Embodied Categorisation for Vision-Guided Mobile Robots [1]

Nick Barnes†* and Zhi-Qiang Liu†‡

†*Department of Computer Science and Software Engineering*

The University of Melbourne, Victoria, 3010, AUSTRALIA

‡*School of Creative Media, City University of Hong Kong, Kowloon, Hong Kong*

Abstract

This paper outlines a philosophical and psycho-physiological basis for embodied perception, and develops a framework for conceptual embodiment of vision-guided robots. We argue that categorisation is important in all stages of robot vision. Further, that classical computer vision is unsuitable for this categorisation, however, through conceptual embodiment, active perception can be effective. We present a methodology for developing vision-guided robots that applies embodiment, explicitly and implicitly, in categorising visual data to facilitate efficient perception and action. Finally, we present systems developed using this methodology, and demonstrate that embodied categorisation can make algorithms more efficient and robust.

Key words: Computer Vision, Robots, Classification, Situatedness, Embodiment.

> I suddenly see the solution of a puzzle-picture. Before there were branches there; now there is a human shape. My visual impression has changed and now I recognize that it has not only shape and colour but also a quite particular 'organization'.
>
> L. Wittgenstein [1]

* nmb@cs.mu.oz.au, Fax: +613 9347 8122

[1] The work is partially supported by Australian Research Council Discovery Project Grant DP0209969, Hong Kong Research Grants Council (RGC) Project No. CityUHK #9040690-873, and Strategic Development Grant (SDG) Project No. #7010023-873 from City University of Hong Kong.

1 Introduction

Many contemporary computer vision algorithms consider the perceiver to be a passive entity that is given images, and must process them to the best of its ability. Purposive Vision, Animate Vision, or Active Perception emphasises that the relationship between perception and the perceiver's physiology, as well as the tasks performed, must be considered in building intelligent visual systems [2]. By physiology we refer to fundamental aspects of vision and interaction, such as being able to fixate, move the viewpoint, as well physically interact with the object. However, such research often ignores categorisation and is mostly concerned with early vision (e.g., optical flow and segmentation). To date, there has been little consideration of the relationship between perception and the perceiver's physiology involving explicit categorisation and high-level vision. In this paper, we are concerned with developing effective computational vision for physically embodied entities (robots). As such we investigate the role of categorisation in vision, and of the embodiment of the perceiver in categorisation. We argue two main points: that categorisation is useful in vision; and, that for an entity that is physically embodied, its embodiment can and should play an important role constructing categorisation systems for the entity. We argue that categorisation is useful at all stages of vision for robot guidance and discuss philosophical aspects of what is required to apply categorisation and high-level vision when the perceiver has a real body, e.g., an autonomous mobile robot. This approach is supported by recent physiological and psycho-physical findings.

In some approaches to computer vision, including the classical approach exemplified by Marr [3], the conception is of a general all-purpose vision. We present an argument from the philosophical literature that such an approach is limited in what it can achieve due to an incorrect understanding of categorisation. However, the philosophical concept of embodied categorisation offers a way forward for computer vision for robot guidance. Embodiment, for humans, is the theory that abstract

meaning, reason, and imagination have a bodily basis [4]. We then examine categorisation in human perception from a philosophical view point, and consider recent physiological studies, finding that categorisation appears at all levels of human perception, and is fed back from later to earlier stages of visual processing. We argue that models are the means by which conceptual intervention can be achieved at the earliest stages of computer vision processes, and that these facilitate feedback from later stages to earlier. However, we also note that contextual information can be fed directly into early vision processes via models in computer vision.

Embodied categorisation can yield advantages: robots with embodied categorisation systems can act successfully based on data that fundamentally underdetermines the robot's environment. Also, we are able to simplify complex models, save unnecessary computation, and make the system more robust by eliminating sources of error.

This paper introduces the relevant philosophical concepts and considers classical vision as outlined by Marr, arguing that the categorisation aspect of classical computer vision is inconsistent with fundamental aspects of what is known about human categorisation. We consider contemporary computer vision and conclude that although the majority is not inconsistent in this manner, the categorisation assumptions of classical vision are present in some work. We then examine embodied categorisation for vision in humans from both a philosophical and physiological perspective. We define embodiment in mobile robots and a methodology for developing embodied vision-guided mobile robots. Finally, we present systems that were developed under this methodology, and other work that illustrates benefits of an embodied approach.

1.1 Embodiment

Embodiment, for humans, is the theory that abstract meaning, reason and imagination have a bodily basis [5]. Embodiment is formed by the nature of our bodies,

including perception and motor movement, and of our physical and social interactions with the world. Lakoff [4] considers categorisation to be the basic element of thought, perception, action and speech. To see something as a kind of thing involves categorisation. We reason about kinds of things, and perform kinds of actions. [2] The human conceptual system arises from conceptual embodiment, and is a product of experience, both direct bodily experience, and indirect experience through social interaction. If human thought, perception and action are based on categorisation, and this categorisation is embodied, then we should consider the proposition that embodied categorisation may be a useful paradigm for robot perception.

2 Marr's Classical Computer Vision Paradigm

Marr's approach [3], begins with images, which are transformed by segmentation into a primal sketch, and then composed into a 2 1/2 D sketch. From this, the system infers what objects are in the real world (Figure 1). The paradigm aims to "capture some aspect of reality by making a description of it using a symbol". It describes a sequence of representations that attempt to facilitate the "recovery of gradually more objective, physical properties about an object's shape." [3]

This approach assumes that the visible world can be described from a raw image. We use Marr's paradigm to exemplify an approach aimed at producing a general vision. This approach rejects any role for the system's embodiment or physiology: it must cater for all embodiments, purposes, and environments. As such, there must be either a single, uniquely correct categorisation for all objects in the world, or vision must enumerate all possible descriptions and categories of visible objects. This idea

[2] We do not argue that actions are discrete, just that there are types of actions, such as grasping with a precision or a power grip [6]. There may be a continuum of actions of grasping with a power grip for different objects.

that computer vision provides an objective description of the world that agents can reason about can be seen within some groups of the agents research community. While some aspects of reasoning can be considered independently of embodiment, our view is that embodied categorisation must be the basis for some aspects of agent reasoning when the agent is connected to a physical world by computer vision.

3 Why a general computer vision is ill-suited to robot guidance

The classical approach has been the basis of considerable development in computer vision. It offered a framework for breaking up vision into a series of processes that facilitated modular development with separate areas of research. The usefulness of hierarchical layers of processing has been demonstrated by numerous successful vision systems, and researchers continue to make valuable contributions in particular areas. Beyond acknowledging a valuable contribution, we are not concerned with a full critique of this approach. Marr's classical conception is ill-suited to robot guidance because it requires that objects in the world are objectively subdivided into categories, thus computer vision system should be able to image a scene, and list of its contents without any consideration of the purpose of the classification. This is an issue not so much of algorithms, but of how they are applied.

We do not question the value of vision algorithms that attempt to be as successful as possible with minimal assumptions. Nor do we object to purely data-driven vision algorithms. However, there are limitations that are apparent in tasks involving classification at all levels (e.g., object recognition, interpretation of line drawings, colour classification etc.). Any general computer vision faces difficulties, because there is no uniquely correct description of the world, and the list of possible descriptions is infinite for practical purposes.

Dupré [7] argues that there is no unique, natural categorisation. If a general vi-

sion program existed, then it must be able to uniquely classify all living organisms. However, Dupré details examples where two standard schemes of biological classification, biological classification of organisms into taxa and everyday language, do not coincide. In taxonomy an organism is classified into a hierarchical series of taxa, the narrowest of which is species. If an objective classification of natural kinds existed, then there should be only one unambiguously correct taxonomic theory, and this should coincide with everyday language. One may allow differences based on human eccentricities, but distinctions made on the basis of functional necessity must coincide. Similarly, a single general objective computer vision system would have to serve the needs of everyday language and scientific classification. Dupré argues that this is not the case, that within the world there are countless legitimate, objectively grounded ways of classifying objects in the world. These often cross-classify one another in indefinitely complex ways.

Dupré gives many examples. We outline just two. Taxonomically, several species of garlic and onions are in the same genus, i.e., taxonomy makes no distinction between them. However, clearly for cooking the difference is functionally important, and therefore everyday language makes a distinction. Secondly, the class angiosperms (flowering plants) includes daisies, cacti and oak trees, but excludes pine trees. The distinction as to whether the plant develops its seeds in an ovary has little application outside biology. Conversely the gross morphological feature of being a tree has no place in scientific taxonomy. However, for an autonomous car, the difference between a daisy and a large oak tree is considerable, when it blocks the road in front of the car. Dupré gives many other examples of cases where people draw distinctions for pragmatic reasons, based on real properties, that are not recognised by taxonomy. Naturally, everyday use and scientific terms do correlate in many circumstances, but clearly this is frequently not the case. Note that this does not mean there is no good way of classifying biological organisms. Rather, there are many good methods of classification, sometimes equally good, and which one is

useful in a particular case depends on the purpose of the classification.

If there is no unique classification of living things, clearly there is no unique classification of all things in the world. Further, other types of objects suffer similar ambiguities. Enumeration of the almost infinite number of possible classifications of an object is intractable. Consider an image of 'a computer', including a monitor. For some purposes the monitor will be a component, and for others it will not, and the monitor may be classified as a television, electronic components, or even a chair, etc. With no unique classification of objects, and a large number of possible classifications, computer vision cannot practically classify all objects that may appear in a scene without considering the purpose of classification. This raises the question of how objects should be categorised if there is no uniquely correct method. The philosophy of the mind gives a possible solution for vision-guided robots: embodied categories. *Objects should be classified by the way the robot relates to them.*

3.1 Difficulties in low-level computer vision

In this paper we are arguing not just that categorisation of high-level objects is not unique, but that embodied categorisation is useful across all levels of vision. We will now examine low-level vision, particularly human colour categorisation. There is a continuous range of visible colours that can be described by hue, saturation, and brightness (or other schemes). Underlying hue and saturation is the wavelength of the light reflected from a surface. Valera, Thompson and Rosch [8] describe three different cone cells in the human eye, whose overlapping photopigment absorption curves have peaks of 560, 530, and 440 nanometers. Excitory and inhibitory processes in post-receptoral cells can add or subtract receptor signals. Clearly there are physiologically real aspects to colour. When we view colours in isolation there is a close correspondence between the wavelength of light reflected from visible surfaces

and the colour that we perceive. However, in a complex scene, the light reflected locally is not sufficient to predict perceived colour. There are two additional phenomena: approximate colour constancy, where the perceived colour remains constant despite large lighting intensity changes; and, simultaneous colour contrast, where the same reflected wavelength can be seen as different colours depending on its surrounds [8]. Thus, we cannot consider the colour of objects in isolation, but must consider visual context.

Colour categorisation is also partially culturally specific. Valera *et al.* [8] point to research by Berlin and Kaye that found [9] that when several languages contain a term for a basic colour category, speakers virtually always agreed on what was the best example of the category. However, boundaries between colour categories varied for different language groups, and the perceptual distance between colours was not uniform. The boundaries of colour are partly defined by culture (a form of embodiment under Lakoff's definition). Thus, for a computer vision system to attempt to give human classifications for colour, different sets of classifications would be required for different languages groups. While computer vision can perform feedback and enumeration for classification, this does demonstrate the presence of categorisation in low-level vision, and that this categorisation may be embodied.

Further, in classifying colour objects, we would ideally like a computer vision system to take object colour into account. However, the system cannot perceived the colour of the object, only of the reflected light, which also depends on incident light. In computer vision, this is often managed by colour calibration, with human intervention to label colours. Certainly this mapping can be resolved from knowledge of incident light properties, or the visual appearance of part of the visible scene (e.g., a set of reference colours). However, in order to discover the mapping between object colour reflectance properties and apparent colour, we require *a priori* information. Thus, although the reflectance properties of objects are real, how the colours will

appear depends on the environment, and how they should be categorised is not simply defined. *Embodied classification applies to early vision.*

3.2 Phenomena and noumena

Another barrier to objective computer vision classification is the perceptual process itself. Bennet [10] offers an analysis of Kant's [11] distinction between phenomena and noumena. The word phenomena covers all the things, processes and events that we can know about by means of senses. Statements about phenomena are equivalent to statements about sensory states, or the appearance of things. From these, Kant distinguishes noumena as anything that is not phenomenal, something which is not a sensory state, and cannot be the subject of sensory encounter. Noumena are sometimes equated with the 'things in themselves'.

For this paper, the important distinction is that there are objects, processes, events, etc., that exist in the world, but as humans, we do not have access to objects themselves, rather to sensory states pertaining to objects. We can see implications of this above, although an object has reflectance properties, what is perceived is only the reflected light, and is dependent on lighting in the environment. For computer vision, a system cannot perceive everything about an object, as it does not have access to the object but only to its own sensory states. Regardless of the quality of sensors, complete information about an object can never be directly perceived.

Real sensor limitations restrict access to the environment properties even further. For example, a laser range finder returns an indication of the distance to the closest obstructed point on its path, for a finite array of beams, each of finite size. If a number of neighbouring sensor values that return the same distance, many robotic systems will assume this indicates the presence of an obstacle. Given an embodied system and a purpose, this may be a reasonable assumption. For example, for a

large robot navigating to a goal it is reasonable to assume that the path is blocked by some real physical entity that lead to these perceptions.[3] However, if we want a general objective scene description, we can assume no such thing. For example, if the sensor reading was caused by small tree branches that were aligned with the beams, small robot may fit between the gaps.

To summarise, sensors do not describe what exists in the world, only what they are able to perceive of it. Thus, we can reasonably assess if a sensor system is adequate for a purpose, but we cannot have a single all-purpose sensor.

We conclude that embodied categorisation is used at multiple levels of human vision, and that there cannot be a general vision system to handle all required categorisations. This does not say that particular vision algorithms are in any way deficient, just that the application of such algorithms under the assumption of a general vision is unsuitable for aspects of some problems, such as visual guidance of robots.

3.3 More recent computer vision

In this section, we examine three types of computer vision that are relevant to robotic guidance: some that do take an embodied framework; some research that is well-suited to an embodied approach; and, some recent work that takes the classical general vision approach. We do not attempt to review vision for mobile robot guidance as an extensive review has recently been published [13].

Active perception research emphasises the manipulation of visual parameters to acquire useful data about a scene [2]. For example, computation can be reduced by controlling the direction of gaze, which provides additional constraints, simplifying early vision [14]. Within active vision, researchers have linked perception directly

[3] A tour group may pose as a wall specifically to confuse guide robots [12].

to action, such as visual servoing, (e.g., [15]), where the control of robotic actuators is connected in control loops directly with features extracted from images.

Model-based methods match sets of features, which are derived from an image, to candidate values or value ranges. A match suggests that a particular structure or object is visible. In this way, the model specifies a description of an object in terms of features it can recognise. This interpretation of model-based representation does not assume there is a *uniquely* correct description for the visible part of the world. Model-based vision is often associated with tasks that explicitly categorise (e.g., object recognition). However, model-based vision can also apply to early visual tasks such as tracking. Drummond and Cipolla [16] render an articulated 3D model into the scene, allowing quite precise recovery of the world position of the tracked object. In contrast, Mansouri [17] examines a minimal model for tracking, assuming only intensity consistency and shape consistency (with deformation) in tracking the region of interest. Both are sound algorithms, however, for the cost of a model of the tracked object, Drummond and Cipolla gain accuracy and robustness.

The geometric viewpoint [18] formalises projection-based vision mathematically, including multiple view geometry (e.g., [19]) and visual motion of curves and surfaces [20]. Some research combines the geometric approach with robust statistics (e.g., see [19]). Studies of general visual geometry do not directly consider perceiver physiology. However, research into fundamental properties of imaging should not be regarded under the classical paradigm as there is no commitment categorisation theories of objects. Indeed such work supports a geometric approach to constrained viewpoint analysis, and so is highly applicable for an embodied approach.

Model-based computer vision is often disembodied. For example, aspect graphs, as discussed in [21], represent a series of generic views of an object. The views are geometrically derived, based on what theoretically may appear, without explicit consideration of what actually can be perceived. This can result in millions of views

being required to represent a complex object [22]. Matching may be a lengthy process, even with hierarchical indexing [23]. This is generally referred to as the *indexing problem* [22]. In Artificial Intelligence the problem of determining what knowledge is relevant to a particular situation is called the *knowledge access problem* [24].

However, some aspects of classical vision still appear in some recent papers. For instance, Shock Graphs [25] are used for content-based image retrieval. The indexing method gives more weight to larger and more complex parts, and models objects by their silhouettes. While both these ideas may be useful heuristics given no information about constraints for a particular image set, and may be particularly good for certain image classes, they may also be a degenerate choice. It cannot be assumed that one set of classification heuristics is suitable for all problems.

4 Applying Embodied Concepts in Human Vision

4.1 A philosophical viewpoint

There is a distinction that can be drawn between *viewing* a scene and *perceiving* something about the scene. Wittgenstein [1] uses the phrase *noticing an aspect* to describe the experience of noticing something about a scene for the first time, e.g., viewing a face and seeing (noticing) a likeness to another face. After noticing an aspect, the face does not change, but we see it differently. Noticing an aspect is not interpretation in a high-level sense as there is no conscious falsifiable hypothesis as to object identity made at this stage, although it may be made subsequently. Consider Figure 2 from [1], and the four descriptions in the caption: each provides a different suggestion of what the same diagram may be. By considering each one separately, we are able to *see* the diagram as one thing or another. This is not an interpretation about what the diagram represents. In seeing the diagram as an inverted open box

we perceive the diagram in a particular way, but do not necessarily consciously hypothesise that it is as such, although we may do so subsequently.

There are two ways to view Figure 3. After seeing one aspect, a mental effort is needed to see the other. In seeing one aspect, we are not necessarily saying this object *is* a cube with the top line on the front face. Wittgenstein notes that in seeing something *as* something for the first time, the object appears to have changed. We may have noticed an organisation in what we see that suggests a structure of what is being looked at. For instance, we may determine that two lines previously considered to be separate are actually a single line. A way of understanding seeing-as is to consider what a group of lines or features may be a picture of. For instance, Figure 2 could be a picture of any of the things that are described in the captions.

This type of seeing is conceptual. Human vision does not simply provide a list of objects to which reason can apply concepts and draw interpretations. Some form of categorisation may occur at multiple levels at an early stage in the visual process, such as identifying basic features, such as edges, and identifying grouping of features as described above. Perceptual mechanisms may classify underlying structures using concepts before a conscious hypothesis about object identity is formed.

4.2 A physiological viewpoint

Jeannerod [6] examines neuroscientific evidence pertaining to visual action, particularly reaching and grasping, tasks that are often considered to rely largely on early vision. Reaching and grasping are fundamental to human action and so may offer an insight into how some aspects of human categorisation have arisen, and thus may be useful in developing algorithms for robotic interaction. Jeannerod examines prehesion, preshaping of the hand in preparation to grasp while reaching. While the hand moves towards the object, the fingers and thumb shape based on factors in-

cluding object size and the type of grip required. The grip required can be classified as either a 'precision' or 'power' grip. In the precision grip, generally, the thumb and one or more fingers are in opposition, whereas in the power grip the fingers are flexed to clamp against the palm. The precision grip is for activities requiring fine motor control such as writing with a pen, while the power grip is stronger and not well-suited to fine motion interaction, but is used for tasks requiring force (e.g., hammering). Both grips can be used alternatively for almost every object. The intended task determines the type of grip used. What we consider important here is that basic visuo-motor tasks such as grasping that might not be thought to be categoric, require a categorisation of the task for which the object is to be used.

Further, Jeannerod presents a subject who, due to brain injury, is unable to shape her hand to reflect object size when reaching for unfamiliar objects. However, if the object is familiar, the subject is able to preshape with a level of accuracy typical of unimpaired subjects. Here categoric data from object recognition is assisting in basic reaching tasks. Specific biological evidence suggesting feedback from higher areas of visual processing to lower has been noted in the visual system also [26].

Another subject has difficulty naming objects. In seeing an iron, the subject is unable to say what it is, and mistakes what it is used for, but is able to indicate that it is used by moving it back and forth horizontally. Jeannerod comments that although identification of the object's attributes is preserved, such as attributes that are relevant to object use, the subject could not identify the object. They were unable to bind the perceptual attributes together in a way that allowed them to access its semantic properties. We see here multiple levels of categorisation.

In binocular rivalry [27] conflicting stimuli are presented to each eye. Frequently, subjects report being aware of one perception, then the other, in a slowly alternating sequence. A number of neurons in the early stages of the visual cortex that were generally active in association with one of the stimuli were active when that

stimulus is consciously perceived. However, a similar number were excited when the stimulus was visible, but not perceived. At the inferior temporal cortex, after the information has moved through other stages of the visual cortex, almost all neurons responded vigorously to their preferred stimulus, but are inhibited when the stimulus was not experienced. This suggests that the information from each eye moves through early stages of the cortex, before being suppressed in later stages. Here we see neural processes relating to multiple levels of categorisation, and some form of categorisation occurring early in the visual cortex.

The physiological evidence and Wittgenstien's insight show that there are multiple levels of processing and categorisation involved in perception, and that categorisation can be used early in the visual process. The process of categorisation is certainly not entirely bottom up, with some level of feedback evident. It also shows that high-level classification can assist in what might be considered to be early vision tasks.

4.3 Models play an analogous role in computer vision

In computer object recognition, categorisation can occur at multiple stages. Take a classical example: interpreting a line drawing image. There must be working hypothesis (model) as to what constitutes an edge pixel. Extracting edge pixels often leads to a set of broken edges. The line drawing of Figure 2 may be just one way of filling in the gaps. We may then decide on a working hypothesis that the edges correspond to a box where the concavity is below the two larger surfaces. Data driven algorithms for interpreting line drawings (e.g., [28]) cannot resolve such ambiguous structures, some form of model is required. Finally, after a model has been used to interpret the basic structure there may be many possible classifications that are consistent with that structure. A hypothesis must be made about object category. We may now decide that our box is consistent with a battery charger or a shoe

box. Other evidence, such as "we are in a shoe shop" may lead us to decide that it is a shoe box. Also, there may be feedback from categorisation at a higher level to lower levels. Here we see that multiple levels of classification are required within computational vision systems, in a manner analogous to the human visual system.

This forms an analogy with the process of "seeing as" as described by Wittgenstein. There may be ambiguity at some stage of visual processing about how image structure should be perceived. Ambiguity can be resolved by applying models set by feedback from later stages of visual processing, or directly by applying a model of the situation immediately based on known contextual information. In either case a model is a means by which conceptual intervention can be applied at any stage of visual process. In a similar manner the physiological studies described earlier showed that object identification could resolve difficulties in grasping.

Conceptual intervention may be necessary at the earliest stages of computer vision and can be applied through models.

Categorisation is also necessary for a vision system to guide a robot through a non-trivial environment. Simple categories like 'free-space' and 'obstacle' may be sufficient in some cases. Brooks [29] noted that a robot has its own perceptual world that is different from other robots and humans. The perceptual world is defined by embodiment. Robots may need embodied categories that are different from human categories to deal with their sensory world.

5 Embodiment of Vision-Guided Robots

We may select categories and features to be used in a robot vision model from the many possible categories and features using the constraints of the robot's embodiment. Brooks [30] considers that the key idea of embodiment is to ground regress of

abstract symbols. Being grounded is essential for real-world interaction, however, a robot's embodiment also constrains the relationships that it will have with objects and the environment. Thus, representations, both symbolic and non-symbolic are not only grounded, they can also be defined structurally in terms of robot embodiment and the impact of embodiment on environment interaction. A vision-guided mobile robot acts upon the world in a causal manner, and can perceive results of its actions. It is embodied in that it has a particular physical existence which constrains its interaction and perception. Some of these constraints are outlined in Figure 4.

It could be argued that a robot is also embodied in its software, for instance, if a vision-guided robot uses only edges, then it will be unable to distinguish objects that have similar basic structure with differing surface shape. This has been deliberately excluded here as it blurs the distinction between embodiment of entities that physically interact with the real world, and agents that exist only in a virtual world. This paper specifically addresses physical entities.

5.1 Embodiment, task and environment

Dreyfus argues that context is crucial to the way human intelligence handles the knowledge access problem. A global anticipation is required which is characteristic of our embodiment [24]. Searle [31] describes this as the background. The background is the enormous collection of common-sense knowledge, ways of doing things, etc., that are required for us to perform basic tasks. The background cannot be made explicit as there is no limit to the number of additions required to prevent possible misinterpretations, and each addition would, in turn, be subject to interpretation. With respect to different backgrounds, any visual scene has an almost infinite number of true descriptions. For example, a house could be 'my house', 'a charming Californian bungalow', or as in Figure 5. The fact that the robot has

physical embodiment means that it has an associated context, which incorporates purpose (task), spatial context (environment), and naturally, a temporal context. This context can apply to mobile robots in the same way as for humans.

5.1.1 The role of the task

Situation theory can be seen as anchoring parameters, such as how entities are categorised in the situation in which they occur [32]. In robotics research, the term 'situated' has often been used in the sense that entities in the world can be seen by a robot in terms of their relationship with the robot [30]. For example, objects may be classified as "the-object-I-am-tracking" or "everything-else" [33] rather than having a category that has meaning beyond the robot's interaction. In terms of conceptual embodiment, the task defines a particular perspective and helps designate the facts that are relevant and those that can be ignored. For a robot viewing the house mentioned above, different categories may be appropriate given different tasks. For example, in Figure 5, the same house may be categorised as an obstacle, or 'Uncle Bill's House' depending on the task. If it is an obstacle, a path around it may be all that is important, however, if it is a house that the robot needs to enter, it may need to know more, such as where the door is.

5.1.2 The role of the environment

The simple categorisation mentioned above is based on a blob segmented from a uniform background. This may be adequate for the environment the system inhabits, but may not be for others. The environment constrains a robot's possible experience of the world, the events that can occur, and the scenes and objects that may appear. It is known that humans recognise objects more quickly and accurately in their typical contexts than when viewed in an unusual context (e.g., fire-hydrant in a kitchen) [34]. The most appropriate conceptual model varies with the environment.

Note that although a robot needs to take the environment into account it does not mean that a system is restricted to a single environment. It is easy to imagine a mechanism for recognising a change of environment, e.g., moving from an office environment through a door into the outdoor world. The robot could then change to a different perceptual model, different behaviours, even different sensors.

5.1.3 Bringing embodiment, task and environment together

All three aspects of embodiment interact to determine classifications. Consider the interaction in Figure 8. Here, the task defines that a particular object is the object of interest, however, whether the other objects in the scene are obstacles, or can be put in the category of objects that can be ignored ("everything-else") is more complex. It depends on the physical embodiment of the robot (i.e., are the objects large enough to block the robot's path, and are they high enough above the ground for the robot to pass safely underneath them). It also depends on the task, as to whether the robot will be required to move close enough to the object that it could block the robot's task. The physical embodiment of the robot defines what may be an obstacle, and then, if the object blocks the path that the robot must take to complete the task, it can be considered to be an obstacle.

5.2 Where do categories come from?

Embodiment places constraints on the categories that a particular robot may have, but there are still many (maybe infinite) categorisation systems that would be equally good for the embodiment. For the purpose of the methodology described in this paper, the designer of the system chooses the categories. In the case of humans, categories must be learned from embodiment: physical, environmental, social and cultural. Robot category learning is difficult. There are examples of learning in

robots that are non-categorical (e.g., visuo-motor control [35]) and implicitly categorical, (e.g., systems that learn to avoid obstacles [36]). Further, localisation and navigation systems can be trained to associate features with locations (e.g., [37]). However, it could be argued that this is only mapping features to categories in a categorisation system that the robot was given, and falls well short of human-style high-level classification. We do not further proceed into this area in this paper.

6 A methodology of embodiment for visually-guided mobile robots

We now propose a methodology for applying embodied concepts to develop model-based visual systems for mobile robot guidance. The focus is on classification and perception, considering how a robot with a specific structure can perform effectively in the context of an environment and task. The aim is to show how the form and content of a conceptual model for a system can be constructed to take advantage of a robot's embodiment. We present vision-guided robot systems that have been constructed by the authors under the methodology of embodied categories, and highlight some other systems in the literature that have applied similar principles.

This methodology suggests how researchers can go about constructing categorisation for visual guidance systems. This is not intended to be a definitive way of proceeding: there may be many other equally good or better ways. However, the methodology elucidates a general approach, describes a possible process, and serves to aid the reader in understanding the role of embodiment in mobile robot categorisation. The first part directs analysis to the appropriate aspects of embodiment in constructing a useful categorisation for the system, and can be seen in Figure 6.

Note that other interactions are possible, such as that suitable discriminating features may not be available, leading to a different system of classification, or even redefining the task based on what is possible rather than desirable. Also, changes to

the physical embodiment may be necessary. For example, adding new sensor modalities that can better exploit differentiating features, modifying the robot itself (e.g., making it smaller) so that it can perceive more detail about objects of interest, or even simplifying the sensors if the required views simplify object discrimination. With these points noted, our suggested methodology is as shown in Figure 7.

Consideration of issues across stages, and iteration between stages is also necessary. For example, one may consider different possible feature sets given the difficulty of constructing the necessary hardware. Also, the task(s) and environment(s) may underconstrain the embodiment, allowing introduction of constraints on embodiment to simplify interaction, and/or the views and thereby the requirements for stage 5.

Note that this methodology contains an implicit partial commitment to view-based representation. [4] There is some psychophysical evidence to support the theory that humans make use of view-based object representations. Bulthoff and Edelman [39] found that if two views of unfamiliar objects were learned, recognition performance was better for views spanned by the training views than for other views.

The remainder of this paper attempts to clarify how the methodology may work in practice, and clarify the principles. We present three systems developed using the methodology, and other systems that exemplify the principles of embodied concepts.

6.1 Object recognition for robot circumnavigation

In [40,21], we presented a system where an embodied approach was used to redefine traditional viewer-centred representations. This enabled a robot system to identify and navigate around known objects, and gain specific computational and recognition advantages. The embodiment of the robot allowed model-based representations to

[4] Not necessarily appearance-based such as [38], e.g., Section 6.1 coming up.

be simplified and optimised for the task, environment, and physical embodiment, and hence made more practical. Figure 9 shows images taken by the robot as it navigates around a power supply, with a cluttered background.

Stage 1: Defining categories. Three categories are important for this task: the object of interest, obstacles, and free-space. The robot is required to move around the object on the ground plane, continuous fixation on the object is important so that searching is not required.

Stage 2: Identify features. Edge features were adequate to discriminate the required objects for this system in the required environments. However, edge features may be unambiguous for a particular view, but identification will be more certain after examining a number of views around the object. To recognise obstacles, a simple method of detecting edges on the floor was used, with any strong edges assumed to be an obstacle, similar to the method of [41].

Stage 3: Determine physical embodiment. A ground-based robot is adequate for the motions required, a pan/tilt platform is necessary for independent fixation. Given that the camera is not looking where the robot is going, a second forward-looking camera is required for obstacle avoidance.

Stage 4: Identify views. The robot is ground-based, and has fixed camera height, thus, if the object sits on the ground, all views of the object are within a plane. Objects of similar size to the robot are unlikely to be viewed from underneath or above. Further, finite camera resolution, combined with the fact that the robot body may overhang the camera lens will prevent the robot from viewing the object from very close proximity. If the task is to navigate around the object, only a coarse model is needed. If task is to interact/dock, detailed models of some surfaces may be required. The views from which the robot will observe the object were determined by a combination of projective geometry and images taken from the possible paths.

Stage 5: Build models. We chose to use view-based edge models of objects. As the possible viewspace of the object is restricted, the storage of a full 3D model is redundant. The *scale problem* [22], at which scale features should be modelled, is problematic when viewspace is unconstrained. The constrained viewspace as discussed above leads to a finite range of scales over which features can be observed. Combining this with the finite camera resolution the scale is effectively defined in building object models for this system. The level of detail required can be directly quantified by the robot taking images of the object from the required path.

The path of the robot is continuous, and so can be indexed by order of appearance. Once an object is recognised, the robot knows which view to expect next. Given that the object is stationary, the robot's next view is caused by the robot's action (with associated uncertainty). We refer to this as causal indexing, that is indexing our representation by the interactions that the robot has with the object. Thus for the particular case of robot navigation, we have a solution to the indexing problem.

Stage 6: Look at interactions. In a cluttered scene, a unique match is difficult to guarantee (consider the possibility of a mirror in the background). However, as the robot is moving around the object, the system exploited several views of different surfaces, fusing the matches with odometric information, which makes mismatches less likely. This is facilitated by causal indexing.

A brief description of the matching process will help to clarify benefits of embodied classification. To match the object, we take the previous match position, and subsequent odometry information, and estimate which view is most likely to appear. The predicted view has a set of edges with restricted orientations due to the constrained viewpoints, so only edges in this range need to be extracted initially. We then find possible candidate edge matches based on orientation, pre-sorting to reduce the total number of match candidates. Binary features (involving two edges) are then used to find candidate view matches. Finally, we evaluate the small number of remaining

candidate matches, partly based on geometric verification. A candidate match is back projected into the scene to obtain an estimate of relative object position and orientation. This is compared to the position estimated from odometry and the previous match. Combining motion information into matching in this manner reduces the number of mismatches, and reduces the effect of mismatches. If the previous match was correct and the current match is incorrect, the object position estimate cannot be far from the true location. This gives graceful degradation.

Finally for interactions, we may decide that an environment is particularly cluttered and so mismatches will be frequent. Thus, we may consider using a robot with high odometric accuracy to facilitate narrow tolerances on the geometric verification.

6.2 Using Log-polar Optical Flow And Fixation For Docking

In [42,43], we presented an algorithm that is able to control robot heading direction to dock at a fixated point. Fixation may be independent of heading direction control, and joint angle information is not required, only log-polar optical flow is required. This approach is different to the previous as, other than some constraints, the environment is unknown, and so the robot embodiment is only partially constrained. Also, the visual processing used is low-level.

Stage 1: The only categories necessary are whether the current heading direction is left-of or right-of the fixated target. This information is used directly to control adjustments to heading direction in a perception/action loop. **Stage 2:** No particular environment is considered *a priori*. Independent fixation was assumed. Fixation may require environmental constraints, but this can be considered independently. **Stage 3:** The robot is required to move on the ground-plane and maintain fixation on the object. The robot's method of locomotion is not constrained, but the robot must have a means for pan and tilt to allow fixation independent of motion.

Stage 4: The ground-based robot motion (physical) and fixation on the object (task) place constraints on the optical flow field which simplify its interpretation. **Stage 5:** Given these constraints, the log-polar sensor separates the motion field such that the component due to motion along the fixation direction only appears in the radial flow. The rotational flow is due entirely to motion perpendicular to the fixation direction. Thus, rotational flow can be used directly to infer the direction of the required adjustment to heading direction. See [43] for a full derivation. This type of action categorisation is typical of active perceptual systems, (e.g. [44]).

There is one final ambiguity: given the same heading, world points further from the robot than the fixation point result in rotational flow in the opposite direction than for points that are closer. This can be resolved using environmental constraints, e.g., if the target is on the floor, then all points below it in the image are closer to the robot than the target. Thus, the robot can estimate heading direction based on the sign of the rotational flow, and control its direction for docking in a closed perception/action loop without camera calibration or knowledge of joint angles.

6.3 Docking Based On Fixation And Joint Angles

We developed a docking system for legged robots that was applied in the four-legged league of robocup [45]. This system fixated on the object of interest, and controlled two variables: heading direction and approach speed. The robot should move to be close to the target and stop when it arrives. Whatever interactions are required can then be performed. **Stage 1:** There are four action categories: turn left, or turn right to head towards the target, move forward (has not yet reached the target), or stop (at the target). **Stage 2:** The target object is on the ground plane, otherwise same as the previous algorithm. **Stage 3:** Same as previous algorithm. However, the robot camera must be higher than the fixated object to facilitate control from

joint angles. **Stage 4:** Again the task involves fixating on the target object. This time, however, we make explicit use of the joint angles. We also make use of the fact that the robot's head is elevated above the ground, and assume that it is somewhat higher than the fixated target, which is assumed to lie on the ground. Note that, by definition, the fixated object lies on the optical axis of the camera. **Stage 5:** If the robot turns to reduce the camera pan angle to zero it will be heading directly towards the object. Also, the tilt angle is proportional to the distance from the target (related by $\tan\theta$), see Figure 10. For large distances this relation will not be accurate, however, when the robot moves close to the fixation object, the measure will allow discrimination of at-target and not-at-target categories.

Our motion model permits rotation and translation to be treated independently from a control point of view. Above we have two independent perceptual variables for the control of rotation and translation. Thus, control can be implemented simply as two interpolated lookup tables, one of pan angle, and one of tilt angle.

Consider the alternative (non-embodied) approach. Calculate the distance to the target by modelling the ball size, calibrating the camera, and using an inverse perspective transform. Fixation arises out of robot embodiment, without it the target may not be centred in the image, so we must also consider image position to estimate the ball position. Then we need to transform to body coordinates, and calculate control parameters for required motion. This clearly requires far more computation time than image based fixation and two table look-ups, and is less robust. We may introduce errors in the transforms, and due to these errors, and the increased computation time, we may loose the ball from camera view. We believe that the embodied approach has lead to a system that is more computationally efficient, robust, and simpler to construct. The approach here is simple and intuitive, and demonstrates a point which is central to the argument of this paper. Constructing robotic perceptuo-motor interaction that is categoric through use of an embodied

methodology can naturally lead to solutions that are superior in computation and robustness when compared to disembodied systems.

6.4 Extraction of Shape

The final example is a shape module used with circumnavigation system discussed above. This was not developed under the methodology, but illustrates the benefits of feeding back hypothetical classifications to gain better results in early vision processing. The module [46] combines knowledge about the basic object structure (given that we have a hypothetical object classification) with knowledge from edge matching. This knowledge is applied in order to add constraints and simplify tasks in shape-from-shading, making tasks solvable that are ill-posed otherwise. Also, computation time advantages can be gained through environment (or task domain) knowledge by better initialisation of surface models for an iterative fitting process.

6.4.1 Other active vision systems

Finally, some active vision algorithms from other researchers also illustrate the benefits of explicitly considering embodiment. Most methods for detecting obstacles consider the image projection of the ground plane and map the image position of detected obstacles directly to robot body centred coordinates. This is not possible unless the specific geometry of the robot is considered. However, in an embodied approach where the robot's view geometry is considered and the ground is largely planar, this projection is quite robust.

The methods exploit the visual appearance of their environment. Some methods assume a floor with a constant textureless surface (e.g., [41]), possibly using higher-level interpretation to deal with exceptions. Other methods assume sufficient texture for optical flow (e.g., [47]) or stereo (e.g., [48]). Obstacles distort the flow pattern

that would be expected to arise from the relative motion of the ground plane.

Techniques based on constrained projective geometry determined by robotic embodiment are common in active perceptual-based systems, such as divergent stereo and other systems for docking (e.g., [49]). In *divergent stereo*, robots navigate along corridors remaining centred between textured walls. Using the fact that the robot is moving along the ground, and has cameras pointing sideways at the walls, the system can move to equalise the optical flow [44,50], which centres it in the environment. The actual appearance of optical flow in this type of situation is different for particular robots, however the method can be effective for a variety of different robots, with control parameters particular to the robot.

7 Conclusion

Philosophical, physiological and psycho-physical research shows that human vision is reliant on categorisation, and that this categorisation is embodied categorisation. We have explored the role of categorisation in robotic vision and the role of embodiment in this categorisation. A physically embodied robot is present in an environment, and typically engaged in tasks. The physical embodiment of a robot, and its tasks and environment constrains the relationship that the robot has with entities in the world. Specifically, it constrains how the robot can perceive and interact with other entities. These constraints can be used as a basis for robot classification and object models. The construction of classification and models based on embodiment is referred to here as conceptual embodiment. As discussed, the classical general formulation of computer vision is inadequate for guiding mobile robots. By the application of conceptual embodiment, low-level vision techniques can be made more efficient and robust, and high-level model-based vision techniques can be made effective for robot guidance. Consideration of embodiment can lead to the development of algorithms

for problems that are otherwise ill-posed, and can produce systems that are more computationally efficient and more robust.

This paper makes two principal recommendations: that categorisation is useful at all stages of visual processing; and, that for vision guided robots, categorisation should be embodied. We have presented a methodology for developing embodied systems, and presented several systems that have been developed based on this methodology, as well as examining other research that has demonstrated benefits from taking embodiment into account.

References

[1] L. Wittgenstein, Philosophical Investigations, Blackwell:Oxford UK, 1996.

[2] Y. Aloimonos, Introduction: Active vision revisited, in: Active Perception, Lawrance Erlbaum Assoc., Hillsdale, NJ, 1993, pp. 1–18.

[3] D. Marr, Vision : a computational investigation into the human representation and processing of visual information, W.H. Freeman, NY, 1982.

[4] G. Lakoff, Women, Fire, and Dangerous Things, University of Chicago Press, 1990.

[5] M. Johnson, The Body in the Mind, University of Chicago Press, 1987.

[6] M. Jeannerod, The Cognitive Neuroscience of Action, Blackwell, Oxford, UK, 1997.

[7] J. Dupré, The disorder of things : metaphysical foundations of the disunity of science, Harvard University Press, Cambridge, Mass, 1993.

[8] F. J. Valera, E. Thompson, E. Rosch, The Embodied Mind: Cognitive Science and Human Experience, MIT Press, 1993.

[9] B. Berlin, P. Kay, Basic Color Terms: Their Universality and Evolution, Berkley:University of California Press, 1969.

[10] J. Bennet, Kant's Analytic, Cambridge University Press:Cambridge, England, 1966.

[11] I. Kant, Critique of Pure Reason, Orion Publishing Group:London, England, 1994.

[12] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun, Experiences with an interactive museum tour-guide robot, Artificial Intelligence 114 (1-2) (1999) 3–55.

[13] G. N. DeSouza, A. C. Kak, Vision for mobile robot navigation: a survey, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (2) (2002) 237–267.

[14] D. H. Ballard, Animate vision, Artificial Intelligence 48 (1) (1991) 57–86.

[15] S. Hutchinson, G. D. Hager, P. I. Corke, A tutorial on visual servo control, IEEE Trans. on Robotics and Automation 12 (5) (1996) 651–670.

[16] T. Drummond, R. Cipolla, Real-time visual tracking of complex structures, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (7) (2002) 932–946.

[17] A. R. Mansouri, Region tracking via level set pde's without motion computation, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (7) (2002) 947–961.

[18] O. Faugeras, Three-dimensional computer vision : a geometric viewpoint, MIT Press, Cambridge, Mass, 1993.

[19] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, Cambridge, UK, 2000.

[20] R. Cipolla, P. Giblin, Visual Motion of Curves and Surfaces, Cambridge University Press, Cambridge, UK, 2000.

[21] N. M. Barnes, Z. Q. Liu, Vision guided circumnavigating autonomous robots, Int. Journal of Pattern Recognition and Artificial Intelligence 14 (6) (2000) 689–714.

[22] O. Faugeras, J. Mundy, N. Ahuja, C. Dyer, A. Pentland, R. Jain, K. Ikeuchi, Why aspect graphs are not (yet) practical for computer vision, in: Workshop on Directions in Automated CAD-Based Vision, 1991, pp. 97–104.

[23] J. B. Burns, E. M. Riseman, Matching complex images to multiple 3d objects using view description networks, in: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 1992, pp. 328–334.

[24] H. L. Dreyfus, What Computers Still Can't Do: A Critique of Artifical Reasoning, The MIT Press:Cambridge, Mass., 1994.

[25] D. Macrini, A. Shokoufandeh, S. Dickinson, K. Siddiqi, S. Zucker, View-based 3-d object recognition using shock graphs, in: Proc of the 16th International Conference on Pattern Recogntion, Vol. 3, 2002, pp. 24–8.

[26] D. C. V. Essen, C. H. Anderson, D. J. Felleman, Information processing in the primate visual system: an integrated systems perspective, Science 255 (1992) 419–424.

[27] R. Blake, N. K. Logothetis, Visual competition, Nature Reviews. Neuroscience 3 (1) (2002) 13–23.

[28] K. Sugihara, Machine Interpretation of Line Drawings, MIT Press, Cambridge, MA, 1986.

[29] R. A. Brooks, Achieving artifical intelligence through building robots, Tech. Rep. 899, MIT Artificial Intelligence Laboratory (1986).

[30] R. A. Brooks, Intelligence without reason, Tech. Rep. 1293, MIT Artificial Intelligence Laboratory (April 1991).

[31] J. R. Searle, The rediscovery of the mind, MIT Press, Cambridge, Mass., 1992.

[32] T. Winograd, Three responses to situation theory, Tech. Rep. CSLI-87-106, Center for the Study of Language and Information, Stanford University, Ventura Hall, Stanford, CA, 94305 (1987).

[33] I. D. Horswill, R. A. Brooks, Situated vision in a dynamic world: Chasing objects, in: AAAI 88. Seventh National Conference on Artificial Intelligence., 1988, pp. 796–800.

[34] I. Biederman, Perceptual Organisation, Lawrance Erlbaum Assoc., Hillsdale, NJ, 1981, Ch. On the Semantics of a Glance at a Scene, pp. 213–253.

[35] G. Metta, F. Panerai, R. E. S. Manzotti, G. Sandini, Babybot: an artificial developing robotic agent, in: Proc. SAB 2000, Paris, France, 2000.

[36] C. Gaskett, L. Fletcher, A. Zelinsky, Reinforcement learning for a vision based mobile robot, in: Proceedings of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems.IROS2000, Vol. 1, Takamatsu, Japan, 2000, pp. 403–9.

[37] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, D. Schulz, Probabilistic algorithms and the interactive museum tour-guide robot minerva, International Journal of Robotics Research 19 (11) (2000) 972–99.

[38] Y. Matsumoto, M. Inaba, H. Inoue, Visual navigation using view-sequenced route representation, in: Proc. IEEE Int. Conf. on Robotics and Automation (ICRA '96), Vol. 1, Minneapolis, Minnesota, 1996, pp. 83–88.

[39] H. H. Bulthoff, S. Edelman, Psychophysical support for a two-dimensional view interpolation theory of object recognition, in: Proceedings of the National Acadamy of Sciences of the United States of America, Vol. 89, 1992, pp. 60–64.

[40] N. M. Barnes, Z. Q. Liu, Knowledge-Based Vision-Guided Robots, Physica-Verlag, Heidelberg, New York, 2002.

[41] I. Horswill, Visual collision avoidance by segmentation, in: IROS '94. Proceedings of the IEEE/RSJ/GI Int. Conf. on Intelligent Robots and Systems. Advanced Robotic Systems and the Real World, 1994, pp. 902–909.

[42] N. M. Barnes, G. Sandini, Active docking based on the rotational component of log-polar optic flow, in: W.-H. Tsai, H.-J. Lee (Eds.), ACCV Proceedings of the Asian Conference on Computer Vision, 2000, pp. 955–960.

[43] N. M. Barnes, G. Sandini, Direction control for an active docking behaviour based on the rotational component of log-polar optic flow, in: European Conference on Computer Vision 2000, Vol. 2, 2000, pp. 167–181.

[44] J. Santos-Victor, G. Sandini, Embedded visual behaviours for navigation, Robotics and Automous Systems 19 (3-4) (1997) 299–313.

[45] G. Baker, N. M. Barnes, An integrated active perceptual behaviour for object interaction, Tech. Rep. 2001/23, Univeristy of Melbourne, Vic, 3010, Australia (2001).

[46] N. M. Barnes, Z. Q. Liu, Knowledge-based shape from shading, Int. Journal of Pattern Recognition and Artificial Intelligence 13 (1) (1999) 1–24.

[47] J. Santos-Victor, G. Sandini, Uncalibrated obstacle detection using normal flow, Machine Vision and Applications 9 (3) (1996) 130–137.

[48] Z. Zhang, R. Weiss, A. R. Hanson, Obstacle detection based on qualitative and qantitative 3d reconstruction, IEEE Trans. on Pattern Analysis and Machine Intelligence 19 (1) (1997) 15–26.

[49] J. Santos-Victor, G. Sandini, Visual behaviours for docking, Computer Vision and Image Understanding 67 (3) (1997) 223–38.

[50] K. Weber, S. Venkatesh, M. Srinivasan, Insect inspired behaviours for the autonomous control of mobile robots, in: M. V. Srinivasan, S. Venkatesh (Eds.), From living eyes to seeing machines, Oxford University Press, Oxford, 1997, pp. 226–248.

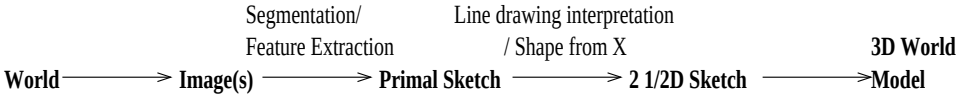


Fig. 1. Marr's model of computer vision.

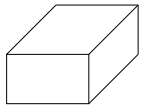


Fig. 2. Wittgenstein's line-drawing could be described as: a glass cube; an inverted open box; a wire frame in a box shape; or, three boards forming a solid angle.

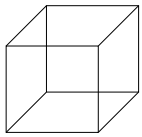
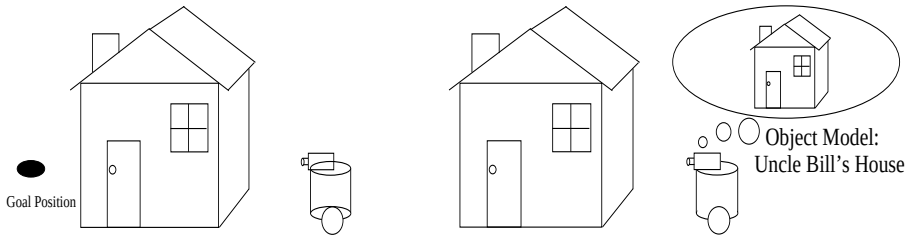


Fig. 3. The Necker cube. Is the top horizontal line or the bottom horizontal line actually on the front face of the cube?

(1) Robot movement is constrained by its physical bulk. A particular robot's structure restricts its ability to fit into places, to bring its sensors close to objects, and constrains the operations it may perform.
(2) Robot movement is constrained by kinematics and a movement control system. Although some robots are omni-directional, many have finite turning circles, and are only able to move in certain ways. In cluttered environments this leads to the *piano movers problem.*
(3) The types of surfaces and environments a robot can traverse are restricted. Many robots, due to lack of ground clearance, structural robustness, or motor strength, are restricted to indoor operations, and even may have difficulties with rough floor surfaces, inclines, and stairs. For outdoor robots, much of the earth's surface is not traversable by wheeled vehicles. Although, legged robots can travel on a greater range of surfaces, they have other limitations.
(4) Most robots travel on the ground, and hence never perceive objects from some viewpoints, or act upon them from certain angles. Also, depending a robot's size, small obstacles can sometimes be ignored. Similarly, overhead shelves do not obstruct robots that are shorter than them.
(5) Robot perception is constrained by sensor limitations. For instance, cameras require sufficient light (e.g., visible, infra-red), and contrast to distinguish an edge between two objects. Digital cameras only perceive features at a particular scale. Features smaller than one pixel are indiscernible, while objects larger than the pixel array cannot be perceived from a single viewpoint. The effect of camera resolution on scale depends on the robot-to-object distance and camera zoom. Further, odometry always has errors, so position is seldom precise.

Fig. 4. Embodied constraints on a robot.



(a) (b)

Fig. 5. Different tasks may require different categorisations for the same object. (a) Task: move to goal position. Description: an obstacle. (b) Task: Find 'Uncle Bill's house'. Description: 'Uncle Bill's house'.

(1) The **task(s)** defines what categories, and actions/motions are required.
(2) In the **environment(s)** of robot operation, some set of features must be adequate to discriminate the categories.
(3) The required actions/motions of the robot and the features which the robot needs to discriminate categories place constraints on the **physical embodiment** of the robot.
(4) Once the robot's **physical embodiment** is determined, it can be considered along with the **task(s)** to define which views of the environment and objects in it that the robot may encounter and need to identify.
(5) Finally, to find which features are adequate to discriminate categories, consider their appearance from the views encountered in the robot's **environment(s)**.

Fig. 6. Analysing robot embodied categories.

(1) Analyse the **task(s)** to identify the required perceptual categories, and actions/motions.
(2) Analyse the robot operating **environment(s)** to identify possible sets of features that differentiate the required categories.
(3) Analyse required actions/motions, the physical positions from which the robot may observe discriminating features, and the sensors required to detect these, to determine appropriate physical **robot embodiment**.
(4) Analyse possible interactions the robot may have, examining both **physical embodiment** and **task(s)** to identify the views from which the robot will be required to discriminate the required perceptual categories.
(5) Analyse the instances of the categories of stage 1, from the views found in stage 4, and find a set of features, and representations for those features, that are effective and efficient in recovering the required information in the **environment(s)** where the **task(s)** is to occur.
(6) Analyse interactions that emerge to look for category changes that may be required, and restart at stage 1 if new categories are found.

Fig. 7. A methology for developing embodied categories.

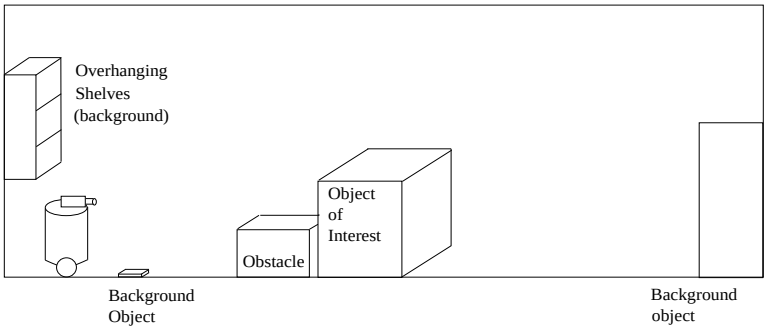


Fig. 8. The task defines the object of interest.

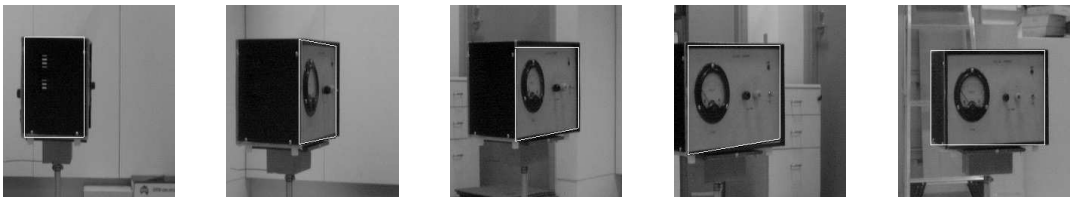


Fig. 9. Images taken while cirumnavigating a power supply. Reprinted with permission from [40] (page 200, Fig. 9). ©Physica-Verlag Heidelberg, 2002. All rights reserved.

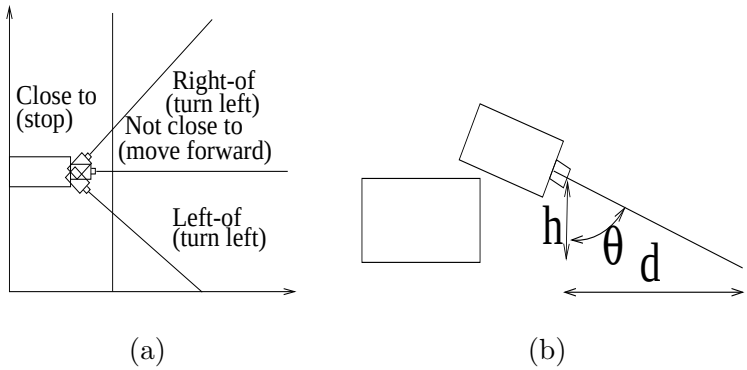


Fig. 10. Action categories for the joint angle based docking system.