A Logical Account of Perception Incorporating Feedback and Expectation

Murray Shanahan
Imperial College
Department of Electrical
Engineering,
Exhibition Road,
London SW7 2BT,
England.
m.shanahan@ic.ac.uk

Abstract

This paper presents the theoretical foundations of a vision system in which the effects of high-level reasoning percolate all the way down to influence the low-level processing of raw sensor data. This is achieved through the mechanisms of feedback and expectation. The main contribution of the paper is to present a formal framework, based on the abductive interpretation of sensor data, that incorporates the ideas of feedback and expectation in a way that marries them to logical reasoning. To enable this, two alternative measures of explanatory value are defined.

1 INTRODUCTION

Work in cognitive robotics has typically adopted one of two views of perception. According to the first view, perception is considered as a black-box process from the perspective of high-level cognition, whereby raw data is turned into fluents on demand, following the execution of a sensing action [Scherl & Levesque, 1993], [Levesque, 1996]. Formal accounts based on this view concentrate on the difficulties associated with formalising the effects of knowledge producing actions, and the task of planning with such actions. According to the second view, perception is a passive process, whereby the robot's model of the world is updated as a side-effect of its physical actions [Shanahan, 1996], [Shanahan, 1997]. In formal accounts based on this view, abduction is used to supply possible explanations of incoming sensor data, and the results of this abductive process are assimilated into the robot's representations.

In both approaches to perception, the flow of information is one-way — from raw sensor data to high-level representations. By contrast, it has long been recognised that in biological brains, the flow of information between low-level perception and high-level cognition is bidirectional [Cavanagh, 1999]. One of the most prominent mechanisms for achieving this bidirectionality is *expectation.* Low-level perceptual cues suggest interpretations of raw sensor data that line-up with past experience. This leads to the expectation of certain features in the environment, which in turn feeds back to low-level perceptual systems, making them more sensitive to those features. Initial expectations are then either confirmed or disconfirmed, and a stable model of the environment is built up.

From an engineer's point-of-view, it makes good computational sense to adopt this biologically inspired technique. Take vision, for example, which is the main subject of the present paper. A robot's visual system delivers a large amount of raw sensor data. In the unidirectional approach, this mass of sensor data is processed using a battery of methods for extracting features, such as edges and patches of uniform texture, which are then used to segment the image, picking out the foreground objects. This is a highly computationally intensive business. In a cluttered scene, a large number of features will be thrown up, which need to be sifted and sorted. Moreover, the process is highly sensitive to noise. Shadows, poor lighting, highlights, and surface patterns can all render the output of the system useless.

The benefits of incorporating a two-way flow of information include a reduction in computation and increased robustness. A reduction in computation is achieved because the sensitivity of low-level image processing routines can be initially set low, so that only a small number of highly prominent features are passed up to the next highest level of processing. Prominent features act as cues leading to the expectation of other features in the visual field. These expectations are fed back down to the low-level perceptual system, resulting in a selective increase of sensitivity confined only to those areas of the visual field that are potentially interesting. Increased robustness also results, first because of the initial insensitivity of the low-level processing, and second because the feedback loop soon

eliminates from consideration features leading to expectations that fail to be confirmed.

Similar arguments to these were used to motivate pioneering work in *active vision* [Aloimonos, *et al.*, 1987], [Ballard, 1991]. Research on active vision emphasises, among other things, the computational benefits of selective head or camera movements to fix on portions of the scene of interest while filtering out the rest. In a sense, the rationale behind the present paper is the same. The chief differences are that the active component of the system presented here is not at the level of physical movement, but at the level of software adjustments to low-level image processing routines. However, in more general terms, the paradigm offered here encompasses a spectrum of possibilities, all involving the following three steps of hypothetico-deductive reasoning.

1. Generate competing hypotheses to explain the sensor data.
2. Determine the consequences (expectations) of those hypotheses.
3. Carry out actions that will confirm or disconfirm these expectations, thus ruling out one or more of the competing hypotheses.

A confirming/disconfirming action might take several forms. At one end of the spectrum, it could be the software adjustment of a parameter of a low-level image processing algorithm, such as edge detection. At the other end of the spectrum, it could be a knowledge producing action, or knowledge producing plan, such as "Find a telephone directory and look up John's number" [Scherl & Levesque, 1993]. Traditional active perception, in which the actions in question are small-scale physical adjustments of the sensory apparatus, can be thought of as lying somewhere in the middle of this spectrum. Although the examples presented here are all of the first type, the theoretical treatment is intended to be generic.

Specifically, this paper presents a formal, logic-based account of visual perception incorporating a two-way flow of information between low-level image processing routines and high-level reasoning. The formalisation is a modification of the abductive model of perception presented in [Shanahan, 1996]. In this modified account, feedback and expectation are incorporated via the preference relation that selects between competing hypotheses. To facilitate this, two measures of the explanatory value of a hypothesis are proposed and compared — an *ad hoc* measure based on some obvious design criteria, and a more principled probabilistic measure. This general account is applied to a specific vision task, namely the recognition of cuboidal objects in a cluttered, noisy scene.

2 VISUAL PERCEPTION AS ABDUCTION

According to the model of robot perception put forward in [Shanahan, 1996], the task of robot perception is characterised roughly as follows. Given a stream of sensor data represented by the conjunction Γ of a set of observation sentences, find one or more explanations of Γ in the form of a logical description Δ of the locations and shapes of hypothesised objects, such that,

$$\Sigma \wedge \Delta \models \Gamma$$

where Σ is a background theory describing how objects in the world impact on the robot's sensors. The form of Δ, that is to say the terms in which an explanation must be couched, is prescribed by the domain. Typically there are many Δs of the permitted form that might explain a given Γ according to this definition. So a preference relation is defined for ordering them. This preference relation will be the vehicle for introducing expectation and feedback in the next section.

While this basic form of abduction is adequate for very rudimentary forms of robot perception — with bump switches or infra-red proximity sensors, for example [Shanahan, 1996] — its limitations soon become apparent when we try to apply it to a richer sensory modality like vision. To begin with, we need to relax the constraint, implicit in the above definition, that an explanation must be found for *all* the sensor data, that is to say for the whole of Γ. Snapshots of the robot's visual field contain a large number of features, and it would be inappropriate for the perceptual process to try to build a model of the world that accounts for every one of them. On the contrary, we want a perceptual system to ignore irrelevant data and pick out only those objects in the scene that are pertinent to the robot's goals or desired behaviour.

More important still, in the context of the present paper, is the fact that the basic abductive account leaves no room for the sort of two-way flow of information whose benefits were outlined in the introduction. It's especially galling that basic abduction has no means for allowing high-level declaratively represented expectations to inform low-level perceptual mechanisms when we consider that one of the chief attractions of an abductive treatment is that it is expressed in high-level, declarative terms. (For another attempt to reconcile abduction with top-down perceptual processing, see [Josephson & Josephson, 1994], Chapter 10.)

To see how this situation might be remedied, let's consider a motivating example. The image at the top of Figure 1 is the output from one of the stereoscopic cameras mounted on the head of an upper-torso humanoid robot. The image at the bottom shows the result of applying a standard edge detection algorithm, using the Sobel operator, with a high threshold. Let's

look more closely at the block circled in white in the unprocessed image. Our aim here is to devise a mechanism that will exploit the cues present in the image to conclude that a block of the right shape and size is out there, and which won't be distracted by the fact that the block is composed of two differently coloured halves, one of which is almost lost in the shadows.

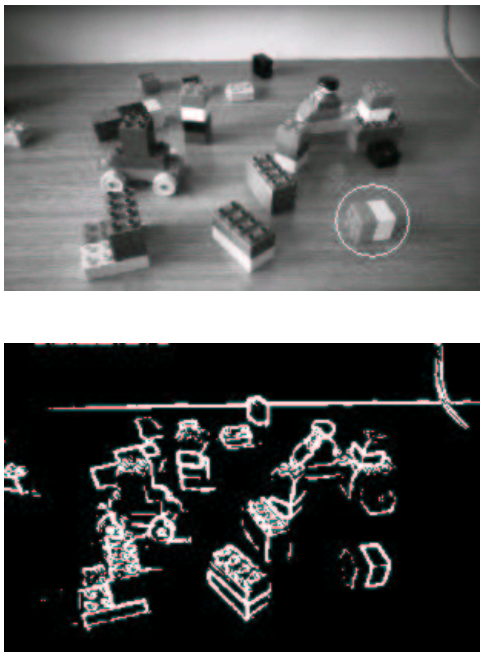

Figure 1: Some Raw Visual Data

More precisely, the problem is this. When a standard region finding algorithm is applied to the edge data in Figure 1, it comes up with the set of lines shown in bold in Figure 2, namely AB, AD, BC, DE, CF, and EF. Other lines, such as GC and HE are visible to the human eye, but the region finding algorithm is blind to them. If we turn up the sensitivity of the edge detector sufficiently to make these lines visible to the region finder, the number of spurious, unwanted lines thrown up as well — caused by the grain of the wooden table-top, for example — is so large that it becomes computationally infeasible to attempt to interpret them all, sorting out the useful data from the rest.

On the basis of this meagre and misleading data, how are we to find the true outline of the block? An abductive approach to perception initially seems like a good idea here. The correct hypothesis — let's call it H_1 — that the visible lines are caused by a block whose base is the line HF, is indeed sanctioned by the abductive definition. But so is the alternative hypothesis — which we'll call H_2 — that the visible lines are caused by a block whose base is the line EF. Moreover, according to almost any plausible preference relation we can think of for ordering multiple explanations, the second, incorrect hypothesis comes out on top, due to the influence of the line AD.

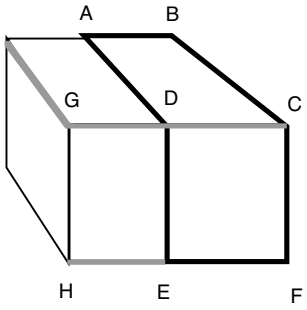

Figure 2: An Ambiguous Block

The main innovation of this paper is to supply a modified abductive treatment of perception in which the expectations generated by each hypothesis feed back into the preference ordering for choosing between them. In the present example, hypothesis H_1 generates the expectation (among others) that the line DC extends beyond D. Hypothesis H_2, on the other hand, generates the expectation of a vertical line extending downwards from A. These expectations retrospectively adjust the preference ordering for multiple explanations at the next level down in the hierarchy of sensor data processing, namely the process of finding edges and lines. This translates into a highly selective increase in the sensitivity of the edge detector at precisely those places where edges are expected. With this increase in sensitivity, the expectations of hypothesis H_1 are confirmed, and those of hypothesis H_2 are disconfirmed, leading to a revised ordering of the hypotheses in which H_1 comes out on top.

3 EXPLANATORY VALUE AND EXPECTATION

In this section, a formal description of abductive perception with feedback is presented. The language and consequence relation of standard first-order predicate calculus are used throughout. First, we have a basic definition of explanation, inherited from previous work on the topic of abductive perception [Shanahan, 1996]. A background theory Σ, a set of atomic *abducible* formulae, and a set of atomic *explainable* formulae are assumed. A *data set* is a conjunction of explainable formulae.

Definition 3.1. Given a background theory Σ, a conjunction ϕ of abducible formulae is an *explanation* for a data set Γ if,

$$\Sigma \wedge \phi \models \Gamma$$

and $\Sigma \wedge \phi$ is consistent. □

In general, the perceptual system of a robot will be many layered. That is to say, low-level features such as pixels will be interpreted in terms of higher-level features such as lines, which in turn will be interpreted in terms of surfaces, which are finally interpreted in terms of solid shapes. However, in the formal account that follows, these will be compressed into a single layer. The generalisation to multiple layers is straightforward. (See [Josephson & Josephson, 1994], Chapter 10 for an abductive account of layered perception.)

In the following account, we suppose the presence of a mass of sensor data. The focus here is visual perception, and the mass of data in question is taken to be the set of edges detected in a single snapshot of the visual field. Dynamic aspects of the data are neglected in the present paper, but obviously motion cues are a valuable source of additional information.

Some form of attention mechanism is taken for granted in the scheme described below. In effect, this means that a small portion of the visual field is isolated, such as the circle in Figure 1, and the task is to find the best explanation for the edges within that area. (This is obviously analogous to the visual attention mechanism of the human brain, whereby eye saccades and head movements cause the eye's foveal area to alight on objects of interest.) If the sensor data of interest is described by a data set Γ, the challenge is to determine the set of potential explanations for Γ, and to order them.

A variety of ways to order alternative explanations of the same sensor data might be employed. To facilitate the introduction of expectation and feedback, the method of choice here will be to assign a numerical score, in the range 0 to 1, to each hypothesised feature that explains Γ. This score is intended to reflect the *explanatory value* of the hypothesis. One measure of explanatory value is proposed in this section, and an alternative based on probability is put forward in Section 6. The set of features is partitioned into *primitive* and *non-primitive*, where primitive features are the raw sensor data themselves, and a non-primitive feature is an abducible formula, representing, for example, the presence of a block in a certain location. The initial explanatory value of a primitive feature is a function of the raw sensor data, scaled to fit the range 0 to 1. The initial explanatory value of a non-primitive feature is a weighted average of the explanatory values of those features it explains.

Definition 3.2. The *initial explanatory value* $V_0(H)$ of a non-primitive feature H is equal to,

$$\left(1-\frac{1}{n+1}\right).\left(\frac{1}{n}\sum_{\psi\in\Psi}V_0(\psi)\right)$$

where,

- Ψ is the set of all ψ such that H is an explanation for ψ and $V_0(\psi) > \theta_1$,
- n is the cardinality of Ψ. □

The idea of the weighting factor is to give the highest value to the feature that explains the most lower-level features. However, the effect of this weighting diminishes as the number of explained features goes up. The role of the threshold value θ_1 is crucial. Only those lower-level features that themselves have a high enough explanatory value are worth explaining. In terms of the example of the previous section, pixels with too low a Sobel value are ignored.

Now we introduce expectation and feedback, which is the means by which those features too faint to be considered initially get taken into account without swamping the system with computational demands. The initial explanatory value of a feature is increased to the extent that it fulfills the expectations of a higher-level feature, and is reduced to the extent that its own expectations fail to be fulfilled.

Definition 3.3. Let k be any integer greater than zero, and H be any feature. The *unmoderated* k^{th} *explanatory value* $V_k^-(H)$ of λ is equal to,

$$V_{k-1}(H) - k.Q.\left(1-\frac{1}{m+1}\right).\left(\frac{1}{m}\sum_{\phi\in\Phi}V_{k-1}(\phi)\right) + \omega$$

where

- Φ is the set of all ϕ such that H is an explanation for ϕ and $V_{k-1}(\phi) \le \theta_2$,
- m is the cardinality of Φ, and
- if
  - $V_{k-1}(H) > \theta_2$, and
  - there exists an α which is an explanation for H such that $V_{k-1}(\alpha) > \theta_1$.

  then $\omega = R$, else $\omega = 0$. □

The constant Q dictates the degree to which an unfulfilled expectation reduces a feature's explanatory value. Similarly, the constant R dictates the degree to which a feature's explanatory value is enhanced if it does fulfill some expectation. The constant θ_2 represents the threshold of explanatory value below which a feature is considered to be effectively absent and therefore to disconfirm higher-level features that expect it. The behaviour of algorithms based on the definitions given here is quite sensitive to the values of these constants, as discussed below.

Definition 3.3 can yield values greater than 1 or less than 0. The following definition imposes the corresponding ceiling and floor.

Definition 3.4. The k^{th} *explanatory value* $V_k(H)$ of a non-primitive feature H is defined as follows.

$$V_k(H) = \begin{cases} V_k^-(H) \text{ if } 0 \le V_k^-(H) \le 1 \\ 1 \text{ if } V_k^-(H) > 1 \\ 0 \text{ if } V_k^-(H) < 0 \end{cases} \quad \square$$

The repeated application of feedback generally takes the explanatory value of each feature to a fixpoint. In dynamical systems terms, this is an attractor basin, analogous to the attractor basins in, for example, the state space of a Hopfield net that might similarly be used to recognise patterns in sensor data on the basis of expectation and past experience. In contrast to a neural net, however, the present system enjoys the advantages of being able to carry out symbolic reasoning with declaratively represented knowledge.

Definition 3.5. The *final explanatory value* $V_\infty(H)$ of a non-primitive feature H is equal to $V_m(H)$ where m is the smallest integer such that,

$$V_m(H) = V_{m+1}(H) \quad \square$$

Further investigation is required into the conditions under which there exists a final explanatory value for a feature. However, none of the (limited) experiments carried out so far has unearthed an unstable combination of logical formulae and numerical values.

Definition 3.6. The *explanatory value* of a set of features is the mean explanatory value of the set. $\square$

Now, the preference relation used to order competing explanations for the same item of sensor data mirrors the assignment of final explanatory values to those explanations.

4 EXPECTATION AND FEEDBACK IN ACTION

The definitions above tell us very little about what conditions that cause a feature's explanatory value to ascend to a fixpoint and what conditions cause it to descend to a fixpoint. Further investigation of this issue is required, but initial experimentation has shown that a vision algorithm based on the definitions of the previous section is effective — in the sense that the explanatory values move in the required directions in response to feedback — when θ_1=0.4, $\theta_2 = \frac{2}{3}\theta_1$, R=0.1, and Q=1.0.

The following example illustrates the way expectation and feedback can distinguish between competing hypotheses. The parameters θ_1, θ_2, R, and Q are set to the above values. Suppose a snapshot of a robot's sensor data includes the primitive features F_1 to F_6, with the following initial explanatory values.

$V_0(F_1) = 0.36$ $\quad$ $V_0(F_2) = 0.63$
$V_0(F_3) = 0.81$ $\quad$ $V_0(F_4) = 0.72$
$V_0(F_5) = 0.81$ $\quad$ $V_0(F_6) = 0.20$

Now suppose we have two non-primitive features H_1 and H_2 such that,

$\Sigma \models \neg [H_1 \wedge H_2]$
$\Sigma \wedge H_1 \models F_1 \wedge F_2 \wedge F_3 \wedge F_4$
$\Sigma \wedge H_2 \models F_3 \wedge F_4 \wedge F_5 \wedge F_6$.

In other words, H_1 and H_2 are competing hypotheses, each of which explains some features of the sensor data. However, the initial explanatory value of H_2 is slightly higher than that of H_1.

$V_0(H_1) = 0.54$
$V_0(H_2) = 0.58$

So, at first glance, H_2 looks a better bet than H_1. However, the two hypotheses have different expectations. In particular, H_1 leads to the expectation of F_1, while H_2 gives rise to the expectation of F_6. Neither F_1 nor F_6 were prominent enough to contribute to the initial explanatory values of the hypotheses. But with the application of feedback, these differing expectations start to make a difference, and the two hypotheses swap places. In particular, we have,

$V_1(H_1) = 0.58$
$V_1(H_2) = 0.56$

From this point on, the two hypotheses start to diverge. This divergence reflects the fact that the initial explanatory value of F_1 is sufficient to fulfill the expectation of H_1, while the initial explanatory value of F_6 is so low that it disconfirms H_2. The final explanatory values of the hypotheses end up at opposite ends of the spectrum.

$V_\infty(H_1) = 0.80$
$V_\infty(H_2) = 0.00$

It might seem counter-intuitive that the final explanatory value of H_2 is zero. After all, this hypothesis surely has some value. In effect, this outcome is an artefact of the way thresholds are used. The result of applying feedback is to amplify the effects of confirmation and disconfirmation until certain hypotheses are dismissed altogether. The method of applying feedback with the probabilistic measure of explanatory value presented in Sections 6 and 7 is less polarising.

This is obviously a simple example, and with a more complex network of logical relations, the dynamics of the system is correspondingly more complicated. However, it should be clear that the example maps easily to the benchmark vision problem from Section 2.

The next section offers a more formal treatment of that benchmark problem.

5 VISION REVISITED

In this section, we take a closer look at the benchmark vision problem of Section 2. The main task is to formalise it as an abductive explanation problem, to which the definitions of the previous section are applicable. To this end, a series of axioms is presented describing blocks in terms of surfaces, surfaces in terms of lines, and the appearance of lines in terms of visible edges.

The first axiom says that, under the right conditions, a linear feature in 3D space gives rise to a visible edge (according to the Sobel operator) in the robot's visual field. The linear feature might correspond to a spatial discontinuity, such as the side of an object, or to a colour discontinuity, such as a shadow across the surface of an object. This axiom is used to reason abductively from data items in the robot's two-dimensional visual field (edges) to spatially-located features of the robot's workspace.

$$[\text{Line}(w) \wedge \text{FromTo}(w,p1,p2) \wedge \neg \text{Occluded}(w) \wedge \text{p3=Project}(p1) \wedge \text{p4=Project}(p2)] \rightarrow \exists e\, [\text{Edge}(e) \wedge \text{FromTo}(e,p3,p4)] \quad (1)$$

The formula Line(w) represents that a linear feature w exists in 3D space. The term Project(p) denotes the point in the robot's visual field onto which point p in 3D space projects. The formula Edge(e) represents the presence of an edge e in the robot's visual field. The formula FromTo(x,p1,p2) indicates the end points, p1 and p2, of the line or edge x. These will be 3D co-ordinates in the case of a line (in the robot's workspace), and 2D co-ordinates in the case of an edge (in the robot's visual field).

The formula Occluded(w) holds if some object lies between w and the robot's viewpoint. Several other kinds of exception, in addition to occlusion, would merit inclusion in a fuller axiomatisation, such as poor lighting conditions, a faulty camera, and so on. In Sections 6 and 7, there will be a requirement for explicit noise and abnormality terms in formulae like Axiom (1).

The next two axioms are used to reason abductively from lines to surfaces and their properties. The two axioms correspond to the two possible explanations of a linear discontinuity — either it's the edge of a solid object or it's a surface feature such as a shadow. The formula Bounded(s,w1,...,w4) means that the rectangular surface s is bounded by the four lines w1 to w4. The formula Marked(s,w) means the surface s has a linear feature w.

$$[\text{Surface}(s) \wedge \text{Bounded}(s,w1,w2,w3,w4)] \rightarrow [\text{Line}(w1) \wedge \text{Line}(w2) \wedge \text{Line}(w3) \wedge \text{Line}(w4) \wedge \text{Parallel}(w1,w2) \wedge \text{Parallel}(w3,w4)] \quad (2)$$

$$[\text{Surface}(s) \wedge \text{Marked}(s,w)] \rightarrow \text{Line}(w) \quad (3)$$

Finally, Axiom(4) below takes the abductive process from surfaces to cuboidal blocks. The formula Faces(b,s1,...,s6) means that the block b has the six faces s1 to s6, each of which is a rectangular surface.

$$[\text{Block}(b) \wedge \text{Faces}(b,s1,s2,s3,s4,s5,s6)] \rightarrow [\text{Surface}(s1) \wedge \text{Surface}(s2) \wedge \text{Surface}(s3) \wedge \text{Surface}(s4) \wedge \text{Surface}(s5) \wedge \text{Surface}(s6) \wedge \text{PParallel}(s1,s2) \wedge \text{PParallel}(s3,s4) \wedge \text{PParallel}(s5,s6)] \quad (4)$$

These and similar axioms form the basis of a visual perception system that combines high-level logical reasoning with feedback and expectation to interpret scenes that would present a challenge to a conventional image understanding system. The knowledge inherent in Axioms (1) to (4) is analogous to a 3D model in a conventional machine vision system, and the abductive interpretation of the sensor data corresponds to the conventional process of matching a model to the data from a given scene [Jain, *et al.*, 1995, Chapter 15]. However, under the present scheme, it's possible to augment this knowledge with declarative sentences, such as "all the red blocks are shorter than all the blue blocks" or "one of the green blocks is hidden". Combined with a suitable inference mechanism, such knowledge can be used to influence low-level processing.

To see how the abductive process works, let's return to the simple example of Section 2. Suppose we have the following set Γ of items of low-level sensor data (Figure 3).

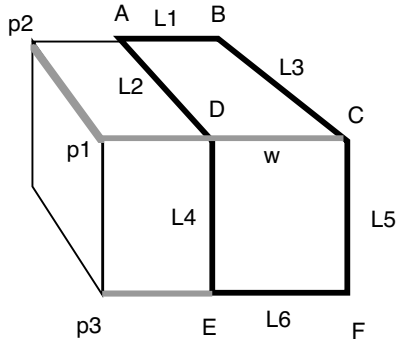

Figure 3: Hypothesised Lines and Points

$\text{Edge}(L_1) \wedge \text{FromTo}(L_1,A,B)$
$\text{Edge}(L_2) \wedge \text{FromTo}(L_2,A,D)$
$\text{Edge}(L_3) \wedge \text{FromTo}(L_3,B,C)$

Edge(L_4) ∧ FromTo(L_4,D,E)

Edge(L_5) ∧ FromTo(L_5,C,F)

Edge(L_6) ∧ FromTo(L_6,E,F)

Let the set of abducible formulae be all those of the form Block(b), Faces(b,s1,...,s6), or Marked(s,w). Now, let H_1 be,

Block(B_1) ∧ Faces(B_1,$S_{1,1}$,…,$S_{1,6}$)

where, for some $S_{1,i}$ and $S_{1,j}$ in $\{S_{1,1},\ldots,S_{1,6}\}$, we have,

Bounded($S_{1,i}$,L1,L2,L3,w)

Bounded($S_{1,j}$,w,L5,L6,L4)

for some hypothesised line w. Now, let H_2 be,

Block(B_2) ∧ Faces($B_{2,1}$,$S_{2,1}$,…,$S_{2,6}$)

where, for some $S_{2,i}$ and $S_{2,j}$ in $\{S_{2,1},\ldots,S_{2,6}\}$ we have,

Bounded($S_{2,i}$,L_3,w_1,w_2,w_3)

Bounded($S_{2,j}$,L_5,w_1,w_4,w_5)

for hypothesised lines $w_1,\ldots,w_5$, such that,

FromTo(w_1,C,p_1) FromTo(w_2,p_1,p_2)

FromTo(w_3,p_2,B) FromTo(w_4,F,p_3)

FromTo(w_5,p_3,p_1)

for hypothesised points p_1 to p_3, all distinct from points A to F. Next, let H_3 be,

Marked($S_{2,i}$,L_2).

Finally, let H_4 be,

Marked($S_{2,j}$,L_4).

It can be straightforwardly shown that H_1 is an explanation for Γ, as is the conjunction of hypotheses H_2 to H_4. The latter combination is, of course, correct, while H_1 is incorrect. However, when data from the actual vision system is plugged in, the initial explanatory value of H_1 is much higher than that of the correct combination. With the application of expectation and feedback, using the definitions of Section 3, this situation is reversed. The incorrect hypothesis H_1 leads, through Axiom (4), to the expectation of a vertical surface, one of whose edges is AD. The failure of this expectation brings about a fall in the explanatory value of H_1. On the other hand, the initially unpromising hypothesis H_2 leads to the expectation of a line extending from B to A and beyond, to a point p_2. This expectation is confirmed, causing the explanatory value of H_2 to rise. In the end, the final explanatory value of the correct combination of hypotheses exceeds that of the incorrect hypothesis, which is the desired result.

6 A PROBABILISTIC MEASURE OF EXPLANATORY VALUE

The measure of explanatory value proposed in Section 3 forms the basis of a workable system for perception incorporating expectation and feedback. However, from a theoretical perspective, the proposal looks *ad hoc*, because the numerical values defined lack any sort of semantic justification. In this section, an alternative measure of explanatory value is devised, based on probability. As we'll see, when compared to the more *ad hoc* measure, the probabilistic version has both advantages and disadvantages.

Suppose we have a set F of abducible formulae. In the absence of further information, all formulae in F are assumed to be independent and to have equal prior probability p. Note that, in the absence of further information, the prior probability $P(H)$ of any conjunction $H = f_1 \wedge \ldots \wedge f_n$ of abducible formulae is p^n.

Now suppose we acquire some sensor data, represented as a formula Γ. Using abduction, we find the set of possible explanations of Γ is $\{H_1 \ldots H_m\}$, where each H_i is a conjunction of abducible formulae. Suppose the set of hypotheses is mutually exclusive. So we have,

$$H_1 \oplus \ldots \oplus H_m.$$

Using basic probability theory, we can now assign an explanatory value to each candidate explanation H_i, corresponding to the posterior probability that it is true, given the above disjunction. From Bayes' Theorem we have,

$$P(q \mid q \oplus r) = \frac{P(q \oplus r \mid q).P(q)}{P(q \oplus r)}$$

which simplifies to,

$$P(q \mid q \oplus r) = \frac{P(q)}{P(q) + P(r)}.$$

Now, by setting q equal to any hypothesis H_i and r to the rest of the hypotheses, we can calculate the posterior probability of H_i given Γ, under the assumption that $H_1 \ldots H_m$ is the set of all the possible explanations of Γ. A related formula is derived by Poole [1993] for abduction with probability. Poole's equation yields the conditional probability $P(q|\Gamma)$, and is equivalent to the present formula under the same assumption.

Now suppose Γ comprises n separate items of sensor data, each requiring explanation. In general, a hypothesised object will only explain a portion of the sensor data of interest. So a complete explanation will comprise a number of hypothesised objects supplemented with a number of *noise terms* to "explain away" the rest of the sensor data [Poole, 1995], [Shanahan, 1997].

To allow for noise terms, the background theory Σ must be supplemented with some additional formulae. For example, as well as Axiom (1) from Section 5, we might have,

[Noise(w) ∧ FromTo(w,p1,p2) ∧
p3=Project(p1) ∧ p4=Project(p2)] →
∃ e [Edge(e) ∧ FromTo(e,p3,p4)].

Let the set F of abducibles comprise a set F_a of noise terms each with probability p_a and a set F_b of propositions positing spatially-located objects, each with probability p_b. Formulae in F_b explain multiple items of sensor data while formulae in F_a explain only 1 item of sensor data. What we're interested in is how the probability (and thus the explanatory value) of a hypothesis grows with the number of items of sensor data it explains.

Suppose the disjoined set of explanations of Γ obtained by abduction is $H_1 \oplus ... \oplus H_m$, such that,

$$H_1 = [f \wedge f_{k+1} \wedge ... \wedge f_n]$$

where f is drawn from F_b and each f_i is drawn from F_a. In other words, H_1 explains k out of n items of sensor data, and the rest have to be explained by noise terms. We have,

$$P(H_1) = p_b p_a^{n-k}$$

Let R be $H_2 \oplus ... \oplus H_m$. Then we have,

$$P(H_1 \mid H_1 \oplus R) = \frac{p_b \cdot p_a^{n-k}}{p_b \cdot p_a^{n-k} + P(R)}$$

or alternatively,

$$P(H_1 \mid H_1 \oplus R) = \left(1 + \frac{P(R)}{p_b p_a^{n-k}}\right)^{-1}$$

Note that the posterior probabilities of H_1 to H_m sum to 1, as we would expect. To simplify the analysis, let's assume from now on that each hypothesis posits exactly one object. So each data item unaccounted for by that object is explained away with a noise term. Plugging in the appropriate formula for $P(R)$, this leads to the following probabilistic measure of initial explanatory value, which is the counterpart to Definition 3.2. The set of data items to be explained will be those in the area of interest whose value exceeds some threshold. We'll designate this threshold θ_1, since it corresponds to θ_1 in Section 3.

Definition 6.1. Given a data set Γ, the *probabilistic explanatory value* $VP(H,\Gamma)$ of a hypothesis H is equal to,

$$\left(1 + \frac{\sum_{\phi \in \Phi} p_a^{S_\phi}}{p_a^{S_H}}\right)^{-1}$$

where

- Φ is the set of all hypotheses that explain Γ, and
- S_ϕ is the number of terms in hypothesis ϕ drawn from F_a. □

Note that the p_b term always drops out. (This is only the case if each hypothesis posits the same number of objects.) Let's apply this definition to the example of Section 4. Let θ_1 be 0.4. Since both hypotheses explain 3 out of 4 items of sensor data, regardless of the value of p_a, we get,

$$VP(H_1,\Gamma) = VP(H_2,\Gamma) = 0.5.$$

Let's pick a different example. Suppose the set of explanations is $\{H_1,H_2\}$, and that H_1 and H_2 respectively explain 1 out of 4 and 2 out of 4 data items without recourse to noise terms. In other words, H_1 includes 3 noise terms while H_2 includes 2 noise terms. Plugging these values into the formula, we get,

$$VP(H_1,\Gamma) = \left(1 + \frac{1}{p_a}\right)^{-1}$$

and,

$$VP(H_2,\Gamma) = \left(1 + p_a\right)^{-1}.$$

So, for example, if we let p_a=0.1, the posterior probability (explanatory value) of H_1 is = 0.09 and the posterior probability of H_2 is 0.91 On the other hand, if we let p_a=0.5, the respective values we get for H_1 and H_2 are 0.33 and 0.67. These examples show that the probabilistic measure of explanatory value does assign higher values to hypotheses that explain more data. So it meets the most important criterion for a measure of explanatory value. But by inspecting the two formulae (the probabilistic version and the more *ad hoc* formula in Definition 3.2), we also see the many differences between the two measures.

- The probabilistic measure takes account of the number of unexplained data items rather than the proportion of data items explained. So, all other things being equal, it assigns the same explanatory value to a hypothesis that explains 99 out of 100 data items as to a hypothesis that explains 1 out of 2. In this respect the *ad hoc* measure has more appeal.
- The probabilistic measure, unlike the *ad hoc* measure, isn't weighted by the value of the sensor data explained. So it doesn't distinguish hypotheses that explain prominent sensor data

from those that explain poor quality sensor data. The assumption behind the probabilistic measure is that all the data items in Γ are present and all require explanation.

- The probabilistic measure takes account of the whole set of competing hypotheses when assigning an explanatory value to each member of that set. The *ad hoc* measure assigns the same value to a hypothesis that is a unique explanation of the data as to a hypothesis that is one among ten competitors. In this respect, the probabilistic measure is to be preferred.
- The probabilistic measure takes into account the probability of a noise term. According to the probabilistic measure, the more noisy the data, the narrower the gap between a hypothesised object that explains a lot and one that explains only a little. This makes sense, so in this respect again the probabilistic measure wins out.

7 ABDUCTION, PROBABILITY, AND FEEDBACK

The question now, of course, is how to incorporate the probabilistic measure of explanatory value into a scheme that exploits expectation and feedback. With the *ad hoc* measure, the explanatory value of a hypothesis is increased if it fulfills the expectation of another hypothesis, and decreased if its own expectations are unfulfilled (Definition 3.3). With the probabilistic version, a different strategy will be adopted, whereby the set Γ of data items to be explained is increased, as a result of feedback, to include the outcome of testing the expectations of all the candidate hypotheses. The tricky part is to augment Γ in such a way as to include the *absence* of expected data items as well as their presence.

First, the notion of a data set is widened to include negated formulae. These will be used to represent unfulfilled expectations.

Definition 7.1. An *augmented data set* is a conjunction of the form $(\neg)\psi_1 \wedge ... \wedge (\neg)\psi_n$, where each ψ_i is a data item. □

The definition of an explanation now has to be modified to cater for augmented data sets. The important thing to note here is that there is no requirement for an explanatory hypothesis to entail a negated formula in the data set, but simply to be consistent with it.

Definition 7.2. Let Γ be an augmented data set. Given a background theory Σ, a conjunction ϕ of atomic abducible formulae is an *explanation* of Γ if, for all positive formulae ψ in Γ,

$$\Sigma \wedge \phi \models \psi$$

and, for all negative formulae $\neg\psi$ in Γ,

$$\Sigma \wedge \phi \not\models \psi$$

and $\Sigma \wedge \phi$ is consistent. □

For this definition of explanation to be effective, non-monotonicity needs to be introduced. Specifically, the expectations of a hypothesis must be given the status of defaults. In the same way that a noise term can be introduced to fill in gaps in the explanation, an abnormality term can be introduced to override a default expectation that was unfulfilled. But when it comes to calculating the explanatory value of a hypothesis, the inclusion of an abnormality term, like the inclusion of a noise term, is costly, as we'll see shortly.

Many of the formulae in the background theory Σ need to be rewritten to turn them into default rules. For example, Axiom (1) from Section 5 becomes,

$$[\text{Line}(w) \wedge \text{FromTo}(w,p1,p2) \wedge \neg \text{Occluded}(w) \wedge \neg \text{Abnormal}(w) \wedge p3=\text{Project}(p1) \wedge p4=\text{Project}(p2)] \rightarrow \exists e\, [\text{Edge}(e) \wedge \text{FromTo}(e,p3,p4)].$$

Default reasoning can now be effected through circumscription. Each hypothesis is circumscribed, minimising the Abnormal predicate. No further details will be given here, and from now on, all hypotheses should be assumed to be implicitly circumscribed in this way. Because only atomic abnormality formulae are made abducible, the mathematics of this is straightforward, always yielding simply the completion of the Abnormal predicate. In real-world terms, the abnormality predicate represents a whole raft of circumstances that might explain the non-appearance of an expected edge, such as poor lighting, lack of contrast between foreground and background, occluding objects, and so on.

The set F of abducible formulae has to be expanded to include abnormality terms. In particular, let's assume the subset F_a of F now comprises both noise terms and abnormality terms, each with prior probability p_a.

We're now in a position to define an augmentation operator that adds both fulfilled and unfulfilled expectations to a data set. First we make precise the notion of an expectation.

Definition 7.3. Given a background theory Σ, the set of expectations $Exp(H)$ of a hypothesis H comprises all data items ψ such that,

$$\Sigma \wedge H \models \psi$$ □

Now we define the augmentation operator, and introduce two new threshold values, θ_2 and θ_3 (Figure 4). Intuitively, the meaning of these thresholds is as follows. If the value assigned to a data item is above θ_1, then that data item is assumed to be present, as before.

If it falls below θ_3, then it's assumed to be absent. And if it falls between θ_2 and θ_1 then it's assumed to present if it was expected.

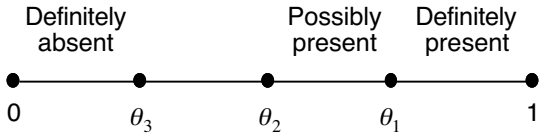

Figure 4: Data Item Thresholds

Definition 7.4. If Γ is a data set, then the augmented data set $Aug(\Gamma)$ is $\Gamma \cup P \cup N$, where,

- P is the set of all ψ such that $\psi \in Exp(H)$ for some H where $VP(H,\Gamma) \geq \theta_1$, and the value of $\psi \geq \theta_2$, and
- N is the set of all formulae of the form $\neg\psi$ such that $\psi \in Exp(H)$ for some H where $VP(H,\Gamma) \geq \theta_1$, and the value of $\psi < \theta_3$. □

In the example of Section 4, if θ_1=0.4, θ_2=0.35, and θ_3=0.30, then $Aug(\Gamma) = \{F_1, F_2, F_3, F_4, F_5, \neg F_6\}$. So H_1 contains only 1 noise term (for F_5) to explain 6 data items, while H_2 contains 2 noise terms (to account for F_1 and F_2) and one abnormality term (to account for $\neg F_6$). If we let p_a=0.1, this gives

$VP(H_1,Aug(\Gamma))$=0.99

$VP(H_2,Aug(\Gamma))$=0.01.

More realistically, if we let p_a=0.5, we get,

$VP(H_1,Aug(\Gamma))$=0.8

$VP(H_2,Aug(\Gamma))$=0.2.

Either way, hypothesis H_1 emerges as the clear winner, as we would expect. Note that $Aug(Aug(\Gamma))=Aug(\Gamma)$. In other words, further applications of the *Aug* operator have no effect in this case.

8 CONCLUDING REMARKS

The methods described above have been incorporated in an implemented vision system for deployment on an upper-torso humanoid robot which is currently under construction at Imperial College (Figure 5). The robot has two arms, each with three degrees-of-freedom, and a pan-and-tilt head with stereoscopic vision. (The exploitation of depth information from the stereo cameras is one of many topics for further investigation.) The robot's forearms are mechanically constrained to be parallel to the workbench in order to cut down on degrees-of-freedom and simplify the control issues. In its first incarnation, the robot will be equipped with simple prods, not grippers or hands, at the ends of its forearms.

The robot's task is to survey a table-top cluttered with objects. Most of the objects will be unfamiliar, but some will be cuboidal. The robot's perceptual system needs to pick out these objects, and form a representation of their locations and dimensions. The robot, or a demonstrator, will nudge the objects, and the robot must track familiar objects, increasing the quality of its representations as the objects are viewed from new angles and previously unseen parts become visible. Initial experiments using the vision system described are promising.

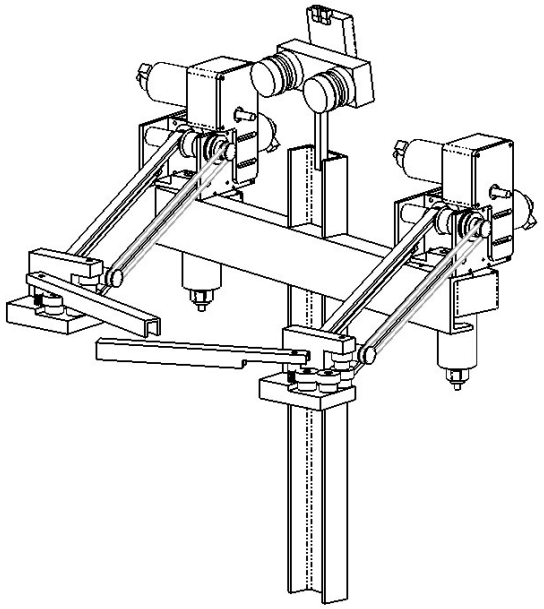

Figure 5: LUDWIG the Humanoid Robot

Several avenues of further research are being followed. These include the use of more sophisticated forms of high-level reasoning in the perceptual process, in order to deal with occlusion and object persistence [Randell, *et al.*, 2001]. This necessitates taking account of the ongoing actions of the robot and other agents [Shanahan, 2000], ensuring that, as in work on animate vision, the dynamics of the robot's interaction with the world enhance, rather than handicap, the perceptual capabilities of the robot.

Acknowledgements

This work was carried out as part of EPSRC grant GR/N13104 "Cognitive Robotics II". Thanks to Yiannis Demiris, Seçil Ozen, Dave Randell, Paulo Santos, Ray Thompson, and Mark Witkowski. Thanks to the anonymous referees for their insight and generosity.

References

[Aloimonos, *et al.*, 1987] J.Aloimonos, I.Weiss, and A.Bandyopadhyay, Active Vision, *Proceedings 1st International Conference on Computer Vision*, pp. 35–54.

[Ballard, 1991] D.H.Ballard, Animate Vision, *Artificial Intelligence*, vol. 48 (1991), pp. 57–86.

[Cavanagh, 1999] P.Cavanagh, Top-Down Processing in Vision, *The MIT Press Encyclopedia of the Cognitive Sciences*, pp. 844–845.

[Jain, *et al.*, 1995] R.Jain, R.Kasturi, and B.G.Schunk, *Machine Vision*, McGraw-Hill, 1995.

[Josephson & Josephson, 1994] J.R.Jospehson and S.G.Josephson, *Abductive Inference: Computation, Philosophy, Technology*, Cambridge University Press, 1994.

[Levesque, 1996] H.Levesque, What Is Planning in the Presence of Sensing? *Proceedings AAAI 96*, pp. 1139–1146.

[Poole, 1993] D.Poole, Logic Programming, Abduction and Probability: A Top-Down Anytime Algorithm for Estimating Prior and Posterior Probabilities, *New Generation Computing*, vol. 11 (1993), pp. 377–400.

[Poole, 1995] D.Poole, Logic Programming for Robot Control, *Proceedings IJCAI 95*, pp. 150–157.

[Randell, *et al.*, 2001] D.Randell, M.Witkowski and M.Shanahan, From Images to Bodies: Modeling and Exploiting Spatial Occlusion and Motion Parallax, *Proceedings IJCAI 2001*, pp. 57–63.

[Scherl & Levesque, 1993] R.Scherl and H.Levesque, The Frame Problem and Knowledge Producing Actions, *Proceedings AAAI 93*, pp. 689–695.

[Shanahan, 1996] M.P.Shanahan, Robotics and the Common Sense Informatic Situation, *Proceedings ECAI 96*, pp. 684–688.

[Shanahan, 1997] M.P.Shanahan, Noise, Non-Determinism and Spatial Uncertainty, *Proceedings AAAI 97*, pp. 153–158.

[Shanahan, 2000] M.P.Shanahan, Reinventing Shakey, in *Logic-Based Artificial Intelligence*, ed. J.Minker, Kluwer Academic Press (2000), pp. 233–253.