Modes of Attention and Inattention for a Model of Robot Perception

Mark Witkowski **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 and compares several aspects of attention and awareness in the context of uniform field image (robotic) vision sensors and foveal (human) natural perception. It builds on a theory of abductive perception using feature clouds, a formal definition for a robot perceptual system, and proposes a unified model for bottom-up and top-down attention. It highlights some shortcomings in existing bottom-up models and presents a uniform solution to them. 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

It is one of the many persistent paradoxes of the human visual system that it both provides a precise mechanism, sensitive to minute changes and detail in a visual scene, while at the same time is capable of surprising lapses of perceptual awareness.

The first is manifest in our ability to inspect objects for flaws or small deviations from a norm, and to recognise individuated objects by very small differences from a class of otherwise similar instances. One such example would be our apparently innate ability to distinguish between and identify thousands (and more) of human faces.

The second is manifest in apparent lapses of "attention", when objects or events in full view appear to be completely overlooked, even when they seem to be of high significance to the observer, or be highly unusual and would normally be drawn into explicit awareness. 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).

Such lapses must be taken in the context of "normal" visual attention (e.g. Itti and Koch, 2001; Posner and Petersen, 1990, for reviews). From the evidence available, visual attention is not one process, but a portmanteau of activities, notably that attention may be both exogenously, sense, driven and endogenously, task, driven.

Models of exogenous ("low-level") attention have been proposed by e.g. Itti *et al*. (1998), and endogenous attention by Stark and Choi (1996). Robot based models of attention in perception have been proposed by, for example, Brazeal *et al*. (2000), Khadhouri and Demiris (2005), Shibata *et al*. (2001), and Vieira Neto and Nehmzow (2005).

The work described here is a part of the on-going Cognitive Robotics research programme at Imperial College London. In particular it develops our approach to the use of abductive reasoning for robot perception (Shanahan, 2002; Shanahan and Randell, 2004; Randell and Witkowski, 2006) in which we use *first order logic* to model core aspects of the processes of perception.

The goal of our research is to create theories of perception within this logical framework, and then to use these theories to reason about the consequences of design choices and to specify implementations based on these principles. Our stance might be broadly characterised as a model based *hypothetico-deductive* approach combined with *abductive inference* (reasoning from observations to possible causes). This combination of sensor and model driven processes accords with empirical data from visual psychophysics (e.g. Rock, 1981) and is related to notions of active vision (e.g. Aloimonos *et al*., 1987).

We make a strong *Assumption of Embodiment*, that the robot exists in a volumetric, material, world in which material things in that world give rise to changes in detectors (sensors) possessed by that agent. The task of the agent or robot is then to establish a single coherent explanation or interpretation for those detector effects in the context of the internal, conceptualised, background model that identify consistent interpretations of the data.

The first part of this discussion paper presents a summary overview of the approach; we present some more detailed aspects of the formalism we adopt in

sections 2 and 3. A detailed treatment may be found in Randell and Witkowski (2006).

The second part of the paper is given over to a discussion of issues of attention and inattention, in the context both of the model presented and previous models of attention. It then considers the role of a transient memory trace in understanding various modes of inattention, as reported in human perception experiments.

In the first part, section 2 introduces abductive perception, *feature clouds*, the underlying object description form used, and *feature detectors*, the interface between the physical and logical domains. Section 3 details the qualitative measures used to guide and inform the perceptual process. Section 4 describes this perception process in the context of the abductive framework. Section 5 describes the properties of a robot-centric memory trace. In the second part, section 6 considers how the dual roles of exogenous and endogenous attention may be modelled with this approach, and highlights some difficulties with existing models. Section 7 briefly considers inattentional blindness in the context of the model.

2 Part 1: Abductive Perception

The treatment of perception we develop 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 explain the observations. As expected, and implied here, the set of alternative explanations generated are rarely unique, we need an additional framework 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 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 world and the result of the robot's actions on it. 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 means 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). Greek letters are used to indicate sets of sentences and meta-logical predicate variables.

In order to measure how well our 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 an abduced 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 are discussed in more detail below.

Our representational 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.

We define and factor out particular subsets of sentences in Σ, and Δ as follows. Within the logic, these sentences take the form of *well-formed formulae* (*wffs*), which are sequences of symbols adhering to the formation rules (syntax) of the language. 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 preferred explanation (Δi), and (iii) Δp, which is the current robot-centric description of the world.

2.2 Feature Detectors

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

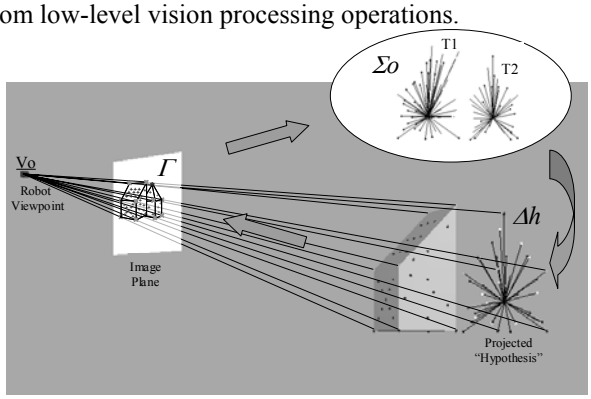

Figure 1: The main perceptual cycle

We assume 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. Detected

features must be reliably localizable onto the image plane (as indicated in figure 1), but need not be invariant on rotation or scale. We recognise that objects in the world may 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 2 for a notional "corner" feature from different viewpoints (above, level and below, representing instances from a set of all possible viewpoints) and at different resolutions, assuming Gaussian blurring towards the periphery. We refer to these variations as the *appearances* of a feature. Appearances are mapped directly (but non-uniquely) to the feature(s) they describe by abductive inference.

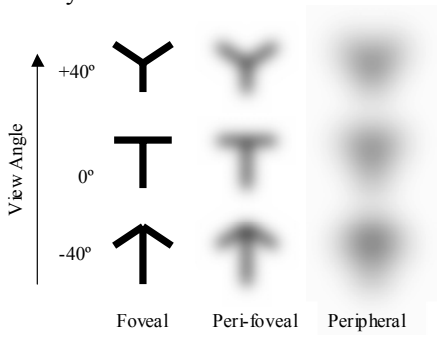

Figure 2: Appearances

All detected features take the form: *Appearance(name, vector, Type, [ ])*. Each different, *named*, appearance associated with a detector is assigned a class *Type* and a *vector* location on the image plane, which will serve to identify the source of each Γ assertion within the system.

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, low-level features may be composited into *compound features*, which increases their discriminatory capacity.

2.1 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. It may be contrasted to other model-based representations in visual perception-based applications such as generalised cylinders (Marr, 1982) or superquadrics (Chella *et al.*, 2000).

The cloud is partitioned into subsets of features that are pre-assigned to a set of inferred, 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 organised tree-structure where the subdivision of a host body 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 volumetric, and so bounded by opaque surfaces, represented as surface patches between the feature points.

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 1 shows feature cloud visualisations (Σo) from our Webots (www.cyberbotics.com) simulation, and a pair of objects, one with its surface representation (as viewed), one without.

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 between objects and their features, and between features and their appearances. Given an object, this is decomposed into object parts, or features; features are further decomposed into feature parts (compound features) or appearances or the empty list as a terminal node. *Type* is a class identifier for each object, feature or appearance.

(A1) $\Phi(x_0,\underline{v}_0,Type_0,[x_1,...,x_n]) \rightarrow$
$\Psi_1(x_1,\underline{v}_1,Type_1,[...]) \&...\& \Psi_n(x_n,\underline{v}_n,Type_n,[...])$,
where:
if Φ*=Object, then*
$\forall i\ i=1$ *to* n: Ψ_i*=Object or,*
$\forall i\ i=1$ *to* n: Ψ_i*=Feature, and*
if Φ*=Feature, then*
$\forall i\ i=1$ *to* n: Ψ_i*=Feature, or*
$\forall i\ i=1$ *to* n: Ψ_i*=Appearance,*
else, $\forall i\ i=1$ *to* n: $\Phi=\Psi_i$*=Appearance*

Axiom A1: Encoding the Feature Cloud

Vectors (always shown italic underlined, $\underline{v}$) locate the *centroid* (notional position) of any object, feature or appearance ($\underline{v}_1$ … $\underline{v}_n$) with respect to the centroid of its supervenient feature or object ($\underline{v}_0$), or of a viewpoint (actual or notional) in space. Using this scheme the position of any sub-part of an object may be determined relative to any other subpart by straightforward vector summation. Every logical operation between objects and their parts implies a corresponding vector operation. 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.

3 Measures

In this section we define the three functions: *dv(Type)*, *ro(Type)* and *ev(x,v)* used to compute the degree to which any hypothesised object, *x*, as seen from viewpoint *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: $Object(x_i,v_i,Type_i,...)$, $Feature(f_i,v_i,Type_i,...)$, and $Appearance(a,v,Type,...)$ 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).

3.1 Distinctiveness Value

The distinctiveness value of a feature *Type*: *dv(Type)* measures the proportion of feature instances of type, *Type*, encoded in Σ_O against all the features of any type, also encoded in Σ_O. This is an *a priori* measure. Here we are measuring the proportion of individual feature variables $f_1,...,f_n$, of a given *Type*, encoded in Σ_O (i.e. in *wffs* of the form: $Feature(f_i,v_i,Type,...))$ against feature variables of any feature type. Let $S(F)=\{f_i|Feature(f_i,\underline{v},F,[...]) \in \Sigma o\}$, *and* $S=\{f_i|Feature(f_i,...) \in \Sigma o\}$, *and let* $|S|$ denote the cardinality of set *S*. Then $dv(F) = 1 - (|S(F)| / |S|)$.

3.2 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:

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

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

3.3 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 Γ (Ω) to all the features of the matched object (Ξ).

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

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, and that the evaluation is independent of viewpoint.

4 The Perceptual Cycle

Figure 1 illustrates the main perceptual cycle, which is summarised here.

(1) Pre-process the image to identify all *detected* (section 3.2) appearances. Record them in the Γ Structure.

(2) Evaluate 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.

(3) 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 hypotheses. This is the abductive step – selecting hypotheses from partial evidence.

(4) Identify four or more non-coplanar matches between features in Γ and features of corresponding *Type* in a single object model. Our implementation uses 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. Write this set of *wffs* (representing all the features for that object) to the hypothesis space Δh for each hypothesised object. Update embedded vector fields to reflect this new projection from the robot's viewpoint. Each feature type (that is not self-occluded, or any other hypothesised object) is then *expected* (section 3.2) in the image plane at a place determined by this projection vector. This is the deductive (or prediction) step.

(5) Evaluate *ev* for each hypothesised object to determine the extent to which it explains the sensor data for the projection area it occupies on the image plane. An object hypothesis is strengthened by corroboration where a feature appearance is both *expected* and *detected*; and weakened where there is mis-match between prediction and sensor data. *ev* is updated according to the *dv* value of the feature being compared. This is the corroboration step.

(6) 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.

(7) Repeat from (3) until all the sensor data elements of Γ have a coherent explanation, and where all the ground hypotheses in Δh satisfy the domain constraints (Σc).

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

This combination of "positioned" object models and the raw sensor data appearances allows higher-level task modules to interrogate the perceptual system, either in terms of the original *sensor data*, or in interpreted terms, as *sense data*. For a foveal system, any apparent detail in the periphery of the scene must be a "reconstruction" – an "illusion" – derived from the model-based data.

The interpretations in Δi are, however, unordered, or at best sorted according to *ev*, the degree of evidence for each. This is very much the "engineering" solution. Part two of this paper considers how contextual issues may be used to change this process to reflect the functional needs of the robot.

5 The Panorama

The current interpretation, Δi exists solely in the present, when the detector stream (I) 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 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.

The panorama 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. 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. An *anticipatory transform* predicts the position the interpretations in Δp would appear in the next time step. As the information in Δp is already encoded spatially, a simple 3D matrix transform may be applied to the vector components 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:

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

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. (iii) The transform derived from imager displacement, as typified by the SLAM class of algorithm.

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). Type (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 (Davison, 2003). It might be presumed that 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, 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 some evidence to suggest 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 the 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 Part 2: Modes of Attention

It has long been demonstrated that attention has a dual nature. In part exogenous, driven by external events that direct the flow of awareness to specific areas or things within the perceptual field. In part endogenous, the object of awareness being selected internally by some other cognitive process within the agent, which focuses or modulates the perceptual system to those ends. In this section we argue that the Abductive Perception Model, with its hierarchical definition of object, part and feature provides a clean model for this dual nature of attention.

Itti *et al*. (1998) present a *saliency* based model of low-level attention, in which multi-scale, low-level feature extraction (intensity, edge orientation, colour, etc.) is uniformly applied to an image. This is used to build a combined feature map, and 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 scanned in saliency order by the process of attention. Once attended to, a temporary process of *inhibition of return* suppresses the salience of the last place, so the next most salient location is visited, and so on.

True models of the high-level processes of attention are less well represented and are generally incomplete. Stark and Choi (1996) present a model that emulates, using a Markov based approach, the eye-gaze path of a simulated human observer. Navalpakkam and Itti (2002) extend the Itti and Koch model with task based relevance.

Brazeal *et al*. (2000) suppress or intensify attributes of the attentive feature map to reflect changing "social" drives.

In the Abductive model, both low-level, feature driven and high-level, model-driven, attention may be emulated by manipulation of the interpretation ordering parameters (*ro* and *ev*) through the distinctiveness value (*dv*). This arises as a natural consequence of the hierarchical nature of object definitions in Σo. Recall that the rank order (*ro*) value of an object is used in the standard model to order the hypothesis testing process and that the explanatory value (*ev*) determines the attention order. This process is ordinarily based on the information content of the features (for *dv*) and the degree to which a model is supported (*ro*) by the prevailing sensor evidence, Γ.

6.1 Exogenous attention

Exogenous attention is detector, and hence feature driven. Where the standard, engineering, approach in the abductive model emphasises the informational distinctiveness to determine interpretation order, an attention based system would bias this order of interpretation according to the functional significance to the robot of the feature or of the object it implies.

Consider first a function *attendfeature(F,n)*, which raises the effect of the *dv* value of a particular feature type, *F*. This scheme is achieved by defining a *significance multiplier* (*sm*) associated with each feature type, such that all *dv(F)* terms in sections 3.2 and 3.3 are substituted by (*dv(F)*sm(F,n)*). 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 rank order (*ro*) value of each object in which the feature appears, whenever that feature type is detected. 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.

Conventionally, inhibition of return is applied retino-centrically to each instance of the feature detector, and 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, while the significance multiplier remains a property of the feature and its detector(s), inhibition is more naturally expressed as a property of object interpretations recorded in Δp. Inhibition of return consequently tracks with the interpreted objects. Expanding the definition of each object (axiom A1), coupled to its vector estimation, generates a *cone of inhibition*: the inhibited area when projected back onto the imaging plane. Application of the anticipatory transform (section 5) ensures that this cone remains directed at the area inhibited, regardless of rotation and translation motions of the robot and its imager. No special inhibition of return factor need be posited. Inhibition is inversely proportional to the recency of the object being attended to, as indicated by the image time, *t* (section 5).

6.2 Endogenous attention

Now 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 normally do. This is endogenous, model driven attention.

Endogenous attention is classically investigated as a visual search task, in which the participant is asked to search an image for an embedded specific but known target (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 3 (after Wolfe, 1994).

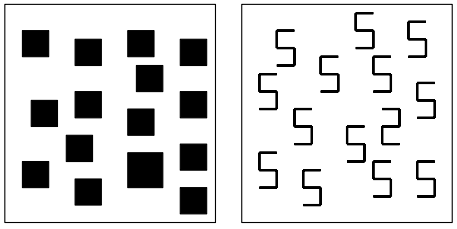

Figure 3: Search tasks – find the "odd one out"

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 targets. However, on the right hand test, eye gaze saccades show a marked (self-terminating) searching strategy, and search time rises linearly with the number of targets presented.

Uniform field and foveal/saccadic imagers give rise to different predictions. In the uniform field model, each detected feature is considered and explained by the appropriate model, and, as all elements of Γ are at the same resolution, every instance of the target model(s), as enhanced, are considered in one time frame, and, in both cases, the target identified.

In a foveal system, off-centre targets are represented by low spatial resolution feature detectors (figure 2, right). Where these characterising features for the different target types are distinct, as would be the case for figure 3 (left), attention, and hence gaze, may be directed immediately. Where the characterising features are not distinguishable at the peripheral resolution (as might be the case for figure 2, right), attention must be directed sequentially, such that the detailed foveal model is brought over the candidate targets in order.

Several things are clear from this. First, attention remains directed to the task, as other items in peripheral vision are ignored. Second, it clearly demonstrates that

low-level 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.

7 Modes of Inattention

This section discusses some interesting apparent lapses in attention. 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 model for its applicability to robots. There may be other modes of inattention, but we are apparently, as yet, blind to them.

7.1 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, the observer apparently completely unaware that the visual change occurred. There are several related effects, each demonstrated with ingenious experiments.

A classic CB experiment is described by Simons and Levin (1998), in which 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 then the conversation continued. On questioning approximately half the participants reported that they 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).

Change blindness has also been observed on a shorter time scale. Several experimenters (e.g. Blackmore *et al*., 1995) have noted that changes to images that would normally be noticed easily are missed if they are made during a saccade, or while a "blocking" patch was displayed over the area of change.

At the other end of the scale, experiments in which apparently 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).

Changes in the structure of scenes are apparently not routinely detected as a consequence of a detailed comparison between the current interpretation of the viewed scene (as Δi) and elements stored in a visual working memory (as Δp). The notion of a visual memory in humans is contentious (e.g. Simons and Rensink, 2005). Wolfe (1999) has dubbed the CB 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 tense.*" This appears largely so, the immediate veridical sensation of sight disappears immediately the eye or imager is obscured – as Δi is lost.

However, we have argued for a temporary, robot-centric, memory structure (Δp) on the grounds that it brings computational (seeding hypotheses into Δh) and functional (it provides a sense of situational awareness) advantages; in addition to its role in the inhibition of return. The contents of the structure, while temporary, are not transient, and so may not equate directly to the prevailing notion of "visual memory".

Such a memory at least explains those instances where changes *are* noticed, change blindness is not an absolute phenomena. Simons and Levin (1998) reported levels of noticing varied greatly. It might be argued, therefore, that the level of abstraction of items written into Δp is high, but not uniform. Matching is Type to Type, recording only that it was of Type "face" in Δp guarantees an inability to detect the substitution, while a record of hair colour or other facial characteristics recorded from the object model enables change detection.

It appears from the image change experiments that human perception relies heavily on highly specialised *rate of change* detectors to draw attention to change. These can only reasonably function during fixations (the 200-400mS periods of eye stability between saccades). Where their operation is disrupted by masking (Blackmore *et al*, 1995), or the rate of change falls below the detector threshold (Simons *et al*., 2000), no attention is triggered, and shift of attention not precipitated – save by the less effective memory comparison route.

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

7.2 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).

We consider this as an extreme form of endogenous attention, in which the value of the *attendobject* multiplier (section 6.2) is raised to such an extent that system is made to expend all its perceptual resources processing a single object model, to the exclusion of all others. Where only a single object is attended to in this way the inhibition of return mechanism is suspended, there is nothing else to "return" to. If this represents a true interpretation of the phenomenon it might better be

described as *hyper-attentional neglect*, as no other objects are even considered.

8 Summary

We have presented a summary description of our formally described, if largely theoretical, approach to robot perception based on abductive reasoning and a feature cloud representation of object models. This has then been used to motivate a discussion of how apparently small modifications to the scheme can be used to parallel a range of attention and inattention phenomena experimentally observed in the human perceptual system.

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

References

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

Blackmore S.J., Brelstaff, G., Nelson, K., Troscianko, T. (1995). Is the Richness of Our Visual World an Illusion? Transsaccadic Memory for Complex Scenes, *Perception*, **24**, pp. 1075-1081.

Breazeal, C., Edsinger, A., Fitzpatrick, P., Scassellati, B. and Varchavskaia, P. (2000). Social Constraints on Animate Vision, *IEEE Intelligent Systems*, **15**-4, pp. 32-37.

Chella, A., Frixione, M. and Gaglio, S. (2000). Understanding Dynamic Scenes, *Artificial Intelligence*, **123**(1-2), pp. 89-132.

Davison, A.J. (2003). Real-Time Simultaneous Localisation and Mapping with a Single Camera, Proc *ICCV-03*, 8pp.

De Menthon, D.F. and Davis, L.S. (1995). Model-Based Object Pose in 25 Lines of Code, *Int. J. Computer Vision*, **15**, pp. 123-141.

Itti, L. and Koch, C. (2001). Computational Modelling of Visual Attention, *Nature Reviews: Neuroscience*, **2**, March 2001, pp. 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**, pp. 1273-1276.

Khadhouri, B. and Demiris, Y. (2005). Compound Effects of Top-down and Bottom-up Influences on Visual Attention During Action Recognition, proc. *IJCAI-05*, pp. 1458-1463.

Levin, D.T., & Simons, D.J. (1997). Failure to detect changes to attended objects in motion pictures, *Psychonomic Bulletin and Review*, **4**, pp. 501-506.

Lowe, D.G. (2004). Distinctive Image Features from Scale-Invariant Keypoints, *Int. J. Computer Vision*, **60**(2), pp. 91-110.

Mack, A. and Rock, I. (1998). *Inattentional Blindness*, MIT Press.

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

Most, S.B., Simons, D.J., Scholl, B.J., & Chabris, C.F. (2000). Sustained Inattentional Blindness: The Role of Location in the Detection of Unexpected Dynamic Events, *PSYCHE*, **6**-14.
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, pp. 453-461.

Posner, M.I. and Petersen, S.E. (1990). The Attention System of the Human Brain, *Ann. Rev. Neuroscience*, 13, pp. 25-42.

Randell, D.A. and Witkowski, M. (2006). Abductive Visual Perception with Feature Clouds, Proc. *KR-06*, pp. 352-361.

Rock, I. (1981). *The Logic of Perception*, MIT Press.

Schmid, C., and Mohr, R. (1997). Local Grayvalue Invariants for Image Retrieval, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **19**(5), pp. 530-534.

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

Shanahan, M.P. and Randell, D.A. (2004). A Logic-based Formulation of Active Visual Perception, Proc. *KR-04*, pp. 64-72.

Shibata, T., Vijayakunar, S., Conradt, J. and Schaal, S. (2001). Biomimetic Oculomotor Control, *Adaptive Behavior*, **9**(3-4), pp. 189-207.

Simons, D.J., & Chabris, C.F. (1999). Gorillas in our Midst: Sustained Inattentional Blindness for Dynamic Events, *Perception*, **28**, pp. 1059-1074.

Simons, D.J., & Levin, D.T. (1998). Failure to Detect Changes to People During a Real-world Interaction, *Psychonomic Bulletin and Review*, **5**, pp. 644-649.

Simons, D.J., Franconeri, S.L., & Reimer, R.L. (2000). Change Blindness in the Absence of Visual Disruption, *Perception*, **29**, pp. 1143-1154.

Simons, D.J. and Rensink, R.A. (2005). Change Blindness: Past, Present, and Future, *Trends in Cognitive Sciences*, **9**-1, pp. 16-20.

Stark, L.W. and Choi, Y.S. (1996). Experimental Metaphysics: The Scanpath as an Epistemological Mechanism, in Zangemeister, W.H. *et al*. (eds.) *Visual Attention and Cognition*, Elsevier Science, pp. 3-69.

Vieira Neto, H. and Nehmzow, U. (2005). Automated Exploration and Inspection: Comparing Two Visual Novelty Detectors, *Int. J. Advanced Robotic Systems*, **2**-4, pp. 355-362.

Wolfe, J.M. (1994). Visual Search in Continuous Naturalistic Stimuli, *Vision Research*, **34**, pp. 1187-1195.

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