FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D5.3.1
Report on requirements for perceptual System for learning concepts and navigation tasks

Due date of deliverable: 31/12/2004
Actual submission date: 31/01/2005

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

Lead contractor for this deliverable: KTH

Revision: final
Dissemination level: PU

Executive Summary

The report provides a summary of some of the key issues involved in the design of a robot system for the Home-Tour Scenario. The main emphasis is on the perceptual requirements to enable mapping and annotation of the environment. The components involved in "human augmented mapping" includes: tracking / following of humans, interpretation of human instructions, and geometric mapping & object recognition. For each of the components a brief literature study is provided with a condensed assessment of the different options available for construction of a system.

Role of cognitive mapping in Cogniron

A fundamental component of any mobile system that is to be considered cognitive is spatial mapping of the environment. To be of utility to a user there is a need to generate a "joint" representation between the user and the robot. Such a representation enables the robot to maintain localized and to utilize user provided "labels" to navigate to destinations and recover information / access to places, structures and objects that have been specified by the user. The present report outlines the options available in terms of endowing the system with perceptual competencies that will enable the robot to perform mapping and user based annotation of the environment to generate a common representation of an environment.

Relation to the Key Experiments

The report specifically focuses on the different methods that can be used for "human augmented mapping" as part of the Hour Tour Scenario (Key Experiment 1).

Contents

Requirements for Perceptual Systems for
Learning of Concepts and Navigation Tasks

Henrik I Christensen (KTH), Roland Siegwart (EPFL),
Elin Anna Topp (KTH) and Ben Kröse (UvA)

1   Introduction

A fundamental competence to be endowed to the Cogniron home-tour scenario is the ability to auto-matically acquire maps of its environment and methods for learning of spatial concepts. For a robot to be useful to non-expert users there is a need to provide it with enough information to build a shared representation of the environment, which can both be used by the user to instruct the robot to locate particular objects and also go to particular locations. At the same time the robot must have an inter-nal representation that will enable it to achieve such tasks. In the present document the perceptual competencies to achieve such a functionality will be discussed, and it will be outlined how the func-tionality may be achieved in terms of existing methods reported in the literature. For the purpose of the present discussion the "Home-Tour" scenario will provide the demonstrator context of the actual demonstration of such a functionality.

Throughout the document the acquisition of knowledge about the environment will be considered in the context of supervised learning, where a user interacts with the system so as to "augment" the model of the environment and specify locations/objects that must be acquired by the robot. Consequently fully autonomous acquisition of models of the physical layout and detection of novel objects is not considered in the present report.

To enable a robot to perform acquisition of models of objects and environments it is necessary to endow the robot with competencies related to:

1. Tracking people to interpret intentions

2. Mapping of the environment

3. Interaction for concept learning and task specification

Consequently the perceptual requirements are directly related to

1. detection and tracking of people

2. mapping the environment while performing exploration (e.g. Simultaneous Localisation and Mapping (SLAM))

3. detection of specified objects

4. detection of specified locations

5. interpretation of commands from the user

In addition to the perceptual processes the system must have facilities to reason about space, objects and commands so as to tie these inputs into a common knowledge representation that can be used for dialogue generation, supervision and navigational tasks.

In the present report an overview of approaches to deliver the required perceptual competencies is provided with a brief review of the relevant literature. Initially in section 2 the basic scenario is

outlined in more detail. In Section 3 tracking and monitoring of human activity is discussed as a basis for following a user and for positioning of a robot with respect to an instructor. In Section 4 interpretation of instructions is discussed. The main emphasis is here on simple static situations rather than continuous interpretation of human behaviour. In section 5 the topic of mapping is discussed both in terms of basic geometric maps and as part of human augmentation of maps. Based on these discussions a prediction of the capabilities to be endowed to initial systems is outlined in section 6 together with pointers to what may be expected for future generations of systems.

2 Scenario Context

The scenario envisaged for the augmentation of the environmental maps can best be described in terms of the initial use of a robot. Imagine that your new Cogniron robot has arrived from "Cogs-R-Us". After the system has been unpacked and basic setup performed, there is a need to give the robot a tour of your home. The details are specified in the definition of the Home-Tour Scenario (Key Experiment 1)

> In this experiment, a robot discovers a home-like environment and builds up an understanding of it and of artefacts in it as taught by humans. The process is open-ended, i.e., it has no completion; the robot continues to learn as it faces new situations. A human shows and names specific locations, objects, and artefacts, to the robot. The robot can engage in a dialogue in the case of missing or ambiguous information. This scenario will enable to demonstrate the capacity of dialogue, of continuous learning of space and objects. It illustrates the possibility for the robot to take initiatives for completion of its knowledge (active perception, manipulation of objects to build a sensory-motor representation, etc.)[1]

The result of the exercise is captured in the cartoon by Larson, shown in figure 1.

"Now! …. That should clear up a few things around here!"

Figure 1: The Larson cartoon on annotation of the environment

In more concrete terms the robot is required to have functions for:

1. Estimation of the position of people in the proximity of the robot. The exact field of view for detection might not be critical, but in general one would expect a detection area that covers $180°$ and up to 2-3 meter in distance (Hall, 1966).

---

[1]From the Cogniron – "Description of work" document

2. Following of people throughout a domestic/laboratory setting incl. passage of doors and some degree of following in the presence of other autonomous agents such as other people or pets. The following behaviour will have to take basic social rules into account, i.e. driving behind the user while following, but facing a user when receiving instructions.

3. A functionality to map the environment, and maintain an estimate of its own position. As there is a need to also recognize objects and places the map might have to be in $2\frac{1}{2}$D or full 3D. The map must include both metric information and labels for entities specified by the user.

4. Interpretation of user instructions either in terms of gestures, a graphical user interface, and/or using speech. As part of in particular object specification there is a need for "shared attention" to enable identification of the object indicated by the user. There is also a need to generate feedback to the user in terms of the object/place specification and as part of resolution of ambiguities.

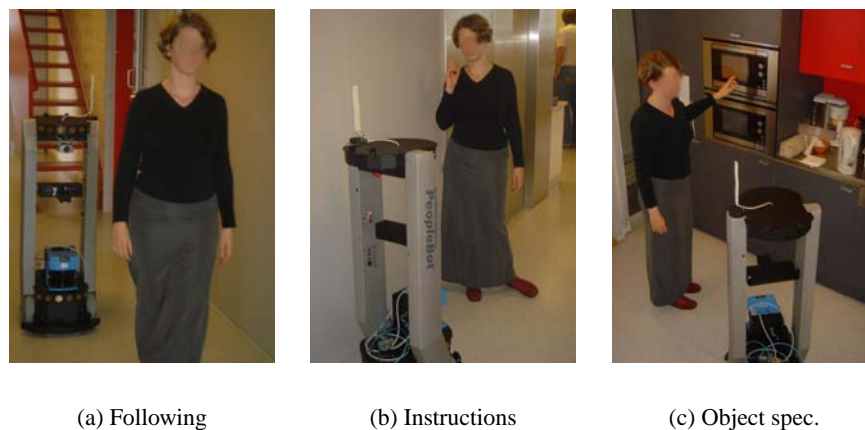Some example situations to be handled are shown in figure 2. As mentioned earlier the basic compe-

| (a) Following | (b) Instructions | (c) Object spec. |

Figure 2: Situations to be covered in the scenario

tencies or rather their sensory processes are covered in the following sections.

3   Tracking and monitoring humans

Sensory based tracking and monitoring human activities has been a popular topic of research for more than 3 decades. The work can be divided into two main categories:

1. Tracking of instrumented humans

2. Unconstrained tracking and monitoring.

Recently the introduction of cheap RFID tags has allowed design of systems for easy tracking of people as for example reported by Intel (Philipose et al., 2003). The RFID technology has developed significantly and already now it can be used for basic activity recognition, but the accuracy is limited (and so is the range). At present the accuracy for a mobile robot application is typically 20-30 cm, which might be inadequate for general tracking or person following.

Other technologies such as WiFi have also been used by companies such as Appear Networks (IST Price Winner 2003). However again the accuracy is limited to about 1 m. Another common problem with all of these solutions is the need to instrument the user, which might be considered intrusive and it might not be generally acceptable to many people.

In terms of non-instrumented tracking of people there has been significant research. The two main approaches that are considered today are based on laser range scanners and computer vision. In addition there is also work on sound based localisation, but as the technique is considered of limited interest in this context, it will not be considered any further.

Research on tracking and monitoring of people using computer vision dates back to the thesis of Badler (1975) and more recently the thesis of Hogg (1983, 1984). Over the last 20 years there has been very significant research on people tracking and today there are standard data-sets for benchmarking of algorithms. The early work (and also more recent work) used a kinematic model of the human that is composed of 2 cylinders for each leg, 2 cylinders for each arm, and a central trunk with a head on top. Other studies have used conics (Drummond & Cipolla, 2001; Bernardo & Perona, 1998), superquadrics (Gavrila & Davis, 1995), and deformable models (Kakadiaris & Metaxas, 1995) for the 3D modelling. Using such models it is possible to extract the pose of the person and perform simple modelling of the motion of a person.

In various studies arms and/or head motion have also been studied, a good overview can be found in (Essa, 1999; Sidenbladh & Black, 2001). Much of the work on human tracking has been performed in the context of behavioural analysis or for interpretation of specific gestures, as part of computer interfaces. To a lesser extend such systems have been used in a robotics context for tracking / following of people. One of the reasons for this is that following implies driving at a short distance behind a person. If the person is to be entirely in the field of view the camera would have to have a field of view of at least $120°$ which is an extreme value for many purposes. Using a 1/4" camera chip this would require a focal length of 4.5 mm or less. In addition as part of following of people they would be seen from the back, where face and hands might not be visible. Trying to track people across variation in clothing etc. might not be realistic. However, as part of interaction for instruction the use of vision might be invaluable.

There has been significant work on gesture recognition for commanding a robot. Early work by Swain(Kahn et al., 1996; Firby et al., 1998) demonstrated how single objects could be identified by static gesture recognition from color segmentation. In a similar vain Cipolla et al demonstrated basic human-robot interaction for grasping operations (Cipolla et al., 1994, 1996). For detection and tracking of gestures there has also been significant work on gesture detection based on hidden Markov Models such as (Starner & Pentland, 1995; Oliver et al., 1998). The recent face and gesture recognition conferences provide a good overview of work in this area. Today there are efficient methods for detection of gestures, which can be utilized as part of the interaction and monitoring of human activities. Using a combination of motion and colour with HMMs appears to be a suitable basis for such systems.

In terms of people tracking there has also recently been significant work on use of laser range scanners, which in particular has been driven forward by the availability of relatively inexpensive scanners from SICK Optics. Some of the most widely cited examples of people tracking are from Schultz et al. (2003) and the thesis by Bennewitz (2004). These approaches adopt a particle filtering approach to tracking of people and can, in environments with good separation between people, track several entities. One challenge with these systems, which typically are positioned 20-30 cm above the floor, is that people might be wearing a skirt or trousers. The appearance is rather different for the two cases, which call for different models of the user for each case (Bruce & Gordon, 2004). In addition a dark skirt can be close to undetectable.

Respecting the limitations of laser scanners it is possible to provide tracking systems that operate at 5-30 Hz and have a localisation performance that is on the order of 10 cm. The laser system is thus well suited for robot following of single persons when travelling through open spaces. In addition, with the given performance it is possible to position the robot with respect to the user, though other modalities such as vision have to be used for detecting front/back of a person.

As part of the monitoring and tracking of people there is further the need to consider the motion control for following a person. Several studies of interaction with people have been reported in the literature. Nakauchi & Simmons (2000) report on a system for entering a line at a conference in which there is a close proximity to other users. Here the robot has to determine the end of the line and align with other users. Althaus et al. (2004) report on a system in which group dynamics is studied so as to form natural distances to other people in a group during informal discussions. The control involves entering and exiting the group and alignment with other actors in the group. Passage of people in a hallway has been reported in Yoda & Shiota (1996, 1997).

As part of the following the robot must have basic capabilities for interpretation of the human intentions, as there is a need to perform mode-switching to change from following to instruction. This requires identification of the "intentions" of the user as for example reported by Bennewitz et al. (2002). For the purpose of behaviour modelling both laser data (with adequate tracking performance) and vision data could be used.

The characteristics of the perceptual modalities for tracking/following of people are briefly summarised in Table 1

Modality	RFID	Vision	Laser
Estimate	2D	2D/3D	2D
Accuracy	$\approx$ 30 cm	10-20 cm	$\approx$ 10 cm
Instrument	Yes	No	No
Speed	30 Hz	1-10 Hz	5-30 Hz
Robustness	+++	+	++
Generality	-	++	+
Estimate	Position	Posture	Position

Table 1: Perceptual modalities for human tracking/following

4 Interpretation of Human Instructions

In the interpretation of human instructions there are four strategies to choose from:

1. The instructions can be performed entirely using a GUI based interface

2. The instructions can be based on a spoken dialogue system

3. The instructions can be based on a combination of gestures and speech

4. The instruction can be performed with instrumentation of the user using markers and/or wearable interfaces (e.g., data-glove).

Each of these interfaces have their own characteristics.
It is important here to realise that the system as a minimum must have facilities for:

- Specification of a particular location (name, and position)

- Specification of a particular object

  - Objects as static structure in the environment (e.g., a door, a window, the refrigerator, the elevator) which is specification of the objects that could be in a position which the robot cannot visit. I.e. the robot might not be able to reach a window, but it can perceive the position of the structure

  - Objects that are movable / mobile, say cups, chairs, books, etc. The robot is here required to recognize the object. Alternatively the robot is required to acquire enough information to enable recognition of the object outside of the present context.

- In a longer term perspective the robot might also be required to acquire dynamic competencies, such as opening a door, pushing a button, etc. The acquisition of such skills and/or tasks is beyond the scope of this report. The problem of skill and task acquisition is studied in other workpackages.

When a graphical interface is used, the perceptual requirements are minimal as the input typically is given through direct point and click actions, which can be made unambiguous. A tablet type interface can be designed for easy interaction, but the learning curve for naive users might be an issue to consider.

Using a spoken dialogue interface there is a possibility to design an interface modality for basic annotation of the environment. I.e. through careful design of a dialogue behaviour situations such as "this is the kitchen" can be handled. The type of specification is here "this is {a ∥ the} <room>". The dialogue can, at least for simple cases, be designed using a regular grammar. Speech technology is gradually getting to the point when recognition rates better than 80% are realistic. This is in particular true for systems with a head mounted microphone. The Sphinx systems from CMU is a good example of the type of system that can be used for such interfaces (Lamare et al., 2003). Other examples of speech recognition systems include the Esmeralda system from Bielefeld (Fink et al., 1998) and the commercial system from Nuance. The systems all appear to have similar performance, in particular for recognition beyond very limited vocabularies. Given that the recognition rate is less than perfect for all systems, there is a need to ensure that the dialogue behaviour includes handling of non-perfect recognition.

The speech interface is useful for specification of locations that the robot is at. When used alone, it is difficult to use such a method for specification of objects, as there is a need to specify a spatial location wrt. the robot. One might use specifications such as "the object to the right is the elevator", but it is considered unlikely that an acceptable generality can be achieved, and the spatial algebra needed for such an approach might be too complex for general environments.

The combination of speech and gestures is better suited for such interfaces. The speech modality can here be used for the entry of specification, and it is possible to disambiguate ambiguous specifications such as "this is my cup", through use of pointing gestures. Such interfaces have successfully been designed by a number of groups as reported by Kahn et al. (1996), Cipolla et al. (1994), Fink et al. (1996), Cassell (1998), and Christensen et al. (2001). In many of these interfaces the on-board robot camera is used for detection of the hands (and possible face) using colour models (Störring, 2004) and the hand motion is recognised using a Hidden Markov Model. Using a motion threshold static gestures can also be identified, and analysed. Through combination with speech it is possible to design more comprehensive dialogues for interaction with places and objects.

One problem in the detection of objects is the need to interpret the direction of a pointing gesture, which requires 3D modelling or inference of the object from pose/size information. In some cases, when tied to recognition, this might be relatively simple; e.g. "this cup" (where a cup is a known category) might pose little ambiguity, whereas "this object is an elevator" (where elevator is an unknown object type) poses a bigger challenge. Steels (2001) has reported grounding of semantic description for a limited set of objects, and this might be a viable approach when the disambiguation is not too difficult, i.e. in uncluttered scenes.

As part of specification of objects there is a need to include recognition of such objects. Recently there has been significant progress on methods for recognition of objects. Using statistical learning and relatively simple figure ground detection it is possible to recognize a large number of objects across pose and varying backgrounds. Examples of such methods include Lowe (1999), Leibe & Schiele (2003), Nilsback & Caputo (2004), Fergus et al. (2003), and Wallraven et al. (2003). In particular work on Support Vector Machines has enabled implementation of modules for real-time recognition of objects, as reported for example by Roobaert (2001).

Another approach to design of human specification is through instrumentation of the user. Through use of data-gloves and simple exo-skeletons it is possible to directly acquire information from the user in term of pointing gestures, and grasping of objects. Such interfaces are especially suited for acquisition of motion skills and analysis of grasping tasks, but can also be used for instructing robot about spatial layout. Such interfaces have been used by a number of researchers as reported for example by Dillmann & Kaiser (1997); Ehrenmann et al. (2001). Given the need to instrument the user it might not be an optimal solution in particular for novice users.

The options available for interpretation of human instructions are summarised in Table 2.

Modality	Location	Structure	Object
GUI	++	++	++
Speech	+	−	−
Speech+Vision	+++	+++	+++
Glove etc	+	+	−

Table 2: The perceptual modalities available for human instructions

5 Perception for Cognitive Mapping / Navigation Tasks

Mapping of the environment is a well known topic in robotics. It has been studied for at least 30 years and there is an abundance of literature. However, the inclusion of semantics/labels into the mapping is not as widely studied. One of the exceptions has been the work by Kuipers (1977, 1978, 2000). For the cognitive mapping of spaces there is a need for inclusion of knowledge at a number of different levels for basic metric information, over topology to inclusion of information about objects and structures in the environment, which includes addition of labels to places and recognition of objects in terms of a directory of objects, that might have associated semantic information. The inclusion of all this information could for example be achieved through the Spatial Semantic Hierarchy (SSH) by Kuipers, though it must be mentioned that the SSH this far only has been studied for 2D environments.

Acquisition of 2D geometric maps of the environment can easily be achieved using ultra-sonic sensors (Leonard & Durrant-Whyte, 1992; Wijk & Christensen, 2000; Castellanos & Tardós, 1999) and laser-ranging (Thrun et al., 1998; Konolige & Gutmann, 1999; Tomatis et al., 2001; Guivant & Nebot, 2002). Using sonar based maps enable localisation with an accuracy of about 10-30 cm even for large

scale environments. One problem with ultra-sonic sensors is that the discrimination of environment can be difficult and sensing is often limited to 2-5 Hz due to physical constraints. The laser scanning methods have a larger range and can operate at speeds upto 30 Hz. Typically the localisation accuracy is on the order of a few centimeters. Using robust statistics it is possible to use the laser methods even in highly dynamic environments.

The most frequently used features for 2D laser based mapping are lines. However, direct scan-matching/alignment is also widely used as a direct sensory representation Gutmann & Schlegel (1996). Today standard software is available off-the-shelf for mapping and localisation. Over the last few years an abundance of papers have been published and there is a rich literature on representations and estimation methods. An archive of papers is available at www.cas.kth.se/SLAM.

Recently there has been an increased interest in use of vision for localisation and mapping. The earliest work on integrated vision based mapping and localisation was presented by Davison (1998). The lack of adequate recognition methods for handling of large scale databases has, however, hampered the research. The recent progress on vision methods has generated a number of new techniques as for example presented by Se et al. (2002!). Early work was entirely based on recovery of 3D structure from motion or stereo, but recent work has included mapping and localisation using a single camera (Davison, 2003).

Another problem that has hampered the use of vision for 2D/3D mapping is the limited field of view of regular cameras. For recovery of 3D position over time there is a need to track features during motion, which requires a large field of view that in turn results in poor spatial localisation. The compromise is difficult. In parallel there has been significant development on use of mirror systems and special optics to provide omni-directional cameras. A comprehensive coverage of the use of omni-directional cameras can be found at www.cis.upenn.edu/ kostas/omni.html. An example of use of omni-cams for localisation and mapping can be found in Leonardis et al. (2002).

In contrast to explicit geometric representations of objects or spaces, *appearance-based* representations have become popular Murase & Nayar (1995); Pourraz & Crowley (1998). In these approaches, the environment is modeled as an 'appearance map' that consists of a collection of sensor readings obtained at known poses (positions and orientations). The advantage of this representation is that the pose of the robot can be determined directly by comparing the sensor with those in its appearance database. The appearance database can contain linear features derived from omnidirectional images Kröse et al. (2001), and the method can be well combined with Monte Carlo methods into a probabilistic localization schema Vlassis et al. (2002). The disadvantage is that the number of stored feature vectors scales badly with space, and that many annotated samples for training are needed.

The metric mapping of the environment is well suited for robot localisation. In many of the systems the basic estimation of position and map updating is based on an Extended Kalman Filter (EKF), which has an inherent complexity of $O(N^2)$ where N is the number of map features. Consequently, there is a need to limit the size of local maps to ensure real-time operation.

An alternative to purely metric mapping is represent the environment in terms of a topological map. Using either complete maps or analysis of visibility has enabled design of methods for (qualitative) localisation and mapping. Choset & Nagatani (2001) presents on a strategy for automatic generation of topological maps from generalised Voronoi graphs. The effort has also been combined with metric maps for hybrid mapping (Lisien et al., 2003). In a similar vain it is possible to detect significant changes in local geometry and use these to generate a map of topologically distinct regions. By coding local appearance in terms of simple geometric primitives it is possible to represent regions as a "finger-print" map. Significant changes in the finger-print then indicate changes in geometry/topology which enable topological mapping, as reported by Lamon et al. (2001); Tapus & Siegwart (2005).

Through combination of topological maps with local metric maps it is possible to generate represen-

tations that are scalable to large scale environments (Kuipers et al., 2004). Guivant & Nebot (2001) demonstrate how the innovation matrix for the EKF can be divided into local regions that are updated separately. The "local" regions are updated in the innovation process and then at critical points a globally consist map is generated by re-conciliation across local regions. The coding of topology is here implicit and discovered as part of the mapping through detection of lack of correlation between different part of a map. A similar ideas has been used in the work by Folkesson & Christensen (2004), where "star" nodes represent local geometric maps. A more direct approach to hierarchical mapping is presented is the ATLAS framework (Bosse et al., 2004, 2003), where both a topological map and purely metric maps are combined to achieve consistent maps of large scale environments. Most of these studies have been based on use of a laser-range scanner.

As part of the mapping and localisation it is equally important to be able to perform recognition of places, which have been specified by the user. This can for example be achieved through continuous localisation using a geometric SLAM method that runs continuously as a background process. Alternatively, places can be explicitly recognised, using either visual recognition (Porta & Kröse, 2004; Tell & Carlsson, 1999; Se et al., 2002!) or finger-prints from vision or laser ranging (Tapus et al., 2005).

As part of the annotation of space there is also a need to recognize objects that are specified by the user. For the basic labelling a catalogue of object categories can be used as a starting point, as outlined in Section 4. However, for unknown objects or structures there is a need for on-line acquisition of a model. Some of the methods from statistical learning could in principle be used for this, as has been attempted for example by Tell & Carlsson (1999) and Lowe (2004). Such an approach does, however, require that the world/objects have sufficient texture / surface structure to generate a rich model of the object. In addition the problem of figure-ground segmentation, the separation of the object from the background, still poses a problem in many cases. The potential set of objects that can be handled must thus be chosen with care.

A problem to be considered here is the issue of 3D geometry. Some of the objects that a user might point to could be located in a 3D context, such as on a table or shelf. To fully understand the context of objects there is a need for recovery of some 3D structure. This has only be attempted in very limited contexts, and there are no general techniques available to achieve this. In principle one could attempt to recover a full geometric model of the environment, this is, however, not a realistic option, which 30 years of research in vision has clearly demonstrated. Association of high and approximate pose might be an option, but the generation of such $2\frac{1}{2}D$ information is still a major challenge (Björkman & Eklundh, 2004). Another issue to consider is that many of the methods that have been proposed for representation of the world in terms of geometry, topology and labels have only been evaluated for 2D worlds.

Realistically the mapping and annotation of a world representation will have to rely on a model that is similar in spirit to the Spatial Semantic Hierarchy proposed by Kuipers (2000). In such a framework there is a need to integrate full geometry (beyond 2D), topological information about the overall layout of the environments, and models of objects, structures and places. To achieve this a new type of mapping and localisation systems is needed. Considering the characteristics summarised in Table 3 a strategy to move forward might be based on a combination of laser range scanners, omni-directional cameras, and a binocular active vision system.

In terms of construction of a system as outlined in the Spatial Semantic Hierarchy (Kuipers, 2000) a viable strategy is to have a graph representation with labels and objects that is integrated with hybrid topological and metric representations. The recognition of objects and structures is performed using vision and statistical learning. The topological mapping is performed using omni-directional vision and/or laser ranging. The metric mapping of local (topologically distinct) regions can be performed

Modality	Sonar	Laser 2D	Laser-Topo	Vision 2D	Vis. Rec.
Map	2D Pts	Lines+Pts	Graph	Places	Disc. 3D
Estimate	2D Pose	2D Pose	Place	3D Pose	3D Pose
Complexity	$O(N^2)$	$O(N^2)$	$O(NlogN)$	$O(N^3)$	$O(N^2)$
$\approx$ Bytes/m^2	300	100	100	50k	20k

Table 3: Perceptual Modalities for Mapping and Object Annotation

using laser ranging using a mixture of line and point features embedded in a graphical SLAM estimator. Using such a structure it is considered likely that the requirements of the representation can be accommodated with a localisation performance better than 5 cm.

6 Summary/Discussion

In the present document the various perceptual tasks to be achieved for acquisition of maps of the environment including both the geometric layout, the recognition of places, structures and objects, and annotation with user provided labels. The particular context for the discussion has been the "Home-Tour" scenario. As part of this there is a need to perform tracking / following of a user and other agents in the environment, interpretation of user instructions for the annotation and mapping of the environment as the tour progresses. A brief review of relevant methods from the literature has been presented and an overall evaluation of the methods has been provided.

In terms of setup of a system for experimental studies a reasonable compromise seems to be a system with methods for:

- Laser ranging+vision (simple pose of person) for detection and following of the user and other people in the vicinity of the robot. For engagement in dialogues there is a need for a limited interpretation of user behaviour to switch from following to face-to-face interaction.

- Speech + Gesture for interpretation of user instructions. The speech system will rely on a head mounted microphone to enable a reasonably recognition rate and suppression of environmental noise. The gesture system will as a minimum have to include facilities for pose estimation of pointing gestures.

- Laser ranging + Omni-Directional Vision + Active Vision for mapping of the environment in terms of metric layout, topology and recognition of objects/structures in the environment. The need for partial recovery of 3D structure is an open problem that will have to be considered as part of the research.

The present document has not been an attempt to outline a final solution to the problem of environmental mapping and object recognition. It has been an attempt to briefly outline the available options and point to the required perceptual competencies that are to be provided to enable the continued study of cognitive mapping in the Home-Tour scenario.

References

Althaus, P., Ishiguro, H., Kanda, T., Miyashita, T., & Christensen, H. I. (2004, April). Navigation for human-robot interaction tasks. In *Proceedings of the IEEE international conference on robotics and automation* (Vol. 2, p. 1894-1900).

Badler, N. I. (1975). *Temporal scene analysis: Conceptual descriptions of object movements*. Unpublished doctoral dissertation, Dept. of Computer Science, University of Toronto, Canadian Thesis on Microfische 33080.

Bennewitz, M. (2004). *Mobile robot navigation in dynamic environments*. Unpublished doctoral dissertation, Alberg Ludwig University, Dept of Computer Science, Freiburg.

Bennewitz, M., Burgard, W., & Thrun, S. (2002, May). Learning motion patterns of persons for mobile service robotics. In *ICRA*. Washington, DC.

Bernardo, L. G. E. D., & Perona, P. (1998). Reach out and touch (motion learning). In *IEEE Intl. Conf. on Automatic Face and Gesture recognition* (pp. 234–238).

Björkman, M., & Eklundh, J. (2004, Sept). Attending, foveating and recognizing objects in real world scenes. In *British machine vision conference – bmvc04*.

Bosse, M., Newman, P., & Leonard, J. (2003). An ATLAS framework for scalable mapping. In *Proc. icra-03* (Vol. 1, p. 1899-1906).

Bosse, M., Newman, P., Leonard, J., & Teller, S. (2004, Dec.). Simultaneous localisation and map building in large scale cyclic environments using the Atlas framework. *Intl. Jour of Robotics Research*, *23*(12), 1113–1141.

Bruce, A., & Gordon, G. (2004, May). Better motion prediction for people tracking. In *Proc. of ICRA*. New Orleans.

Cassell, J. (1998). A framework for gesture generation and interpretation. In R. Cipolla & A. Pentland (Eds.), *Computer vision for machine interaction* (pp. 191–215). Cambridge University Press.

Castellanos, J. A., & Tardós, J. D. (1999). *Mobile robot localization and map building: A multisensor fusion approach*. Kluwer Academic Publishers.

Choset, H., & Nagatani, K. (2001, April). Topological simultaneous localisation and mapping: Towards exact localisation without explicit localisation. *IEEE-TRA*, *17*(2), 125–137.

Christensen, H. I., Kragic, D., & Sandberg, F. (2001). Vision for interaction. In G. Hager, H. Christensen, F. Klein, & H. Bunke (Eds.), *Intelligent robot systems*. Heidelberg, DE: Springer Verlag.

Cipolla, R., Hadfield, P., & Hollinghurst, N. (1994, December). Uncalibrated stereo vision with pointing for a man-machine interface. In *Iapr workshop on machine vision application*. Tokyo.

Cipolla, R., Hollinghurst, N., Gee, A., & Dowland, R. (1996). Computer vision in interactive robotics. *Assembly Automation*, *16*(1).

Davison, A. (1998). *Mobile robot navigation using active vision*. Unpublished doctoral dissertation, Univ of Oxford, Oxford (UK).

Davison, A. (2003, October). Real-time simultaneous localisation and mapping with a single camera. In *Proc. international conference on computer vision, nice*.

Dillmann, R., & Kaiser, M. (1997). Issues in skill acquisition via human demonstration. In H. Bunke, H. Noltemeier, & T. Kanade (Eds.), *Modelling and planning for sensor based intelligent robot systems*. World Scientific.

Drummond, T., & Cipolla, R. (2001). Real-time tracking of highly articulated structures in the presence of noisy measurements. In *Iccv* (pp. 315–320).

Ehrenmann, M., Zöllner, R., Knopp, S., & Dillmann, R. (2001). Sensor fusion approaches for observation of user actions in programin by demonstration. In *Multi-sensory fusion.* Karlsruhe.

Essa, I. A. (1999). Computers seeing people. *AI Magazine*, *20*(2), 69–82.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Cvpr.*

Fink, G., Jungclaus, N., Kummert, F., Ritter, H., & Sagerer, G. (1996). A distributed system for integrated speech and image understanding. In *International symposium on aritifical intelligence* (p. 117-126). Cancun, Mexico.

Fink, G. A., Schillo, C., Kummert, F., & Sagerer, G. (1998, September). Incremental speech recognition for multi-modal interfaces. In *Ieee 24th conf. on industrial electronics* (pp. 2012–2017). Aachen.

Firby, J., Prokopowitz, P., & Swain, M. J. (1998). The animate agent architecture. In P. B. D. Kortenkamp & R. Murphy (Eds.), *Artificial intelligence and mobile robots* (pp. 243–276). Menlo Park, CA: AAAI Press/The MIT Press.

Folkesson, J., & Christensen, H. I. (2004, April). Graphical slam - a self-correcting map. In *Icra-04.* New Orleans.

Gavrila, D. M., & Davis, L. S. (1995). Towards 3-D model-based tracking and recognition of human movement: a multi-view approach. In *Proc. international workshop on automatic face and gesture recognition* (pp. 272–277).

Guivant, J., & Nebot, E. (2002, May). Improving computational and memory requirements of simultaneous localization and map building algorithms. In *Ieee intl. conf on robotics and automation* (pp. 2731–2736). Washington DC.

Guivant, J. E., & Nebot, E. (2001, June). Optimization of the simultaneous localization and map-building algorithm for real-time implementation. *IEEE Transactions on Robotics and Automation,*, *17*(3), 242–257.

Gutmann, S., & Schlegel, C. (1996). Amos: Comparison of scan-matching approaches for self-localization in indoor environments. In *1st euromicro conf on adv. mobile robotics.*

Hall, E. (1966). *The hidden dimension*. New York: Doubleday.

Hogg, D. (1983). Model-based vision: A program to see a walking person. *Image and Vision Computing*, *1*(1), 5-20.

Hogg, D. (1984). *Interpreting images of a known moving object*. Unpublished doctoral dissertation, School of Cognitive and Computer Sciences, University of Sussex, Brighton, UK.

Kahn, R., Swain, M., Prokopowicz, P., & Firby, R. (1996). Gesture recognition using the Perseus architecture. In *Cvpr–96* (pp. 734–741). San Francisco, CA.

Kakadiaris, I., & Metaxas, D. (1995). 3D human body model acquisition from multiple views. In *ICCV* (pp. 618–623).

Konolige, K., & Gutmann, S. (1999, November). Incremental mapping of large cyclic environments. In *Intl. symp. on comp intell. in rob and aut. – cira-99* (pp. 318–325). Monterey, CA.

Kröse, B., Vlassis, N., Bunschoten, R., & Motomura, Y. (2001, April). A probabilistic model for appearance-based robot localization. *Image and Vision Computing*, *19*(6), 381-391.

Kuipers, B. (1978). Modeling spatial knowledge. *Cognitive Science*, *2*, 129–153.

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, *119*, 191–233.

Kuipers, B., Modayil, J., Beeson, P., MacMahon, M., & Savelli, F. (2004, May). Local metrical and global topological maps in the hybrid spatial semantic hierarchy. In *Intl. conf on robotics and automation ICRA*. New Orleans.

Kuipers, B. J. (1977, July). *Representing knowledge of large-scale space* (Tech. Rep. Nos. TR-418 (revised version of Doctoral thesis May 1977, MIT Mathematical Department)). MIT Artificial Intelligence Laboratory.

Lamare, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., & Wolf, P. (2003). *The CMU Sphinx-4 Speech Recognition System* (Tech. Rep.). Pittsburgh, PA: Sun Micro Systems/CMU Speech Group.

Lamon, P., Nourbakhsh, I., Jensen, B., & Siegwart, R. (2001, May). Derivign and matching image fingerprint sequences for mobile robot localisation. In *Icra*. Seoul, Korea.

Leibe, B., & Schiele, B. (2003, Sept.). Interleaved object categorization and segmentation. In *Proceedings of british machine vision conference (bmvc'03)*.

Leonard, J. J., & Durrant-Whyte, H. F. (1992). *Directed sonar sensing for mobile robot navigation*. Boston: Kluwer Academic Publishers.

Leonardis, A., Jogan, M., Skočaj, D., & Artač, M. (2002, September). Mobile robot localization using panoramic eigenspace representations. In *East-west-vision 2002 : proceedings* (p. 13-22). Wien, Austria: Österreichische Computer Gesellschaft.

Lisien, B., Morales, D., Silver, D., Kantor, G., Rekleitis, I., & Choset, H. (2003, Oct). Hierarchical simultaneous localisation and mapping. In *Iros-03* (pp. 448–454). Las Vegas, NV.

Lowe, D. (1999, September). Object recognition from local scale-invariant features. In J. Tsotsos (Ed.), *Intl. conf. on computer vision* (pp. 1150–1157). Corfu, Greece.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *Intl. Jour of Compputer Vision*, *60*(2), 91–110.

Murase, H., & Nayar, S. K. (1995). Visual learning and recognition of 3-d objects from appearance. *Int. Jrnl of Computer Vision*, *14*, 5-24.

Nakauchi, Y., & Simmons, R. (2000, October). A social robot that stands in line. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vol. 1, p. 357-364).

Nilsback, M. E., & Caputo, B. (2004). Cue integration through discriminative accumulation. In *International conference on computer vision and pattern recognition.*

Oliver, N., Rosario, B., & Pentland, A. (1998). Statistical Modeling of Human Interactions. In *Proc. ieee workshop on the interpretation of visual motion.*

Philipose, M., Fishkin, K. P., Perkowitz, M., Patterson, D., & Haehnel, D. (2003, Dec). *The probabalistic activity toolkit: Towards enabling activity-aware computer interfaces* (Tech. Rep. No. IRS-TR-03-013). Seattle: Intel Research.

Porta, J. M., & Kröse, B. J. (2004, March). Appearance-based concurrent map building and localisation. In F. Groen, N. Amato, A. Bonarini, & B. Kröse (Eds.), *Intelligent autonomous systems 8* (p. 1022-1030). Amsterdam, NL: IOS Press.

Pourraz, F., & Crowley, J. (1998). Continuity properties of the appearance manifold for mobile robot estimation. In *Proc sirs'98.* Edinburgh.

Roobaert, D. (2001). *Pedagogical support vector learning: A pure learning approach to object recognition.* Unpublished doctoral dissertation, Electrical Engineering and Computer Science, Royal Institute of Technology, NADA/CVAP, SE-100 44 Stockholm, sweden.

Schultz, D., Burgard, W., Fox, D., & Cremer, A. B. (2003, February). People tracking with mobile robots using sample-based joint probabilistic data association filters. *Intl. Jour of Robotics Research*, *22*(2), 99–116.

Se, S., Lowe, D., & Little, J. (2002a, October). Global localisation using distinctive visual features. In R. Siegwart & C. Laugier (Eds.), *Iros-2002* (pp. 226–231). Lausanne, CH.

Se, S., Lowe, D., & Little, J. (2002b, January). Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *Intl. Jour. of robotics Research*, *21*(8), 735–757.

Sidenbladh, H., & Black, M. J. (2001, July). Learning the statistics of people in images and video. In *Proc. international conference on computer vision.* Vancouver, Canada.

Starner, T., & Pentland, A. (1995). Visual Recognition of American Sign Language Using Hidden Markov Models. In *International workshop on automatic face and gesture recognition* (pp. 189–194).

Steels, L. (2001, Sept/Oct). Language games for autonomous robots. *IEEE Intelligent Systems*, *16*(5), 16–2.

Störring, M. (2004). *Computer vision and human skin colour.* Unpublished doctoral dissertation, Aalborg University, Faculty of Engineering and Science, Aalborg, DK.

Tapus, A., & Siegwart, R. (2005). Incremental topological mapping with fingerprints for mobile robot navigation. In *Robotics science and systems.*

Tapus, A., Vasudenvan, S., & Siegwart, R. (2005, Jan.). Toward a multilevel cognitive probabilistic representation of space. In *IS&T/SPIE Symposium on electronic imaging.* San Diego, CA.

Tell, D., & Carlsson, S. (1999, june,). View based visual servoing using epipolar geometry. In *Proc. 11th scandinavian conference on image analysis.*

Thrun, S., Fox, D., & Burgard, W. (1998, May). Probabilistic mapping of an environment by a mobile robot. In *Proc. of the ieee international conference on robotics and automation (icra'98)* (Vol. 2, pp. 1546–1551). Leuven, Belgium.

Tomatis, N., Nourbakhsh i., Arres, K., & Siegwart, R. (2001, May). A hybrid approach to robust and precise mobile robot navigation with compact environment modelling. In *Intl. conf on robotics and automation* (pp. 1111–1116). Seoul, Korea.

Vlassis, N., Terwijn, B., & Kröse, B. (2002, May). Auxiliary particle filter robot localization from high-dimensional sensor observations. In *Proc. IEEE int. conf. on robotics and automation* (pp. 7–12). Washington D.C., USA.

Wallraven, C., Caputo, B., & Graf., A. (2003). Recognition with local features: the kernel recipe. In *International conference on computer vision* (pp. I: 257-264).

Wijk, O., & Christensen, H. (2000, December). Triangulation based fusion of sonar data with application in robot pose tracking. *IEEE Transaction on Robotics and Automation*, *16*(6), 740–752.

Yoda, M., & Shiota, Y. (1996, November). Analysis of human avoidance motion for application to robot. In *Proceedings of the IEEE international conference on robot and human communication* (p. 65-70).

Yoda, M., & Shiota, Y. (1997, September). The mobile robot which passes a man. In *Proceedings of the IEEE international conference on robot and human communication* (p. 112-117).