Retinal Sampling, Feature Detection and Saccades; A Statistical Perspective.

N.A.Thacker[1] and E.C.Leek [2]. neil.thacker@manchester.ac.uk
1: Univerity of Manchester, 2: University of Bangor.

Abstract

This paper applies statistical design principles to a simple biological model of human vision so that we can more clearly interpret the apparent role of eye saccades. In doing so we show that many structural features of the biological system (such as the optical geometry of the retina) are strategies for minimising the resources required to construct a working image recognition system. The ideas presented have implications for the construction of artificial (computer) vision systems. The computational model is very closely related to (but not based upon) SIFT, but more strongly based on a consideration of vision as a process of measurment while also linking the idea of multi-scale analysis with biological structure.

Introduction

It is generally assumed that a primary function of eye movements (saccades) is to maximize information processing within the high resolution region of the fovea [19, 20]. The measurement of saccades has been widely used in psychophysical studies in domains such as visual search, reading, scene exploration and interpretation and 2-D pattern recognition [7, 3, 14], and in machine vision studies, for example, of knowledge-based scene analysis and human interactions with complex displays [4, 8] These studies have shown that eye movements can be influenced by a variety of factors, including low- level image statistics, prior knowledge and task requirements. Surprisingly, little is known about the relationship between eye movements and three-dimensional object recognition. The problem is of course complicated by the difficulty of understanding exactly what data is provided by the visual process for the task.

In this paper we examine three fundamental questions. First, can fixation data reveal preferences for specific types of image features? Second, are gaze preferences consistent between stimulus encoding and recognition? Third, are extracted features invariant across tasks and across changes in 3-D viewpoint? In order to interpret the results of this study we must first define the data we beileve to be available from the retina. The first part of this paper therefore makes an argument for a simplified interpretation of biological function.

There are several structural properties of the retina which are well known. In particular the retina is known to have spherical geometry with logarithmic sensitivity to intensity. It is also considered reasonable to assume that the input data has radially varying spatial resolution. The evidence for this comes primarily from observations of the structure of the striate cortex and secondary visual cortex [13] (cortical magnification). This would appear to be in conflict with the known structure of the retina, where although there is such a distibution of sensors, they come in two specific types (the density of illumination

sensors (rods) reduces to zero in the fovea, while the density of colour sensors (cones) are at their greatest). What we need to remember however, is that it is in principle possible to synthesise an intensity measurement from a combination of colour ones. We argue here that the requirement to construct invariant quantites for recognition, combined with the statistical nature of the data, in particular measurement stability and consideration of a quantitative understanding of the information present in the data, allows us to arrive at an hypothesis for a simplified equivalent form. Specifically a set of homogenous regional samples from an exponentially distributed set of scales. As the retina can provide only one set of image samples at any one time, building an internal representation of the world around us requires this sensor to be moved around the scene. This process is supported in the human vision system by the saccades. In this paper we provide evidence which exclude some of the less likely generators of saccadic eye movement.

Computing Invariant Quantities

In order to understand why a detector such as the retina may have evolved it is necessary to consider the fundamental problems associated with visual recognition. Much of the process of visual recognition is understood to be template matching. This is a simple idea which is difficult to make work effectively in practice. Such a process is potentially very memory intensive unless the spatio-temporal relationships are constructed in a way which eliminates unnecessary variation in the input image, such as illumination[1], rotation and scale. This is often referred to as construction of an **invariant representation**. Many researchers in computer vision would also like to build systems with invariance to perspective changes, induced by 3D sensor or object motion. Often this can be locally approximated for small rotations by affine invariance (invariance to linear skew).

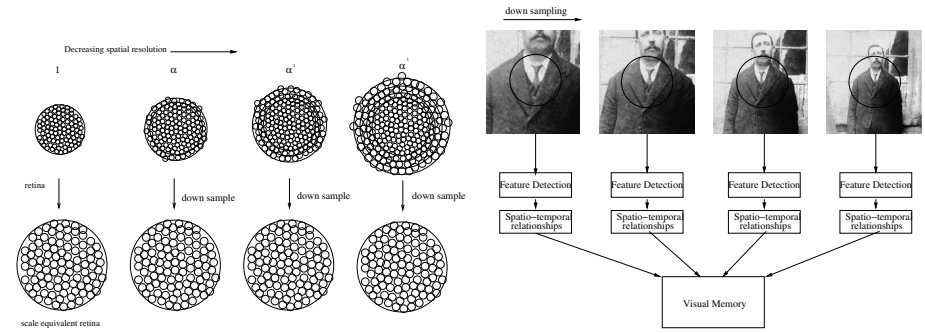
Yet combination of measured values into invariant quantities solves only part of the problem. We must also consider the associated noise characteristics. Scale and illumination invariance provide specific examples of this problem. For a CCD array the simplest way to compute illumination invariant quantities q from measured intensities $h_i \propto I + N(\sigma)$ at location i, is to compute a ratio such as $q = h_i/h_j$. Distributions of these quantities can then be treated as patterns for template matching. However, using error propagation it is possible to show that the noise characteristics of q will then be $var(q) = \sigma^2(1/h_j^2 + h_i^2/h_j^4)$. This introduces spatially varying errors on the computed invariant quantities, which then have to be properly addressed during pattern matching, by for example committing extra memory resources. The simplest way to deal with this issue is to make the invariant quantities have homogenous error by modifying the measurement process, ie: $g_i \propto log(I) + N(\sigma)$, as approximated in the human vision system. Then $q = g_i - g_j$ and $var(q) = 2\sigma^2$. One way to view this is to say, if we cannot deal with the variability introduced into invariant quantities by measurement noise then the extra information potentially gained by having a uniform sensor (h) provides no additional benefit to the alternative measurement system (g). Thus logarithmic sensitivity to light would appear to be a sensible strategy for an illumination invariant recognition system and one likely to be found in a system which has been optimised to minimise computational resources.

As with the illumination example provided above, potentially a relationship constructed from features detected at one scale can have different statistical reliability (re-

[1] I use this term here to refer to a simple linear scaing of overall intensity, and not the more general illumination variation which can occur in real scenes.

peatability error) to the same feature constructed from data at a higher scale. In the limit we can scale an object by such a factor that an object projects entirely onto one sensor. Although this is an extreme example, it is an illustration that data from a fixed sized sensor has a finite limit of information. Though the mean of any invariant quantity might remain fixed, the variance around that mean will change as a function of image scale. This will prevent reliable and efficient scale invariant recognition, which we might describe as **variable scale sensitivity**. If we wish to perform recognition with a pattern matching approach, simply constructing invariant geometric quantities is therefore not enough, as we need also to take account of the scale varying error on computed relationships.

The physical structure of the eye might appear unnecessarily complicated in comparison to a simple colour CCD array. To begin with the retina is a curved (approximately hemi-spherical) surface whereas a CCD array is flat. The optical model for a conventional electronic device is close to a pin-hole model. The geometric imaging process is described as perspective projection. Under such a model objects viewed at the centre of the field of view appear different to those at the edge. A spherical imaging surface on the other hand is rotation invariant and will produce an equivalent focused image of an object for any position in the field of view. This can be considered a simple form of perspective invariance. However, such a property pre-supposes a uniform sampling of the image, which is manifestly not the case for the human vision system.



(a) (b)

Figure 1: Computing scale invariant samples in the retina (a) and Scale invariant recognition (b).

One way to avoid the problem of variable scale sensitivity is by designing our image sensor so that it delivers a scale invariant measurement. We don't need to believe that this is exactly the biological solution at this stage, but if we cannot produce a scale invariant measurement system then we certainly cannot expect to produce any other scale invariant quantities[2]. As it is easier to consider the issue of scale invariance in cartesian geometry, we start on a planar sensor in what we will call the fovea where we assume the data has approximately homogenous spatial sampling [3]. We wish to synthesise a new sensor now at a lower resolution ($1/\alpha$). This process is illustrated in Figure 1a. The inner values of this new sensor can be computed from the fovea using a process of down-sampling. The most likely computational form for this process is a Gaussian convolution (or its irregular

[2]We are not talking here about the approximate invariants generally used in computer vision, but fully invariant quantities obtained by constructing a sensor so that such things are computable.

[3]As discussed above, in the biological system this process is complicated by the presence of two distinct types of light detector, rods and cones.

sampled equivalent) as this is the only scale invariant sub-sampling process, due to the central limit theorem. The outer values of the new region are unavailable in the fovea and therefore require additional sensors around the edge of the previous active region.

The new ring of sensors need only to have a spatial resolution which is a factor α larger than the ring of original samples on the edge of the fovea. To have better spatial resolution is potentially wasteful with resources. Down sampling once more, in order to construct the next equivalent sensor, we can again subsample the inner regions and we will need a new layer of measurements from areas now α^2 larger than the original. As this process continues we gradually construct a sensor surface, with ever decreasing spatial resolution. The mathematical process we have just described for positioning each ring of new sensor locations is

$$r_n = r_0(1 + k \sum_i^n \alpha^n)$$

Outside of the fovea, the radial function which characterises this process is obtained if the spacing of sensors is exponential outside of the fovea, (ie: a logarithmic sensor). This model differs from previous interpretations [13] [4], in that it accomodates the uniform sample density in the fovea as an intrinsic part of the calculation of scale invariant shape represenatation.

Making the adjustment from a planar sensor to another geometry, such as locations as angles ϕ on a spherical sensor, is mainly a case of applying the appropriate transform [5].

In conclusion, the equivalent homogenous sampling interpretation for feature detection on the retina now allows us to exploit the sampling properties of a spherical optical geometry. In particular, the spatial distribution of uniform sensors in the fovea can be used to compute the locations that sensors need to be placed when the same object is viewed at a larger scale in order to recieve identical input data. Such a scheme provides a partial invariance to projective deformation which is absent in a conventional electronic device.

An Overview of Statistically Based Feature Detection

We would like to take methods from computer vision as candidates for feature extraction in the brain. The topic of computer vision represents a large body of literature and in most cases the difference in behaviour and capabilities of the methods are heavily influenced by intended use. There are consequently as many feature detectors as there are applications. However, the general principles involved for much of this work can be characterised as template matching and interest operators. Both of these are physiologically viable methods for scene analysis. Template matching is analogous to the concept of receptive field patterns, while interest operators are more akin to the hypothesised mechanisms related to the possible role of micro saccades. The following section makes an argument for interpretation of these approaches as different aspects of the same underlying principle for information extraction.

Scaled features can be computed at a range of spatial scales by processing regular equivalent sensors with a fixed set of feature detectors, or (more likely) by combining

[4] Where an adjustment to the logratihmic form $log(r) \rightarrow log(r+a)$ is made which breaks the pure scale invariance property exactly where we might expect it is most needed, in the fovea.

[5] Actually, this argument to apply exactly we need the initial definition of the sample on the fovea to be equivalent to an homogenous sampling on a spherical surface not a planar one. The difference here is negligible for a fovea of small angular extent.

sub-sampling with the process of feature detection. The equivalent sample images need never be explicitly computed, but the resulting output of the system is equivalent to if they had been. Scale invariance can then be achieved by applying a selection process to choose the one scale most suitable for representation of the local image patch (Figure 1b). The selected representation at a single scale is then compared to stored visual memory. If the viewed object changes scale, by for example moving towards the viewer, then the scale at which selected features are detected will change so that the same, scale invariant, spatio-temporal relationships are selected for pattern matching. This process is effectively what is advocated in the Scale Invariant Feature Transform (SIFT) [10].

The best known of the template based approaches include the Canny edge detector [2] and the Difference of Gaussian operator [12]. These methods apply combinations of image convolution operations in order to enhance selected features. Connected or isolated feature points can then be identified as maxima on these enhancement images, in this case step edges and ridge locations respectively. The main difference between these two approaches is that although a step edge detector will respond to all but the finest of ridge features, Difference of Gaussian processing will not enhance step edges [6] (the most common manifestation of an object boundary) and has an output which is more strongly dependant upon feature scale and illumination.

In order to take advantage of the inherent properties of illumination invariance, feature detection processes need to be computed as combinations of differences between measured values g. For many features defined in computer vision, such as conventional first derivative based edge detectors, this is clearly the case.

$$e \;=\; G(\sigma) \otimes \sqrt{(g_{i+1} - g_{i-1})^2 \;+\; (g_{j+1} - g_{j-1})^2}$$

where $G\otimes$ represents a Gaussian convolution of width σ. Interestingly e^2 is a smoothed local estimate of the Fisher Information associated with local image plane orientation.

Although using a multitude of template feature detectors, all matched to distinct feature types, is a possible algorithmic solution for the extraction of object structure, this approach raises an important question; What is a valid mechanism to combine the responses from the different detectors? This must be done in a way which provides generalisation for recognition of shape, not only for changes in illumination and object scale, but also over possible responses to changes in scene content (for example arbitrary possible backgrounds at object boundaries). Interest operators provide an alternative which is capable of detecting many characteristic feature types with one simple computation. The simplest interest operator would be a local variance estimate of image signal, which when applied at a range of spatial scales, is also useful as a descriptor of texture [5]. For example;

$$v \;=\; G(\sigma) \otimes (g - < g >)^2 \quad \textit{where} \quad < g > \;=\; G(\sigma) \otimes g$$

which can be considered as either the etimate of signal to noise associated with local image variation, or the inverse of Fisher Information associated with a mean value. This simplifies to;

$$v \;=\; G(\sigma) \otimes g^2 \;-\; (G(\sigma) \otimes g)^2$$

[6] The response to a step edge from a DoG filter (as advocated in SIFT), is exactly zero at the position of the edge. Although it takes large positive and negative values on either side, the locations of the maxima are systematicaly shifted as a function of the Gaussian kernel size and therefore cannot be considered as spatially consistent.

Notice that these calculations embed directly Gaussian convolutions, which we have already stated are required for scale invariant resampling.

Another popular feature detector in computer vision is the corner detector and one approach uses the concept of an interest operator which is often based on the idea of auto-correlation. Corner detection can also be performed with templates, but is significantly more difficult than edge detection due to additional variation in orientation and corner shape [7]. The Harris corner detector [6] defines corner locations using the second order spatial variation of an auto-correlation around a point. However, auto-correlation can be interpreted as a log-likelihood for the degree of match between the original local image patch and a shifted version. In adddition the matrix of second order behaviour is the second term in a Taylor expansion, so it can also be interpretted as the second derivative with respect to image location. As the Cramer-Rao bound is the second derivative of a log likelihood, this is the Fisher Information for spatial localisation.

In summary we now have three definitions of feature detector based upon quantitative measurements which define positions of maximum information; local variance for spatially varying intensity, edge strength for orientation, and interest operators for spatial location. In fact, if we consider feature detection as a template based approach supported in the biology by receptive fields, then grey level scale, orientation and location are the only measurable quantities possible. It makes sense to suggest that if restrictions on processing capacity (for example finite connectivity) results in the need for the brain to identify a subset of features in order to solve quantitative tasks, then those which make the largest quantitative contribution (ie. those which maximise some aspect of Fisher information) are the ones it should be using and the ones we should be basing any model of scene interpretation upon.

Finally, illumination invariance of these measures and also for colour is entirely reliant upon logarithmic sensitivity to light. As with spherical optical geometry, this is a property which is lacking in a conventional electronic sensor. It is becoming increasingly obvious that when it comes to getting a simple solution to visual analysis tasks, the biological sensor has characteristics which make a lot of sense. Indeed, the analysis of data from a conventional colour CCD array will be difficult by comparison.

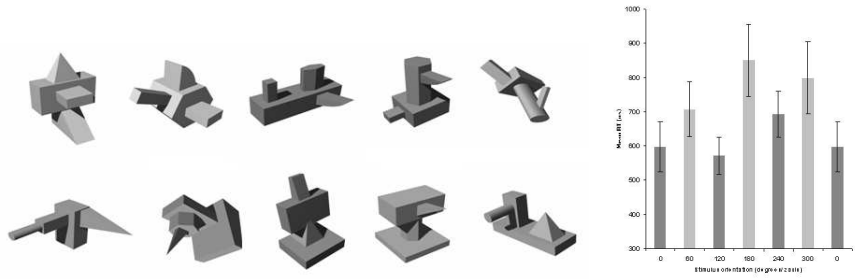
Assuming that the process of shape recognition is based upon conjunctions of detected features, then the above description of a multi-scale feature detection process eliminates the need for scale invariance. If we also eliminate the possibility of full 3D rotation invariance (on grounds of in-homogenous error characteristics), the required invariances for a shape representation are therefore translation and rotation within the sensor "plane". An ideal representation of shape would be one which supported the reconstruction of the shape up to an unknown position and orientation. Such a representation has been previously described as "complete". Although simple regional histograms of local image orinetation (as used in SIFT) are not complete [16], the property was established over a decade before for the representation scheme referred to as "pairwise geometric histogrames" [15]. This approach provides an encoding of local shape as a 2D frequency distribution of relative angle against perpendiular distance.

Methods: Investigating Patterns of Eye Movement

The simplest hypothesis for the role of saccades is that we move our eyes in order to build up a high resolution measurement of the scene. This hypothesis can be immediately ex-

cluded by observing real eye movements, which do not uniformly scan the potential view field, but seem instead to be drawn to particular visual features, movement or objects. A more sophisticated hypothesis for the role of saccades in visual exploration is that we saccade to areas which are expected to have useful spatial information for the interpretation of shape or structure [19]. We would therefore expect the eye to saccade to those features which are most useful for this task. As we have only low resolution data available in the periphery of the retina, we must assume that this is somehow used to predict the most useful places for fixation. Although we may not know precisely what the human vision system does, we suggest here that we can take the standard feature detection processes as characterised by interest operators and template matching approaches as indicative of those features which would be useful for the purpose of extracting image structure. We can then see to what extent the saccadic process targets locations in images which contain structural features in order to examine our initial hypothesis.

In brief, participants (N = 24; Mean age = 22.67) first viewed sets of six novel 3-D objects each containing one principal component and three sub- components or volumetric parts (see Fig 3a). 12 objects (6 targets and 6 distractors) were presented from three different viewpoints (0, 120, 240 degrees) each for 10 seconds while eye movement patterns were recorded. Following the Learning Phase, participants performed a recognition memory task in which they had to discriminate learned from unfamiliar objects, presented either at practiced (0, 120, 240 degrees) or novel orientations (60, 180, 300 degrees) in depth. Behavioural responses (accuracy and Reaction Times (RT)) were recorded. Eye movement data were recorded on a Tobii ET17 remote eye tracking system running at a data acquisition rate of 50 Hz. Experimental stimuli were viewed from a distance of 60 cm at a screen resolution of 1280 x 1024.



(a) (b)

Figure 2: Novel objects (a) and reaction times for recognition (b).

Results

Accuracy of target detection in the Test Phase was very high (range 80- 94.17 %). As expected, targets were detected more accurately at the practiced viewpoints, F (1, 23) = 26.01, $p < .001$. Mean RTs for correct trials are shown in Fig 3(b). A 2 (Familiar vs Familiarity) x 3 (Viewpoint) repeated measures ANOVA showed that RTs were faster for practiced over unfamiliar viewpoints, F (1, 23) = 13.73, $p < .001$. The main effect of Viewpoint was not significant. There was no interaction.

Analyses of eye movements were conducted by initially pre-processing raw gaze data by applying spatial and temporal filters to remove micro-saccades and drift. Fixations were defined as eye movements that remain within the same circular region (diameter 60

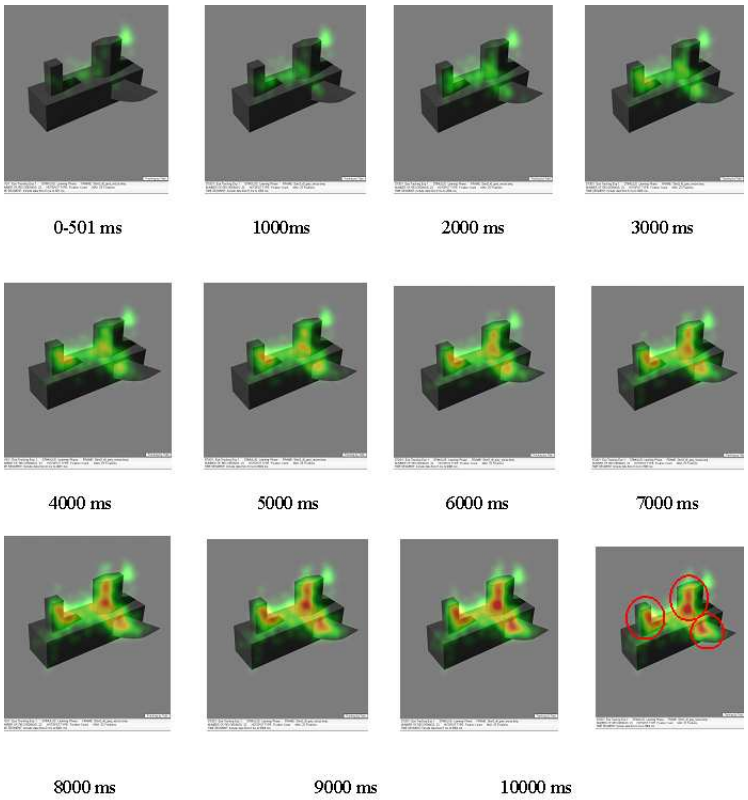


Figure 3: Time development of fixations.

px) for a minimum of 100 msecs. Filtered data were used to compute fixation frequency across participants for each stimulus. Figure 4 shows a time series fixation frequency plot for all participants overlaid onto the original stimulus image. The data are grouped into 10 epochs corresponding to the first 500 msecs post stimulus onset, and then for each 1000 msecs thereafter. The data show that participants rapidly fixate on image regions which appear to correspond to salient 3-D image segment points around object sub- components. Figure 5(a) shows an analysis of the consistency of fixations across changes in the 3-D viewpoint for two of the test stimuli. This shows that participants consistently search for and fixate the same 3-D image segmentation points across viewpoint, despite changes in the low-level image properties of these locations (e.g., vertex types). Figure 5(b) shows an analysis of the consistency of fixations between the Learning and Test Phases for two items. As in the previous analysis, participants fixate the same image regions between phases.

Conclusions

This paper has sought to explain saccadic eye movement within a framework which includes some of the more obvious structural features of the human vision system. It seems to be possible to account for many observed properties, including logarithmic intensity sensitivity, and spherical optical geometry, in terms of construction of invariant repre-

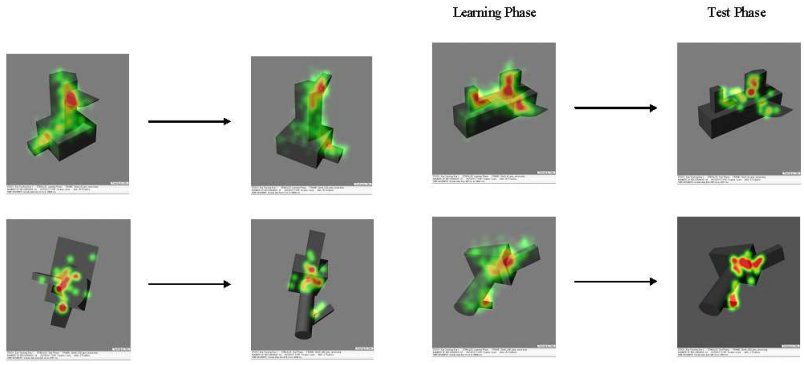


Figure 4: Changes across viewpoint (a) and consistency between learning and test phases (b).

sentations which take account of measurement error. The kinds of algorithms generally developed in the area of computer vision, with regular input lattices and at fixed scales, may seem a world away from irregular sampling of the retina and saccades. However, it seems possible to replicate the process of scale analysis by simply processing at multiple scales and selecting one result. This opens the possibility of applying insights from image analysis to interpretation of visual biological.

Ultimately the brain must use detected features for the construction of shape representations. The brain will need to analyse the incoming spatio-temporal data to extract compact descriptions for the purpose of accurate prediction and categorisation. Analysis of the statistical nature of the data tells us that it is not possible to construct a representation which is invariant to every form of variation produced during image formation. However, invariances are key to the construction of efficient vision systems, as the more we can correctly generalise from data we have already learned and understood, the easier it is to interact with our environment. The development of invariant recognition processes could be invoked as an implicit target during the process of human evolution, thereby explaining the kind of structure we see on the retina today.

Our study supports the following conclusions: (1) Human visual biology is consistent with a simple structural hypothesis (based on multi-scale samples) for the construction of invariant recognition systems which opens the way for interpretation of retinal data using conventional machine vision approaches, (2) Fixational eye movement patterns during 3-D object recognition are not random, but rather structured and highly consistent among observers. (3) There is remarkable consistency in the patterns of fixational eye movements shown for 3-D novel objects across both changes in viewpoint and between learning and test phases. (4) These locations show evidence of 'top down' selection and are **not** the low level features generally constructed for machine vision.

These conclusions are not dependent upon the particularly simple nature of our stimuli, and tell us that tracked features are a long way from the input visual data, in terms of processing. They imply a high level representation of 3D structure which is already available prior to eye movement. The most striking observation is that fixated locations are often on surfaces, which in our data at least contained no low level information. At first sight this may seem to be at odds with a feature based analysis of shape. However,

these conclusions can be reconciled with a feature based analysis if we take a view based approach to recognition, whereby sets of features within a focal region are used for shape representation such as a learned set of geometric co-occurences (such as the PGH). We can have every reason to believe that fundamental understanding of the problems involved in extracting shape information from conventional images is potentially of direct relevance to understanding high level processing in human vision. This being the case, locations of saccadic fixation should contain valuable information which can help identify these high level processes. This is an avenue we intend to explore further.

References

[1] A.P.Ashbrook, N.A.Thacker, P.I.Rockett and C.I.Brown, 'Robust Recognition of Scaled Shapes Using Pairwise Geometric Histograms.', proc., BMVC 95 Birmingham, 503-512, July 1995.

[2] J. Canny., A computational approach to edge detection., IEEE Transactions on Pattern analysis and Machine Intelligence, 8(6),679-698, 1986.

[3] A.T. Duchowski, Eye Tracking Methodology: Theory and Practice. Springer. London, 2003.

[4] J.M. Findlay and I.D. Gilchrist, Eye guidance and visual search. In G. Underwood (Ed.), Eye Guidance in Reading and Scene Perception. Oxford. Elsevier, 1988.

[5] R.M. Haralick, Statistical Image Texture Analysis, Handbook of Pattern Recognition and Image Processing, Ed. T.Y Young and K.S. Fu, Academic Press, Orlando, 247-279, 1986.

[6] C.Harris and M.Stephens., "A Combined Corner and Edge Detector" Proceedings of the Fourth Alvey Vision Conference. 147-151, August 1988.

[7] J.M. Henderson, C.C. Williams, M.S. Castelhano, R.J. Falk, Eye movements and picture processing during recognition. Perception and Psychophysics, 65, 725-734, 2003.

[8] R.S. Johansson, G. Westling, A. Backstrom, J.R. Flanagan, Eye-hand coordination in object manipulation. Journal of Neuroscience, 21, 6917-6932, 2001.

[9] A.J.Lacey, N.A.Thacker and N.L.Seed. 'Smart Feature Detection Using an Invariance Network Architecture', proc., BMVC 95 Birmingham, 327-336, July 1995.

[10] D.G.Lowe, Distinctive Image features from Scale-Invariant Key-points, Int. Jou. Comp. Vis, 2004.

[11] S. Martinez-Conde, S.L. Macknik and D. H. Hubel,, The Role of Fixational Eye Movements in Visual Perception, Nature Reviews Neuroscience, 5(3), 229-238, March 2004.

[12] T. Peli and D. Malah., A study of edge detection algorithms., Computer Graphics and Image Processing, 20, 1-21, 1982.

[13] E.L.Schwartz, Spatial Mapping in the Primate Sensory Projection: Analytic Structure and Relevanve to Perception, Biol. Cyber., 25, 181-194, 1977.

[14] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, C. Zetzsche, Knowledge-based scene analysis with saccadic eye movements Human vision and electronic imaging IV; Proceedings of the Conference, San Jose, 520- 532. 1999.

[15] N.A.Thacker, P.A.Riocreux, and R.B.Yates, 'Assessing the Completeness Properties of Pairwise Geometric Histograms", Image and Vision Computing, 13, 5, 423-429, 1995.

[16] N.A.Thacker and J.E.W.Mayhew, 'Designing a Network for Context Sensitive Pattern Classification.' Neural Networks 3,3, 291-299, 1990.

[17] N.A.Thacker, I.A.Abraham and P.Courtney, 'Supervised Learning Extensions to the CLAM Network.' Neural Networks Journal, 10, 2, pp.315-326, 1997.

[18] N.A.Thacker, F.Ahearne and P.I.Rockett, 'The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.' Kybernetika, 34, 4, 363-368, 1997.

[19] L. Walker and J. Malik. Sequential information maximisation can explain eye movements in an object learning task. Journal of Vision, 4, 744a. 2004.

[20] A.L. Yarbus, Eye Movements and Vision. New York. Plenum Press, 1967.