FP6-004250

CoSy

Cognitive Systems for Cognitive Assistants

Integrated Project

Information Society Technologies

D.10.1

Initial report on simplest scenarios and their requirements

Due date of deliverable: 1/12/2004
Actual submission date: 15/7/2005

Start date of project: September 1st, 2004   **Duration:** 48 months

Organisation name of lead contractor for this deliverable:
University of Birmingham

Revision: final
Dissemination Level: PU

Executive Summary

The objective of workpackage 10 is to begin to integrate components from the different disciplines in the project into two prototype systems driven by two complex scenarios: the Explorer and the PlayMate. Because research in this project is scenario driven this report breaks down early work on both the Explorer and the PlayMate scenarios into a series of component (or sub-)scenarios. In this report we describe the earliest of these. Tackling some of these will require building systems that integrate information from several modalities, and combine action and perception.

The outline of the report is as follows. We briefly describe each early scenario and summarise early executed or proposed experiments for each one; we also discuss some of the requirements on a system arising out of these scenarios; and present detailed templates for some scenarios (included as an annexe). The report concludes with some issues to be addressed during later stages of the project.

Role of Simplest Scenarios and their Requirements in Cosy

This workpackage will result in initial robots that bring together multiple kinds of functionality in well-integrated working systems. Since integrated systems are complex this report describes how we have initially broken the Explorer and PlayMate scenarios down into simpler scenarios (or tasks). These have been chosen in a forward chaining manner, working from what we know as a way to achieve initial integration. This is separate from the parallel process of chaining backwards from more detailed versions of complex scenarios to simpler tasks. This process of backward chaining to discover the other sub-problems we will solve will be an important contribution of the project, since we believe that it reduces (but doesn't eliminate) the risk of solving unnecessary sub-problems.

Relation to the Demonstrators

Since the demonstrator systems will integrate the various components this report should make clearer how those components can fit together.

Simplest Scenarios and Their Requirements

1 Scenarios for the Explorer

Before we outline the scenarios we are working on we recap on the two proposed over-arching scenarios: the Explorer and the PlayMate. The goals for the Explorer scenario over the four years of the project are introduced in the workplan as follows:

> *[Exploration/Mapping of Space – The Explorer] A fundamental competence of a cognitive agent is the ability to recognise its own interaction with the environment and to be able to generate an internal model of its environment as a prerequisite for operation and interaction with the environment. In this scenario we address the issue of self-recognition and insertion into the environment. This involves modelling of perception, interaction and recognition of the embodiment. Once such a competence is available the agent can move about in the environment and generate an internal model of the environment. Without a complete/comprehensive method for categorisation of objects/structures in the environment the agent must be assisted by a tutor to recognise / disambiguate structures in the environment. The emphasis of this scenario is thus on spatial models and self-awareness.*

Over the first eighteen months the more specific aims are:

> *For the Explorer scenario we shall set up an experimental platform that has facilities for mobile manipulation with a suite of sensors (stereo vision, haptic feedback, speech recognition / synthesis). Through monitoring of sensory information in combination with generation of actions it is possible to achieve self recognition as studied in WP3. In addition the system must have basic facilities for navigation through the environment and mapping of space. This work will be based on existing methods available within the consortium [23, 44, 51]. For the mapping of space and detection of objects there is also a need to deploy methods for recognition and categorisation. Component methods are available within the consortium, [31]. For the interaction with users a text and speech interface will be used. A key issue during the early phase of the project is to integrate the facilities into an operational platform. As a range of different platforms are used across the involved partners there is a critical need to integrate the work with the effort on architectures and to consider the design of a Hardware Abstraction Layer (HAL) that allows easy transfer of results across institutions.*

2 Scenarios for the PlayMate

The goals for the PlayMate scenario over the four years of the project are introduced in the workplan as follows:

> *[Models of objects and concepts – The PlayMate] For interaction with the environment, reasoning about scenes, and communication with other agents it is essential that the system has facilities for acquisition of models of objects, events and structures. This involves both static objects and dynamic phenomena. The number of objects present in a natural environment calls not only for recognition, but for categorisation of objects to place them in a spatio-temporal hierarchy. For operation and interaction with objects*

pure RE-cognition will facilitate a limited complexity of interaction and there is thus a need to recognize/classify affordances as a more generic level of modelling of structures and objects. In this scenario the emphasis will be on categorisation of objects, events and structures. Again the agent will cooperate with other agents in its environment to accomplish its missions/goals. In addition the agent will actively explore the environment to "discover" new objects and structures.

In the first 18 months there are a more specific set of aims:

For the early integration as part of the PlayMate scenario, many of the basic competencies from [the explorer] scenario ... can be used. There will thus be sharing of methods for mapping, basic mobile-manipulation control, recognition/categorisation, and dialogue systems. In addition there is, however, a need to have more advanced facilities for manipulation of objects to allow for interaction with objects for discovery of a simple set of affordances. Moreover, whereas many of the tasks in the Explorer scenario can treat the robot as a point moving around in a two dimensional space, the PlayMate will constantly be confronted with structured 3-D objects and the need to produce and perceive 3-D motions involving both translation and rotation. Further, the task of constructing complex objects from simpler ones may have something in common with the task of constructing complex routes from simple route-fragments in the Explorer, but the details will be very different. The differences are exacerbated by the fact that some of the time part of the robot is itself one of the perceived complex 3-D structures in motion, e.g., when it is grasping and moving an object. Because one of the styles of learning used in this workpackage will be tutor driven learning the use of language to allow communication between the robot and the tutor will be important. In the PlayMate scenario we also expect to sketch the integration of very early results from workpackage 5 on modelling objects and their affordances with those from packages 8 and 9 on communicating about actions, objects and locations.

Unfortunately there are few if any standard libraries of manipulation skills that can be used for this purpose. Consequently the initial set of manipulation skills will be limited to pickup, pushing, and transfer of simple block type objects. It is important to point out that the project will not be limited to block type objects, but such objects are adequate for initial integration and early experiment on discovery of affordances. At a later stage more 'natural' objects, such as occur at a "tea party" will be introduced.

3   A Robot at a Tea Party: a long term motivating scenario?

In this and the following sections we outline some of the specific scenarios for the PlayMate and the Explorer to be tackled in the early stages of the project. For the first period of the project these are:

1. Where am I? Knowing where you are.

2. What is it? Categorising objects.

3. What happened? Recognising actions.

4. Where is it? Spatial relationships between objects.

We can see how these are relevant to our long term goals by seeing how they could contribute to a unifying scenario, or task. Imagine a robot which has to learn a sequence of actions performed by a human on a set of objects, e.g. while laying out items for dinner, the robot may have to pick up a mug and place it on a saucer with the handle near the hand position of the user given where the user will sit. The action sequence changes the spatial relationships of the objects. The robot is shown the task, while the action and the goal is described by the human tutor. The robot then copies the tutor, achieving what a human would regard as a similar end state. The robot needs to extract from the visual and speech input what the goal of the activity is, and what parts of the activity are important to reproduce. The robot will then be asked to perform the same task, with different objects in equivalent roles (e.g. "put the ball on the plate next to the mug"). It is important to note that we will not implement this precise scenario, it is simply a way to motivate and tie together the scenarios we outline in the following sections. Indeed some of the abilities we discuss will remain beyond the state of the art for some time to come.

To perform the task the robot must be able to extract information about spatial relationships that hold between objects in the scene. This means that it must have visual routines for identifying and categorising objects, segmenting them from the background, and calculating their relative locations. In addition it must be able to connect the description of the scene it generates from vision to that from speech input from the human tutor, and to ensure that these are consistent. Second it will be necessary for the robot to understand references to objects, in order that it can follow instructions, and where necessary take non-linguistic information into account in order to resolve ambiguous references. Third, the robot will be required to recognise the sequence of actions performed by the human tutor, and map those onto actions it can itself perform. This will require that it is able to recognise where part of the action ends and the next part begins. It will also need to be able to understand descriptions of those actions and their purpose as described by the human tutor, e.g. ``I put the red cup on the red saucer''. Human descriptions like this explicitly give some of the goal, i.e. the relation of the cup and the saucer, but they don't tell the robot everything — ideally we would expect the cup to be placed on the saucer so that the handle is graspable by a human sitting in front of it, so either to the left or right side, not in front of the cup's body relative to the user, or behind it.

Such a robot will also require the ability to recognise places, and to navigate from place to place, following instructions from a human, such as ``Fetch the plates from the kitchen.''. The robot might have to ask questions to obtain information about their location in more precise terms, before navigating there, recognising both the room, the correct area within the room, not to mention the items themselves.

Other requirements also arise from this scenario. Having learned the conventional sequence of actions, and the constraints on the goal state from the human tutor the robot must then carry out the correct generalisation of the action sequence when instructed — e.g. laying out similar items in a new place setting, perhaps on the other side of the table. Following instructions will require that the robot is able to map the description of the action to a sequence of its own actions. Finally the robot must also be able to reason about what the objects are to be used for by the humans, as this will place constraints on the task. If the robot has a model of the fact that cups are grasped by their handles, then it will be able to set a place correctly for a person. In general for more sophisticated kinds of manipulation the robot will have to be able to predict what the effects of particular actions are on different types of objects (e.g. balls and cans roll, blocks don't), and be able to infer what features (visually) of an object allow it to behave in that way.

We argue that this unifying scenario — laying a table for a meal — based on copying actions, fetching items from other places, and following instructions to lay out objects at for a robot tea party constitutes a grand challenge for robotics. The task is reasonably easy to define precisely, in terms of success and

failure, it is easily motivated in terms of engineering problems that need to be solved for the real world, and we argue that it captures a range of requirements for skills that will significantly push forward the state of the art in robotics, without being so far beyond the current state of the art that no-one has any idea of how to tackle the problem.

For each of the scenarios we give an informal overview of the scenario in the main text. For one scenario we also give a template setting out more details. These templates set out the assumptions and restrictions, so that the set of cases which the robot should be able to handle are well defined. Sometimes in this document we will refer to sub-scenarios, tasks and target interactions. By these we mean the following. A sub-scenario (or equivalently a task) is an intermediate step to one of the scenarios listed here, but one that is small enough that it is not useful to separate it. This may be because it only directly supports one other scenario. A target interaction is a script outlining the sort of qualitative behaviour we expect to see in the scenario, and is therefore a basis for qualitative evaluation.

4 "Where am I?" — Knowing where you are

4.1 Introduction

We want to be able to interact with a mobile robot, such as a service robot, in a variety of ways. Specifically we would like to be able to teach the robot, give instructions, and have simple conversations about places in the world. We can imagine a variety of target interactions, and in this scenario we are working towards these. What does it mean to say *where* I am? In some circumstances a robot may need to recognise that it is in a specific place ("I am in the kitchen"), in others it may need to categorise the place it is in, ("This room is a kitchen"). In some instances we may want to teach the robot about locations by naming them, and in still others we will want the robot to drive that acquisition of labels by asking for them. Thus, there is not just one precise target interaction, but rather a cluster of closely related scenarios and competences arising out of the requirements for them. We describe the ones we are tackling in detail below. They are, in order:

- "Telling you where you are" — Tutor driven exploration.

- "Telling me where I am" — Tutor assisted exploration.

- Recognising indoor places.

- Learning perception-action maps.

- View-based localisation.

- A hierarchy of cognitive maps.

4.2 "Telling you where you are" — Tutor driven exploration of space

The robot operates in a dynamic, possibly only partially known environment. This means that it needs to learn more, e.g. through communication with other agents. In this scenario we explore a tutoring setup, in which a tutor supervises the robot in its exploration of the environment, instructing the robot what a given location is. The challenge is to construct the quantitative perceptual models (*line-based SLAM*), qualitative models based on ontological concepts and described linguistically (*concept-labelled topological graphs*), and the mapping between these models. This mapping is done

synchronously: we assume that the linguistic description applies to the room that the interlocutors are currently situated in.

The type of interaction we are dealing with here is (primarily) tutor-driven. The tutor tells the robot what kind of room they are currently situated in. The tutor may refer linguistically to the room, and may direct the robot's attention to it by giving it e.g. commands to look or go somewhere.

(1) Sit Robot (R) and human (H) in same room; R in corner, H outside field of view of R.

    H.1 "Hello."

    R.2 Turns body and head around, scanning for H; upon finding H, looks at H. "Hello."

    H.3 "Come over here."

    R.4 "Okay.". R drives up to H. When ±60cm from H, stops, and looks up to H.

    H.5 "Follow me." H starts walking.

    R.6 R recognizes H moving. "Okay." R starts following H.

    H.7 H arrives to a new room. "This is the laboratory."

    R.8 R looks around the room. R looks up at H. "Thank you for showing me the laboratory."

Example (1) illustrates several basic aspects of the kind of interaction we are interested in: The robot pays attention to tutor, providing verbal and non-verbal grounding feedback (cf. [42] for the importance of meaningful gaze in human-robot interaction), it can follow up movement commands, and it is capable of acquiring mapping information (while indicating explicitly it is doing so).

4.3 "Telling me where I am" — Tutor assisted exploration of space

In tutor-driven exploration of space, the initiative lies with the human tutor. We can move to a more mixed-initiative setting. For example, the robot explores parts of a space itself and asks for feedback from the tutor, or asks for clarification questions regarding the situation. The following two examples illustrate this. In both cases, the robot tries to determine the possible entrances to a room.

(2) Sit The robot (R) and the human (H) have just arrived in a room that the robot did not know about yet.

    H.1 "This is the laboratory."

    R.2 R looks around the room. R looks up at H. "Thank you."

    R.3 "Can I have a look around?"

    H.4 "Sure."

    R.5 R starts driving around the room, constructing a map and a route graph.

    R.6 R recognizes, from laser data, that there is a doorway to its left.

    R.7 R drives into the doorway. R turns to H. "Is there a door here?"

    H.8 "Yes, that is correct."

Example (2) illustrates the robot requesting the initiative (R.3), acquiring it (H.4), and then moving around the room to explore it (R.5). While exploring, the robot uses line-based SLAM to constructed a map, and a route graph. The nodes in the route graph can be coloured, to classify them as belonging to the same "space" (e.g. the same room), or as a doorway. Upon recognition, the robot poses a confirmation question (R.7), to ensure it is correct in its classification.

Example (2) examplifies how we can use dialogue to enhance the information in a map. The example below, Example (3), illustrates how the robot can also use dialogue to correct its map.

(3) Sit The robot (R) and the human (H) are walking around the room. The robot creates a route graph, and classifies a node as a door. However, the nodes immediately preceding and succeeding the door node are both classified as belonging to the same room – indicating a possible inconsistency in the map.

R.1 "Excuse me."

H.2 "Yes?"

R.3 "Was there a door just back there?"

H.4 "Where?"

R.5 R drives back to the location corresponding to the node classified as a door. "Here".

H.6 "No, there is no door there."

R.7 "Okay, thank you." R corrects its map, and drives back to the position it was last at, before returning to the misclassified node.

Example (3) illustrates how the robot requests the initiative to ask a question (R.1), and upon acquiring it -tentatively- (H.2) asks a clarification question (R.3). Because the tutor does not understand what the robot refers to (H.4), the robot drives to the location that its question is about (R.5). The tutor can then resolve the reference, and answer the question in the negative (H.6). The robot then uses this information to correct the misclassification of a node in its route graph.

If we are to achieve these target interactions we will need certain basic competences. Being able to learn and talk about places requires that we are able to recognise them from sensory data, and locate ourselves and the place within some kind of representation of space. In particular we focus on building a set of competences that take visual input. We describe the competences, the scenarios we will use to test them, and our approaches to achieving them in the following three sub-sections.

4.4 Recognising indoor places

Imagine that we teach a robot the name of a room. If we wish the robot to recognise the room in the future, to be able to name it, its recognition routines will have to be robust to a number of variations. First, the visual appearance of a place varies in time because of several factors: illumination changes due to different light sources (day-light vs artificial lighting); small changes due to everyday activities (like people using the room, objects being removed from cupboards, chairs being moved around and so on); pieces of furniture can be re-arranged, be removed or even replaced, and so forth. A robust cognitive explorer will need to be able to recognise a place reliably despite these variations in its visual appearance.

We assume an indoor environment with a fixed number of rooms, such as an office or a private house. We will also assume that the system will start to build its representations of the environment from scratch (in other words the robot will have no prior knowledge of place categories). We are trying several different ways of posing the learning problem. In one approach we intend to deal with the problem in a purely supervised manner. In the other approach we will try both supervised, and semi-supervised learning in stages. In the first stage of learning the system will build the initial representation from labelled data (supervised learning); then the system will update and enrich its representation on self-labelled data (semi-supervised learning). The focus will be on two research issues:

1. **Holistic versus landmark based strategies:** It is an open issue whether a scene should be seen and recognized as a whole (the holistic approach), or should be considered as a collection of objects, and therefore be recognized on the basis of a few key landmarks. Both approaches have

been tested in purely visual scenarios, and both have strengths and weaknesses. We will investigate both strategies in our indoor place recognition scenario, first in a controlled setting and using only visual cues, then on a robot integrated with geometric and topological information. Topological regions will be used for generation of local maps that could be estimated using traditional geometry based methods such as Folkesson and Christensen [16, 17].

2. **Holistic and landmarks strategy** A third way to place recognition, seldom explored in the literature, is to use both types of information as input to some kind of cue integration algorithm. We take two approaches to integrating the global and local information. In the first approach we try to extend our previous work on disciminative accumulation for cue integration [37]. In the second approach [36, 35] there are two stages. In the first we extract simple geometric features from laser scans, and landmark objects from images. A supervised learner (AdaBoost) then learns how to combine these. In addition we attempt to improve the final classification by incorporating information about the likely trajectory of the robot. For example, if the classification of the current pose is "kitchen", then it is rather unlikely that the classification of the next pose is "office" given the robot moved only a short distance. To get from the kitchen to the office, the robot has to move first through a doorway or a corridor.

4.5   Learning perception-action maps

We are investigating the problem of spatial insertion- the agent's self-image that models the relationship between perception and action. The aim is to make the agent understand how its control actions influence its sensory input. This knowledge will be used to build a sensori-motor map that the agent can use for vision-based motor control.

The classical approach is to have objects stored in a map by their visual aspect and their associated spatial locations (coded either in absolute or relative coordinate systems). This approach has two major drawbacks. It requires a good estimate of the robot's position, which is very hard, and the maps have to be rebuilt when changes in the environment occur.

We want to achieve flexible and robust robot control without using prior knowledge about the robot and its environment in form of internal models. Our research is inspired by O'Regan's and Philipona's ideas of sensorimotor dependencies, and in particular by an algorithm developed by Philipona [9], [10]. The outcome of the algorithm is a 'Lie bracket' that contains motor commands for performing translations and rotations, respectively. Philipona's algorithm simulated a virtual agent in a virtual environment. We want to use the idea of sensorimotor dependencies on a real mobile robot acting in a real world environment. A possible approach would be to implement Philipona's algorithm directly. A problem with this direct approach is that Philipona's algorithm assumes a convex world, that is a world where the agent's senses capture all the information about the environment. In the real world, the robot is equipped with a camera that has a limited field-of-view and at each point in the environment there will be different parts of the environment that are occluded. Another problem is that it assumes perfect sensing. In reality, the sensory information that the robot gets is noisy, the sensor's properties (camera internal parameters) may change over time and the robot may be moving in conditions of unusual lighting, where the response characteristics of the sensors might be severely modified.

In stead of implementing the algorithm directly, our solution involves estimating Lie brackets that are related to certain objects in the environment. By doing so, we solve the problem of occluded objects in the environment. In practice, this means that we have to partition space into different Lie bracket areas. For example, if the robot is in an office and it has to move to the coffee container in the kitchen,

the robot does not see the coffee container in its start image. To be able to move to the kitchen, the robot has to perform certain Lie bracket motions in the office, other Lie bracket motions in the corridor linking the office to the kitchen and other Lie bracket motions within the kitchen. This means that we will estimate different Lie brackets in different parts of the environment and we have to build an algorithm that senses when the robot crosses the boarder of Lie bracket areas and should change its motor commands accordingly.

The perception-action map will contain information about what Lie brackets to use to get from one place to another in an environment. To be able to estimate the Lie brackets, a place where the robot can move around at random is required. Moreover, there have to be enough features in the environment to estimate the Lie brackets.

4.6   View Based Localisation

Aside from being able to localise using geometric information (from, for example, lasers), and to recognise places using visual information we also wish to be able to buildmaps and localise within them using visual information. In this scenario, and the next one, we therefore employ vision as the primary sensor for acquiring data from the environment, and for estimating the robot's current position [13].

The operator places the robot into a room or an environment and instructs it to move across the room in a certain direction. While moving, the robot acquires images and stores them along with the odometry information from the robot, thus exploring the environment. The operator can then instruct the robot to explore another parallel path in another part of the environment or let the robot decide on which areas are interesting for exploration. If the operator names the places the robot is exploring (as in tutor driven or tutor assisted exploration as described above), the language component can provide the labels for different parts of the environment. The system can select a small set of images showing prominent elements of different parts of the environment according to the labels. These images serve as visual location placeholders.

When the robot has explored the environment to some extent, the operator can instruct it to navigate to one of the goals by naming the destination. Alternatively, the operator can show the robot an image of a scene visible from the desired goal location. The robot then needs to localise itself from its current viewpoint image, and then navigate to the destination viewpoint.

Our approach to this problem is as follows. Our localisation technique [2] relies on matching full image appearances of the current query view to some reference view with known viewpoint parameters. In the exploration phase we store a set of images from viewpoints on a straight trajectory into a spatio-temporal volume. We use an image-based rendering [41, 4] (IBR) technique to produce approximations of the views at arbitrary (virtual) viewpoints [50]. To get a localisation estimate of a query viewpoint, we use IBR to synthesize hypotheses and then match the best hypothesis to the query view. View-based localisation [24, 3, 29] is a simple yet powerful technique for matching images and therefore also for robot localisation. However, to get a good estimate, we need reference views with known parameters as close as possible to the query location. With IBR, however, we need to have the actual views from only a fraction of possible locations, while we are able to approximate or predict the views at arbitrary locations not directly visited during exploration.

4.7   Localisation using a hierarchy of cognitive maps

In this scenario, as in the one above, vision is used as a primary means for acquiring data of the environment and position estimation. The operator guides the robot through the environment. While

moving, the robot acquires panoramic images and stores them along with the movement direction acquired by an onboard electronic compass. The operator can also use language to give labels to certain part of the environment.

When the environment has been sufficiently explored the robot uses the model described in [47] to build a hierarchy of cognitive maps. The robot performs localisation using only an acquired panoramic image and upon instruction from the operator navigates towards the destination.

The model described in [47] implements a hierarchy of cognitive maps based on panoramic images of the environment. The resulting map consists of place cells placed in a topologically consistent metric space. The number of place cells is determined by multiple eigenspace analysis [32] of the stored image sequence. The formation of the cognitive map is achieved by passing subspace representations of panoramic images to a computational model inspired by Hafner [20]. A physical force model is applied to translate the non-metric map to a sparse topological map with metric information using local relative orientations only. Finally, a hierarchy of maps is formed in order to implement different levels of representations.

4.8 Concluding remarks

One of the main themes in the scenario "Where am I?" is how space should be represented. We have a variety of sensory sources — here vision, language and laser scans — and these lead naturally to two broad ways of characterising a place. We can treat the place itself as a single entity (the holistic, or global approach); or we can attempt to characterise it as a collection of landmark objects (a local approach). We are also trying to combine these approaches. Then there is the problem of how we localise within a map where places are represented in such sophisticated ways. If we have a collection of places that have been represented visually, and wish to localise ourselves within that "map", how should we do it? This has led us in turn to consider a variety of representations of how action and sensing are related, in other words the different kinds of maps there are and how they can be learned and used.

5 "What is it?" — Categorising objects

5.1 Introduction

One of the main requirements of a visually enabled cognitive system is the ability to **recognise objects**. A cognitive robot has to be able to use the visual information to detect and recognise or categorise objects in its surroundings. It also has to be able to acquire the knowledge about the objects through learning and dialogue with a tutor. These particular issues are addressed in this scenario.

The first specific research issue that will be emphasised is **incremental learning**. The learned object representations should be continuously updated over time, adapting to changes, considering newly encountered views of the known objects and adding new representations of novel objects. Different **modes of learning** will be explored: *tutor-driven learning* of objects where the initiative will be on the side of the tutor, and mixed-initiative *tutor-assisted learning*. Such a scenario requires a rich interaction between the robot and the tutor, which will be achieved in a user-friendly way through a dialogue. The interaction between **vision and linguistics** will thus play an important role. The last specific research issue addressed in this scenario is **object categorisation**. The robot will not only be able to correctly identify already seen objects, but also to categorise novel objects according to the learned representations of visual object categories.

Before describing several examples, we briefly outline the overall scenario. At the beginning the tutor shows to the robot a few objects (individually) and names them, and the robot learns their representations. Then several objects are put on the table and robot has to point and name them. If an object can not be recognized, the robot asks the tutor about the object identity, and after it receives the answer, it updates its current knowledge (adds the representation of the new object or updates the current representation of the object if it is known). Many novel objects will be then presented to the robot (pottery, toys, etc.). It will have to keep continuously recognising them and updating the current representations. When some novel objects, similar to the already presented ones, will be introduced, the robot will be able to correctly categorise them. In all ambiguous situations, the robot will solve the ambiguity via a dialogue with the tutor.

5.2 "Telling you what it is" – Tutor-driven learning of objects

In the tutor-driven mode of learning, the tutor takes the initiative and teaches the robot about the objects. We explore a setup in which the tutor shows the robot various new objects, and tells the robot what these objects are.

(4) Sit The robot (R) is facing a particular object, which is -thus- salient in its field of vision.

  H.1: "This is an $\langle object \rangle$."

  R.2: R turns its head to look at the object. R constructs a visual model of the object.

  R.3: R turns its head to look at H. "Thank you for showing me the $\langle object \rangle$."

Example (4) illustrates the basic idea. Assuming the 'new' object is visually salient for the robot, the tutor provides a description of the object (H.1). The robot then provides feedback that it has understood the tutor by first non-verbally focusing at the object (R.2), and then by turning to the tutor and replying verbally (R.3). The object may be new or previously already presented. The robot updates the current representations and adds a new one if necessary. The tutor can also rotate the object to present it from different viewpoints (in later stages, the robot will also be able to pick up and rotate the object by itself and observe it from different viewpoints to obtain a more complete representation of the object).

Of course, if the tutor wants to teach the robot a new object which is not currently visually salient, the tutor may direct the robot, e.g., through a spatially referring expression. Example (5) illustrates such a situation.

(5) Sit The robot (R) is facing the tutor (H). To the left of the robot is the object to be learned.

  H.1 "To your left there is an $\langle object \rangle$."

  R.2 R turns to its left. R turns its head to get an object in its center of vision. R constructs a visual model of the object.

  R.3 R turns back to look at H. "Thank you for showing me the $\langle object \rangle$."

The situation may get more complicated if the tutor's reference is ambiguous, or if the robot thought it is looking at a different (known) kind of object. In both cases, the robot would need to be able to pose clarification questions: Example (6) illustrates disambiguation, Example (7) correction.

(6) Sit The robot is told there is an object to its left. When looking left, it discerns two (unknown) objects surrounding a known object X.

  R.1 "Do you mean the object on the left of the X, or the object on the right?"

     H.2 "I meant the object on the left."

     R.3 R turns its head to get an object in its center of vision. R constructs a visual model of the object.

(7) Sit The robot is told there is an object to its left. When looking at the object, the robot recognizes it as an object of type X, different from the type asserted by the tutor.

     R.1 "I am sorry, I thought this was an X."

     H.2 "No, this is a Y."

     R.3 R uses the information and updates the representations accordingly.

5.3 "Tell me what it is" – Tutor assisted learning of objects

There are several possibilities for moving beyond a completely tutor-driven interaction. The following examples illustrate different situations in which the robot takes the initiative to guide learning. Just like the case for the spatial exploration scenario, these examples underline the general idea of seeing learning as collaborative interaction.

(8) Sit The robot (R) is observing the objects in its surroundings. R realizes that it has encountered something it does not know.

     R.1 "Excuse me."

     H.2 "Yes?"

     R.3 "What is this?" R turns the head to look at the object in question.

     H.4 "That is an $\langle object \rangle$."

     R.5 R constructs a visual model of the object.

     R.6 R turns head to look at H. "Thank you."

Example (8) illustrates a case where the robot realizes it is looking at an object it does not know yet, and that the tutor does not seem to intend to tell the robot more about it. The robot may decide that it would like to know what object it is looking at, and as such requests (R.1) and acquires the initiative (H.2). The robot then asks what object it has encountered, making explicit what object it means by looking at it – a form of deictic reference (R.3). The tutor replies (H.4), the robot updates its knowledge (R.5) and thanks the tutor (R.6). A variation of this example is a setting in which the robot systematically explores the environment, looking for all objects it cannot recognise/categorise. In contrast, in the examples below, the robot is explicitly exploring the environment, looking for objects that instantiate the concept it has been taught.

(9) Sit The tutor (H) has just shown the robot an object of type X. The robot (R) now starts exploring the environment for more objects of this type.

     R.1 R notices an object and looks at it.

     R.2 R recognizes the object as being of type X.

     R.3 "Is this an X?"

     H.4 "Yes, that is an X."

     R.5 R updates the current representations.

While scanning the environment, the robot encounters an object which, when paying attention to it (R.1) it recognizes as being an object of type X (R.2). It then asks a confirmation question (R.3), which the tutor answers in the affirmative (H.4): the observed object is also of type X. The robot then updates its knowledge accordingly.

5.4 Concluding remarks

The described examples will serve as general directions defining the tasks the robot will have to be able to accomplish. They are not specified in detail and allow for gradually increasing the level of difficulty considering different problems the robot may encounter. From the computer vision point of view, for instance, detecting an object may be an easy task if the segmentation is given (or the background is uniform), however it can turn into a very complex problem when the background is cluttered or the objects are partially occluded. Also, the object recognition/categorisation can become very difficult by changing the illumination conditions, scale, viewpoints, enlarging the number of objects and categories, varying the visual variability within and between categories, etc. We will start accomplishing the tasks described in the example scenarios in easier settings and then gradually challenge more complex conditions.

We will go beyond the state of the art in several ways. From the visual learning point of view, the incremental way of learning is not a very common approach. Most of the state of the art algorithms follow the standard paradigm, dividing the off-line learning stage and the recognition stage. Most of these approaches are not designed in a way that would enable efficient updating of the learned model, so we will pay a special attention to this problem. We will also explore different modes of learning and underline the role of the embodiment and a physical and a verbal interaction with the tutor. We will also propose new methods for object categorisation, which will go beyond the current state of the art techniques.

We will take different approaches (considering various local and global features, multiple cues, different types of representations (generative, discriminative, combined), hierarchies of categories and representations, etc.) and combine these approaches to achieve the best performance in accomplishing the tasks described in this scenario.

6 "Where is it?" — Spatial relationships between objects.

6.1 Introduction

A basic ability for a robot that is able to manipulate objects in collaboration with a human is the ability to understand verbal references to objects, their properties and their spatial relationships within a scene. In this scenario we will attempt to build a robot that is able to display some understanding of, and have simple conversations about, some the spatial relationships that can hold between objects on a table top.

(10) Sit Robot (R) and human (H) at the same desk; with objects as arranged in Figure 1.

    H.1 "Hello."

    R.2 Turns body and head around, scanning for H; upon finding H, looks at H. "Hello."

    H.3 "Look over there." H points at some items on the desk.

    R.4 Robot turns to look at the objects at which H is pointing.

    H.5 "What is to the right of the cube."

    R.6 "Do you mean the red cube?"

    H.7 "Yes."

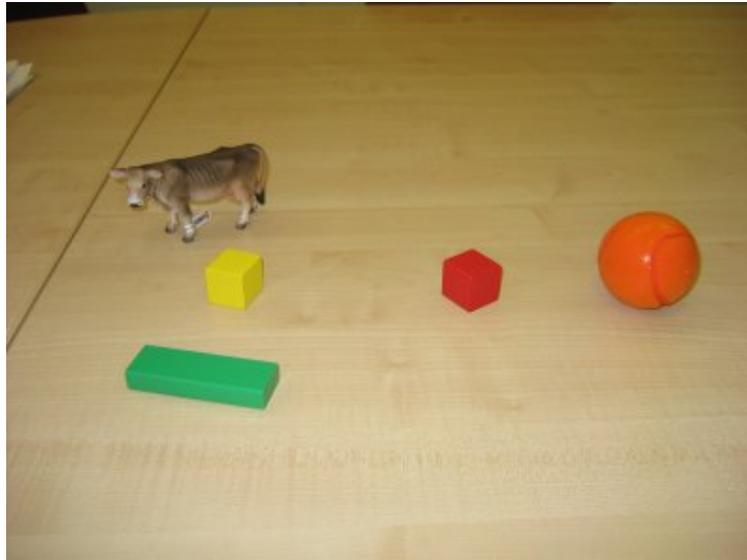   R.8 "There is a ball to the right of the red cube."

Figure 1: A desktop scene.

6.2 Modelling spatial relationships between objects

To answer such a question the robot must be able to model the qualitative spatial relationships in the scene. This requires in turn that we are know how to extract the qualitative spatial relationships from a visual scene and how they can be represented in a way that can be linked with the representations produced by parsing utterances by the human. To satisfy these requirements we will develop language resources for understanding utterances about spatial relationships. We will also develop routines for extracting information about the same relationships from the visual input. The categorisation of these spatial relationships will need to be context sensitive. Humans, for example, modulate their use of the expression ``X is near Y'' according to the other objects in the scene, and the relative size of the objects in the relationship.

The linguistic conception of space is essentially relativistic. In general, the location of an object is described relative to another object, called a landmark. Within English, prepositions are the main class of words used to describe these spatial relationships. The set of English spatial prepositions can be divided into those that predominantly describe static relationships and those that describe dynamic relationships. Static relationships only describe the location of an object relative to a landmark, e.g. "the X 'to the right' of L". Whereas, dynamic relationships describe both the location and path of the object relative to the landmark, e.g. compare the paths in "the X went "along" the road" versus "the X went "across" the road". In this scenario, we are focusing on modelling the semantics of static spatial prepositions. The set of static spatial prepositions can be further divided into topological relations (i.e. prepositions that describe the location as an object as proximate to the landmark, e.g. at, on, in etc.) and projective relations (i.e. prepositions that describe the location of an object as being in a region that is in a particular direction relative to the landmark, e.g. in front of, behind, etc.).

The approach we adopt is grounded in the cognitive linguistic tradition (inter alia, [43, 45]). This approach focuses on the schematic nature of the human conceptualisation of space that underpins spatial language. The concept of a "spatial template" [33] is taken as basis for our semantic models of prepositions. Based on psycholinguistic findings (inter alia, [8, 25]) we have created a contextually

based model of topological proximity [27] projective prepositions [28]. Using these models we have defined a cognitively motivated hierarchy of spatial contexts. This hierarchical approach addresses the issue of combinatorial explosion that effects the construction of relation based contexts. This hierarchy has been used to develop a framework for generating locative descriptions [26].

6.3   Planning to resolve ambiguity in reference

In the target interaction the robot also had to ask a clarifying question to tie the reference made by a speaker to an object to an object the robot can see. One of the problems of human speech is that references to objects are often underspecified linguistically, and that the robot may need to incorporate other information to resolve the reference. Another example is the case below:

(11)   Sit   The objects are arranged as in Figure 6.1.
     H.1   "What is to the left of the red cup?"
     R.2   "Do you mean the large red cup?"
     H.3   "Yes."
     R.4   "There is a green ball to the left of that cup."

Here the tutor has asked the robot about the identity of an object with a particular relationship to another object (a red cup) in the scene. First of all the question makes a reference to a red cup as a landmark, and indirectly — through the landmark — to another object, the green ball. Answering the question requires that the robot is capable of identifying the landmark even though the reference is ambiguous. In this case the human has assumed a frame of reference that is the same as the robot, but it could be that the human is sitting opposite the robot and is using their own frame of reference. Therefore, there are at least two ways of interpreting the reference "to the left of the red cup" depending on the context. In addition the reference to the red cup includes statements about its properties: its type and colour. If the reference to the red cup is ambiguous (here there are two cups with significant

areas of red, each with an object to their left), then the robot must take action to resolve this ambiguity. This could involve checking to see if the human is pointing at the object, or it could involved asking a clarifying question as here. If the second is necessary we will require a system for generating clarifying questions and for deciding which question is most appropriate. In addition the system must be able to incorporate information from either language or vision.

In tackling this scenario since it will not be feasible to construct a model of the entire scene in order to identify or resolve references it will be necessary to build an attentional system to direct the robot's processing. Currently there are systems that can resolve [19] and generate [30] unambiguous references to objects, and that can resolve ambiguous references to objects in a simple way (asking "which one is it?")[39]. There are currently no robot systems that can plan dialogue or other actions to resolve ambiguous references to objects in a more sophisticated way. Our approach is to use techniques from information state MDP planning to address this problem [48]. The system should eventually be able to answer questions about the absolute attributes of objects (e.g. colour), their relative attributes (e.g. the big box), and their spatial relationships to one another (near, to the left of, behind, in, on top of).

7 "What happened?" — Categorising actions

7.1 Introduction

A robot that works with humans needs to be able to learn and reason about activities that it, or human collaborators, can perform with objects. A simple requirement is that such a robot must be able to recognise and categorise actions. This includes biological motions such as walking, and actions involving objects. In the first sub-scenario involving actions we will collect a database of labelled simple actions. By *simple* we mean that the actions have a natural atomic nature, i.e. they are not naturally decomposed into sequences of sub-actions. In later versions of this scenario (beyond the first 12 months of the project) we will tackle complex actions that are composed of sequences of the simple actions we will tackle in this scenario.

The simple actions will be filmed from different view points, and we will produce a system that gives a classification of the type of action in each sequence. We will also attempt to learn such classifications. We will not restrict ourselves to supervised learning, but will also investigate unsupervised approaches. One of the main research issues will be whether to represent the action sequence globally, or whether to segment the images into parts (e.g. into the hand, the background and the object) and produce a state based description. We will consider both types of approach.

We also need the robot to be able to self-identify events in which it is involved. If the effects of its actions are fully observable then this is trivial. It becomes far more challenging, however, when unexpected events occur that are not directly detectable. Such events are usually associated with different types of failure. Our aim in the second sub-scenario for "What happened?" is to be able to infer the type of failure or the cause of an unexpected event.

7.2 Global models of actions

Acquiring knowledge about actions that people perform in a robot's environment requires different levels of precision for different tasks. A robot needs to be able to learn some global representation of actions in the environment, so that it can recognize and distinguish them efficiently, and focus with higher precision only on tasks that it needs to learn to perform.

To enable generalized learning of global action models with a human using one or more objects it may be beneficial not to assume geometry. There have been few attempts of motion-based learning

of actions without assuming a specific geometric model. Most of these approaches are based on modeling of optical flow fields [6, 49, 14], motion history [7] or on the modeling of manifolds of local features [40, 38]. We propose that instead of modeling the motion directly, prediction of local motion features can be exploited as a cue for activity recognition.

Prediction has been used in the literature to model video dynamics and the dynamics of geometrical models. Fablet et al. [15] have used causal probabilistic models to represent video dynamics. Jebara and Pentland [22] have used Gaussian probabilistic models to predict reaction from an interval of action. Agarwal and Triggs [1] have demonstrated that a mixture of regression models can be used as a predictor with a geometrical model. Bissacco et al. [5] have used subspace angles between autoregressive models of skeletal angles to recognize gait.

We will explore global action models based on autoregressive predictors of local spatio-temporal velocity features to enable approximate modeling of arbitrary geometry in motion. Such models should enable generalization of actions performed by different people, even when trained on very short video sequences.

7.2.1 Training and testing scenario

The robot has a general idea where the tutor will perform some kind of action with one or more of a collection of objects but it does not have any information about the number and geometry of the objects. The objective of the robot is to learn to visually distinguish a set of simple actions in order to be able:

- to recognize a vocabulary of simple actions;

- to generalize similar actions among people and different objects;

- to recognize simple constituent actions in longer sequences of actions.

The robot observes the area where an action is about to happen. The "tutor" performs a basic action with one or more objects. The robot has to be able to:

- learn the action;

- identify the action as known;

- identify the action as unknown or at least quantify dissimilarity from known actions.

For the first set of experiments different people as tutors perform different manipulative actions with different objects in different surroundings and label the actions. The scientific objectives are:

- to identify the proper level of generalization of a vocabulary of actions to be able to recognize them in practical scenarios using global motion based representations of actions;

- to develop methods to identify discriminant localities in different actions within sequences of different actions that are similar in particular intervals;

- to investigate the applicability of motion based action representation with little information about the state of the world (basically only the foreground/background mask) and object geometry.

7.3   State based models of simple actions

As stated previously, from the visual observation of a human performing a task, we want the robot to be able to say what class of activity it sees; e.g. pick up an object, put it down, move it to another location, rotate it, etc. Arguably, state-based models such as finite state machines (FSM) [21, 34] and hidden Markov models [18] can be used in addition to motion pattern analysis to account for many of these activities. One weakness of many state-based approaches is that the activity models are carefully handcrafted and often account for single object motion. In this approach to the scenario, we shall attempt the automatic construction of the activity model of multiple objects (i.e., an agent acting on an object).

In a state-based model of an activity, a state represents a primitive motion state of an object and the transitions among states represent the temporal constraints (i.e., all possible dynamics of the activity) Our main challenge is the segmentation of a video sequence into appropriate states and the construction of the models for primitive motions. Primitive motion states can be recognized by tracking some features on the object surfaces and cluster the motion into some distinct classes. The difficulty of the learning task depends on the degree of supervision. We shall assume the activity label and a rough (possibly incomplete) textual description of video content are available (e.g., approach, grab, move, etc.). Such information and other prior knowledge about the domain can simplify the search for an optimal number of states in a FSM and for clustering motion. An activity can be recognized by searching for an FSM that best matches the motion patterns derived from the video sequence.

7.4   What went wrong? Diagnosing faults

The previous two sub-sections have dealt with our approaches to recognising the actions of other agents. In this sub-scenario we will deal with identifying unexpected events involving the robot itself. This problem arises because an autonomous robot has to function in complex environments with unpredictable influences. The components of the robot, like motion-control, navigation, perception, or higher level reasoning, use simplified world models to efficiently carry out their tasks. Failures can occur when assumptions introduced by these models no longer hold. As an illustrative example, take a deflated tyre on a wheeled robot. If the motion-control component assumes wheels of equal size, as is usually the case, the robot loses its mobility partially, or in the worst case completely. It is therefore a basic requirement for an autonomous robot to include fault detection mechanisms that reliably identify unmodelled system states.

We approach the fault diagnosis problem using probabilistic state estimation techniques [12, 11, 46], where the belief about the state of the system is represented by a discrete set of samples. This set is successively updated using the performed actions and sensor measurements to derive the belief distribution after each time step. For different system models that define possible actions and their effects, we can choose the dimensionality of the state samples and the belief updates accordingly. In this formulation, a fault corresponds to a switch from one system model to another. This switch can be executed for a portion of samples to estimate the fault likelihood during operation. In the example above, the system model for unequally inflated tyres includes an additional variable for the difference in wheel diameter and thereby captures the motion dynamics correctly. We consider complex faults like the collision with unrecognized obstacles (e.g. stairs, small boxes, or gaps between floor planks) and defects in tyres and motors. An important question that will be adressed in this context is how much the local environment and the performed actions influence the detectability of certain types of faults.

8 Concluding Remarks

This paper has introduced and summarised the four main scenarios on which we are working within the project during the first 12 months. Within each of these we have described the research issues, and some of the cross-cutting themes. Also, within each scenario, there are a number of requirements or competences that we are working on to achieve the target interactions. For each requirement or competence we are in some cases also exploring more than one approach to the same problem. A running theme through all the scenarios is that of representation, and particularly the choice we have been global and local representations. In some instances we attempt to combine these, whereas in others we investigate them separately.

9 References

References

[1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *ECCV*, volume 3, pages 54–65, 2004.

[2] Matej Artač, Matjaž Jogan, Hynek Bakstein, and Aleš Leonardis. Panoramic volumes for robot localization. To be published at IEEE/RSJ International Conference on Intelligent Robots and Systems, August 2–6, 2005, Edmonton, Alberta, Canada, August 2005.

[3] Matej Artač, Matjaž Jogan, and Aleš Leonardis. Mobile robot localization using an incremental eigenspace model. In *IEEE International Conference on Robotics and Automation, May 11–15, 2002, Washington, D. C.*, volume 1, pages 1025–1030. IEEE, 2002.

[4] Hynek Bakstein and Tomáš Pajdla. Rendering novel views from a set of omnidirectional mosaic images. In *Proceedings of Omnivis 2003: Workshop on Omnidirectional Vision and Camera Networks*, Los Alamitos, USA, June 2003. IEEE Computer Society Press.

[5] A. Bissacco, A. Chiuso, Yi Ma, and S. Soatto. Recognition of human gaits. In *CVPR 2001*, volume 2, pages 52–58, 2001.

[6] M. J. Black, Y. Yacoob, and X. S. Ju. Recognizing human motion using parameterized models of optical flow. In Mubarak Shah and Ramesh Jain, editors, *Motion-Based Recognition*, pages 245–269. Kluwer Academic Publishers, 1997.

[7] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[8] L.A. Carlson-Radvansky and G.D. Logan. The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37:411–437, 1997.

[9] Philipona D., O'Regan K., and Nadal J.-P. Is there anything out there? inferring space from sensorimotor dependencies. *Neural Computation*, 2003.

[10] Philipona D., O'Regan K., Nadal J.-P., and Coenen O.-M. Perception of the structure of the physical world using unknown multimodal sensors and effectors. *Advances in Neural Information Processing Systems*, 2004.

[11] N. de Freitas, R. Dearden, F. Hutter, R. Morales Menendez, J. Mutch, and D. Poole. Diagnosis by a waiter and a mars explorer, 2003.

[12] Richard Dearden and Dan Clancy. Particle filters for real-time fault detection in planetary rovers. In *Proceedings of the Thirteenth International Workshop on Principles of Diagnosis*, pages 1–6, 2002.

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

[14] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *ICCV 2003*, volume 2, pages 726–733, October 2003.

[15] R. Fablet, P. Bouthemy, and P. Pérez. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Transactions on Image Processing*, 11(4):393–407, 2002.

[16] J. Folkesson and H. I. Christensen. Graphical slam - a self-correcting map. In *ICRA-04*, New Orleans, April 2004. IEEE.

[17] J. Folkesson and H. I. Christensen. Robust SLAM. In *IFAC – 5th IAV-2004*, Lisboa, PT, July 5-7 2004.

[18] N. Oliver, A. Garg and E. Horvitz. Layered representations for learning and inferring office activity for multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.

[19] Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.

[20] V. V. Hafner. *Cognitive Maps for Navigation in Open Environments*. Master's thesis, University of Sussex, UK, September 1999.

[21] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden Markov models. In *IEEE Proceedings of the International Conference on Computer Vision*, pages 1455–1462, Nice, France, 2003.

[22] T. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *ICVS '99*, pages 273–292, 1999.

[23] P. Jensfelt and S. Kristensen. Active global localisation for a mobile robot using multiple hypothesis tracking. *IEEE Transactions on Robotics and Automation*, 17(5), 2001.

[24] Matjaž Jogan and Aleš Leonardis. Robust localization using an omnidirectional appearance-based subspace model of environment. *Robotics and Autonomous Systems, Elsevier Science*, 45(1):51–72, 2003.

[25] J. Kelleher and F. Costello. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-SIGSEM Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 2005.

[26] J. Kelleher and G.M. Kruijff. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the 10th European Workshop on Natural Language Generation*, 2005.

[27] J. Kelleher and G.M. Kruijff. A context-dependent model of proximity in physically situated environments. In *Proceedings of the Second ACL-SIGSEM Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 2005a.

[28] J. Kelleher and J. van Genabith. A computational model of the referential semantics of projective prepositions. In P. Saint-Dizier, editor, *Computational Linguistics Dimensions of the Syntax and Semantics of Prepositions*, pages 200–215. Kluwer, 2005.

[29] Jana Košecká and Xiaolong Yang. Location recognition and global localization based on scale-invariant keypoints. In A. Leonardis and H. Bischof, editors, *Statistical Learning in Computer Vision, ECCV 2004 Workshop, Prague, Czech Republic*, pages 49–58, May 2004.

[30] E. Krahmer, E.S. van Erk, and A. Verleg. Graph based generation of referring expressions. *Computational Linguistics*, 29(1), 2003.

[31] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *Proceedings of British Machine Vision Conference (BMVC'03)*, Sept. 2003.

[32] Aleš Leonardis, Horst Bischof, and Jasna Maver. Multiple eigenspaces. *Pattern Recognition*, 35(11):2613–2627, 2002.

[33] G. D. Logan and D. D. Sadler. A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, M. Garrett, and L. Nadel, editors, *Language and Space*. M.I.T. Press., 1996.

[34] A. Galata, A. Cohn, D. Magee and D. Hogg. Modeling interaction using learnt qualitative spatio-temporal relations and variable length Markov models. In *Proceedings of the European Conference on Artificial Intelligence 2002*, July 2002.

[35] Óscar Martínez-Mozos, Cyrill Stachniss, and Wolfram Burgard. Semantic place classification of indoor environments with mobile robots using boosting. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. AAAI, 2005.

[36] Óscar Martínez-Mozos, Cyrill Stachniss, and Wolfram Burgard. Supervised learning of places from range data using AdaBoost. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2005.

[37] M.E. Nilsback and B Caputo. Cue integration through discriminative accumulation. In *Proceedings of CVPR-2004*, 2004.

[38] M. Peternel and A. Leonardis. Visual learning and recognition of a probabilistic spatio-temporal model of cyclic human locomotion. In *ICPR*, volume 4, pages 146–149, 2004.

[39] Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(3):1374–1383, 2004.

[40] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, volume 3, pages 32–36, 2004.

[41] Heung-Yeung Shum and Sing Bing Kang. A review of image-based rendering techniques. In *IEEE/SPIE Visual Communications and Image Processing (VCIP) 2000*, pages 2–13, June 2000.

[42] Candace L. Sidner, Christopher Lee, Cory D. Kidd, and Neal Lesh. Explorations in engagement for humans and robots. In *Proceedings of the IEEE RAS/RSJ International Conference on Humanoid Robots*. IEEE, 2004.

[43] L. Talmy. How language structures space. In *Spatial Orientation: Theory, Research, and Application*. 1983.

[44] S. Thrun, D. Fox, W. Burgard, and F. Dellaert. Robust monte carlo localisation for mobile robots. *Artificial Intelligence*, 128(1–2):99–142, May 2001.

[45] B. Tversky and P. Lee. How space structures language. In *Lecture Notes In Computer Science, Vol. 1404, Spatial Cognition, An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, pages 157–176. 1998.

[46] Vandi Verma, Sebastian Thrun, and Reid G. Simmons. Variable resolution particle filter. In *IJCAI*, pages 976–984, 2003.

[47] Aleš Štimec, Matjaž Jogan, and Aleš Leonardis. A hierarchy of cognitive maps from panoramic images. In Allan Hanbury and Horst Bischof, editors, *Proceedings of the 10th Computer Vision Winter Workshop*, 2005.

[48] Jeremy Wyatt. Planning clarification questions to resolve ambiguous references. In Ingrid Zukerman, editor, *Proceedings of the 4th International Workshop on Reasoning in Practical Dialogue Systems*, 2005.

[49] Y. Yacoob and M. J. Black. Parameterized Modeling and Recognition of Activities. *CVIU*, 73(2):232–247, 1999.

[50] Yasushi Yagi, Kousuke Imai, and Masahiko Yachida. Iconic memory-based omnidirectional route panorama navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 1, pages 564–570, Taipei, Taiwan, September 2003.

[51] Guido Zunino and Henrik I Christensen. Simultaneous mapping and localisation in domestic environments. In R. Dillmann, editor, *Multi-Sensory Fusion and Integration for Intelligent Systems*, pages 67–72, Baden-Baden, DE, August 2001.

Annexes

10 Annex 1: Template for sub-scenario on Modelling Spatial Relationships

10.1 Scenario title:

"Where is it?" — Modelling Spatial Relationships.

10.2 Author(s):

Jeremy Wyatt

10.3 Modified by:

10.4 Scenario summary:

The robot will be able to produce representations of the spatial relationships between objects on a table top. The robot will receive information about these from both vision and spoken language. An experiment to check whether these representations are consistent will be performed.

10.5 Motivation:

We want to link representations for vision and language, and at the very least we want representations of spatial relationships grounded in each modality that can be linked or cross-referenced. When dealing with spatial relations this will require producing representations that capture the relations between the objects, and which also display the qualities of accommodation and context sensitivity that humans demonstrate []. Our engineering goal is therefore to produce a system which can understand spoken statements about the spatial relationships between and properties of objects in scene that can also be observed by the robot. This is the first step in producing a robot that is able to converse with a human in this domain.

10.6 Background:

Talk about how humans are context sensitive, so we have relations but they are flexible. Which relations we spot are task dependent. Also the state of the art in robot conversations (Roy,Sidner et al).

10.7 Precursor scenarios:

None

10.8 Follow-on scenarios:

"Where is it?" — Planning to resolve ambiguity in reference.

10.9 Scenario Ontology:

- Requirements and constraints on objects:

  1. the number of objects in view should be variable

  2. the system may not be able to deal with occlusions

  3. there should be a variety of types of shapes adequate to test state of the art visual algorithms for segmentation and recognition. They should also be adequate to support a variety of positive and negative affordances that will be relevant when objects are manipulated in later scenarios. A subset of these will be chosen for this scenario.

  4. the objects can be rotated, but not laid on their side or upside down, this is to make training of object identification software easier. There is also evidence that humans find identification easier if the view of the object is common.

  5. all the objects will be on the base surface (a table), not on one another. This is to aid localisation of the objects on the ground plane using visual information alone.

- The subset of objects we will use in this scenario are balls, blocks, toy cows, mugs, and bowls.

- Requirements and constraints on prepositions in utterances

  1. The system should be able to handle a variety of topological and projective prepositions [27, 30], namely:

     (a) next to, at, by, near.

     (b) behind, in front of, left of, right of.

     These will support understanding references to objects by a human in later scenarios where the robot is given instructions to manipulate those objects, or asked questions about them.

  2. initially in this scenario the frame of reference will always be the robot's, so left of and right of, are relative to the scene. Later we will add functionality to allow the robot to use frames of reference belonging to the speaker or to particular objects.

  3. single objects may be used as reference landmarks (e.g. the mug is to the left of the cow).

  4. initially the system won't be able to cope with groups of objects as landmarks.

10.10 Robot Ontology:

The robot will know about the:

- spatial relationships listed above.

- object types listed above.

- colour and relative size of objects.

- the position of the objects on a known ground plane.

10.11 Pre-requisites:

Deliberative mechanisms: The system will compare the relational information obtained from the vision and language processing sub-systems to check whether the description of the world given by the human matches what the robot sees. **Linguistic capabilities:** Extent of grammar, will be able to parse a variety of utterances making reference to objects, their properties and their spatial relationships. Will produce a relational structure summarising these. In this scenario the robot's responses will be canned to indicate whether or not the description of the world matches what the robot can see. **Kinds of perceptual mechanisms:** The system will need a vision system that is able to identify (but not necessarily categorise) objects and name them independent of orientation. It should be possible to scale the system efficiently to new objects, i.e. the system shouldn't be a set of specific recogition routines.The vision system should also give an estimate of the position of requested objects (at a particular point in the image plane) on the ground place. It should also be able to update in real time, and to produce a list of attributes for each object such as colour, approximate size etc. There should also be higher level perceptual routines that can estimate whether spatial relationships hold between objects that have been observed. **Kinds of action mechanisms:** The robot will have no actions in this scenario. **Other prerequisites for the scenario to work**

10.12 Scenario Scripts:

10.13 Negative Scenarios:

The robot will not able to deal with:

- groups of objects as landmarks.

- objects that are occluded or in non-typical poses.

- objects that are not on the ground plane.

- part-whole relationships.

- prepositions about spatial relationships other than those listed above.

- reference to patterns on object surfaces.

10.14 Kinds of Integration:

The primary integration will be between vision and language.

10.15 Mode of Evaluation:

The system will be evaluated by a series of collections of objects laid out on the table. The object will be moved about the table top by the human tutor, and s/he will make statements about arrangement of objects on the table top. The robot will confirm verbally whether or not those match what it sees. The robot will be scored according to whether it gives the correct response for each case.

10.16 Progress so far:

Code completed Publications [27]

11 Annex 2: Template for sub-scenario on Planning to Resolve Ambiguity in Reference.

11.1 Scenario title:

Where is it?

11.2 Author(s):

Jeremy Wyatt

11.3 Modified by:

11.4 Scenario summary:

In this scenario we will test the ability of the robot to answer questions about objects in the scene.

11.5 Motivation:

If the robot is to be able to carry out a cooperative task with a human involving some objects on the table it must be able answer questions and follow instructions. To this end it is necessary to have a system that is capable of redirecting attention to objects, performing appropriate processing in order to satisfy some task, and to establish an unambiguous common ground for the dialogue and the task. Building a system that can answer questions about the objects and their properties is a good way to develop and test such abilities. It will require us to create paths from vision to language that allow the robot to make utterances based not only on the dialogue but on what it can see.

11.6 Empirical/Scientific context:

Question answering is a very simple form of interaction between the robot and a human. However, it requires that we are able to establish references to objects or relationships that the robot might make in an unambiguous manner. Many human references are underspecified if only the linguistic information is used, and humans rely on information about the task, the context, and visual salience to resolve ambiguous references. In a robot system the unreliability of perception will make the problems of ambiguity still greater. Therefore the primary challenge in this scenario (building on the work already done in the pre-requisite scenario) is for the robot to be able to generate sensible actions that will enable to resolve ambiguous references.

11.7 References:

11.8 Precursor scenarios:

"What is it?" — Spatial relationships between objects.

11.9 Follow-on scenarios:

TBA.

11.10   Scenario Ontology:

The ontology for this scenario will be the same as for the pre-requisite scenario.

11.11   Robot Ontology:

11.12   Pre-requisites:

Reactive mechanisms: Deliberative mechanisms: Metamanagement/meta-semantic mechanisms: Affective states/processes: Linguistic capabilities: The robot must be able to generate questions. This requires the ability to plan dialogue, to model the beliefs of another agent, and to generate speech. **Kinds of perceptual mechanisms:** The robot needs to be able to redirect it's attention in order to identify referents that may be outside it's original field of view. In order for this system to scale the robot must be able to selectively process parts of the scene in order to identify certain properties of particular objects (e.g. type). **Kinds of action mechanisms:** The robot will be able to ask questions. In the later version of the system the robot will also be able to use pointing gestures.

11.13   Scenario Scripts:

Below we list some typical questions, which ask for different types of information. In the list of example questions we identify the sort of information contained within the question that can be used by the robot to identify the referents, and the type of information requested. Following this we give two pictures, each with the list of example questions coupled with possible, correct answers.

11.13.1   Example Questions

1. What is the orange thing?

   - Want: type-based description.
   - Given: unambiguous feature-based reference.

2. Where is the cow?

   - Want: location-based description.
   - Given: unambiguous type-based reference.

3. What is the biggest?

   - Want: unambiguous reference.
   - Given: (global?) feature based query.

4. . What colour is the cube?

   - Want: feature-based description (should trigger clarification dialogue).
   - Given: ambiguous type-based reference (speaker is ambiguously referring to a specific object)

5. What is nearest the ball?

   - Want: unambiguous reference.

- Given: unambiguous type-based reference.

6. What is to the left of the orange ball?

   - Want: unambiguous reference
   - Given: region relative to an unambiguous feature and type reference.

7. Where are the blocks?

   - Want: location-based description
   - Given: quantified type-based description.

8. Is there anything behind the cube?

   - Want: assert or denial, possibly plus unambiguous references
   - Given: region relative to an ambiguous type reference (ambiguity can possibly be resolved by relation)

9. Is there anything orange behind the cube?

   - Want: assert or denial, possibly plus unambiguous references
   - Given: feature query for region relative to an ambiguous type reference.

10. What is behind the cube?

    - Want: unambiguous references
    - Given: region relative to an ambiguous type reference.

11. What is bigger than the red block?

    - Want: unambiguous references
    - Given: unambiguous feature and type reference.

11.13.2 Possible Answers: Picture 1

1. What is the orange thing?

   - It's a ball.
   - The orange thing is a ball.

2. Where is the cow?

   - Behind the yellow cube.
   - At the back.
   - Near the yellow cube.

3. What is the biggest?

   - I don't know.
   - It's either the cow or the ball.
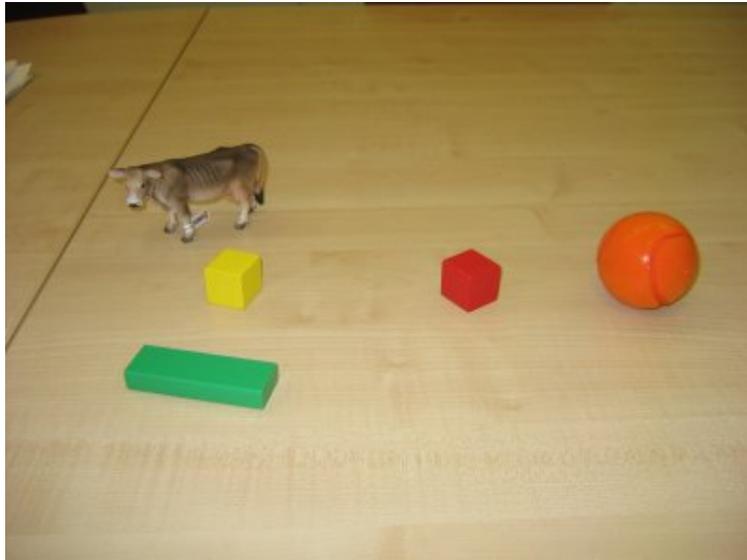
- It's the ball.

4. What colour is the cube?

  - The cube near the cow is yellow.
  - The cube on the left is yellow, the other one is red.
    - R: Which cube?
    - H: It's the cube in front of the cow.
    - R: That cube is red.

5. What is nearest the ball?

  - The red cube.

6. What is to the left of the orange ball?

  - The red cube.
  - Everything is to the left of the orange ball.
  - Some blocks and a cow.

7. Where are the cubes?

  - In the middle.
  - To the left of the orange ball.
  - The yellow one is in front of the cow, the red one is to the right of the yellow one.

8. Is there anything behind the cube?

  - No.
  - Yes, there is a cow behind the yellow cube.
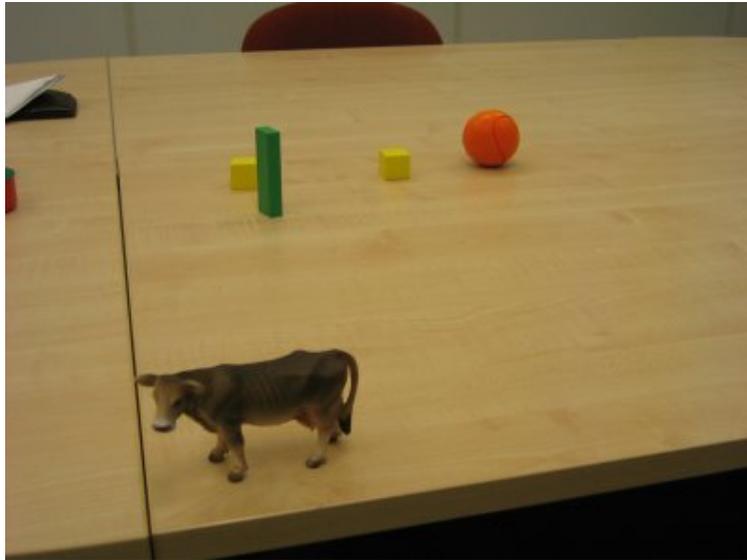
- R: Which cube?
- H: It's the red cube
- R: There is nothing behind that cube.

9. Is there anything orange behind the cube?

   - No. There is nothing orange behind either cube.
     - R: Which cube?
     - H: It's the red cube.
     - R: There is nothing behind that cube.

10. What is behind the cube?

    - The cow.
      - R: Do you mean the yellow cube?
      - H: No the red cube.
      - R: There is nothing behind that cube.

11. What is bigger than the red block?

    - The green block is bigger than the red block.
    - The ball.
    - Everything except the yellow block.

11.13.3 Possible Answers: Picture 2

This picture has changed in at least the following ways:

- The cow's location possibly precludes it from being grouped with the other objects.

- The cubes are now both the same colour.

- The cow now looks bigger.

- The green block now occludes a yellow cube.

- The green block is now taller than the others, and as such appears more prominent.

We've only included the questions which possibly have different answers for the new picture.

1. Where is the cow?

    - In the corner.
    - At the front.
    - Near the edge of the table.

2. What is the biggest?

    - The cow looks the biggest.
    - The green block is the tallest.

3. What colour is the cube?

    - They're both yellow.
    - It's yellow.
        - R: Do you mean the cube on the left?
        - H: Yes.
        - R: That cube is yellow.

4. What is nearest the ball?

    - The yellow cube.
    - The yellow cube on the right.

5. What is to the left of the orange ball?

    - The yellow cube.
    - The blocks.

6. Where are the cubes?

    - Behind the cow.
    - To the left of the orange ball.
    - In the middle.

7. Is there anything behind the cube?

    - No.
    - No, but it's behind the green block.
        - R: Do you mean the cube near the ball?

- – H: Yes.
- – R: There is nothing behind that cube.

8. What is bigger than the yellow block?

- The cow.
- The green block is bigger than the red block.
- Everything except the other yellow block.

11.14 Kinds of self-understanding:

The system requires a form of self-understanding in so far as it must be able to represent and reason about the possible bindings of referents within a question to objects in the visual scene.

11.15 Negative Scenarios:

The restrictions on this scenario are as listed in the pre-requisite scenario.

11.16 Mode of Evaluation:

There are at least two different aspects of the performance of the PlayMate which we can evaluate. Given that the system is designed to interact with a human interlocutor, we can evaluate how the interaction proceeds with reference to both the interlocutor and to a third party. Because the PlayMate is also an information processing system, we can evaluate it in terms of how well (e.g. how fast, or how accurately) it processes the information available to it. These two interrelated evaluation metrics will be discussed in the following paragraphs.

In terms of evaluating the PlayMate's question-answering interaction with a human, we can look at two different aspects of the interaction. The first aspect is whether the PlayMate provides the correct answers to the questions asked by the interlocutor. In simple cases this will be relatively straightforward to evaluate. For example, when the question "What is nearest the ball?" is asked about the first picture (Figure 11.13.2) (as all the question in this section will be), the only correct answer can be a reference to the red ball. When evaluating purely in terms of information provided in response to a question, it is important to note that the surface form of the response is not important. The PlayMate could have said "the cube on the right", or "the cube furthest from the cow", both of which can be resolved to the same object in the environment.

When the question asked about the scene contains ambiguous references to objects, or is about an ambiguous quality of the scene, evaluating the information content of the PlayMate's responses becomes correspondingly harder. If the question requires the PlayMate to identify a single object in the scene (e.g. "What is behind the cube?"), or a feature of a single object in the scene (e.g. "What colour is the cube"), then the information content of the replies can be evaluated in two ways. First, we can ask human interlocutor whether the answer provides the information they had in mind when they asked the question. Second, to remove any potential bias that may be introduced by the interlocutor evaluating the PlayMate's response, we can get a third party to inspect both parts of the dialogue and evaluate whether they think that the human received the correct information from the PlayMate. When the answer desired from the PlayMate should reference ambiguous qualities of the scene (e.g. "What is to the left of the orange ball", or "What is bigger than the red block?"), we could use the same approach, although it becomes a lot harder for either evaluator (the interlocutor or the third party), to objectively judge whether the PlayMate's response is correct.

The second aspect of the interaction between a human and the PlayMate that we can evaluate is its *naturalness*. There is a variety of existing work from which we could derive approaches to the evaluation of the naturalness of the system, and metrics for use in such evaluations. We could choose to evaluate the system purely on the naturalness of its linguistic interactions with the interlocutor, or we could expand this to include additional aspects of the PlayMate, including gesture and gaze. Such evaluations will be effected by numerous aspects of the experimental setup, and can vary immensely with different human subjects, so evaluations done in this manner should be carefully controlled (as should all evaluation).

In addition to evaluating the interaction between the PlayMate and an interlocutor, we can also evaluate the information processing architecture of the PlayMate. In order to be precise about this, we must attempt to separate the characteristics of the PlayMate's information processing architecture (i.e. the architecture we designed to satisfy the requirements of the scenario) from the characteristics of the implementation of the PlayMate (i.e. how we have implemented the design, including programming approaches and implementation substrate). Before the information processing architecture is evaluated, the implementation should be evaluated to determine how well it actually reflects the original design. This should allow any artifacts of the implementation of be separated from the evaluation of the design proper.

The design of the information processing architecture can be empirically evaluated in at least two ways. One way is to evaluate the design of aspects of the architecture using above the previously described approaches for evaluating the interaction between the PlayMate and a human. We can consider a number of examples of this: evaluating the naturalness of the overall interaction when using different designs for components in the architecture (e.g. context models or ambiguity resolution approaches); evaluating the naturalness of landmark objects selected for use in referring expressions as visual attention varies (with reference to feature cues or spatial cues); and evaluating whether the correct information is supplied in response to questions when selective attention is added to the architecture design.

In addition to this, the performance of components in an information processing architecture (and the interactions between groups of such components) can be empirically evaluated in order to demonstrate how they are influenced by various aspects of the overall design. Examples of possible evaluations include the following: the time taken to categorise an object in a scene can be evaluated with or without visual attention and priming; the complexity involved in constructing and querying a scene-graph can be evaluated when this done with and without contextual information; and the complexity of reference resolution can be evaluated with reference to the number of objects allowed into processing by the attentional mechanisms present in the architecture. There are at least two different aspects of the performance of the PlayMate which we can evaluate. Given that the system is designed to interact with a human interlocutor, we can evaluate how the interaction proceeds with reference to both the interlocutor and to a third party. Because the PlayMate is also an information processing system, we can evaluate it in terms of how well (e.g. how fast, or how accurately) it processes the information available to it. These two interrelated evaluation metrics will be discussed in the following paragraphs.

In terms of evaluating the PlayMate's question-answering interaction with a human, we can look at two different aspects of the interaction. The first aspect is whether the PlayMate provides the correct answers to the questions asked by the interlocutor. In simple cases this will be relatively straightforward to evaluate. For example, when the question "What is nearest the ball?" is asked about the first picture above (as all the question in this section will be), the only correct answer can be a reference to the red ball. When evaluating purely in terms of information provided in response to a question, it is important to note that the surface form of the response is not important. The PlayMate could have said "the cube on the right", or "the cube furthest from the cow", both of which can be

resolved to the same object in the environment.

When the question asked about the scene contains ambiguous references to objects, or is about an ambiguous quality of the scene, evaluating the information content of the PlayMate's responses becomes correspondingly harder. If the question requires the PlayMate to identify a single object in the scene (e.g. "What is behind the cube?"), or a feature of a single object in the scene (e.g. "What colour is the cube"), then the information content of the replies can be evaluated in two ways. First, we can ask human interlocutor whether the answer provides the information they had in mind when they asked the question. Second, to remove any potential bias that may be introduced by the interlocutor evaluating the PlayMate's response, we can get a third party to inspect both parts of the dialogue and evaluate whether they think that the human received the correct information from the PlayMate. When the answer desired from the PlayMate should reference ambiguous qualities of the scene (e.g. "What is to the left of the orange ball", or "What is bigger than the red block?"), we could use the same approach, although it becomes a lot harder for either evaluator (the interlocutor or the third party), to objectively judge whether the PlayMate's response is correct.

The second aspect of the interaction between a human and the PlayMate that we can evaluate is its *naturalness*. There is a variety of existing work from which we could derive approaches to the evaluation of the naturalness of the system, and metrics for use in such evaluations. We could choose to evaluate the system purely on the naturalness of its linguistic interactions with the interlocutor, or we could expand this to include additional aspects of the PlayMate, including gesture and gaze. Such evaluations will be effected by numerous aspects of the experimental setup, and can vary immensely with different human subjects, so evaluations done in this manner should be carefully controlled (as should all evaluation).

In addition to evaluating the interaction between the PlayMate and an interlocutor, we can evaluate the information processing architecture. In order to be precise about this, we must attempt to separate the characteristics of the PlayMate's information processing architecture (i.e. the architecture we designed to satisfy the requirements of the scenario) from the characteristics of the implementation of the PlayMate (i.e. how we have implemented the design, including programming approaches and implementation substrate). Before the information processing architecture is evaluated, the implementation should be evaluated to determine how well it actually reflects the original design. This should allow implementation details to be separated from the evaluation of the design proper.

The design of the information processing architecture can be empirically evaluated in at least two ways. One way is to evaluate the design of aspects of the architecture using above the previously described approaches for evaluating the interaction between the PlayMate and a human. We can consider a number of examples of this: evaluating the naturalness of the overall interaction when using different designs for components in the architecture (e.g. context models or ambiguity resolution approaches); evaluating the naturalness of landmark objects selected for use in referring expressions as visual attention varies (with reference to feature cues or spatial cues); and evaluating whether the correct information is supplied in response to questions when selective attention is added to the architecture design.

Finally, the performance of components in an information processing (and the interactions between groups of such components) can be empirically evaluated in order to demonstrate how they are influenced by various aspects of the overall design. Examples of possible evaluations include the following: the time taken to categorise an object in a scene can be evaluated with or without visual attention and priming; the complexity involved in constructing and querying a scene-graph can be evaluated when this done with and without contextual information; and the complexity of reference resolution can be evaluated with reference to the number of objects allowed into processing by the attentional mechanisms present in the architecture.