

[5] TD Grove. Attention directed iconic object matching. Master's thesis, Dept. of Artificial Intelligence, University of Edinburgh, 1995.

[6] J Hick. Gestalt theory. In RL Gregory, editor, *The Oxford Companion to the Mind*, page 288. Oxford University Press, 1987. reference book entry.

[7] J Jennens. Quasi-invariant iconic object recognition. Master's thesis, Dept. of Artificial Intelligence, University of Edinburgh, 1994.

[8] Tony Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *Int Journal of Computer Vision*, 11(3):283–318, 1993.

[9] D Marr. *Vision.* W.H. Freeman and Company, 1980.

[10] J Matas, R Marik, and J Kittler. Illumination invariant colour recognition. In *Proceedings of the British Machine Vison Conference*, Birmingham, September 1995.

[11] H Murase and SK Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[12] DW Murray, KJ Bradshaw, PF McLauchlan, ID Reid, and PM Sharkey. Driving saccade to pursuit using image motion. *International Journal of Computer Vision*, 16(3):205–228, 1995.

[13] A Pentland, B Moghaddam, and T Starner. View-based and modular eigenspaces for face recognition. Technical Report 245, MIT Media Lab Vismod, 1993.

[14] PN Rajesh and DH Ballard. An active vision architecture based on iconic representations. *Artificical Intelligence*, (78):461–505, 1995.

[15] RPN Rao and DH Ballard. Object indexing using an iconic sparse distributed memory. Technical Report TR 559, Computer Science Dept., U. Rochester, 1995.

[16] AL Yarbus. *Eye Movements and Vision.* Plenum Press, 1967.