A model of modes of attention and inattention for artificial perception

Mark Witkowski and David Randell
Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, U.K.

{m.witkowski, d.randell}@imperial.ac.uk

Abstract. This paper considers several aspects of natural visual attention and its link to wider notions of awareness, natural and artificial, in the context of foveated vision. It builds on a theory of abductive perception; a formal definition for an artificial or robot perceptual system, using objects represented as feature clouds. It proposes a broad, but unifying approach to several aspects of visual attention in the light of this, including autonomic eye gaze movements, aspects of secondary and covert attention, and exogenous (sense driven) and endogenous (task driven) attention. Modes of attentional lapse, commonly referred to as inattentional blindness and change blindness, are also discussed in the context of the model presented.

1 Introduction

This paper addresses the task of providing the groundwork for a logically formulated and computationally motivated model of foveal-based attention in visual perception. Notions of sensory attention, preferential emphasis on, or selection of, one or a few items of sensory information from a much larger stream of data, appear ubiquitously across the animal kingdom (Bushnell 1995, Zentall and Riley 2000) and it takes many forms in human perception. While much has already been written on the subject of human visual attention, see e.g. Duncan (2006), Itti and Koch (2001), Pashler (1998) or Posner and Petersen (1990) for general reviews across several discipline boundaries; few have sought to embed attention within a formal framework.

Specifically, we develop a unifying approach to perception and visual attention, extending prior work in *abductive perception* (Shanahan 2002, Shanahan and Randell 2004, Randell and Witkowski 2006). We address several aspects of human visual attention in the light of this work, including generation of autonomic scanpaths of eye gaze movements, aspects of secondary and covert attention, and of exogenous (sense driven) and endogenous (task driven) attention. We also consider modes of inattention: change blindness and inattentional blindness; situations were there are apparently surprising lapses of attention.

This paper is primarily concerned with the role of semantically meaningful *objects* in the generation of attention strategies. The notion of object-based attention has arisen in the literature (Duncan 1984, Kahneman *et al* 1992, Pylyshyn 2001, Scholl 2001). Scholl (2001) provides a detailed discussion of the issues involved, highlighting the difficulties in defining what should constitute an object relative to attention and raising the question as to whether notions of objects are formed pre-attentively or post-attentively. We begin with an *a-priori* notion of "object" represented as a feature cloud and argue from the model developed that visual attributes of such objects are processed pre-attentively and that attention is therefore a *late-selection* phenomenon (e.g. Pashler 1998). We also make a strong *assumption of embodiment*, that objects have volume and occupy physical space about the viewer. The world about us is naturally and unavoidably perceived in depth (where flat or iconic "objects" must similarly be projected to a physical surface).

In order to address issues of attention we develop a notion of attentive *preference* across objects, defining the significance or salience of those objects to the perceiver. We also

consider the role of an object memory structure (which we refer to as the *panorama*) and its role in *inhibition of return* (Taylor and Klein 1998) and change detection (Rensink 2002).

We will initially consider shifts of overt visual attention. This is indicated by autonomic eye-gaze movements, referred to here as the *scanpath* (Stark and Choi 1996). We take the view that objects may be accurately identified and categorized only when located within the central foveal regions of the retinal imager, whereas, due to the reduced resolution, only partial identification is possible in the visual periphery. We assume that this estimation in the periphery suffices to provide a "gist" (Rensink 2000) or characterisation of the overall scene, but that more detailed (peri-foveal) inspection is required to confirm the general identity of an object. Foveal inspection is also required to perform detailed tasks, such as reading normal sized text. We will also consider notions of secondary attention, where the observer is apparently cognitively aware of several visual items at once, but which will not necessarily be visited by the gaze-path. We also briefly consider the role of volitional gaze movements, *secondary attention* and *covert attention* (Horowitz *et al* 2006).

In part visual attention is task or activity dependent, controlled by higher level cognitive processes that select what is to be viewed preferentially according to the current role (endogenous attention), for instance, as indicated by visual search tasks (e.g. Treisman and Gelade, 1980, Wolfe, 1994). It is also well established and within the human common experience that areas of high spatial contrast, colour saturation, and certain types of movement in the peripheral field will attract both gaze and attention (exogenous attention). Models of exogenous attention have been proposed by e.g. Itti *et al* (1998) and endogenous attention by Stark and Choi (1996). Hochstein and Ahissar (2002) present a "reverse hierarchical model" incorporating both aspects.

Itti *et al* (1998) present a *saliency-based* model of exogenous (sense-driven) attention, which uniformly applies multi-scale feature extraction (intensity, edge orientation, colour, etc.) to an image. This is used to build a combined feature map. From this mapping various feature combinations are assigned "salience" according to the application under consideration. A process of competition is applied to isolate places of maximal saliency and these points are then visited in saliency order to emulate the process of visual attention. Once attended to, a temporary process of inhibition of return (IOR) suppresses the salience of the last attended location, so that the next most salient place is visited and so on.

Models of task-driven processes of attention are less well represented and are generally incomplete. Stark and Choi (1996) present a model that emulates the eye-gaze path of a simulated human observer, using a Markov state model based approach. Navalpakkam and Itti (2002) extend the Itti and Koch model with task based relevance. Breazeal *et al* (2000) suppress or intensify attributes of the attentive feature map to reflect changing "social" drives, emulating some aspects of cognitive control.

More general models of human visual and attention processes are represented by Rensink's (2000) *coherence theory*, and Pylyshyn and Storm's (1988) notion of FINSTs (*fingers of instantiation*), based on findings that humans can track about five visual objects, apparently covertly. Equally, more specific aspects of the attention process have been modelled also, such as the *feature integration theory* of Treisman & Gelade (1980) seeking to explain aspects of visual search. It is notable however, that while there is considerable evidence that many modes of attention are present there is little consensus as to which, or indeed whether some combination, provides the better overall explanation of the range of attention phenomena that may be postulated. The approach adopted here takes a broad view of attention, postulating mechanisms that impinge on many aspects of the overall problem.

This paper also considers apparent lapses of "attention", when seemingly highly significant or unusual objects or events in full view appear to be completely overlooked, where they would normally be drawn into explicit awareness by the visual attention system. Such lapses have been known for many years, and have often been ascribed to avoidable fault on the part of the observer. Such issues are, of course, of particular significance where potentially dangerous, possibly everyday, activities are being performed, such as driving a car, riding a motorcycle or flying an aircraft.

More recent research has highlighted the fact that these lapses of perceptual attention are highly repeatable phenomena, and depend on the circumstances the observer is in. They can be placed into (at least) two distinct categories: Inattentional Blindness (Mack and Rock 1998, Simons and Chabris 1999, Most *et al* 2000) and Change Blindness (Levin and Simons 1997, Simons and Levin 1998, Simons *et al* 2000; Rensink 2002 for review).

Our interest arises from the use of formally defined models of perception and attention models in a *cognitive robotics* context, applying observations about human and natural perception to robotics within a formal framework. The formal model of perception described in this paper builds on and adapts the feature cloud representation of visual objects of Randell and Witkowski (2006), itself an extension of a scheme of abductive perception proposed by Shanahan (2002) and developed in Shanahan and Randell (2004). Both use a logical language expressed in first-order predicate logic to describe the physical objects in a robot's world and specifically the consequence of the world interacting with the robot's sensors. The primary purpose of this paper is not to extend the formalization previously given, but rather to apply the formal model to investigate and discuss aspects of attention. We retain elements of the notation to illustrate the approach but rely on textual descriptions whenever possible. Several aspects of the description call on vision processing techniques properly the domain of computer vision and we identify these where appropriate. Other robot motivated models of attention in perception have been proposed by, for example, Aziz *et al* (2006), Breazeal *et al* (2000), Khadhouri and Demiris (2005), and Vieira Neto and Nehmzow 2005.

In our model, objects in the world – and their sub-parts – are encapsulated in a hierarchical (qualitative) symbolic description, coupled to a (quantitative) vector based representation embedded within that symbolic description. Vectors determine the relative positions of every part of the object. Lower nodes in this hierarchy, *features*, are matched directly with the visual characteristics of the object or surface at a place in the visual field. Features form the link between the symbolic object description and the sensing mechanism. It is recognized that the characteristics at a single place on the surface of an object (giving rise to the feature description when impinging on the sensor) may vary according to viewing conditions, notably the distance and angle at which the feature is viewed. Each feature is further mapped to a (non-empty) set of terminal *appearance* nodes, each mapping directly to a specific sensor *detector* state. Detectors are tuned to respond to specific patterns in the visual field. The vectors embedded in each of the appearance node descriptions provides an estimate of the relative pose of the feature, and so may be used to infer the approximate pose of the whole object relative to the current viewpoint.

This paper assumes a foveal visual system, figure 1 (left). In the peripheral area, features are identified with a low spatial resolution (equating to low spatial frequency distribution). Only within the area of the fovea are features detected with a high resolution. Some angular portion of the visual field is given over to the central foveal area, denoted *r* in figure 1, equivalent to the singer figurine's face (*r* is approximately 15º in the human visual field, but may be set arbitrarily in the model). In addition, the human eye has a smaller area (about 1.5º) of greatly increased resolution, the *fovea centralis* (*r'* in figure 1) One task of any attention mechanism is to direct the gaze so that the foveal area is centred on the area of current interest. This has been likened to a spotlight, preferentially "illuminating" areas of heightened interest (Itti and Koch, 2001). Eye-gaze movements may therefore be seen as a strong indicator of both the choice and ordering of targets selected for visual attention when viewing a scene. Equally, this selection should not be directly equated with notions of visual or attentional *awareness*; a person may look directly at a target object apparently without it being registered by or materially affecting any cognitive process.

On eye-gaze scanpaths

Human eye-gaze is not a smooth track, but normally proceeds in a series of rapid eyeball motions, *saccades*, interspersed with periods of relative stability, *fixations*[1]. The motion of the eye during each saccade is essentially ballistic, a planned movement to a pre-determined place, which can achieve a rotational speed of 600º/sec. or more (Becker 1991). Saccades are also accompanied by rotation of the head if they would exceed 15-20º. Fixations typically last 200-400ms. It is generally considered that perception only occurs during fixations and is suppressed during each saccade (as well as during blinks) to give an illusion of constant vision (but see Brockmole *et al* 2002 for a detailed discussion). Eye-gaze motions are easily measured (Duchowski 2003). Figure 1 (right) shows a human eye-gaze trace for an example image (23 seconds). Fixations are shown with an 'F'. The gaze trace in figure 1 (right) was recorded with our LC Technologies, Inc. (www.eyegaze.com) eye-gaze tracking system.

Figure 1: Simulated foveal area (top-left); human eye-gaze path (top-right and below)

Autonomic scanpaths display a characteristic, if variable, pattern (Hayhoe and Ballard 2005, Henderson 2003, Just and Carpenter 1976, Kowler *et al* 1995, Noton and Stark 1971, Rao *et al* 1997, Stark and Choi 1996, Torralba *et al* 2006). When presented with a variety of static image types, the visual search performed is highly selective on the image surface and gaze scans around a number of selected locations in rapid succession (e.g. Yarbus 1967), for example, figure 1, bottom left, 0.0-4.1 seconds. There is no systematic search over the entire image field (unless the observer is engaged in some task that requires this, such as reading text). Instead, a limited number of selected, apparently salient, places of the image are visited in turn. These locations most likely correspond to objects of significance within the image (we return to notions of significance later).

Then, typically, the path is repeated, often in the same or similar order (figure 1, bottom, middle, 4.1-10.9 seconds. As the viewing progresses these places are re-visited frequently, though new areas of interest may be introduced. There is no discernable overall tendency to minimize the scanpath, long saccades being interspersed with short ones. During the scanpath period, some areas, particularly faces, are selected for detailed inspection (figure 1, bottom right, detail, 5.9-8.9s). It has been long established that the overall pattern of the scanpath can be substantially modified according to tasks set the observer (Yarbus 1967) and that the

[1] There are other modes of eye movement as well, such as pursuit, in which the gaze follows a moving target and nystagmus, which compensates for observer motion (e.g. Tatler and Wade 2002).

modification is seemingly instantaneous once the task is established (but see Hodgson *et al* 2004 for a detailed investigation). Sun and Fisher (2003) present a scanpath emulation based on colour segmentation at multiple levels of resolution; Rao *et al* (1997) emulate scanpath search strategies; Lee and Yu (1999) take an information theoretic approach; Torralba *et al* (2006) consider how object context and location affect the scanpath.

Visual attention, in this limited sense, is largely continuous – the gaze shifts continue throughout periods of normal vision. We take the eye gaze scanpath only as an indicator of visual attention (see also Hoffman 1998). Visual attention is not, however, to be taken as synonymous with cognition or awareness – it is not an end in itself. It is clear that visual attention impinges on and modifies awareness; equally it is clear that cognitive processes can prime and direct the visual attention mechanism according to tasks. Within the visual field a number of other, non-foveal, items may also be brought into partial cognitive awareness (covert or fringe awareness) and these may in turn become candidates for subsequent foveal visits (Horowitz *et al* 2006).

It should also be noted that the visual (and sensory) component of attention accounts for only a proportion of human awareness, which is directed to different aspects of mental activity throughout the day on a continuing basis (James 1890, Hurlburt *et al* 2002).

Paper outline

Section 2 describes the background to and the principles behind the theory of abductive perception, which forms the underlying rationale for the pre-attentive perceptual process. Section 3 provides a formal definition and accompanying description of the feature cloud volumetric object representation, which bridges detectable sensor items (features and their appearances) to conceptual object types. Section 4 considers the necessary properties for feature (sensor) detectors within the formalism. Section 5 describes the essential properties of the panorama structure, which records object information in a robot-centric representation, acting as a (combined) transient visual buffer and conceptual memory, managing, among other things, the information required for inhibition of return. Section 6 describes the pre-attentive perceptual process, matching sensor items to form object hypotheses. Section 7 describes the attentional process model, covering the use of preference ordering to give and maintain attentive salience over objects and the role of salience partitions (7.1), the underlying scanpath cycle, incorporating the perceptual cycle (7.2) and the management of preferences (7.3). Section 8 discusses each of the identified attention modes in terms of attentional preference ordering and partitioning, inhibition of return, detector input and task relevance. Section 9 considers how these aspects contribute to inattentional phenomena. Appendix 1 gives a formal definition of preference and appendix 2 provides definitions of the three measures that regulate the pre-attentive perceptual process.

2 Theory of Abductive Perception

The treatment of perception and attention developed here exploits a mode of inference known as abduction. This works from a set of observations and generates a set of alternative conjectures, that if true would best explain the observations. As expected, and implied here, the set of alternative explanations generated are rarely unique, an additional framework is needed to interrogate and test these conjectures and select those that best explain the available evidence.

To this end we adopt a hypothetico-deductive model, where, when given a set of possible explanatory hypotheses via abduction, the set of predictions that follow are then used to prune the hypothesis space, according to how well the predictions are confirmed (or refuted) when compared to the original sensor data. The abductive perception approach models visual perception as a combination of bottom-up, sense driven, and top-down, expectation driven processes. As such it accords well with a substantial body of empirical and experimental data from visual psychophysics (e.g. Rock 1981) and from neurophysiological evidence of a two-way flow of information (e.g. Hochstein and Ahissar 2002, Lee and Mumford 2003) and

whose roots may be traced to the mid-19th century (Helmholtz 1865). The approach is also related to that of active perception (Aloimonos *et al* 1987, Ballard 1991).

The formal model assumed here is an extension of that proposed by Shanahan (2002). Here we use a logical language expressed in first-order predicate logic to describe the visual world. The language is used to construct sets of sentences that describe the given background theory (Σ), the interpreted sensor data (Γ), and the set of abduced hypotheses that, if true, explain that sensor data (Δ). Formally, this is woven together and represented by the logical schema: $\Sigma \cup \Delta \models \Gamma$, which states that Γ is a logical consequence of Σ and Δ. Hence, given Γ and Σ, we construct Δ (by abduction), and then use deduction to test the consequences of those hypotheses (using the hypothetico-deductive model). Within the logic, all sentences take the form of *well-formed formulae* (*wffs*), sequences of symbols adhering to the formation rules (syntax) of the language and where Greek upper-case letters are used to indicate sets of sentences (e.g. Σ, Δ and Γ).

We define and factor out particular subsets of sentences in Σ and Δ as follows. First Σ (being the background theory) is separated out into (i) the generic descriptions of objects (encoded as feature clouds) which we call Σo, and (ii) sets of constraints, such as those embodying various commonsense properties of the world (Σc). Amongst these are the "commonsense" beliefs that two volumetric bodies may abut but not overlap (i.e. they may not share a volume in common) and that an opaque body occludes from view anything directly behind it from a viewpoint. Similarly, Δ is divided into (i) the set of alternative interpretations of the world (Δh), (ii) the currently single preferred explanation (Δi), and (iii) Δp, which is the current robot-centric description of the world.

In order to measure how well generated hypotheses explain the data, a set of numerical measures are introduced. These include a *distinctiveness value* (*dv*) that measures the rarity value of individual features encoded in Σ, an *explanatory value* (*ev*) that measures how well a hypothesized object of a given predicted position and pose matches the sensor data, and a *rank ordering* (*ro*) that ranks the likelihood that a particular object type explains the sensor data. These measures are defined formally in appendix 2.

The feature cloud model encapsulates both symbolic and numerical information. The former allows us to exploit established symbolic automated reasoning methods (both for abduction and deduction) while the latter mirrors these operations in the application of linear transformations on sets of vectors encoded in our 3D model descriptions.

The new abductive strategy for processing sensor data is illustrated in overview in figure 2 and proceeds as follows. Firstly, detectors operating on the image plane (the *field of view*, FOV, denoted ABCD in the lower part of the figure) identify and extract features (from their appearances) and these populate the sensor data structure Γ. Next, distinctive elements of the sensor stream are matched to object descriptions in Σo to generate a set of candidate hypotheses in a hypothesized space Δh that may explain the sensor data (shown in the upper part of figure 2. Based on this partial evidence, each inferred host-object is hypothesized to occupy a specific place in the space about the robot. Then using deduction coupled with the manipulation of 3D linear-transforms (as prediction) the ramifications of this projection are expanded. The feature cloud is used to determine the expectations of other features associated with the hypothesized object(s) in question. The sensor data is then re-consulted to refine the explanatory value by determining the extent to which each hypothesis is supported by the observed data; these are represented by short left facing arrows in the upper part of the figure. On the basis of this, hypotheses with low explanatory value are rejected (and sensor data items they would have explained are released, though still in need of an explanation). The process is repeated until all the sensor data elements of Γ have a coherent explanation, and where all the ground hypotheses in Δ satisfy all the constraints applicable to the domain. This sequence of events is revisited in section 6.

A cognitive agent cannot deny or disprove sensor data, it can only interpret it in a manner consistent with the use it intends to make of the data. Where no coherent explanation can be found for the sensor data, the agent may: (i) reject or ignore the data as, for example, stemming from sensor noise, or arising from inconsistency with the assumed domain model

or (ii) update the existing domain model to accommodate the new previously unassimilated sensor data (Σo). The latter case both serves to extend the range of objects known to the agent and may be used as a cue to attend to the "novel" object. Similarly, a group of features interpreted as a known object, but with significant deviations from the model, may also be candidates for detailed inspection. Leading, in turn, to an updated model or differentiation of the model into two separate ones.

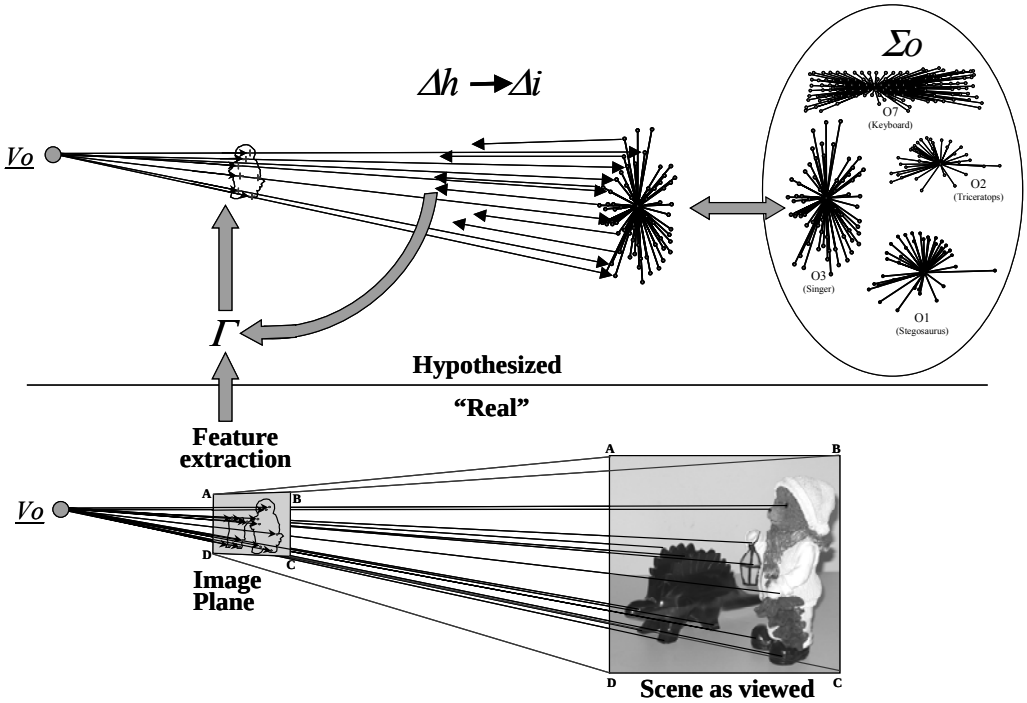

Figure 2: The main perceptual cycle

3 Feature Clouds

A feature cloud is a data structure that encodes a heterogeneous and spatially distributed set of sensor-detected features, where each feature is individually mapped to a *position vector* within a local coordinate frame. Feature clouds may be seen as a form of *annotated point cloud* in which the structure of the object is encoded as many points about a local coordinate frame, some (or all) of which are labelled with information about how that place on the surface of the object will impact on the observer's visual sensors (the features). Point cloud representations have found favour recently in both computer graphics and engineering (e.g. Linsen, 2001).

This point-based approach may be contrasted to other model-based representations in visual perception-based applications such as generalized cylinders (Marr 1982), superquadrics (Chella *et al* 2000) or geons (Pirri 2005) in which the shape of the modelled object must be deformed and fitted to the sensor data. Feature clouds directly encode the sensor (detector) related properties of the modelled object and are localisable (both in the 2D image plane and the 3D world space of the robot) to an arbitrary desired precision and do not rely on shape or surface matching.

The cloud is partitioned into subsets of features that are pre-assigned to an assumed set of volumetric regions. Such regions correspond to our informal notion of a physical object or its parts. The whole takes on the form of a hierarchically organized tree-structure of such regions, where the subdivision of a host object into sub-parts proceeds until all their named features and their viewpoint-dependent appearances eventually appear as terminal leaf-nodes.

Although described by sparse points all objects are considered to be bounded by opaque surfaces, represented as surface patches between the feature points[2].

Each feature cloud is represented by sets of *vector pencils;* where each pencil comprises a set of straight-line segments intersecting at a single point – the *centroid*. Figure 2 shows feature cloud visualizations (Σo) of number of exemplar objects. Pencils mapping features to their appearances are interpreted as *lines of sight* fanning out into space. As each viewpoint also acts as the origin of another vector pencil whose end-points potentially locate a set of features, the overall geometrical form of an object with its set of features and their viewpoint indexed manifold appearances can be likened to a stellated polyhedron with the vertices mapping to the view points.

Axiom A1 encodes the hierarchical decomposition of objects into object-parts, their features and appearances. *Type* is a class indentifier for each object, object-part, feature and appearance. In this definition, objects, object-parts and features can all have associated appearances[3]. This reflects the underlying foveal-based vision model assumed here, where at different levels of visual resolution an object may have no detectable (i.e. resolvable) object-parts or constituent features, though it can have an overall detectable appearance at any particular level of resolution.

(A1) $\Phi(x0,\underline{v0},Type0,[x1,...,xn]) \rightarrow$
$\Psi 1(x1,\underline{v1},Type1,[...]) \;\&...\&\; \Psi n(xn,\underline{vn},Typen,[...])$,
where:
if Φ*=Object, then* $\Psi i \in$ *{ObjectPart,Feature,Appearance} else*
if Φ*=ObjectPart, then* $\Psi i \in$ *{ObjectPart,Feature,Appearance} else*
if Φ*=Feature, then* $\Psi i \in$ *{Feature,Appearance}, else*
$\Phi = \Psi i$*=Appearance*

Axiom A1: Encoding the Feature Cloud

All elements (expressed as *wffs*) of this hierarchy uniformly take the form: Φ *(name,* $\underline{vector}$*, Type, Part-list)*, where Φ substitutes for *Object*, *ObjectPart*, *Feature* or *Appearance*. Each of these elements is uniquely identified by its *name* and is assigned a class *Type*. Vectors (always shown italic underlined, $\underline{v}$) locate the *centroid* (notional position) of any object, feature or appearance ($\underline{v_1} \ldots \underline{v_n}$) with respect to the centroid of its supervenient feature or object ($\underline{v_0}$), or of the observer's viewpoint (actual or notional) in space. Using this scheme the position of any sub-part of an object may be estimated relative to any other subpart by straightforward vector summation. Every logical operation between objects and their parts implies a corresponding vector operation.

The *Part-list* enumerates each of the component items (*ObjectPart*, *Feature* or *Appearance*) at the next level in the description hierarchy. Each appearance is directly associated with a detector is assigned a class *Type* and a $\underline{vector}$ estimating the pose and position of the appearance from the current viewpoint, which together serve to identify the physical source of each sensor data assertion within the system and an estimate of its position in 3-space relative to the current viewpoint. Appearances have no sub-parts. The hierarchical nature of this definitional form allows objects to be represented and reasoned with at multiple levels of detail, and at any arbitrary level of precision, using the embedded numerical vectors.

[2] Surfaces are not explicitly represented in the feature cloud, but are factored out and treated separately. We model object surfaces as a Deluanay triangulation of an arbitrary topology manifold surface in 3D space. Volumetric solids consequently take the form of polyhedra whose faces are ultimately decomposed into a finite set of triangles. See Randell and Witkowski (2006) for further discussion and formal definitions.

[3] Axiom A1 given here modifies and contrasts with the axiom (also designated "A1") and logical model adopted in (Randell and Witkowski 2006; Witkowski and Randell 2006), in which the hierarchical model restricts appearances to attaching to features. This effectively limited the previous definition to a uniform field (i.e. non-foveated) visual sensor.

4 Feature Detectors and Appearances

Central to the abductive process here is the deployment of "detectors", devices that make assertions into Γ when the specific conditions they are tuned to occur in the visual field of view (FOV). In the scenario we describe here, these detectors are assumed derived from low-level vision processing operations.

Figure 3 (left) illustrates the effect of appearances of features. A single feature (for instance, the corner of the larger cube) appears differently from alternate viewpoints, and each is assigned a separate appearance type. Large cones in figure 4 (left) arranged radially from features represent the range for a given series of detectors. Each represents a pencil of viewpoints from the feature centroid. They may share a single appearance, or be divided into multiple appearances for any given range of distances from the feature. These then represent volumetric lobes in the space about the object. Each such lobe gives rise to one appearance definition. The small cones indicate individual appearances of detected features (as projected onto a notional image plane) from specific viewpoints.

Note that it is the properties of the detector that determine the angular and positional range of the appearance type. Detectors (and hence appearances) may be designed to work over a wide range of viewpoints, or be highly specific to a small volume of space. The scope of detectors (and hence appearances) may overlap, may be ambiguous – in that distinct features may give rise to the same appearance. Features may not be detected at all from some viewpoints, either because the detectors are not sensitive to it, or by virtue of self-occlusion.

Figure 3 (right) illustrates the appearances for some aspects of the corner detectors when viewed in the peri-foveal and foveal regions of the visual field. Note that the three distinct appearances of the features become effectively indistinguishable at the periphery of the visual field.

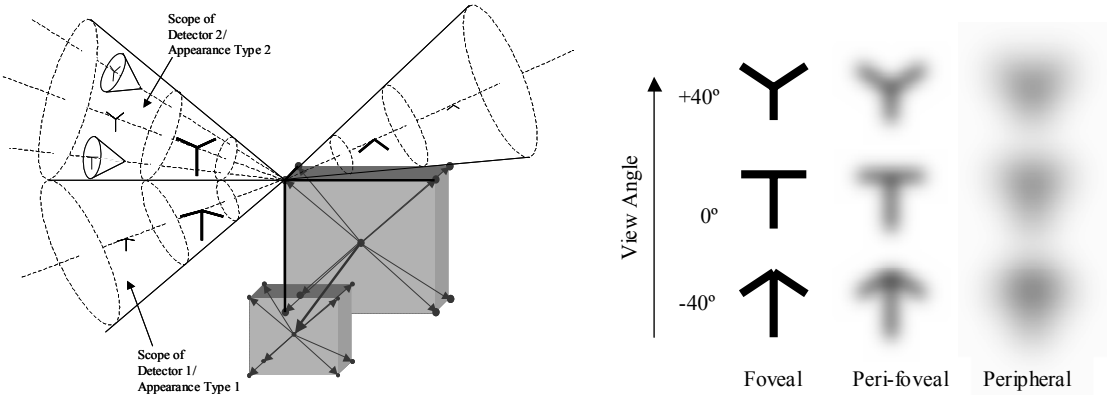

Figure 3: Left: Appearances of a feature (corner) from different viewpoints; right: Appearances of the feature in the foveal, peri-foveal and peripheral regions

We admit a wide range of detector types to be assimilated within the logical framework described. Such operation might include line finding or edge detection routines, or, more usefully, the detector will respond to some distinctive, though not necessarily a unique, property of the physical object being observed (see the work of Lowe 2004; and Schmid and Mohr 1997 for recent advances in this area). Lowe, in particular, has developed powerful feature encoding techniques that allows for the fast storage and retrieval of arbitrary complex patterns drawn from real images, building on earlier notions of *steerable filters* (Rao and Ballard 1995). Such detectors may function over several octaves of image resolution and be insensitive to feature rotation. There is no stipulation here of the form detectors should take beyond that they are representative of the objects, parts or features that cause them, that they may be assigned a type and may be localised in the FOV with a centroid to support the vector representation. They need not be invariant on rotation or scale. There is also no suggestion here that artificial and natural detectors need operate on directly comparable principles.

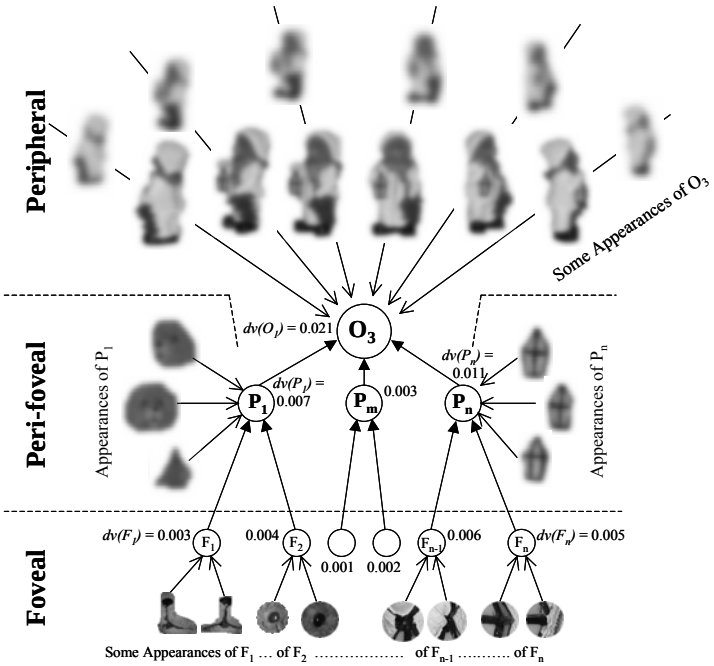

Figure 4: An Object ("O_3") hierarchy and its appearances at different levels of resolution

Recall that appearances are mapped directly (but non-uniquely) to the feature(s), parts and objects they indicate by abductive inference. In this respect, individual sensor data items may be said to afford (by analogy with Gibson, 1979) their own explanation through their description as appearances. Physical objects in the world give rise to different effects on the detectors under differing circumstances, such as when viewed from different distances, from varying angles, under different conditions or different areas of a foveal imager. This effect is illustrated in figure 4 for a single physical object at three notional levels of resolution. The top layer shows possible appearances at the peripheral resolution (shown here as Gaussian blurring, for illustration) from different viewpoints – six stages of rotation at two different distances, representing instances from a set of all possible detectable viewpoints. The peri-foveal layer shows a small number of sub-part appearances and the foveal layer a number of exemplar high-resolution detectors. Note that the detector images shown are illustrative only, the actual encoding of the detector is by type, a symbolic entity. Figure 4 also shows the effect of propagation of the *distinctiveness values* (*dv*) for each type up the hierarchy.

Note that some easily extracted feature types, such as line segments, are common to very many objects, and, as such, provide little discriminatory power. However, by virtue of the hierarchical definition of the feature cloud, simple features may be composited into *compound features*, which increases their discriminatory capacity.

Objects are not necessarily only characterized by static feature detectors. Many everyday (animate) objects are detected by distinctive motion profiles and may equally serve to trigger a hypothesis of the associated object type by abductive reasoning[4]. These play an essential part in notions of exogenous attention. The reliable detection of both change and recognition by motion represent substantive challenges to the computer vision community (e.g. Cédras and Shah 1995, Radke *et al* 2005). We will also consider the role of non-object based detectors, those that respond primarily to areas of high colour saturation, luminance or change.

5 The Panorama

The current interpretation, Δi exists solely in the present. When the detector stream (Γ) changes or is interrupted the current interpretation collapses and must be rebuilt. Yet it seems clear we both have and need a memory of visual events and percepts. This is the role of the

[4] Even more broadly, one might also consider cross-model forms of detection, sounds, for instance, which may be both localisable and distinctive of an object type.

panorama (denoted Δp) in the Abductive Perception model. At the conclusion of each perceptual cycle, elements of the current interpretation Δi are transferred to Δp for retention. All elements of Δi are in viewer-centric coordinates and these are transferred directly to Δp.

The panorama structure is intended to model the "sense of space" about the self, generating and maintaining a "situational and spatial awareness" in terms of objects in the robot's immediate surroundings and to provide a stable ego-centric "platform" into which the current field of view (FOV) may be projected. This structure allows the robot, for instance, to reason about its surroundings without direct perception of them. It allows a robot with a rapidly moving visual field to establish a baseline set of hypotheses in Δh (on efficiency grounds) by repopulating the area of the image-plane with items from Δp.

The notion of a panorama appears to be particularly relevant to humans, who have both a foveal eye and whose eyes saccade constantly, in establishing a stable, viewer centric, percept of their immediate environment.

Elements of Δp are transformed to maintain a constant notion of "forward/left/right" centred about the observer's egocentric viewpoint following any motion. We propose an *anticipatory transform* that predicts the position the interpretations in Δp would appear in the next time step. As the information in Δp is already encoded spatially about the current viewpoint, a simple 3D matrix transform may be applied to the vector components to place them in an appropriate place following any movement by the robot or its gaze system.

As an exemplar of the notional form used, Let $f_1,...,f_n$ be a set of linear transformations (deployed as matrix operations) s.t. $f: V \rightarrow W$ where V and W are vectors spaces. And let $t_1,...,t_n$ be a totally ordered set of time points (i.e. image frames). Each *wff* of the form: $\Phi(x,v,Type,[...])$ is now re-worked as follows: $\Phi(x,v,Type,[...],t)$. Let $Tf(\Delta p)$ be the transformation function f applied to the set of *wffs* in Δp, s.t. $Tf: A \rightarrow B$, where A and B are sets of *wffs*. Then the updating of the vectors in Δp is defined as follows:

$$Tf(\Delta p) = \{\Phi(x,\underline{v_i},Type,...,t+1) \mid (\Phi(x,\underline{v},Type,...,t) \;\&\; f_i(\underline{v},\underline{v_j})) \rightarrow \Phi(x,\underline{v_i},Type,...,t+1),\ \text{where: } \Phi(x,\underline{v},Type,...,t) \in \Delta p \}$$

The anticipatory transform, *Tf*, has the effect of re-writing each vector embedded in every statement in Δp between image time points t and $t+1$. *Tf* may be determined by at least three different, but computationally well established, routes: (i) the anticipated consequences of an initiated motor action – trivially computed in the case of a mobile robot from the x,y,θ displacement begun, less obviously so for a multi-degree of freedom humanoid, (ii) displacement detected directly from motion sensors or odometry, and (iii) a transform derived from imager displacement, as typified by the SLAM (Simultaneous Localization and Mapping) class of algorithm (e.g. Davison 2003).

The advantage of option (i) being that the computation can be conducted during the motion. Pre-determined ballistic eye saccades are also well modelled in this way (e.g. Shibata *et al* 2001). Option (iii) is most appropriate where the compound effects of many degrees of freedom must be considered, but it does require an active imager and current SLAM based techniques are not generally well suited to systems exhibiting both broad and rapid saccadic movements. It might be conjectured that an analogue of each of these methods finds application in human perception, depending on the prevailing circumstances.

The question arises as to what (and at what level of detail) should items from Δi be transferred to Δp for retention after the interpretation cycle. Individual objects are naturally ordered in Δi by their explanatory value (though we shall argue later for an attentive salience of preference ordering), directly reflecting their level of evidential support. A sub-set of these items are selected for attention at the highest level of description (i.e. *wffs* of the form *Object(...)*), and reported to the higher cognitive layers. There is evidence (section 8) that humans select one as the locus of attention and record several more sub-attentively.

Level of evidential support seems a poor measure on which to select for attention. The next section describes how this ordering and selection process may be modified to meet the robot's functional needs. We would further argue that it is not appropriate to transfer descriptive sentences at the feature or appearance level into Δp, as these are potentially indistinguishable from direct sensor data, but rather as an abstraction, taken from the higher levels of the object description.

6 The Perceptual Process

Figure 2 illustrates the main perceptual cycle, which is further considered here. The original definition of this cycle (Randell and Witkowski, 2006) assumed a uniform field sensor and that the perceptual process will be completed in one imaging timeframe. The treatment here will modify that assumption to account for the three levels of resolution, foveal, peri-foveal and peripheral, previously identified for an idealized foveal imaging system. This process is applied exhaustively to all detected features in the image. This is equivalent to the notions of pre-attentive processing (e.g. Theeuwes 1993) or late selection for attention (Pashler 1998). It establishes the conditions required to: (a) process sensor items within the peri-foveal region in detail, (b) to identify areas in the peripheral region that are indicative of salient objects to be attended to in the near future, and (c) to inspect specific features in detail in the central foveal area. Section 7, describing the attention process, spreads across several perceptual episodes representing the flow of overt visual attention, as indicated by eye/imager gaze movements, and reflecting a serialisation of the perceptual process.

The cycle starts with a retinal mapping of features (Γ), using a small selection of distinctive features (and their appearances) to hypothesize objects than might explain those features and seeks to corroborate (or refute) those hypotheses by projecting the expected locations of other features associated with the conjectured object and comparing these with the sensor data.

For each separate image frame Γ will contain a mix of appearances, corresponding to the peripheral, peri-foveal and foveal imager areas. Superimposed over this will be the area (swept volume) of the transformed panorama (Δp), used primarily in this instance to maintain a record of recently attended items and their locations relative to the viewpoint of the observer in a viewer-, as opposed to retino-, centric coordinate frame.

The following steps summarize the (pre-attentive) perceptual process using the abductive perception model:

(1) Complete anticipatory transform on Δp, seed Δh with the currently overlapping FOV.

(2) Pre-process the image to identify *detected* (section 4) appearances. Record them in the Γ Structure.

(3) Evaluate the rank order *ro* of each object, identifying those object models in Σo that are supported by the current evidence in Γ. This is a preliminary "recognition by parts" step, establishing candidate objects for treatment as hypotheses.

(4) Match sets of unexplained but distinctive elements (according to *dv(F)*) from Γ to candidate (according to *ro* order) object descriptions in Σo to generate object hypotheses. This is the abductive step – selecting hypotheses from partial evidence.

(5) For features detected in the peri-foveal region a more detailed match between sensor data and object models may be attempted to both obtain a high confidence identification of the visual object and corresponding explanation of the sensor features, as follows:

(5.1) Identify four (or more) non-coplanar matches between features in Γ and features of corresponding *Type* in a single object model and postulate a projection into the 3D space represented by the FOV[5].

(5.2) Write this set of *wffs* (representing all the features for that object) to the hypothesis space Δh for each hypothesized object; rewrite every embedded vector in the object definition reflecting the new projection from the robot's viewpoint. This is the deductive (or prediction) step.

(6) Evaluate the *ev* for each hypothesized object to determine the extent to which it explains the sensor data for the projection area it occupies on the image plane. This is the corroboration step.

(7) Retract untenable hypotheses of low explanatory value from Δh. All *wffs* relating to the hypothesis are retracted and any sensor data features it might have accounted for are released, requiring further explanation.

(8) Repeat from (3) until all significant sensor data elements of Γ have a coherent explanation (i.e. that the sensor data can be accounted for as being matched to features of a hypothesised object), and where all the ground hypotheses in Δh satisfy the domain constraints (Σc). Discard (as noise) any un-interpretable data items or create new object description (section 2) from group of unexplained data items.

(9) Transfer explanation of the sensor data (Δh) to the interpretation space (Δi). The net effect of this processing is to place a single explanation in Δi for the incoming data, as *wffs*, and their (reconstructed) poses.

(10) Replace all objects in Δp in the area of the FOV with those from Δi, but retain any IOR labels on objects *according to location*. Remove direct sensor detail from any elements of Δp and successively remove detail (to full deletion) from older Δp elements not currently in the FOV. Finally, increment the process time *t* (section 7.2).

In peripheral areas any single distinctive feature, characteristic of an object type may serve to establish the hypothesis of an instance of that type of object in the image plane. Equally, objects in the peripheral area that subtend a larger area, or that are partially occluded, may be hypothesized from combinations of their characteristic parts. Objects sampled at low-resolution (e.g. located in the peripheral region of the imager) take the form of peripheral "impressionistic" hypotheses. These will later drive the saccadic scanpath/visual attention strategy. These hypotheses are transferred to Δi as a plausible interpretation of these areas of the image. Step (4) is therefore intended to be largely equivalent to Rensink's (2002) notion of a "gist" or "quick and dirty" assessment of the image properties for this area of the image plane.

Step (10) implies that old visual data is overwritten by the current interpretation in Δi (i.e. with no active visual memory, Wolfe 1999) and that the record of older objects outside the FOV are stripped of all visual data, retaining only a conceptual record within the working space of the robot.

The net effect of this process is to present a hypothesis-based explanation of the visual scene into Δi in terms of Σo object descriptions. This combination of broadly estimated and hypothesized peripheral objects and detailed "positioned" object models in the central area, combined with the raw sensor data appearances allows higher-level task modules to interrogate the perceptual system, either in terms of the original *sensor data*, from Γ, or in

[5] In trials, we used the DeMenthon and Davis (1995) POSIT (Pose from Orthography and Scaling with ITerations) method to determine the "pose" of the object model, as though it were projected back into the image space of the robot. Of the many algorithms that have been developed to solve this geometrical correspondence problem, POSIT (Dementhon and Davis 1995) and SoftPOSIT (David *et al* 2004) are of particular note. In general, non-unique solutions to the image point to object pose problem can be obtained from a minimum of three matches (Haralick *et al* 1994), however POSIT assumes a match of ≥4 non coplanar registration points and that their relative geometry is known. This requirement is relaxed in SoftPOST. See also (Horaud *et al* 1997, Hu and Wu 2002) for real-time vision and robotics applications.

object interpreted terms in the manner of *sense data* (e.g. Huemer 2004), from Δi. For a foveal system, any apparent detail in the periphery of the scene must therefore be a "reconstruction" – an "illusion" – derived from the hypothesized model-based data in Σo.

The object at the foveal centre will be considered as the locus of overt visual attention for the current cycle and as such will be passed to the (notional) higher levels of cognitive processing. We suggest that several additional items that are worthy of attention are also transferred to the higher levels, as items of covert attention. A mechanism for this is presented in the next section. Recall (from section 1) that elements of visual attention are not necessarily "adopted" or acted upon by the higher levels.

In the Itti and Koch (2001) model, inhibition of return is applied retino-centrically to each instance of the feature detector. This presupposes a fixed viewpoint. This cannot be the case for a mobile or humanoid robot, or any system with saccadic (or directed) eye movements, in which the image appears to translate across the image or retinal plane with each motion. Under these assumptions, inhibition of return is more naturally expressed as a tagged property of object interpretations recorded in Δp. Inhibition of return consequently tracks with the interpreted objects. Expanding the definition of each object or sub-part (axiom A1), coupled to its vector estimation, generates an *area of inhibition* on the imager plane: indicating the inhibited volume represented in Δp as projected back onto the imaging plane. Application of the anticipatory transform (section 5) ensures that this area remains directed at the object volume inhibited, regardless of rotation and translation motions of the robot and its imager.

7 The Attentional Process Model

In order to achieve the characteristic "sparse attend" eye-gaze model described earlier we adopt a three part attentional strategy. We continue with the assumption that peripheral detectors are sufficiently characteristic of the objects they are appearances of to establish a reasonable working hypothesis about the objects that cause them. Further, it will be assumed that peri-foveal detectors have sufficient resolution to make robust recognition of the objects they describe, when the perceptual cycle (section 6) is applied to the available features and their appearances.

Finally, it will be assumed that foveal gaze is required for detailed corroboration of object identity (and effective perception of detail, such as reading text). Where an observer may be expected to have many individuated models of otherwise similar object types, such as faces, we may expect the foveal gaze phase to be concentrated on the features. This will serve to confirm the specific identity, to update the model to accommodate on-going change (as is the case with human faces), and to acquire (learn) new variants of the object model as needed.

7.1 Preferences over objects

The order in which objects are attended to exploits various preferences defined on object-types, objects and their respective features. These preferences are defined on the sets of names and predicates defined within our formal language. Preference orderings are defined on all object types ($P\succ_{pref}$) from Σ_O, types of detected objects ($A\succ_{attend}$), detected instances of objects ($OI\succ_{instance}$) and on the component object-parts and features of objects ($SP\succ_{subpart}$). We refer to $\succ$ as the *preference order operator*. An extended definition of the preference order operator is given in appendix 1. The following preference orderings are defined over object types (in Σ_O), object instances and object-parts:

(i) $P\succ_{pref} = <O_1, O_2, \ldots, O_{n-1}, O_n>$, where $O_i \in \Sigma_O$
(ii) $A\succ_{attend} = <<O_a, O_b, O_c, O_d, O_e> , <O_f, \ldots , O_n>>$, where $O_i \in P\succ_{pref}$
(iii) $OI\succ_{instance} = <O_a \mid << o_{a1}, o_{a2}, \ldots , o_{ak}> , < o_{ak+1}, o_{ak+2}, \ldots , o_{an}>>>$
(iv) $SP\succ_{subpart} = <O_a \mid << S_1, S_2, \ldots , S_k> , < S_{k+1}, S_{k+2}, \ldots , S_n>>>$

Definition (i) provides a preference ordering over all the objects currently known to the system. Given a set of all the object types in Σ_O: $O=\{O_i|Object(x,y,O_i,...)\}$, *where:*

$Object(x,\underline{y},O_i,...) \in \sum_O$, we can order the elements of $O=\{O_1, O_2, ..., O_{n-1}, O_n\}$ by some preference measure *pref*, indicating the current predisposition of the system to prefer one object type over another, such that $P \succ_{pref} = <O_1, O_2, ..., O_{n-1}, O_n>$ meaning that on the $\succ_{pref}$ measure, O_i is preferred to $O_{j,}$ where $i<j\leq n$. This ordering need not be constant, but will change to reflect the task being carried out by the system. For instance, in a navigation or driving task, object types that represent hazards or obstacles will be moved to the from of the list, while during a search task the object or class of objects being sought will be moved to the front of the list, relegating other object definitions not directly required. In section 8.5 we propose a specific method of achieving this preference ordering by increasing the significance of individual features associated with each object.

Definition (ii) divides the previous overall preference list ($P \succ_{pref}$) into two parts, those elements that are supported within the current field of view (the head) and those that are not supported within the FOV (the tail). Definition (iii) considers all the instances (*o*) of a single object type (*O*) in the current FOV. By splitting this list into two parts, only some of those instances (in the list head) will be considered, before attention moves on to some other aspect of the scene (as defined by the next object type in the $A \succ_{attend}$ list). Definition (iv) decomposes a single object type (*O*) into its sub-part types (*S*) or specific object instance (*o*) into its individual sub-parts (*s*). This is required to allow detailed inspection by the scanpath of the parts of an object, as indicated in figure 1.

Note the <<H>,<T>> and <O|<<H>,<T>>> partition constructs for $A \succ_{attend}$, $OI \succ_{instance}$ and $SP \succ_{subpart}$, where H is the head and T the tail of the list. In the former (ii) this represents a simple splitting of the list into two parts (the overall preference order within the two parts is unaffected). The latter indicates an expansion of an object type (O_a) into its instances currently in the FOV (iii), or its subparts (iv). We also use the head of the list to factor out primary (overt) and secondary items of attention in the model: in this case the first element of the list represents the autonomic candidate for overt attention (pre-attentive) and therefore the destination of direct foveation. Elements from the remainder of the list head are then candidates for covert attention.

In each case the cardinality (length) of the head of the list will vary according to the mode of attention we wish to emulate and this also determines the length of the cyclic scanpath. Where many elements are admitted to the list head, attention is lightly focussed with many candidates for covert attention and attention is highly susceptible to interruption by exogenous events. As the number of elements is reduced attention becomes increasingly focussed with fewer secondary items and attains greater resistance to interruption. We also consider the case where the list head comprises a single object type or, in the extreme, a single object instance (section 9.2).

The scan-path order is determined via a structural decomposition of the elements of preference lists by successively applying (i) through to (iv). This process maps detected objects in the FOV to the root nodes of their object descriptions, and uses their object descriptions (as hierarchical structures in Σ_O) to determine a node-node traversal path through the set of detected objects and their constituent parts and features. As the whole perceptual process is reactive, the traversal of the tree during this cycle can at any time be modified by IOR, which may release the inhibition associated with a previously detected object at a specific location.

One might suggest substituting *dv* as the *pref* index value for the $P \succ_{pref}$ list (i) in a robotics context. Distinctive objects serve to explain more of the sensor data, leaving less to be explained with subsequent processing effort. In many ways, though, this is unsatisfactory. Attentive preference across objects should directly reflect the observer's functional requirements in relation to the objects, the uses they will be put to – and the cognitive tasks the observer is undertaking at any given time.

This preference ordering is itself covert, in general a person is not aware of the preference ordering being adopted by their visual attention system. We assume that no overt act of will is required to establish a new preference ordering in relation to an established task (task

shifting) and that individual items may be selected voluntarily for preferential attentive processing. In a robot, these will be inherent in, or inferred from, the task being performed.

Where a static image is being viewed, we may expect this order of the $A\succ_{attend}$ list (ii) to remain largely constant, while no task is requested of the observer. Under static image conditions the scanpath will follow the general order defined by the object types in the list head, and repeat once the IOR effect has dissipated. When the observer is moving we may expect the elements of this list to change between image frames as objects appear in view and others leave the frame.

In the case of the $OI\succ_{instance}$ list (iii), the preference order is appropriately defined by the level of evidential support (*ev*) for individual objects *o* of type O_a at each of the given image locations *1…n*, such that instances where there is strong support are preferred and so investigated or confirmed before the others. Alternative ordering strategies, such as object proximity, might also be considered. In this case, the list head length determines the number of individual object instances that will be inspected before the object type is inhibited and the next most preferred object visited.

For the $SP\succ_{subpart}$ list (iv), as with the object preference list, the sub-part/feature order can be directly related to the *dv* value (the "default model", see section 8.5) of the part in question or, more reasonably, relate to the functional significance of the part relative to the whole (i.e. the eyes and mouth to a face). This sub-division into sub-parts may be applied recursively through the object description hierarchy until the feature level is encountered, but we do not consider this case further.

7.2 The scanpath cycle

This section describes the basic scanpath model. At the start of the cycle we assume that the previous saccadic imager movement has just been completed, that a new image has just been presented and that the destination of the movement was an object instance indicated as being of the highest attentional preference (o_p). Note that the pre-attentive hypothesis that an object in the peripheral area is of a particular preferred type is no guarantee that it will be so on closer examination at the fovea. The record (by labelling in Δp) of recently attended items will serve the function of inhibition of return (IOR). We define a function *ior(o)* which labels any object instance *o* (or part or feature, *s*) in Δp with the most recent time of inhibition *t*. The symbol τ will be used to represent a time increment (typically some number of perceptual cycles) defining the temporal extent of the inhibition of return effect. Values of $(ior(o) - t) > 0$ indicate an object with active inhibition, zero or negative values indicate spent or previously uninhibited values.

We define four further functions: (1) *saccade(o)*, which rotates the imager to align with the current (off-fovea) image plane location of object (or subpart) instance *o* or *s*. (2) *attend(o)*, which labels the *wff* describing *o* (or *s*) as the item of overt visual attention for this cycle. (3) *pre_attend(o)* labelling *o* (or *s*) as the pre-attentive item. (4) *sub_attend(o)* labelling *o* (or *s*) as (covertly) sub-attentive.

The attentive process may be summarized as follows:

(1) Complete the perceptual cycle (section 6, resulting in an interpretation Δi, indicating the current object of highest attentional preference o_p at the fovea) yielding a set of detected object types in list S_1 ($\{O_1,...,O_n\}$) and their associated objects in list S_2 ($\{o_1,...,o_n\}$).
(2) Process attended object (o_p) at fovea. If object o_p is to be inspected in detail (on the basis of novelty, ambiguity, complexity) perform sub-steps, else (3)
(2.1) Expand the current object at fovea o_p into its constituent sub-parts $s_1 \ldots s_k$ (i.e. $< o_p \mid < s_1, s_2, \ldots, s_k >, < \ldots >>$) according to the sub-part ordering $SP\succ_{subpart}$)
(2.2) Select first sub-part s_i without any inhibition of return ($ior(s_i) - t \leq 0$ in Δp)
(2.3) Send identity of sub-part s_i to cognitive layer as overt attend item ($attend(s_i)$)
(2.4) Set IOR for attended sub-part s_i to current time *t* plus inhibition increment ($ior(s_i) = t + \tau$)

(2.5) Initiate saccadic movement to location of s_i on image plane (*saccade*(s_i))
(2.6) Continue inspecting sub-parts of o_p in detail? If yes, then (5)
(3) Set IOR for attended object o_p to current time t plus increment ($ior(o_p) = t + \tau$)
(4) Send identity of object o_p to cognitive layer as overt attended item (*attend*(o_p))
(5) Sort object types in S_1 according to current attend list ($A \succ_{attend}$) to form ordered list (S_3) of object types: $S_3 = <<O_1, \ldots, O_n>, <\ldots>>$
(6) Select and report items for covert attention: While (O_i from head S_3)
(6.1) Expand $< O_i \mid < o_{i1}, o_{i2}, \ldots, o_{ik} >, < \ldots >>$ (by $OI \succ_{instance}$ from S_2)
(6.2) Select object o_{ij} instance in this list with minimum inhibition of return ($min(ior(o_{ij}))$ in Δp) and send to cognitive layer as the pre-attentive item (*pre_attend*(o_{ij}))
(6.2.1) If no s_i initiate *saccade*(o_{ij}) in image plane (o_{ij} becomes o_p next cycle)
(6.3) Next n-1 o_{ij} where $min(ior(o_{ij}) - t)$ in Δp *sub_attend*(o_{ij})
(7) Return to (1), starting the next perceptual cycle

This basic cycle captures the essential aspects of the scanpath strategy indicated by previous studies (e.g. Stark and Choi 1996) and illustrates a number of interesting properties. We use the text that follows to further nuance the scanpath behaviour.

Step (1) performs the underlying perceptual cycle (section 6), returning a partial description of the scene in Δi in terms of object hypotheses (S_2), expressed as *wffs*, and their corresponding types in S_1. These will later be sorted by the preference criteria. The object at the fovea is designated o_p.

Step (2) emulates inspection within a single object. In this case a number of smaller saccadic movements are made within the boundary of the attended object. At this stage the pose and estimated size of the object has been determined and the specific locations of features ($s_1 \ldots s_k$) should be accurately predicted. Step (2) proceeds in sub-stages. First (2.1) the foveated object is decomposed into sub-parts (object-parts or features) according to the prevailing preference order and the oldest (least recently inhibited) element is selected (2.2). The part (s_i, not the object) is labelled as the object of visual attention (2.3) and maximally (i.e. most recently) inhibited (2.4). The saccade to the selected part is initiated (2.5). Inhibition on the part ensures further close inspection will select the next preferred sub-part, if required (2.6) following the next perceptual cycle. This early onset of the saccade is broadly consistent with findings in human studies that attention to parts within a single object is faster than between objects (e.g. Duncan 1984, Vecera *et al* 2000).

There are several conditions under which such detailed inspection is warranted. First, the detailed construction of new or novel object models for deposit into Σ_O, or to accommodate new information about an existing object model. Note that novel or recently introduced objects do not necessarily precipitate attentional shift in humans, though such events are invariably associated with luminance transients – which do (e.g. Franconeri *et al* 2005, Theeuwes 1995). Second, to seek to reduce any mismatch between an existing model and the data presented and, third, to reduce ambiguity of interpretation[6].

Step (3) sets the inhibition label for the current attended object (o_p), as reported to the cognitive layer (step 4). Step (5) builds the attention candidate orderings from the object type list S_1. Step (6) expands each element of the ordered object type list S_2 into its localized instances (o_{i1}, etc.) The first of these is labelled as pre-attentive and (if there was no part based saccade), the imager is rotated to the vector direction of the object. This object instance will be o_p in the next cycle. Finally, a small number (n-1) of remaining uninhibited object instances from the expanded object lists are also labelled as sub-attentive for the cognitive layer.

[6] Each (single) hypothesized interpretation of each object in Δi may be labelled with an indication of the degree of uncertainty in Δh just prior to the selection. We note in passing that such labelling might be interpreted at the cognitive levels as a "mindsight" (after Rensink 2004) event, the perception that the object interpretation is in some way suspect, but without any indication as to why.

Note that this description only refers to autonomic (involuntary) eye movements, as determined by the preference list ($P \succ_{pref}$), although the object preferences and so the attention list ($A \succ_{attend}$) may be either the default or task based. We suggest that for each list, or sub-list, an upper bound should be established limiting the number of instances of a given type that will be inspected before the next item in the list is selected. This is the list head length. For illustration, we will assume this bound to be four or five[7]. So that, if there are more than, say, five instances of object O_a, then only the first five are inspected before O_b is selected. If the object to be inspected in detail, then only the first five or so features are inspected[8], and so on.

7.3 Managing attention preferences

We assume that each individual observer will develop a attentive preference strategy over time, according to their interests and daily needs. In general, we note that people have a strong propensity to attend to the human form and in particular to facial features. There is doubtless a shared core of object types that remain high on the human attentional preference order. Equally, attentive preference order ($A \succ_{pref}$) is not intended to be unchanging or unchangeable. It seems attention preference order is easily temporarily manipulated according to task or need. This context change can be both rapid and comprehensive. A driver may attend to one set of object types at one moment, leave his car and attend to quite another as a pedestrian immediately after. Equally, the participant in an attention experiment may attend preferentially to the object of the experiment for its duration and subsequently never do so again. It also seems highly likely that the long term, underlying, preference order is malleable, perhaps reflecting the frequency with which each object definition is intentionally raised to the front of the preference order. It is within our common experience that practice (i.e. intentional looking) will lead to automatic noticing of such objects of interest. For instance, an ornithologist may notice birds in a scene, where others, and he previously, would not have attended to them.

8 Discussion: Aspects of Attention

In this section we discuss four related aspects of natural attention in the context of the formal model of perception and attention just described. These four aspects are: (1) autonomic scanpath generation and overt (fixated) attention, (2) secondary and covert attention, (3) the role of exogenous (sensor-driven) attention, (4) endogenous (volitional task-based) attention. The next section will consider two forms of inattention, change blindness effects and inattentional blindness effects. The representation of perception employed in this paper is inspired by natural systems but is then highly abstracted away from the underlying neural mechanisms that mediate animal and human attention processes. The primary purpose is to identify underlying principles that might be applied to an artificial or robot model, inspired by, but not slavishly emulating, natural attention. However, we will indicate where we have noted interesting parallels or significant differences.

The notion of "visual attention" covers a vast portmanteau of effects, which have been extensively researched in a wide range of diverse disciplines. The discussion here is necessarily limited to a sub-set of this diversity. Each is analysed in terms of the object perception approach, the management of preference lists (in particular, the content and length

[7] Such limits are open to debate and must be established according to need. At the object level we note that some investigations indicate that inhibition of return is effective over a period of approximately 1.5 seconds (e.g. Taylor and Klein 1998) and each saccade/fixation pair is approximately 300mS. This accords with established eye-gaze data (Noton and Stark 1971) and, less directly, with Multiple Object Tracking (MOT) tasks (Pylyshyn 2001), although part inspection appears to extend the scanpath cycle (figure 1 and text).

[8] Note again that at the object-part or feature level a minimum of four separate image based position estimates of parts or features are required to perform the POSIT pose reconstruction (DeMenthon and Davis 1995) to project the object model into Δh.

of the list head) to predispose the perception system to attend to specific object types, and the inhibition of return (IOR) mechanism.

8.1 The Autonomic Scanpath and Overt Visual Attention

This section considers the autonomic control of eyegaze movements. The order of the gaze scanpath in the model is determined by the current preference ($P\succ_{pref}$) list and its active subset, the attention list ($A\succ_{attend}$). In turn, these preference orderings are manipulated by the cognitive layer according to current tasks and needs. If no task is specified the preference orderings revert to their defaults. No further stipulation need be made by the cognitive layer to control attention, the scanpath is determined by the cycle given in section 7.2. At each fixation, a single typed object or object-part from the foveal area is identified as of primary or overt visual attention. This single *wff* is tagged as the primary locus of visual attention and reported to the cognitive layer. Recall that the cognitive layer need not take any action on the basis of this notification, which is automatic and continuous.

We see this autonomous mode as normal for both humans and robots. Foveal attention is properly directed at semantically significant objects that appear within the visual field, in a priority order established by the preference list. Inhibition of return recorded in the panorama emulates a cyclic condition for the scan, with conditions defined for detailed inspection of objects in the visual field. The intention here is to emulate a situation where the observer has a "gist" of the scene comprising broad hypotheses, but about which foveal attention is directed to give a detailed, but narrow, view of each place that is attended to in a sequential manner.

It may be more appropriate to consider the item selected in step 6.2 of the attention cycle (o_{ij}) as the true locus of overt visual attention, as it is the location that will be confirmed by foveal inspection in the next perceptual cycle, and it is therefore pre-attentive (Theeuwes *et al* 1998). Where the effect of attention is to initiate motor action, this gives the robot one clear saccade/fixation cycle (say 200-300ms, by analogy with human perception) in response time advantage in advance of the confirmed attention. Where foveation subsequently disconfirms the hastily prepared hypothesis, the action may be cancelled.

Autonomic mode is suspended for brief (or possibly somewhat extended) periods under specific conditions, some of which are described further in the following sections. Certain exogenous signals (e.g. detected abrupt changes of luminance, or high saturation detectors, etc., section 4) override this object based autonomic control (section 8.3), directing the gaze point to the estimated source of the detected signal. Equally, specific voluntary or planned movements of gaze, initiated directly by the cognitive layer also take precedence over autonomic control. Voluntary control operates independently of inhibition of return (or nearly so, e.g. Klein and MacInnes, 1999), whereas exogenous (Itti and Koch, 2001) control is apparently always subject to inhibition of return.

8.2 Secondary and Covert Attention

We define secondary attention as the notion that several items (as *wffs* in the model) are passed to the cognitive layer at each perceptual cycle, in addition to the primary locus of attention. As there can only be one point of overt attention on the scanpath at each instant, these points are, by definition, independent of the fixation point. Note, however, that we separate out the single pre-attentive item (steps 5 and 6 in the scanpath cycle, section 7.2), which becomes the candidate location for the next fixation point in the immediately following perceptual cycle. The notion of covert attention reflects the possibility that the cognitive layer might inspect the Δi structure[9] as a volitional activity to bring any item currently interpreted in the visual field into attention.

[9] Although it might be argued that such inspection is more properly conducted within the panorama Δp, particularly in the light of the debate about visual memory (e.g. Simons and Rensink 2005, Wolfe 1999, considered further in section 9.1).

In addition to the single primary point of visual attention, we model a secondary attentional process in which a (selectable) number of object instances in the visual field from the head, but not at the head, of the attention list A are labelled as attentionally significant. These *wffs* are also reported to the cognitive layer. It is assumed that these preferentially evoke overt attentional behaviour if acted upon by the cognitive layer, but are in any case likely candidates for overt attention by virtue of their ranking in the attention list. They are not subject to any inhibition of return effect, so a single sub-attentive object may be retained over a number of attentive cycles without apparent diminution so long as it remains in the head of the object instance ($OI\succ_{instance}$) list.

Pylyshyn and Storm's (1988) seminal Multiple Object Tracking (MOT) task experimental procedures point directly to the presence of these secondary attentional artefacts. Participants are shown a screen of multiple identical objects, some of which are made to blink at the beginning of the experiment indicating they are to be tracked. The objects then move smoothly, but unpredictably, about the display area. At a given time one of the objects is indicated with a pointer and the participant indicates whether that object was to be tracked. Participants generally perform substantially better than chance on these tasks where the number of tracked objects is five or fewer[10]. The experiment points to an ability to maintain a sub-attentional track of multiple non-foveal objects. This is modelled by installing the n identified objects to be tracked as instances at the head of the ($len(OI\succ_{instance}) = n$) list for a single nominated object type ($len(A\succ_{attend}) = 1$), which are therefore monitored by the *sub_attend()* function described earlier.

An equally significant question arises as to whether it is possible to bring into awareness any item at will, whether sub-attended or not, and without gaze shift – covert attention. The model would propose that the description of the whole FOV is available in Δi (or Δp) and that this may be interrogated at will while the point of gaze is held constant under volitional control (although apparently with some effort in the case of a human). Under these circumstances, of course, only the coarse grained hypotheses are available for consideration, but these are generally sufficient for recognition and general placement (via the object appearance description and its associated viewpoint vector representation).

8.3 Exogenous attention

We define exogenous attention as that which is mediated directly by the visual properties of the items in the visual field. Exogenous attention is therefore largely independent of the semantic content of items viewed, not all attentional control is object preference derived. A number of detector conditions have been identified (e.g. Itti and Koch 2001) that are independent of the features associated with object definitions (section 4), although not necessarily independently of object location. These include (but are not necessarily limited to) rapid changes of luminance at a given place in the field of view, areas of high colour saturation and certain forms of movement or change.

These events may override the normal course of overt attention and direct the gaze to saccade to the location of the source within the visual field. Such events are therefore in direct competition with other mechanisms of gaze control[11]. Within the model, such artefacts direct the perceptual cycle to process the item at the centre of the field of view and so the object associated with the source of the sensory event. This form of gaze control is subject to the

[10] It is unclear why human attention appears to be restricted to tracking this number of objects, possibly an evolutionary legacy stemming from our role as a pack hunter in which actions of both the quarry and other member of the group must be tracked to coordinate effective predation behaviour. Equally, it might be argued that the ability to track multiple predators is advantageous, where the roles are reversed. In the model, this value is controlled by the free variable n (step 6, section 7.2). There is no reason to suppose that it might not be set to any convenient value in a robotic or artificial application, allowing an arbitrary number of sub-attentive items to be considered.

[11] Several detailed models of the exogenous/endogenous gaze selection mechanism have been proposed for the *superior colliculus*, the dorsal midbrain structure widely considered to be the seat of the neural basis of saccadic eye movement control (e.g. Grossberg *et al* 1997, Trappenberg *et al* 2001).

inhibition of return mechanism and the object (and so the source of the exogenous sensory event) is labelled as inhibited, as with overt attention.

We use the cardinality of the list head as an indicator of the degree to which exogenous events will pre-empt the autonomic or endogenous modes of attentive behaviour. Where there are many items in the head (unfocussed attention), items of low exogenous salience may be admitted into the object list (i.e. general properties such as colour saturation compete effectively with object preference based features) and drive foveation. Where the length of the list head is reduced (object focussed attention) the salience threshold becomes successively raised, making it increasingly hard for exogenous events to be represented in the attentive process. We describe the mechanism no further here, beyond observing that were numeric values to be associated with the object preference list and an equivalent preference range assigned to the salience measures, the two might easily be merged into a single preference ranking. Opinions vary as to the bounds of interruptability by exogenous visual events (e.g. Theeuwes 2004) or directed volitional gaze shifts (Theeuwes *et al* 1998). The question also then arises as to the effect of admitting only a single object instance (and therefore type) into the list head, which are considered further in section 9.2.

8.4 Endogenous attention

We define endogenous attention here to be that which is directed to performing some task imposed by the cognitive layer. It is a form of overt visual attention, in that it is mediated by manipulation of ranking within the object preference list. Endogenous attention is classically investigated as a visual search task, in which participants are asked to search an image for an embedded specific but known target (e.g. Treisman and Gelade 1980, Wolfe 1994); or to identify "the odd one out" from a set of otherwise identical targets as quickly as possible. Exemplar tasks are shown in figure 5 (after Wolfe 1994).

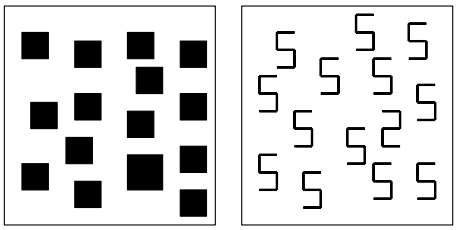

Figure 5: Search tasks – find the "odd one out"

Where these characterizing features for the different target types are distinct at the peripheral resolution, attention and hence gaze, may be directed immediately[12]. Typically, on the left hand test, the desired item is reported almost immediately, gaze saccades to the correct place with little hesitation and search time is broadly independent of the number of test items. It might be argued that the differences in the target to background items are sufficiently large to register at the peri-fovial resolution and attention may therefore be directed to the target immediately. The effect is particularly marked when the target is a different colour to the background items.

Where the characterizing features are not distinguishable at the peri-foveal or peripheral resolution (such as illustrated in figure 3, right), attention must be directed sequentially, such that the detailed foveal model is brought over the candidate targets in order and passed to the cognitive layer for detailed comparison (that is, as a semantic rather than visual properties comparison). On the right hand test of figure 5, saccades show a marked (self-terminating) searching strategy, and search time rises linearly with the number of items presented.

This may be characterized as a single object (two, if the target and distracter are both known), multiple instance task. Inhibition of return is spread over the number of instances,

[12] This is the "pop-up" effect, which is frequently observed when a single distinct (colour or simple shaped) item is embedded among several otherwise identical ones. We suggest that this effect is a form of exogenous attention (not all do, see, e.g. Mack and Rock 1998), with a special purpose mechanism within the perceptual layer directing overt attention, which we do not model here.

allowing each to be visited in turn. Depending on the complexity of the task, each may be passed to the cognitive level for evaluation. Once the target is encountered the attentional preference list may revert to its default and the task is complete.

Several things are clear from this. First, attention remains directed to the task, as other items in peripheral vision are not visited. Second, it clearly demonstrates that inhibition of return is not applied directly to the feature detector, or to any particular area of the image plane. These tasks are contrived, but Wolfe (1994) suggests they are also indicative of the process when applied to more naturalistic tasks.

8.5 Worked example

This section primarily illustrates the effects of exogenous attention ordering through a detailed worked example. For simplicity we will assume that the objects have been (correctly) hypothesised during the peripheral recognition phase described above. To aid the calculations, and make the effects of each change clearer, only a small sub-set of possible object features will be considered. The agent will scan over (overtly attend to) the objects in various orders, reflecting changing attentional priorities arising from: (i) The default information content (distinctiveness) based ordering, (ii) an ordering in which a subset of objects in Σo is advantaged by (artificially) increasing the relative strengths of the *dv* values of all their constituent features, thereby changing the preference order P (section 7.1) and (iii) Task or activity imposed orderings (section 5).

This scheme is achieved by defining a significance multiplier function (*sm()*) associated with each feature type, such that all *dv(F)* terms in Appendix Two section A2.2 are substituted by (*dv(F)*sm(F,n)*).

When applied with *sm()*, values of $n > 1$ enhance the attentive properties of the feature type *F*; $n < 1$ suppress it. Raising any *dv* value in this way consequently raises the preference order ranking of each object in which the feature appears. This raises the priority of the object model and it will be preferentially projected into Δh for corroboration. If corroborated, the *ev* value of the object is also increased. Its place in the attention prioritisation is thereby raised. This is exogenous, data driven attention, analogous to the Itti and Koch model.

Next, consider the effect of the function *attendobject(O, n)*, enhancing the distinctiveness values (*dv*) of all features of an inferred object by the significance multiplier. A high level cognitive system may now select from any of the object descriptions those that are to be considered relevant to its immediate task. Detection of any of the features implied by the object definition and passed upwards via the abductive step now contributes to the enhanced ranking. The greater the multiplying factor, the higher the ranking given equivalent evidence. This pre-disposes the system to select this model on the basis of weaker evidence than it would otherwise.

A small sub-set of feature points from each object definition is selected here and a *dv* value calculated for each resulting type (Table 1). These feature points serve as "fixation" points in the emulation. The emulation calculates the object and feature distinctiveness values and preference orders according to the attention criteria stipulated. The resulting path orderings plotted and overlaid on the (greyed) original image for presentation in Figure 4.

Figure 5a (top-left) shows the "default" gaze preference order, based on the distinctiveness values computed for each of the five objects ($P_a \succ_{pref}$ = <Cat, Singer, Triceratops, Stegosaurus, Frog>). In each case we elect to expand the object into its sub-parts for detailed inspection (steps 2.1-2.6, section 7.2). The gaze order within each object is determined by the computed distinctiveness value of the individual feature types for that object. This order may change between examples as it depends in the relative effects of *sm* on each of the features, as described later in this section. In this simplified model the secondary attention order (section 8.2) is also given by the remaining items in the preference list. Covert attention is not considered here.

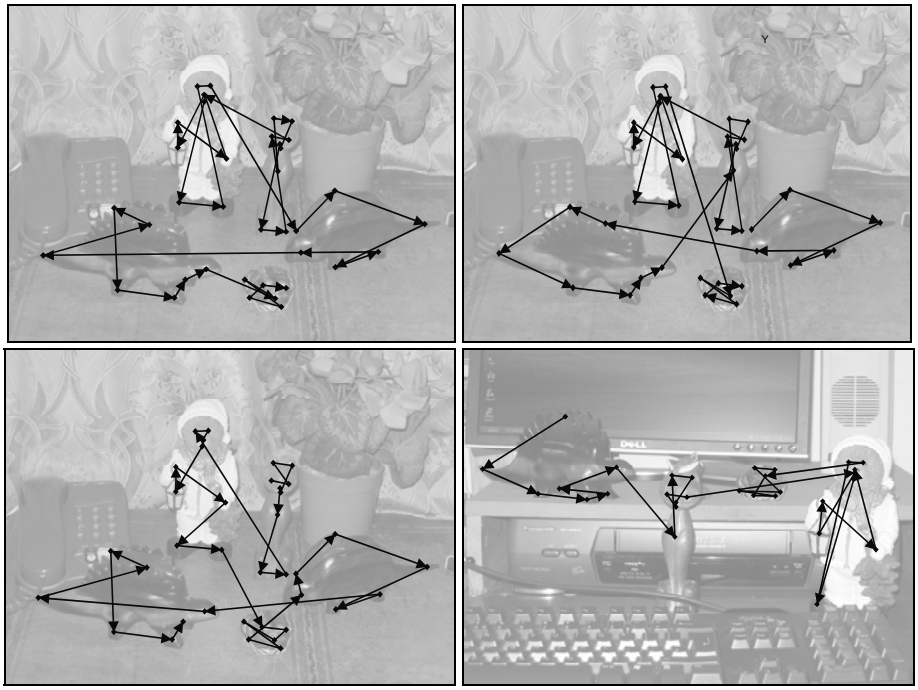

Figure 5a-d: Software emulated gaze patterns

The computed measures used to determine the order in which features and objects are visually scanned is given in Table 1. The columns give the name of the feature/object, its *Type* and *dv(Type)* value, and the ordering for the four images shown in figures 5a to 5d. Empty order cells indicate features that are (self-) occluded or absent. Note the change in the computed ordering between the features in figures 5a and 5b. In this case (and as explained in the text) arising here from a change of saliency, compared with the change in the order in which objects are attended to, but where the search order of features within the same object remain unchanged, as when comparing the simulated scanpaths in figures 5a and 5c.

Figure 5b (top-right) shows the gaze order where the feature type distinctiveness value for feature types plate and horn in the two dinosaur (these feature types being typical of "dinosaurs" in this restricted world) definitions has been multiplied by a *sm* constant factor (x10), so as to guarantee they are raised to the front of the preference list: ($P_b \succ_{pref}$ = <Triceratops, Stegosaurus, Cat, Singer, Frog>). Note that the feature visit order in the selected dinosaur objects is identical to that of figure 5a (the default), but that the internal scan order of the other objects is changed as the modified *dv* values propagate through all object models. In the general case, because feature types are necessarily shared between types of objects, it is not possible to induce all possible object preference orders using this weighted feature approach.

Figure 5c (bottom-left) shows a task based ordering based on an *sm* value of 10 for the singer and its effect on the gaze path. The preference order becomes ($P_c \succ_{pref}$ = <Singer, Cat, Frog, Stegosaurus, Triceratops >) with the singer attended first, as expected.

Similarly, Figure 5d (bottom-right) shows an alternative attention ordering ($A_d \succ_{attend}$ = < Stegosaurus, Cat, Singer, Frog>), based on the conditions described for figure 5b, illustrating the effect of a missing object type (O2) and occluded object sub-parts (f41 and f42 of O5, the cat). The preference list P is unaltered, but the attend list A only contains elements of the preference list supported by the evidence in the image. The feature scan order within each object is not affected between these latter two rankings as this is based on the original *dv* preference ordering.

			Figre 5a			Figure 5b			Figure 5c			Figure 5d		
Index	Name	Type	dv(o)	dv(Type)	Ordering	dv(o)	dv(Type)	Ordering	dv(o)	dv(Type)	Ordering	dv(o)	dv(Type)	Ordering
	o1	Stegasauros		0.14632			0.689095			2.431664			0.689095	
1	f1	Eye	0.14632	0.006494	32	0.689095	0.006196	13	2.431664	0.899814	26	0.689095	0.006196	7
2	f2	Eye	0.14632	0.006494		0.689095	0.006196		2.431664	0.899814		0.689095	0.006196	
3	f3	Foot	0.14632	0.012554		0.689095	0.012256		2.431664	0.142239		0.689095	0.012256	3
4	f4	Foot	0.14632	0.012554	29	0.689095	0.012256	10	2.431664	0.142239	27	0.689095	0.012256	4
5	f5	Foot	0.14632	0.012554	30	0.689095	0.012256	11	2.431664	0.142239	28	0.689095	0.012256	5
6	f6	Foot	0.14632	0.012554	31	0.689095	0.012256	12	2.431664	0.142239	29	0.689095	0.012256	6
7	f7	Tail	0.14632	0.027706	26	0.689095	0.027408	9	2.431664	0.021027	30	0.689095	0.027408	2
8	f8	Plate	0.14632	0.027706	27	0.689095	0.300135	7	2.431664	0.021027	31	0.689095	0.300135	1
9	f9	Plate	0.14632	0.027706	28	0.689095	0.300135	8	2.431664	0.021027	32	0.689095	0.300135	
10	o2	Tricerotops	0.166667	0.166667		1.255491	1.255491		2.329004	2.329004		1.255491	1.255491	
11	f10	Eye	0.166667	0.006494	25	1.255491	0.006196	6	2.329004	0.899814		1.255491	0.006196	
12	f11	Eye	0.166667	0.006494		1.255491	0.006196		2.329004	0.899814	33	1.255491	0.006196	
13	f12	Horn	0.166667	0.088312	20	1.255491	0.906196	1	2.329004	0.081633	36	1.255491	0.906196	
14	f13	Plate	0.166667	0.027706	21	1.255491	0.300135	2	2.329004	0.021027		1.255491	0.300135	
15	f14	Foot	0.166667	0.012554		1.255491	0.012256		2.329004	0.142239	37	1.255491	0.012256	
16	f15	Foot	0.166667	0.012554		1.255491	0.012256		2.329004	0.142239		1.255491	0.012256	
17	f16	Foot	0.166667	0.012554	23	1.255491	0.012256	4	2.329004	0.142239	34	1.255491	0.012256	
18	f17	Foot	0.166667	0.012554	24	1.255491	0.012256	5	2.329004	0.142239	35	1.255491	0.012256	
19	f18	Tail	0.166667	0.027706	22	1.255491	0.027408	3	2.329004	0.021027	38	1.255491	0.027408	
20	o3	Singer	0.300433	0.300433		0.298051	0.298051		4.774273	4.774273		0.298051	0.298051	
21	f19	Eye	0.300433	0.006494	18	0.298051	0.006196	31	4.774273	0.899814	1	0.298051	0.006196	22
22	f20	Eye	0.300433	0.006494	19	0.298051	0.006196	32	4.774273	0.899814	2	0.298051	0.006196	23
23	f21	Muzzle	0.300433	0.088312	10	0.298051	0.088014	23	4.774273	0.899814	3	0.298051	0.088014	15
24	f22	Lamp	0.300433	0.088312	11	0.298051	0.088014	24	4.774273	0.899814	4	0.298051	0.088014	16
25	f23	Hand	0.300433	0.042857	12	0.298051	0.042559	25	4.774273	0.445269	5	0.298051	0.042559	17
26	f24	Hand	0.300433	0.042857	13	0.298051	0.042559	26	4.774273	0.445269	6	0.298051	0.042559	18
27	f25	Foot	0.300433	0.012554	16	0.298051	0.012256	28	4.774273	0.142239	9	0.298051	0.012256	21
28	f26	Foot	0.300433	0.012554	17	0.298051	0.012256	30	4.774273	0.142239	10	0.298051	0.012256	
29	f27	Mouth	0.300433	0.027706	15	0.298051	0.027408	28	4.774273	0.293754	8	0.298051	0.027408	20
30	f28	Nose	0.300433	0.042857	14	0.298051	0.042559	27	4.774273	0.445269	7	0.298051	0.042559	19
31	o4	Frog	0.090909	0.090909		0.088825	0.088825		2.662338	2.662338		0.088825	0.088825	
32	f29	Eye	0.090909	0.006494	37	0.088825	0.006196	37	2.662338	0.899814	20	0.088825	0.006196	29
33	f30	Eye	0.090909	0.006494	38	0.088825	0.006196	38	2.662338	0.899814	21	0.088825	0.006196	30
34	f31	Mouth	0.090909	0.027706	33	0.088825	0.027408	33	2.662338	0.293754	22	0.088825	0.027408	24
35	f32	Foot	0.090909	0.012554	34	0.088825	0.012256	34	2.662338	0.142239	23	0.088825	0.012256	25
36	f33	Foot	0.090909	0.012554	35	0.088825	0.012256	35	2.662338	0.142239	24	0.088825	0.012256	26
37	f34	Foot	0.090909	0.012554	36	0.088825	0.012256	36	2.662338	0.142239	25	0.088825	0.012256	27
38	f35	Foot	0.090909	0.012554		0.088825	0.012256		2.662338	0.142239		0.088825	0.012256	28
39	o5	Cat	0.31039	0.31039		0.307412	0.307412		3.816327	3.816327		0.307412	0.307412	
40	f36	Eye	0.31039	0.006494	8	0.307412	0.006196	21	3.816327	0.899814	11	0.307412	0.006196	13
41	f37	Eye	0.31039	0.006494	9	0.307412	0.006196	22	3.816327	0.899814	12	0.307412	0.006196	14
42	f38	Ear	0.31039	0.042857	2	0.307412	0.042559	15	3.816327	0.445269	13	0.307412	0.042559	9
43	f39	Ear	0.31039	0.042857	3	0.307412	0.042559	16	3.816327	0.445269	14	0.307412	0.042559	10
44	f40	Leg-join	0.31039	0.088312	1	0.307412	0.088014	14	3.816327	0.081633	19	0.307412	0.088014	8
45	f41	Foot	0.31039	0.012554	6	0.307412	0.012256	19	3.816327	0.142239	17	0.307412	0.012256	
46	f44	Foot	0.31039	0.012554	7	0.307412	0.012256	20	3.816327	0.142239	18	0.307412	0.012256	
47	f45	Tail	0.31039	0.027706		0.307412	0.027408		3.816327	0.021027		0.307412	0.027408	
48	f46	Mouth	0.31039	0.027706	5	0.307412	0.027408	18	3.816327	0.293754	16	0.307412	0.027408	12
49	f47	Nose	0.31039	0.042857	4	0.307412	0.042559	17	3.816327	0.445269	15	0.307412	0.042559	11

Table 1: Example tabulated output values for the objects and features depicted in Figures 5a-d.

Human eye-gaze paths have a notable tendency to concentrate on facial representations where they appear. These are revisited frequently, even where they are largely devoid of apparent features (see Figure 1, top right). We might conjecture that the human brain has evolved to differentially attend to these most salient of areas and will record and search for distinguishing features as a priority. An artificial perceptual system might similarly represent facial areas with a high density of feature detail, and artificially raise the salience of these types of features.

It is clear that a single fixation (glance) at an object within the greater foveal area is sufficient for immediate recognition, so the question arises as to why human gaze returns to a face (or any object) for detailed inspection with the *fovea centralis*? We might conjecture that these periods of detailed inspection are related to the building and on-going maintenance of the feature cloud description, making small adjustments and improvements to the representation.

9 Modes of Inattention

The next two sub-sections discuss some interesting apparent lapses in attention. As with much of the human perceptual process, such lapses and anomalies can be highly informative, identifying underlying mechanisms where they were previously hidden from investigative view. Two classes of inattentional lapse have been subject to much experimental investigation recently, change blindness (e.g. Simons and Rensink 2005 for review) and inattentional blindness (Mack and Rock 1998). Each is described briefly, and discussed in the context of the abductive perception and attention model. There may be other modes of inattention, but we are, as yet, blind to them.

9.1 Change Blindness

The phenomenon of "change blindness" (CB) refers to instances where changes in a visual scene, which would normally be expected to initiate a shift of attention are ignored, with the observer apparently unaware that the visual change occurred. There are several related effects, which will be broadly characterized here according to the timescale over which they occur. We will consider two such ranges, one operating over a time range commensurate with a saccade and fixation (~300ms), the equivalent of one perceptual cycle, and another where the changes are measured in seconds. Each is demonstrated with a range of ingenious experiments.

It has been widely reported that significant changes to an otherwise static image that occur during even brief periods of visual disruption are substantially less likely to be reported. Such disruptions can arise as a natural consequence of the human visual system, such as a blink (O'Regan *et al* 2000) or saccade (Blackmore *et al* 1995, Grimes 1996), or can be experimentally induced by interrupting the image field briefly.

The *flicker paradigm* (Rensink *et al* 1997) illustrates this. Participants are shown an image that changes in some semantically material respect (i.e. a significant object has been added or removed by photo-manipulation), but where this change is made during a brief period (e.g. 80ms) when the image is replaced by a neutral (grey) or patterned disrupter field. Without the disruption changes are noticed readily, but this falls dramatically when disruption is introduced. The original and altered images may be shown alternately for an extended period, with each transition masked by the disruption, until the change is reported or the experiment ends.

Given that we have proposed a working memory structure, Δp, it is reasonable to question whether change is detected by a direct and immediate comparison of current object type (in Δi) and the *immediately* preceding object type (in Δp) at the equivalent location. Both representations co-exist at step 10 of the perceptual process (section 6) and comparisons are possible then. If a comparison is to be made, then it is only made between immediately consecutive object types at any given location in Δp. This is consistent with the data for the flicker experiments. If the change is made between successive image "frames" the change is detected, any disrupter event swamping the attention mechanism with multiple changes.

The notion of a visual memory in humans is contentious (e.g. Simons and Rensink 2005). Wolfe (1999) has dubbed the change blindness phenomena *inattentional amnesia*, on the presumption that these changes are not noted because no memory has been retained of the visual stimuli, so no comparison might be made: "*visual representation has no memory. It exists solely in the present ...*". In one sense, this appears largely so, the immediate veridical sensation of sight disappears immediately the eye or imager is obscured – as Δi is lost. The idea that such a structure might exist within the human perceptual system to build a composite visual scene over several saccade/fixation pairings is also largely unsupported by experimental evidence (Hollingworth and Henderson 2002). Hollingworth *et al* (2001) have reported that changes are predominantly detected only at or near the fixation point. This is commensurate with the notion that change detection under these circumstances is the prerogative of comparison at the conceptual level, the current overtly attended item and a previously recorded one. Given that changes are sometimes noted outside the foveal area, secondary items presumably share this property as well.

However, experiments in which seemingly large changes were made to an image, but gradually over an extended time period (10+ seconds), say to the colour of a prominent object or by the gradual introduction/removal of an object, are also largely ignored (e.g. Simons *et al* 2000). It appears from these and similar image change experiments that human perception relies heavily on specialized *rate of change* detectors to draw attention to change and that these may be strongly biased to luminance rather than colour change (Theeuwes 1995). Such change detectors can only reasonably function during fixations and are not activated when the rate of change in the stimulus is low, and are apparently reset by blink or saccadic image motion or where their operation is disrupted by masking (Blackmore *et al* 1995), when no

attention is triggered and gaze shift not precipitated. Change detection under normal (i.e. uninterrupted) circumstances is therefore a form of exogenous attention.

This appears consistent with *mudsplash* experiments (e.g. O'Regan *et al* 1999), in which an uninterrupted salient object change is made – which would normally be noticed – but which is simultaneously accompanied by a flashed "mudsplash", an irregular area of high visual contrast. Noticing of the semantic change is greatly inhibited, the exogenous (change) attention mechanism having directed the overt locus of attention to the mudsplash area.

In the seminal long timescale CB experiment described by Simons and Levin (1998), a passer-by was approached by the experimenter and asked for directions. During the reply, two "workmen" carrying a door passed between the two and the conversation continued. On questioning approximately half the participants reported that they had not noticed that the experimenter had been exchanged for another (not particularly similar) person during the interruption. Low rates of noticing have also been reported for film clips in which actors are substituted or objects change between scene cuts (Levin and Simons 1997).

A conceptual or object level memory of recently attended items passed to the cognitive layer at least serves to explain those instances where changes *are* noticed, change blindness is by no means an absolute or inevitable phenomena. Simons and Levin (1998) reported that levels of noticing varied greatly. It might be argued, therefore, that the level of abstraction of items attended to and reported is high, but not uniform. Matching is type to type, recording only that it was of type "face" guarantees an inability to detect the substitution, while a record of hair colour or other facial characteristics recorded from the object model enables, though does not guarantee, change detection.

The question remains as to why human perception apparently relies so little on visual memory – and whether a robot should do likewise. Change is endemic in visual scenes. To compare details in the current scene with preceding ones invites a flood of items for attention and explanation, so reliance on the immediate interpretation remains the appropriate option unless there are specific, task related, reasons for doing otherwise.

9.2 Inattentional Blindness

Inattentional Blindness is wittily demonstrated in a classic experiment due to Simons and Chabris (1999), in which participants are required to attend to a demanding task of counting the number of ball passes between members of a basketball team. In attending closely to the task, a significant proportion of viewers fail to note the appearance of an actor dressed in a gorilla costume, walk slowly across the scene, stopping centre stage to beat his chest. The effect may also be achieved under laboratory conditions with synthetic images (Most *et al* 2000), in which an otherwise highly prominent distracter pattern is ignored while conducting a demanding attentional task.

These tasks are characterized by a high attentive load to a single or small number of items, a single instance of the ball object in Simons and Chabris (1999). They are appropriately modelled by very short or single item list heads. Under these conditions both competing objects, which would normally be easily attended to, and all but the most salient exogenous events are excluded from the attention process. Inhibition of return is suspended – with only one item in the list head the IOR mechanism is effectively undefined. The secondary attention mechanism is, by definition, unavailable – a form of attentional (as opposed to physiological) "tunnel vision". Equally, as there is no pre-attentive item, inter-object saccades are suppressed and eye movements adopt a pursuit strategy, tracking the single remaining attentive item. Under this interpretation, the phenomena might be better described as *hyper-attentional neglect*, as no other objects but those stipulated by the experimenter and attended to by the participant are considered for the duration of the trial.

10 Summary and Conclusions

We have presented a description of our formally described, if largely theoretical, approach to robot perception and attention based on ideas of abductive and deductive reasoning (section 2) coupled to a feature cloud representation of object models. We have focussed on the notion

of a visual object, embodied in space and represented as a feature cloud (section 3), which gives rise to specific and distinctive effects on feature detectors (section 4). We describe a perceptual process (section 6) that creates hypotheses (Δh and Δi) to "explain" the visual sensor data in terms of individual feature cloud object models (individually represented in Σo). This process is guided by several measures (appendix 2) derived from the semantics of the object definitions. We define a form of transient object memory (section 5) that serves several roles, including inhibition of return. We assume a foveal imaging system, which restricts accurate identification of objects to the centre of the visual image field, but allows partial hypotheses to be formed in the periphery. One aspect of attention defines how the fovea is directed to different objects to serialize the perceptual process, the scanpath.

This is the basis of the attentional process model (section 7), in which object preferences or predispositions (section 7.1) are used to order the importance or salience of objects and so define how they will be visited (section 7.2). We briefly discuss the management of such preference orderings (section 7.3). The resulting system is intended to isolate aspects of perception and attention that are germane to an artificial or robot model. Each design stage has been inspired by biological (primarily human) considerations, but it is not principally a model of human attention. Rather it is an artifice that shares many properties with the system it is based on, but which does not seek to emulate it exactly.

This has then been used to motivate a discussion of how small modifications in the management of these preference orderings, coupled to notions of inhibition of return and a partial memory model, can be used to emulate within the artifice a range of both attention and inattention phenomena (section 9) experimentally observed in the human perceptual system (section 8).

We consider six such attention related phenomena. First, the foveal scanpath (section 8.1), autonomic control of eye movements to model notions of overt attention. Second, notions of covert attention (section 8.2), recognising that we may be aware of more items than just the item of overt attention. Third, exogenous attention (section 8.3), the role of sensor-based events to capture the attention mechanism and direct it independently of autonomic (or volitional) gaze movements. Fourth, endogenous attention (section 8.4), the role of task directed attention to direct the scanpath according to the requirements of a cognitive control layer. Fifth, change blindness, a range of situations in which the attentional system apparently demonstrates unexpected lapses of attention, using the model to discuss how this might be approached. In section 9, we briefly considered inattentional blindness, where seemingly highly conspicuous visual events are unexpectedly ignored due to "hyper" attention to one item.

Our computationally motivated model and discussion here can only engage a sub-set of the plethora of individual phenomena that make up the total animal and human attentional system. We hope that it will act as a catalyst and framework to consider attention as a complex, but highly coordinated bridge between the underlying perceptual systems, both informing and being informed by cognitive control centres. The use of eye-gaze scanpath strategies, largely assumed here, are not conducive to ready analysis and are largely avoided by the psychological attention community, who prefer to concentrate on highly controlled experimental conditions and quantitative measures of performance, such as response time, to conduct their research. While precise in their individual measurements it is often hard to relate the findings under these highly artificial conditions back to the consequences for a broader view of the attentional processes as they might be applied in the synthesis of an artificial model for robot control or to gain an overall understanding of the human attentional system (Sutherland 1998, for instance).

Acknowledgements: This research is supported by EPSRC under grant EP/C530683/1, "Abductive Robot Perception: Modelling Granularity and Attention in Euclidean Representational Space".

Appendix One: Representing Preferences

In this appendix we establish a formal notion of preference ordering defined on a set of types, $Type_1, Type_2, \ldots, Type_{n-1}, Type_n$. Let $\succeq$ be a reflexive, transitive relation over a set of types encoded in Σ_O. Then we say α is preferred to β with respect to (wrt) criterion ϕ, $\alpha \succeq_\phi \beta$ if and only if wrt to ϕ, α is greater than or equal to β. For example, in the case where ϕ is taken as the *dv* measure, then: $\alpha \succeq_{dv} \beta \equiv dv(\alpha) \geq dv(\beta)$. Strict and equal preferences are respectively defined as follows:

$\alpha \succ_\phi \beta \equiv \alpha \succeq_\phi \beta \;\&\; \neg(\beta \succeq_\phi \alpha)$

$\alpha \simeq_\phi \beta \equiv \alpha \succeq_\phi \beta \;\&\; \beta \succeq_\phi \alpha$

The relations $\succ_\phi$ and $\succeq_\phi$, respectively define a total and partial ordering over the type values. We now wish to define a total ordering of type elements and sets of types: $<S_1, S_2, \ldots, S_{n-1}, S_n>$, the elements of each set having equal preference and whose elements between each S_i and S_{i+1} indexed set mean that all the elements of S_{i+1} are strictly preferred to the elements of S_i, thus:

$S\succ_\phi = <S_1, S_2, \ldots, S_{n-1}, S_n>$

where: $\forall\alpha_i \in S_i, \forall\alpha_j \in S_j \;\; (\alpha_i \succ_\phi \alpha_j)$ and $\forall\alpha_i \in S_i, \forall\alpha_j \in S_i \;\; (\alpha_i \simeq_\phi \alpha_j)$

For example, given: $S\succ_\phi = <O_1, \{O_2, O_3\}, O_4>$, shall be interpreted to mean that wrt ϕ, O_1 is strictly preferred over O_2 and O_3, with these strictly preferred to O_4. In the case where an arbitrary preference order is imposed on the elements of S (independent of their respective *dv* values) this will be denoted as: $S\succ_\alpha$, where α is some indexed property. While not necessarily the case, the indexing property used (e.g. α) may be a real number ordering. By convention here, higher positive values will indicate a greater degree of preference.

Appendix Two: Definitions of Measures

In this appendix we define the three functions: *dv(Type)*, *ro(Type)* and *ev(x,v)* used to compute the degree to which any hypothesized object, *x*, as seen from viewpoint $\underline{v}$, explains and predicts currently available sensor data. In each case, we measure *symbolic* information encoded in *wffs* and in particular encoded in *wffs* of the form: $Object(x_i, \underline{v_i}, Type_i, \ldots)$, $Feature(f_i, \underline{v_i}, Type_i, \ldots)$, and $Appearance(a, \underline{v}, Type, \ldots)$ in the logical formalism used. Here, in the interest of brevity, we simply give a brief description of these functions and how they are used. For a full description see (Randell and Witkowski 2006).

A2.1 Distinctiveness Value

The distinctiveness value of a feature *Type*: *dv(Type)* measures the proportion of feature instances of type, *Type*, encoded within the objects in Σ_O against all the features of any type, also encoded in Σ_O. This is an *a priori* measure of the rarity of the feature type. Formally, we are measuring the proportion of individual feature variables $f_1, \ldots, f_n$, of a given *Type*, encoded in Σ_O (i.e. in *wffs* of the form: $Feature(f_i, \underline{v_i}, Type, \ldots)$) against feature variables of any feature type. Let $S(F)=\{f_i | Feature(f_i, \underline{v}, F, [\ldots]) \in \Sigma o\}$, and $S=\{f_i | Feature(f_i, \ldots) \in \Sigma o\}$, and let $|S|$ denote the cardinality of set *S*. Then $dv(F) = 1 - (|S(F)| / |S|)$. The distinctiveness of an object or object-part is derived by propagation from the calculated values of its features (section 4).

A2.2 Explanatory Value

The explanatory value (*ev*), the degree to which a specific hypothesized object *o* in Δi is supported by the sensor data, is defined as:

$ev(o) = (P+Q) - (R+S) / (P+Q+R+S); \;\; -1 \leq ev(o) \leq 1$

where:

$P = [\Sigma | \{fi\}| \times dv(\Phi): [Feature(fi, \ldots, \Phi, \ldots) \;\&\; Expected(fi, \underline{v0}) \;\&\; Detected(fi, \underline{v0})]]$

$Q = [\Sigma | \{fi\}| \times dv(\Phi): [Feature(fi, \ldots, \Phi, \ldots) \;\&\; \neg Expected(fi, \underline{v0}) \;\&\; \neg Detected(fi, \underline{v0})]]$

$R = [\Sigma |\{fi\}| \times dv(\Phi)\text{: } [Feature(fi,\ldots, \Phi,\ldots) \,\&\, Expected(fi,\underline{v0}) \,\&\, \neg Detected(fi,\underline{v0})]]$
$S = [\Sigma |\{fi\}| \times dv(\Phi)\text{: } [Feature(fi,\ldots, \Phi,\ldots) \,\&\, \neg Expected(fi,\underline{v0}) \,\&\, Detected(fi,\underline{v0})]]$

Here P represents the sum of instances where a feature (fi of type Φ) is *Expected* in the object hypothesis projection and is matched (by both type and location) by a *Detected* observation in the sensor data stream. Q denotes the expectation of no feature coupled to no data. R denotes an expectation unmatched by a corresponding feature and S sensor data detected without any matching expectation. P and Q support the hypothesis (tending to +1); R and S tend to refutation (tending to -1). P, Q, R and S are weighted according to the distinctiveness (*dv*) of the expected feature, providing a partial ordering of significance.

A2.3 Rank Order

The remaining measure *rank order ro(Type)* measures the *a posteriori* likelihood that object x of type *Type* defined in Σ_O causally explains the available sensor data in Γ. Rank order measures the number of possible *substitutions* (of terms for variables in *wffs)* that match interpreted sensor data items to individual features in Σ_O, *Type* for *Type*. It is the ratio of all features of the object that have a corresponding appearance in $\Gamma(\Omega)$ to all the features of the matched object (Ξ).

$$ro(Object) = \frac{\Sigma\ |\{f_i\}| \times dv(F_i)\text{: } [Feature(f_i,\ldots,F_i,\ldots) \,\&\, \Omega]}{\Sigma\ |\{f_j\}| \times dv(F_j)\text{: } [Feature(f_j,\ldots,F_j,\ldots) \,\&\, \Xi]}$$

This process is restricted so that at most one interpreted feature f_i belonging to an object definition x, is matched to exactly one appearance of that feature detected in the sensor data stream Γ. Note the features are weighted by their distinctiveness here.

References

Aloimonos J, Weiss I and Bandyopadhyay A 1987 Active Vision *Proc. 1st Int. Conf. on Computer Vision* 35-54

Aziz M Z, Mertsching B, Shafik M and Stemmer R 2006 Evaluation of visual attention models for robots *Proc ICVS 2006* 6pp

Ballard D H 1991 Animate Vision *Artificial Intelligence* **48** 57–86

Becker W 1991 Saccades (ed) Carpenter R H S *Eye movements* Macmillan Press 95-137

Blackmore S J, Brelstaff G, Nelson K and Troscianko T 1995 Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes *Perception* **24** 1075-81

Breazeal C, Edsinger A, Fitzpatrick P, Scassellati B and Varchavskaia P 2000 Social constraints on animate vision *IEEE Intelligent Systems* **15**-4 32-37

Brockmole L R, Carlson L A and Irwin D E (2002) Inhibition of attended processing during saccadic eye movements *Perception & Psychophysics* **64**-6 867-81

Bushnell P J (1995) Behavioral approaches to the assessment of attention in animals *Psychopharmacology* **138** 231–59

Cédras C and Shah M 1995 Motion-based recognition: a survey *Image and Vision Computing* **13**-2 129-55

Chella A, Frixione M and Gaglio S 2000 Understanding dynamic scenes *Artificial Intelligence* **123**(1-2) 89-132

David P, DeMenthon D, Duraiswami R and Samet H 2004 SoftPOSIT: Simultaneous Pose and Correspondence Determination *Int. J. Computer Vision* **59**(3) 259-84

Davison A J 2003 Real-time simultaneous localisation and mapping with a single camera *Proc ICCV-03* 8pp

DeMenthon D F and Davis L S 1995 Model-based object pose in 25 lines of code *Int. J. Computer Vision* **15** 123-41

Duchowski A T 2003 *Eye Tracking Methodology Theory and Practice* London: Springer

Duncan J 1984 Selective attention and the organisation of visual information *J. Experimental Psychology: General* **113**-4 501-17

Duncan J 2006 Brain mechanisms of attention *Quarterly Review of Experimental Psychology* **59**-1 2-27

Franconeri S L, Hollingworth A and Simons D J 2005 Do new objects capture attention? *Psychological Science* **16**-4 275-81

Gibson J J 1979 *The ecological approach to visual perception* Boston: Houghton Mifflin Co.

Grimes J 1996 On the failure to detect changes in scenes across saccades (ed) Akins K *Perception: Vancouver studies in cognitive science* Vol. 5 Oxford University Press 89–110

Grossberg S, Roberts K, Aguilar M and Bullock D 1997 A neural model of multimodal adaptive saccade eye movement control by superior colliculus *J Neuroscience* **17**-24 9706-25

Haralick R, Lee C, Ottenberg K and Nolle M 1994 Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem *Int. J. Computer Vision* **13** 331-56

Hayhoe M and Ballard D 2005 Eye movements in natural behavior *Trends in Cognitive Science* **9**-4 188-94

Helmholtz H v 1865 *Treatise On Physiological Optics* Dover publications (2005, in facsimile)

Henderson J M 2003 Human gaze control during real-world scene perception *Trends in Cognitive Sciences* **7**-11 498-504

Hochstein S and Ahissar M 2002 View from the top: hierarchies and reverse hierarchies in the visual system, *Neuron* **36** 791-804

Hodgson T L, Golding C, Molyva D, Rosenthal C R and Kennard C 2004 Eye movements during task switching: reflexive, symbolic, and affective contributions to response selection *Journal of Cognitive Neuroscience* **16** 318-30

Hoffman J E 1998 Visual attention and eye movements (ed) Pashler H *Attention* Hove: Psychology Press 119-54

Hollingworth A and Henderson J M 2002 Accurate visual memory for previously attended objects in natural scenes *J. Experimental Psychology: Human Perception and Performance* **28**-1 113–36

Hollingworth A, Schrock G and Henderson J M 2001 Change detection in the flicker paradigm: The role of fixation position within the scene *Memory & Cognition* **29**-2 296-304

Horaud R, Dornaika F, Lamiroy B and Christy S 1997 Object Pose: The Link Between Weak Perspective, Paraperspective, and Full Perspective *Int. J. Computer Vision* **22**(2) 173-89

Horowitz T S, Fine E M, Fencsik D E, Yurgenson S and Wolfe J M 2006 Fixational eye movements are not an index of covert attention *Psychological Science* accepted, to appear (preprint)

Hu Z Y and Wu F C 2002 A Short Note on the Number of Solutions of the Noncoplanar P4P Problem *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(4), pp. 550-555

Huemer M 2004 Sense data *Stanford Encyclopedia of Philosophy*, summer 2004, at:
http://plato.stanford.edu/entries/sense-data/

Hurlburt R T, Koch M and Heavey C L 2002 Descriptive experience sampling demonstrates the connection of thinking to externally observable behavior *Cognitive Therapy and Research* **26**-1 117-34

Itti L and Koch C 2001 Computational modelling of visual attention *Nature Reviews: Neuroscience* **2** March 2001 1-10

Itti L, Koch C and Niebur E 1998 A model of saliency-based visual attention for rapid scene analysis *IEEE Trans. Pattern Analysis and Machine Intelligence* **20** 1273-76

James W 1890 *The Principles of Psychology* (Dover Publications, 1950, 2 volumes in facsimile)

Just M A and Carpenter P A 1976 Eye fixations and cognitive processes *Cognitive Psychology* **8**-4 441-80

Kahneman D, Treisman A and Gibbs B 1992 The reviewing of object files: object-specific integration of information *Cognitive Psychology* **24** 175-219

Khadhouri B and Demiris Y 2005 Compound effects of top-down and bottom-up influences on visual attention during action recognition *Proc. IJCAI-05* 1458-63

Klein R M and MacInnes J 1999 Inhibition of return is a foraging facilitator in visual search *Psychological Science* **10**-4 346-52

Kowler E, Anderson E, Dosher, B and Blaser, E 1995 The role of attention in the programming of saccades *Vision Research* **35**-13 1897-916

Lee T S and Mumford D 2003 Hierarchical Bayesian Inference in the Visual Cortex *J. Opt. Soc. America* **A-20**(7) 1434-48

Lee T S and Yu S X 1999 An information theoretic framework for understanding saccadic eye movements *Proc 1st Conf. on Neural Information Systems* MIT Press 7pp

Levin D T and Simons D J 1997 Failure to detect changes to attended objects in motion pictures *Psychonomic Bulletin and Review* **4** 501-6

Linsen L 2001 Point cloud representation *Tech report* Faculty of Computer Science, U of Karlsruhe

Lowe D G 2004 Distinctive image features from scale-invariant keypoints *Int. J. Computer Vision* **60**(2) 91-110

Mack A and Rock I 1998 *Inattentional Blindness* Cambridge, MA: MIT Press

Marr D 1982 *Vision* New York: W.H. Freeman & Co.

Most S B, Simons D J, Scholl B J and Chabris C F 2000 Sustained inattentional blindness: the role of location in the detection of unexpected dynamic events *PSYCHE* **6**-14 at:
http://psyche.cs.monash.edu.au/v6/psyche-6-14-most.html

Navalpakkam V and Itti L 2002 A goal oriented attention guidance model *LNCS-2525* Springer-Verlag 453-61

Noton D and Stark L 1971 Eye movements and visual perception *Scientific American* **224** 34-43

O'Regan J K, Deubel H, Clark J J and Rensink R A (2000) Picture changes during blinks: looking without seeing and seeing without looking *Visual Cognition* **7** 191-212

O'Regan J K, Rensink R A and Clark J J (1999) Change blindness as a result of "mudsplashes" *Nature* **398** (4 March 1999) 34

Pashler H 1998 *The Psychology of Attention* MIT Press

Pirri F 2005 The usual objects: a first draft on decomposing and reassembling familiar objects images *Proc XXVII Ann. Conf. of the Cognitive Science Society* 1773-8

Posner M I and Petersen S E 1990 The attention system of the human brain *Ann. Rev. Neuroscience* **13** 25-42

Pylyshyn Z 2001 Visual indexes, preconceptual objects, and situated vision *Cognition* **80** 127-58

Pylyshyn Z and Storm R W 1988 Tracking multiple independent targets: evidence for a parallel tracking mechanism *Spatial Vision* **3** 179-97

Radke R J, Andra S, Al-Kofahi O and Roysam B 2005 Image change detection algorithms: a systematic survey *IEEE Transactions on Image Processing* **14**-3 294-307

Randell D A and Witkowski M 2006 Abductive visual perception with feature clouds *Proc. KR-06* 352-61

Rao R P N and Ballard D H 1995 An active vision architecture based on iconic representations *Artificial Intelligence* **78** 461-505

Rao R P N, Zelinsky G J, Hayhoe M M and Ballard D H (1997) Eye movements in visual cognition: a computational study *Technical Report 97.1* Department of Computer Science, University of Rochester, March 1997

Rensink R A 2000 The dynamic representation of scenes *Visual Cognition* **7**(1/2/3) 17-42

Rensink R A 2002 Change detection *Ann. Rev. Psychol* **53** 245-77

Rensink R A 2004 Visual sensing without seeing *Psychological Science* **15** 27-32

Rensink R A, O'Regan J K and Clark J J 1997 To see or not to see: the need for attention to perceive changes in scenes *Psychological Science* **8** 368-73

Rock I 1981 *The Logic of Perception* Cambridge, MA: MIT Press

Schmid C and Mohr R 1997 Local grayvalue invariants for image retrieval *IEEE Trans. on Pattern Analysis and Machine Intelligence* **19**(5) 530-4

Scholl B J 2001 Objects and attention: the state of the art *Cognition* **80** 1-46

Shanahan M P 2002 A Logical Account of Perception Incorporating Expectation and Feedback *Proc. KR-02* 3-13

Shanahan M P and Randell D A 2004 A Logic-based formulation of active visual perception *Proc. KR-04* 64-72

Shibata T, Vijayakunar S, Conradt J and Schaal S 2001 Biomimetic oculomotor control *Adaptive Behavior* **9**(3-4) 189-207

Simons D J, and Chabris C F 1999 Gorillas in our midst: sustained inattentional blindness for dynamic events *Perception* **28** 1059-74

Simons D J and Levin D T 1998 Failure to detect changes to people during a real-world interaction *Psychonomic Bulletin and Review* **5** 644-9

Simons D J, Franconeri S L and Reimer R L 2000 Change blindness in the absence of visual disruption *Perception* **29** 1143-54

Simons D J and Rensink R A 2005 Change blindness: past, present, and future *Trends in Cognitive Sciences* **9**-1 16-20

Stark L W and Choi Y S 1996 Experimental metaphysics: The scanpath as an epistemological mechanism (eds.) Zangemeister W H et al *Visual Attention and Cognition* Elsevier Science 3-69

Sun Y and Fisher R 2003 Object-based visual attention for computer vision *Artificial Intelligence* **146** 77-123

Sutherland S 1998 Feature selection (review of Pashler 1998) *Nature* (26 March 1998) **392** 350

Tatler B W and Wade N J 2002 On nystagmus, saccades, and fixations *Perception* **32**-2 167-84

Taylor T L and Klein R M 1998 On the causes and effects of inhibition of return *Psychonomic Bulletin & Review* **5**-4 625-43

Theeuwes J 1993 Visual selective attention: a theoretical analysis *Acta Psychologica* **83** 93-154

Theeuwes J 1995 Abrupt luminance change pops out; abrupt color change does not *Perception and Psychophysics* **57** 637–644

Theeuwes J 2004 Top-down search strategies cannot override attentional capture *Psychonomic Bulletin & Review* **11**-1 65-70

Theeuwes J, Kramer A F, Hahn S and Irwin D E 1998 Our eyes do not always go where we want them to go: capture of the eyes by new objects *Psychological Science* **9**-5 379-85

Torralba A, Oliva A, Castelhano M S and Henderson J M 2006 Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search *Psychological Review* **113**-4 766-86

Trappenberg T P, Dorris M C, Munoz D P and Klein R M 2001 A model of saccade initiation based on competitive integration of exogenous and endogenous saccade control signals *J. Cognitive Neuroscience* **13**-2 256-71

Treisman A M and Gelade G 1980 A feature-integration theory of attention *Cognitive Psychology* **12** 97-136

Vecera S P, Behrmann M and McGoldrick J 2000 Selective attention to the parts of an object *Psychonomic Bulletin and Review* **7**-2 301-8

Vieira Neto H and Nehmzow U 2005 Automated exploration and inspection: comparing two visual novelty detectors *Int. J. Advanced Robotic Systems* **2**-4 355-62

Witkowski M and Randell D 2006 Modes of attention and inattention for a model of robot perception *Proc TAROS-06* 246-53

Wolfe J M 1994 Visual search in continuous naturalistic stimuli *Vision Research* **34** 1187-95

Wolfe J M 1999 Inattentional amnesia (ed.) Coltheart V *Fleeting Memories* Cambridge, MA: MIT Press 71-94

Yarbus A L 1967 *Eye movements and vision* New York: Plenum Press

Zentall T R and Riley D A 2000 Selective attention in animal discrimination learning *J General Psychology* **127**-1 45-66