2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

An Architecture for Object-based Saccade Generation using a Biologically Inspired Self-organised Retina

Sumitha Balasuriya and Paul Siebert
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ
E-mail: sumitha@dcs.gla.ac.uk, psiebert@dcs.gla.ac.uk

Abstract— **Our paper presents a fully automated computational mechanism for targeting a space-variant retina based on the high-level visual content of a scene. Our retina's receptive fields are organised at a high density in the central foveal region of the retina and at a sparse resolution in the surrounding periphery in a non-uniform, locally pseudo-random tessellation similar to that found in biological vision. Multi-resolution, space-variant visual information is extracted on a scale-space continuum and interest point descriptors are extracted that represent the visual appearance of local regions. We demonstrate the vision system performing simple visual reasoning tasks with the extracted visual descriptors by combining the sparse information from its periphery (which gives it a wide field of view) and the high resolution information from the fovea (useful for accurate reasoning). High-level semantic concepts about content in the scene such as object appearances are formed using the extracted visual evidence, and the system performs saccadic explorations by serially targeting 'interesting' regions in the scene based on the location of high-level visual content and the current task it is trying to achieve.**

I. Introduction

The main novel contributions of this paper are (1) computational machinery to extract scale and orientation invariant *interest point descriptors* from a vision system that uses any arbitrary non-uniform sampling pattern, (2) top-down high level object hypothesis based saccadic behaviour in a *space-variant* vision system that uses a biologically inspired retina with a non-uniform self-organised retina tessellation.

The vision of all higher order animals is *space-variant*. Unlike most conventional computer vision systems, in these animals, sampling and processing machinery are not uniformly distributed across the animal's angular field-of-view. The term space-variant was coined to refer to (visual) sensor arrays which have a smooth variation of sampling resolution across their workspace similar to that of the human visual system. In the human retina and visual pathway, visual processing resources are dedicated at a much higher density to the central region on the retina called the *fovea*. The retina regions surrounding the fovea (which will be referred to as the *periphery*) are dedicated increasing less processing resources, with resources reducing with distance from the fovea. There is a smooth, seamless transition in the density of the processing machinery between the central dense foveal region and the increasingly sparse periphery.

Space-variant visual processing is incomplete without an effective, automated means of targeting or fixating the high resolution foveal region of the retina on 'interesting' regions in the visual scene. While information from the high acuity, narrow angle foveal region of the retina is useful for making task-based reasoning judgments, it is mainly the low resolution information from the large, wide angle, peripheral areas of the retina that is useful to decide upon potential locations for future retinal fixations.

Interesting regions (*salient*) locations are determined using *bottom-up* information to help restrict attention of the system to areas in the image where there is activity or entropy among low-level visual features and *top-down* information which enables the saccadic fixations of our system to be influenced by the particular task that the system is attempting to perform.

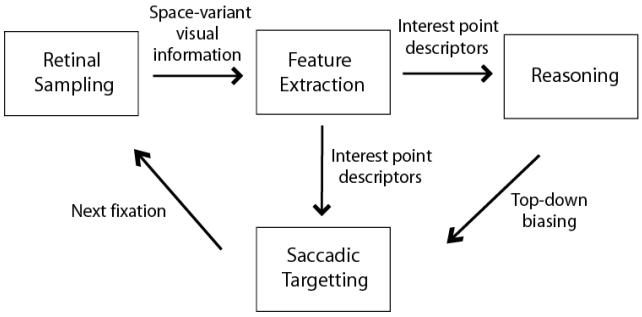

Fig. 1. Feed-forward space-variant architecture for object based saccade generation

II. Related Work

The topological mapping of biological retina afferents to the cortical visual areas has inspired researchers to compute the mathematical projection of visual stimuli at coordinates in the input image to those in an image structure often referred to as the *cortical image*. These analytic projections are often called *retino-cortical transforms*, and the most widely used is the log(z), complex-log or log-polar transform [1]. Other retino-cortical transforms such as the log($z+\alpha$) model [2] which splits

0-7803-9490-9/06/$20.00/©2006 IEEE

the cortical image along the vertical meridian into two visual hemispheres similar to that in biology and mapping [3] which attempt to closely approximate the magnification transfer functions in primates have been proposed.

The construction of computer vision systems that sample images with retinae based on retino-cortical transforms have resulted in singularities or distortions in the sampling of the visual scene. The widely used log(z) transform samples images at a super-Nyquist rate at the fovea region resulting in a cortical image with redundant correlated visual information. As the cortical image, which is the internal representation for visual information processed by higher level reasoning machinery, contains highly correlated visual information, this results in a large waste of system resources negating the efficacy of space-variant vision.

Approaches have been proposed that attempt to address the mentioned problems within the retina's central foveal region by having a separately defined uniform fovea [4] or a discontinuous cortical image [5]. However these create discontinuities in the internal representation, difficulties in reasoning with features that cross the foveal and peripheral representations or result in a sudden change in the topological structure of the retina tessellation between the fovea and periphery not maintaining a seamless merging of the uniform foveal topology to the space-variant peripherally as found in nature.

Biological retinae appear to exhibit an locally, almost uniform hexagonal receptor packing structure in the fovea that transforms seamlessly into an exponential sampling structure (resembling log-polar) towards the peripheral field. The authors have not found any analytic or geometric mapping reported in the literature with a closed form solution that can describe this change in topography of the receptive field centres between the fovea and periphery of a retina tessellation. These reported retino-cortical transform approaches do not address the question of the actual $\mathbb{R}^2$ locations (retina tessellation coordinates) where the computer vision system should extract visual information from an image or video. The work in the literature is based on the projection of the *radial* component of retinal coordinates to a cortical space and does not deal effectively with the angular relationships between the coordinates in the input retinal plane when performing the projection. This failing results in images being over-sampled in the foveal region during the construction of plausible retinae that sample pre-digitised media.

Recently, researchers have constructed plausible non-uniform artificial retinae [6] based on a form of self-organisation [7] which can extract space-variant visual information from images. They demonstrated the detection of stable Laplacian of Gaussian scale-space extrema in the space-variant visual information. Stable, local regions in the visual scene have proved to be very useful in computer vision. These locations referred to as interest points have been effectively used to represent visual content in rectilinear images or video based on the (two-dimensional) local appearance of the support around a detection interest point [8, 9, 10] and an approach similar to Lowe[10] but implemented in scale-space over a space-variant non-uniform tessellation will be used in this paper.

The unification of different visual feature modalities into a single structure representing importance or saliency was addressed in *Feature Integration Theory* [11]. Quoting from the paper [11], "focal attention provides the glue which integrates the initially separable features into unitary objects." Bottom-up computational models based on this approach [12] combined different visual features into a single saliency map which was used to determine the next location for fixation in a winner-take-all manner with inhibition-of-return preventing the system revisiting previously explored visual regions. While this model [12] is currently widely used in computer vision it may be considered lacking as a model for space-variant attention as a space-variant sampling or sensor was not used to extract visual formation for saliency calculation. Therefore their attention mechanism would previously know the visual contents of an unattended region in the scene *before* fixating upon it with a focus of attention spotlight.

The top-down priming of features for object search is reported in the literature using simple low resolution colour histogram cues to drive the saccades of their system in an object search task [13]. This was an early implementation of top-down biasing of a multi-resolution search which did not use a space-variant sensor with sampling density continuity between foveal and peripheral regions and the primitive colour cues used in that study are neither robust nor descriptive. Rao [14] presented work on top-down gaze targeting based on a cortical image generated by the space-variant log-polar transform [1]. Visual content was represented by the responses of regions in the cortical image to Gaussian derivatives and a goal or target image was used to create a saliency map in an object search task. However, the system needed to be manually provided with a 'scaling correction' to reason between visual information extracted in its fovea and its periphery. The top-down search algorithm was not effective without this external scaling correction input. This work also did not use a local, interest point based visual representation which would increase robustness and efficiency, but instead the system processed the whole cortical image for saliency.

Models of visual object and hierarchical grouping based approaches for saliency and attention have been proposed for targeting a space-variant sensor[15]. These were dependant upon the manual generation of top-down visual content such as objects. In this paper the authors will demonstrate a space-variant system centred around a self-organised retina performing fully automated saccade generation based on high-level understanding of the presence of visual objects in the scene.

III. Self-organised Artificial Retina

A self-organisation methodology [7] was used to determine the retina tessellation, i.e. the locations of the centre of receptive field on the retina. In this approach the stimulatory input for self-organising the retina tessellation is derived by applying a composite transformation to the current tessellation itself. For a retina with N receptive fields, each characterised by a two dimensional vector $r_i(n)$, the input stimulus $s_i(n)$ at iteration n is calculated by the following,

$$s_i(n) = T(n)\, r_i(n-1)\,. \tag{1}$$

where $r_i(n - 1)$ is the i^{th} receptive field centre at iteration n - 1 and $1 \leq i \leq N$. To generate a space-variant retina with a uniform fovea the following (ordered) composite transform T is used

1. A random rotation about the centre of the coordinate space between 0 and 2π
2. A dilation (increase in eccentricity from the centre of the coordinate space) of the exponent of a dilation factor which is random between 0 and log(8).
3. A random translation between 0 and *f*, in a random direction between 0 and 2π, where f is associated with the required foveal percentage of the resultant retina.

The self-organisation is initialised with a random retina tessellation configuration and recursively iterated with the described composite transformation T and the following learning rule to find the updated receptive field location vector $r_j(n)$

$$r_j(n) = r_j(n-1) + \alpha(n) \sum_{i \in \Lambda_j(n)} \left(s_i(n) - r_j(n-1) \right) . \quad (2)$$

$$\Lambda_j(n) = \left\{ i : \left\| s_i(n) - r_j(n-1) \right\| < \left\| s_i(n) - r_k(n-1) \right\|, k \neq j \right\} . \quad (3)$$

$\Lambda_j(n)$ contains the indices to the input stimuli $y_i(n)$ to which $x_j(n-1)$ is the closest network vector. $\alpha(n)$ is a learning parameter which controls the stimulation of the network weights. The learning parameter α is linearly reduced (annealed) throughout the self-organisation to increase the speed of convergence of the network weights to a stable configuration. Intuitively, one can visualise the effect of the learning rule as each network weight $x_i(n-1)$ being updated individually by the input stimuli $y_i(n)$ that are closer to that weight than any other in the network.

A retinal tessellation is not yet a retina. To prevent aliasing, visual information must be gathered over a large support region around each retina tessellation coordinate using a receptive field. The standard deviation (and in turn size) of the receptive field was related to the local spatial sampling rate of the retinal receptive fields. Basing a retinal receptive field's size on local node density also results in space-variant retinal receptive fields. As illustrated in Fig. 2, at the foveal region, where visual information is densely sampled, receptive fields will have a narrow spatial support, while large receptive fields will be placed at the periphery with its widely spaced sampling points. All processing machinery in this paper will operate upon the non-uniform space-variant visual information extracted by the self-organised retina.

IV. Processing Non-Uniformly Sampled Visual Information

The retinal receptive field responses extracted by the self-organised retina do not have an associated cortical image data structure [1] for storage for higher level reasoning. The authors stored these responses as a one dimensional vector which will be referred to as an *imagevector*. Allocation into the imagevector was based on a look up table; therefore each location on the imagevector has a consistent spatial relationship with a receptive field on the retina tessellation. These are able to extract higher level visual features such as blurring or contrast information from the imagevector representation.

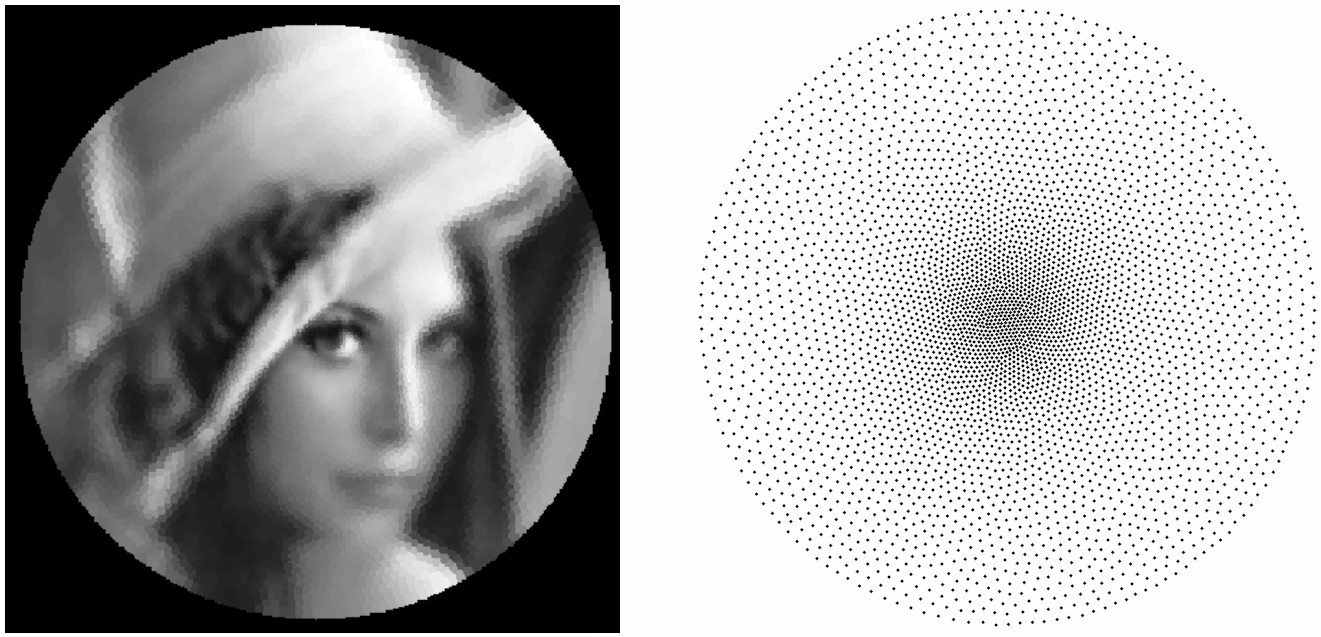

Fig. 2. (Left) Self-organised retina tessellation of a retina with 4096 receptive fields. (Right) Responses of a retina with 8192 Gaussian receptive fields fixated upon the centre of the standard greyscale Lena image. The receptive field responses were visualized in their respective Voronoi regions on the non-uniform tessellation.

The responses from the retina receptive fields were processed by machinery which shall be referred to as *cortical filters*. Nodes in the non-uniform retina tessellation (or any arbitrary sampling grid) v_i are within the support region $\Omega(v_c)$ of the cortical filter centred on v_c (the same or another non-uniform sampling grid) if the following is satisfied

$$v_i \in \Omega(v_c), (v_{i,x} - v_{c,x})^2 + (v_{i,y} - v_{c,y})^2 < r_c^2 . \quad (4)$$

where r_c is the radius of the cortical filter support and x and y are the coordinate of the respective cortical filters. As before, with retina receptive fields, the cortical filter support region was made space-variant by making the size of the support related to the local node density around the cortical filter (v_c). The response of the cortical filter centred on v_c was computed based on the responses of nodes in the imagevector within the cortical filters neighbourhood $\Omega(v_c)$. The response of the cortical filter O(c) centred on v_c is in the form of an imagevector. The number of elements in the output imagevector O could be different to that in the input imagevector R.

$$O(c) = \sum_{m=1}^{M} R(p_c(m)) \times F_c(m), c = 1..N . \quad (5)$$

F_c are the 1xM filter kernel coefficients over the neighbourhood $\Omega(v_c)$ for the cortical filter kernel on v_c. Whereas p_c is the 1xM of indices of elements in the imagevector R with which F_c's are multiplied in the local convolution operation. The filter coefficients F_c are calculated based on a particular filter support profile based on the spatial positions on the field-of-view of the elements p_c in the input imagevector. As the visual information in the input imagevector R may have previously been processed, the kernels coefficients F_c for a particular cortical filter (for example performing low pass filtering) have to reflect the prior processing in this hierarchical feature extraction process.

The ability to operate on visual information stored as imagevectors enables the cortical filter layers to perform image processing operations on the responses extracted by the self-organised retina receptive fields. It is no longer necessary to represent and deal with data as rectilinear frame arrays to perform image processing in a feature extraction vision hierarchy. The authors extracted multi-resolution low pass space-variant information using cortical filter layers with Gaussian support regions on non-uniform self-organised retinae with 4096, 1024 and 256 nodes creating an octave-separated space-variant multi-resolution Gaussian *retina pyramid*. Each layer processed the output of the immediately higher resolution layer and only the 8192 receptive field retina sampled the input image. A Laplacian of Gaussian retina pyramid layer that sampled immediately higher resolution Gaussian retina layers was also constructed to extract space-variant contrast information from the visual scene.

V. Interest Point Descriptor

An interest point descriptor was formed based on the local gradients around stable scale-space extrema. The granularity of the sampling of scale by the Laplacian of Gaussian retina pyramid was made finer [10] by using cortical filters with (five) different effective standard deviations on the same retina tessellation, extracting contrast information spanning the octave scale separation indicated in Fig 3. Scale-space extrema in Laplacian of Gaussian responses scale normalised by the standard deviation squared have found to be stable in visual scenes[9]. To detect scale-space extrema equitably across the scales in the retina pyramid the authors had to also normalise each (unique) Laplacian of Gaussian cortical filter on the non-uniform tessellation with its mean response to random stimuli. The authors detected discrete extrema locations on the non-uniform retina pyramid by finding cortical filters greater or less than all their neighbours in scale and space. The exact location of the extrema on a scale-space continuum near its discrete location is detected by fitting a quadratic across scale (solving for the extrema at $\sigma_{extrema}$) and then a bi-quadratic across space at $\sigma_{extrema}$, giving the accurate location of the extrema ($x_{extrema}$, $y_{extrema}$, $\sigma_{extrema}$). The localisation of the detected extrema was improved by removing extrema generated at edges in the scene using a corner detector on the spatial derivatives of the bi-quadratic.

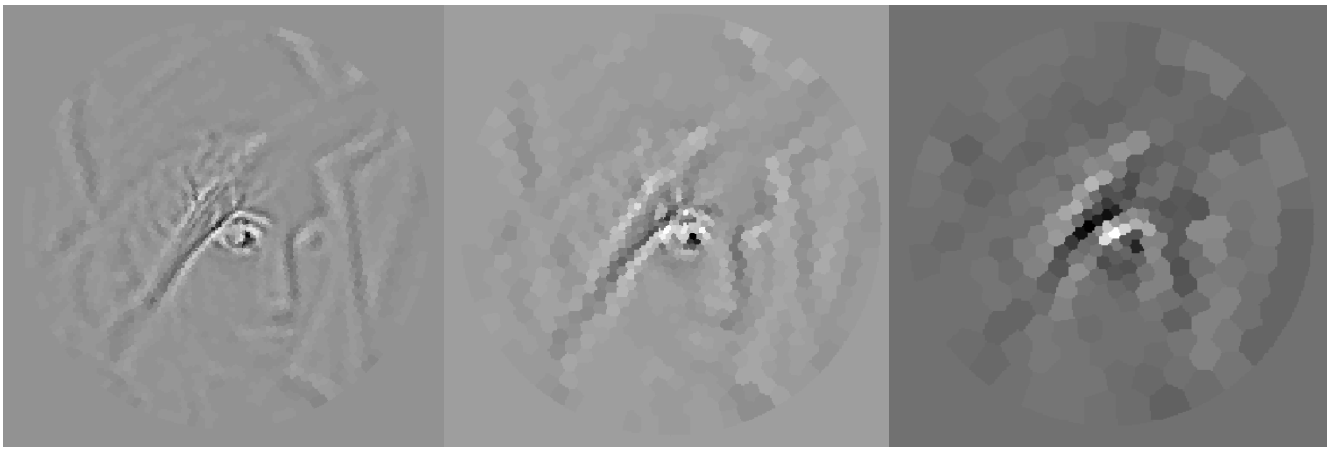

Fig. 3. Responses of Laplacian of Gaussian retina pyramid cortical filters with (from left to right) 4196, 1024 and 256 nodes, fixated upon the centre of the standard greyscale Lena image. Responses were visualised as before using the cortical filter's Voronoi region.

The detected space-variant Laplacian of Gaussian scale–space extrema were used as locations for interest points. The authors defined the support of an interest point based on adjacency in the associated non-uniform retina layer tessellation's Delaunay triangulation. Nodes with a graph geodesic less than or equal to j away from the interest point's associated discrete extrema were considered within the interest point's space-variant support. These support nodes will be denoted as $v_N \in \mathbb{N}_j(v_c)$ where $\mathbb{N}_j(v_c)$ is the set of nodes within graph geodesic j of v_c. A value of j=4 was used in this paper. The responses of the cortical filters v_N at the scale $\sigma_{extrema}$ were determined by solving for contrast information on a the scale-space continuum as before giving $L(I,v_N, \sigma_{extrema})$. Local gradient vectors were computed by also finding $L(I,v_i, \sigma_{extrema})$ were $v_i \in \mathbb{N}_1(v_N)$. Therefore the vertical and horizontal components of the local gradient vectors on the non-uniform retina pyramid around v_c is given below

$$O_x(v_N, I, v_c, \sigma_{extrema}) = \sum_i (x(v_i) - x(v_N)) \frac{L_{norm}(I, v_i, \sigma_{extrema}) - L_{norm}(I, v_N, \sigma_{extrema})}{(x(v_i) - x(v_N))^2 + (y(v_i) - y(v_N))^2} \quad (6)$$

$$O_y(v_N, I, v_c, \sigma_{extrema}) = \sum_i (y(v_i) - y(v_N)) \frac{L_{norm}(I, v_i, \sigma_{extrema}) - L_{norm}(I, v_N, \sigma_{extrema})}{(x(v_i) - x(v_N))^2 + (y(v_i) - y(v_N))^2} \quad (7)$$

where $x(v_i)$ and $y(v_i)$ are functions that return the spatial coordinate of node v_i. The magnitude and orientation of the local gradient at v_N are as follows

$$O_{mag}(v_N, I, v_c, \sigma_{extrema}) = \sqrt{O_x(v_N, I, v_c, \sigma_{extrema})^2 + O_y(v_N, I, v_c, \sigma_{extrema})^2} \quad (8)$$

$$O_{angle}(v_N, I, v_c, \sigma_{extrema}) = \tan^{-1} \frac{O_y(v_N, I, v_c, \sigma_{extrema})}{O_x(v_N, I, v_c, \sigma_{extrema})} \quad (9)$$

The described computations for achromatic local gradients O_x and O_Y were extended to chromatic gradients (simultaneously extracting spatial and chromatic contrast information) by sampling $L(I,v_i, \sigma_{extrema})$ and $L(I,v_N, \sigma_{extrema})$ from separate chromatic contrast channels.

Local orientation gradients at v_N , were aggregated to form the interest point descriptor using a Gaussian weighting to reduce the influence of local gradient vectors at the extremes of the interest point's support. The spatial standard deviation $\psi(v_c)$ of the Gaussian was based on the size of co-located retina receptive fields on the retina tessellation on the Gaussian retina pyramid resulting in a space-variant standard deviation. As the scale-space extrema is detected on a continuous scale-space, receptive field size is modulated with the offset of the scale extrema on the octave on the retina pyramid to determine $\psi(v_c)$ accurately.

A. Orientation histograms

An orientation histogram for the local descriptor was obtained by binning the local gradient vectors $O(v_N)$ from over a discrete set of (eight) orientations θ separated by $\Delta\theta$ (45°). Redundant representation of local gradient vector components was prevented by only binning the positive cosine component of a local gradient vector. Therefore the descriptor orientation histogram at interest point ($x_{extrema}$, $y_{extrema}$, $\sigma_{extrema}$) is as follows

$$H(I, x_{extrema}, y_{extrema}, \sigma_{extrema}, \Psi(v_c), \theta) = \sum_N G(v_N, x_{extrema}, y_{extrema}, \Psi(v_c)) \times O_{mag}(v_N, I, v_c, \sigma_{extrema}) \times \cos(O_{angle}(v_N, I, v_c, \sigma_{extrema}) - \theta), \cos(O_{angle}(v_N) - \theta) \geq 0 \quad (10)$$

The canonical orientation(s) θ_{peak} of the descriptor was found by computing the peaks over the discrete orientations in the descriptor orientation histogram $H(I, x_{extrema}, y_{extrema}, \sigma_{extrema}, \psi(v_c), \theta)$. Separate interest point descriptors (orientated differently) were created for each canonical orientation θ_{peak}. The exact canonical orientation of each descriptor was determined by fitting the quadratic over θ at $\theta_{peak} \pm \Delta\theta$.

Based on the canonical orientation and support region of the descriptor it is possible to divide the descriptor into sub-regions to increase its spatial representation acuity. Nine sub-regions in a rectilinear grid separated by $k\psi(v_c)$ and rotated with the canonically orientation of the descriptor. A value of $k = 0.4$ was chosen for experiments in this paper. If x_{bin} and y_{bin} are the x and y coordinates of the descriptor sub-region *bin*, local gradients were aggregated to form the descriptor sub-region's orientation histogram H_{bin} using a Gaussian centred at each sub-region centre with a standard deviation of $k\psi(v_c)$. As previously, only the positive cosine orientation components of the local gradient vector are binned into the orientation histogram.

$$H_{bin}(I, x_{extrema}, y_{extrema}, \sigma_{extrema}, \Psi(v_c), \theta_{canonical}, x_{bin}, y_{bin}, \theta) = \sum_N G(v_N, x_{bin}, y_{bin}, k\Psi(v_c)) \times G(v_N, x_{extrema}, y_{extrema}, \Psi(v_c)) \times O_{mag}(v_N, I, v_c, \sigma_{extrema}) \times \cos(O_{angle}(v_N, I, v_c, \sigma_{extrema}) - \theta) \quad (11)$$

The interest point descriptor is formed by concatenating all sub-region histograms and normalising to unity magnitude. These features together with the spatial location of the interest point ($x_{extrema}$, $y_{extrema}$), canonical angle $\theta_{canonical}$ and support region size (i.e. scale) $\psi(v_c)$ are used to represent visual content based on the non-uniform feature extraction.

VI. Object-based Saccade Generation

A. Interest point matching

By extracting interest point descriptors using the space-variant retina pyramid it is possible to match unknown descriptors from a visual scene to known descriptors extracted from labelled, known visual content. The χ^2 distance was used as a distance metric to match descriptors and the log likelihood ratio statistic was used to test the match between the descriptors. Based on the log-likelihood ratio it is possible to give a confidence to the match between pairs of interest points, thereby assigning an object label to previously unknown descriptors.

The Hough transform [10, 16] is also able to assign an object scale and pose in the image based on the spatial arrangement of the unknown and matched known interest points. Matches between test and training interest point descriptors are used as evidence that votes into a discrete Hough accumulator space. The problem may be formulated as: if during training descriptor H_{train} was found at location x_{train}, y_{train} at scale $\psi_{train}(v_c)$ and with canonical angle θ_{train}, what rotation R, scaling S and translation T is consistent with finding descriptor H_{test} (which was matched with H_{train}) at location x_{test}, y_{test} at scale $\psi_{test}(v_c)$ and canonical angle θ_{test} in the scene? The authors used the log-likelihood ration statistic between training and test descriptors as the vote into Hough space. The Hough transform was able remove outlier interest point descriptor matches that

were not constant with a stable object hypothesis in the test image as only stable hypotheses will generate a high votes.

As the quantisation of the Hough space is coarse, an affine transformation of the interest point matches which contributed to the stable object hypothesis is used to determine an accurate pose hypothesis for the object in the unknown scene. The confidence of the match between interest point descriptors is used to bias the Gaussian elimination when solving the system for the affine transformation parameters m_1, m_2, m_3, m_4 and t_x, t_y. A system with a single pair of matched interest points is given below.

$$\begin{bmatrix} p(H_{train}\mid H_{test})x_{test} \\ p(H_{train}\mid H_{test})y_{test} \\ \vdots \end{bmatrix} = \begin{bmatrix} p(H_{train}\mid H_{test})x_{train} & p(H_{train}\mid H_{test})y_{train} & & \\ 0 & 0 & & \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & p(H_{train}\mid H_{test}) & 0 \\ p(H_{train}\mid H_{test})x_{train} & p(H_{train}\mid H_{test})y_{train} & 0 & p(H_{train}\mid H_{test}) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} \quad (12)$$

B. Object-based Saccade Generation

In an object search task, top-down saliency information about object position and pose can be used to generate the next fixation location. The authors performed object based saccade generation using two mechanisms. Mechanism I, used the translation parameter from the affine transformation in Equation 12 to generate in a hypothesis location (x_{obj}, y_{obj}) for the object in the scene based on the current fixation. A saliency map S for the scene specific to the object search task is constructed where c_{obj} is the confidence of the object hypothesis.

$$S\left(round\left(x_{obj}\right), round\left(y_{obj}\right)\right) := S\left(round\left(x_{obj}\right), round\left(y_{obj}\right)\right) + c_{obj} \quad (13)$$

Saccadic targeting to a next fixation at the maxima location on the saliency map will result in the system actively searching for the target object based on its current hypothesis. The retina is prevented from revisiting visual content previously sampled with the high resolution fovea by an inhibition-of-return map which suppressed the saliency map with a uniform circular region with size of the retina's fovea.

The location of (x_{obj}, y_{obj}) on the known test image will correspond to the hypothesised *centre* of the target object. If (x_{obj}, y_{obj}) is suppressed by the inhibition-of-return mechanism, the authors generated saliency cued using Mechanism II, based on the top-down spatial locations of *expected* target object features in the unknown scene. These expected feature locations ($x_{expected}$, $y_{expected}$) are generated by transforming *all* the training interest points (x_{train}, y_{train}) using the solved affine transformation parameters (Equation 12).

$$\begin{bmatrix} x_{expected} \\ y_{expected} \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x_{train} \\ y_{train} \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \quad (14)$$

The saliency information from the expected training object feature location is temporarily aggregated into the saliency map based on the spatial support of the training interest point $\psi(v_c)$.

$$S\left(round\left(x_{expected}\right), round\left(y_{expected}\right)\right) := S\left(round\left(x_{expected}\right), round\left(y_{expected}\right)\right) + \psi(v_c) \quad (15)$$

The following demonstration of saccadic exploration (Figure 5) for object based visual search uses the saliency mechanisms I and II to generate top-down attention based saccadic behaviour. The space-variant system was presented with the target scene with multiple objects taken from the SOIL database [17] and given the task of finding the Ovaltine object which the space-variant vision system observed previously using bottom-up attention based saccade generation. The graph in Figure 4 shows convergence of the Ovaltine object's hypothesised pose parameters $m_1..m_4$ from Equation 12 to the ground truth with the object-based saccadic exploration of the scene.

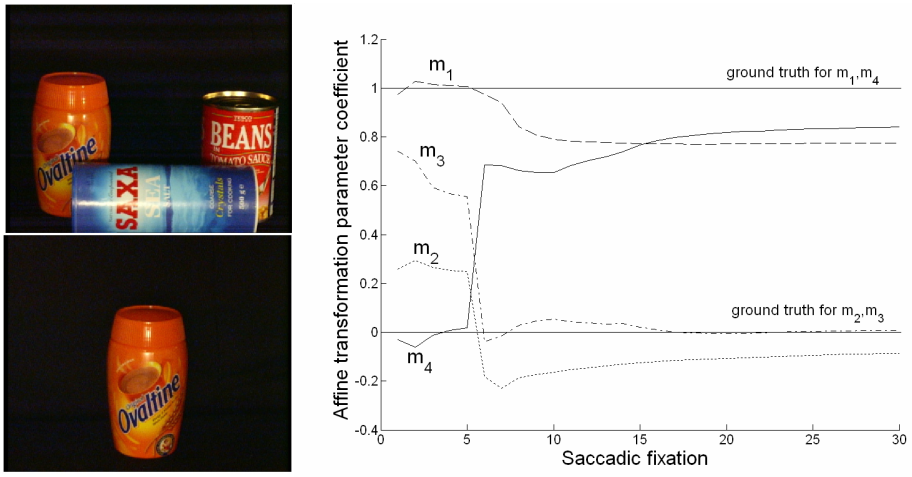

Fig. 4. (Left) Composite scene and training appearance of target object. (Right) Convergence of target object's hypothesised pose parameters to the ground truth with the saccadic exploration of the composite image. The images have been captured under real-world conditions with lighting differences, occlusions and pose variations occurring between the instance of the Ovaltine object in the training appearance view and the test composite scene. The retina pyramid was initially fixated upon the centre of the image and the high-resolution foveal region of the self-organised space-variant retina is only the size of the X on the SAXA salt object.

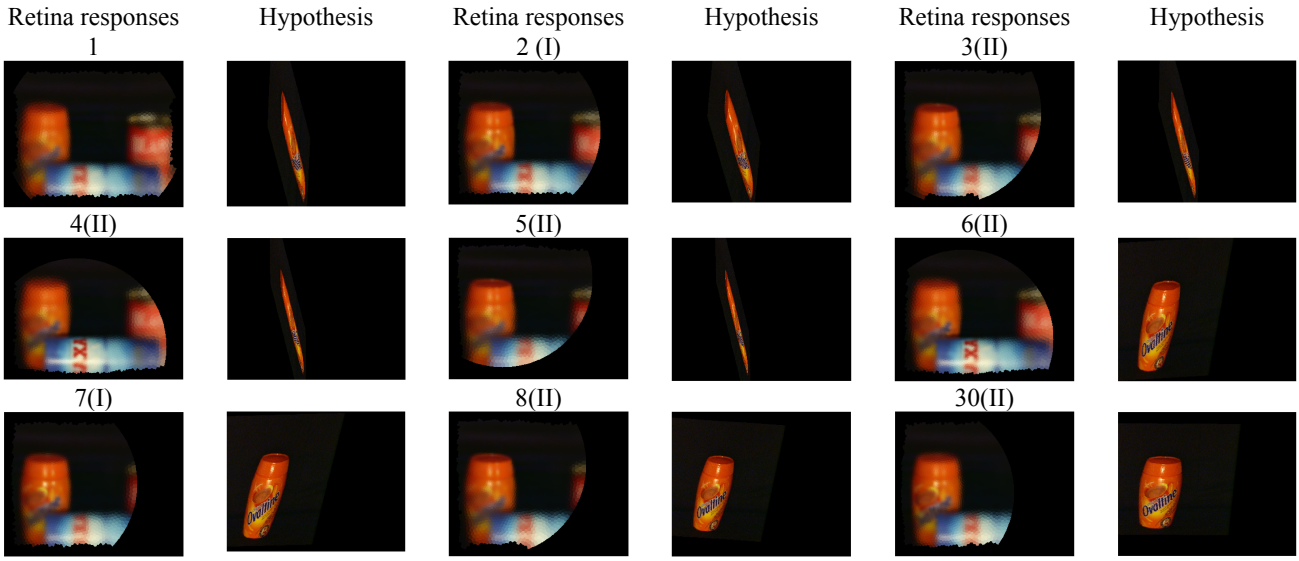

Fig. 5. Visual search for the Ovaltine target object. The responses of the Gaussian retina receptive fields and a rendering of the vision system's hypothesised target object location in the scene. The saccadic fixation number and the saccade generation mechanism that spawned the saccade are listed above each instance. A type I saccade tends to occur when there is a large change in the hypothesised object location (to an unattended region in the scene). Type II saccades explore the spatial locations in the scene where parts of the target object are *expected* to be found.

VII. Discussion

The visual machinery implemented in this paper enables the extraction of interest point descriptors generated from any arbitrary sampling tessellation. The authors showed that interest points could be used to construct object hypotheses based on visual information extracted from the low-resolution periphery of a self-organised retina. Object based saccades would locate the target object appearance and explore areas in the visual scene where parts of the target object are expected to occur based on the hypothesised pose of the object. We believe this to be the first space-variant vision system reported in the literature that is capable of fully automated top-down high-level object hypothesis based saccadic behaviour using interest points from a single unmodified appearance based target.

Acknowledgments

This work is supported by the European Union 6th Framework Program project "Integrated Project Research Area CINE" Project ref: IST-2-511316-IP.

References

1. Schwartz, E.L., "Spatial mapping in primate sensory projection: Analytic structure and relevance to perception", Biological Cybernetics, 25: pp. 181-194, 1977.
2. Schwartz, E.L., "Computational Anatomy and functional architecture of the striate cortex", Vision Research, 20: pp. 645-669, 1980.
3. Johnston, A., "The geometry of the topographic map in striate cortex", Vision Research, 29: pp. 1493-1500, 1989.
4. Bolduc, M. and Levine, M.D., "A real-time foveated sensor with overlapping receptive fields", RealTime Imaging, 1996.
5. Gomes, H., Model Learning in Iconic Vision. PhD Thesis, University of Edinburgh. 2002.
6. Balasuriya, L.S. and Siebert, J.P., "A Biologically Inspired Computational Vision Front-end based on a Self-Organised Pseudo-Randomly Tessellated Artificial Retina", IJCNN, 2005.
7. Clippingdale, S. and Wilson, R., "Self-similar Neural Networks Based on a Kohonen Learning Rule", Neural Networks, 9(5): pp. 747-763, 1996.
8. Schmid, C. and Mohr, R., "Local Grayvalue Invariants for Image Retrieval", PAMI, 19(5): pp. 530-535, 1997.
9. Mikolajczyk, K., Detection of local features invariant to affine transformations, PhD Thesis, Institute National Polytechnique de Grenoble, France, 2002.
10. Lowe, D., "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60(2): pp. 91-110, 2004.
11. Treisman, A. and Gelade, G., "A feature integration theory of attention", Cognitive Psychology, 12: pp. 97-136, 1980.
12. 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, 20(11): pp. 1254-1259, 1998.
13. Swain, M.J., Kahn, R.E., and Ballard, D.H., "Low Resolution Cues For Guiding Saccadic Eye Movements", CVPR, 1992.
14. Rao, R.P.N., "Top-Down Gaze Targeting for Space-Variant Active Vision", ARPA, 1994.
15. Sun, Y., Object-based visual attention and attention-driven saccadic eye movements for machine vision. PhD Thesis Thesis, University of Edinburgh: Edinburgh. 2003.
16. Ballard, D.H., "Generalizing the Hough transform to detect arbitrary patterns", Pattern Recognition, 13(2): pp. 111-122, 1981.
17. Koubaroulis, D., Matas, J., and Kittler, J., "Evaluating colour object recognition algorithms using the SOIL-47 database", Asian Federation of Computer Vision Societies, 2002.