Multimodal Robot/Human Interaction for Assisted Living

Ray Jarvis
Intelligent Robotics
Research Centre
Monash University
Wellington Road
Clayton Vic 3168
+ 61 3 9905 3470

Ray.Jarvis@eng.
monash.edu.au

Om Gupta
Intelligent Robotics
Research Centre
Monash University
Wellington Road
Clayton Vic 3168
+ 61 3 9905 3410

Om.Gupta@eng.
monash.edu.au

Sutono Effendi
Intelligent Robotics
Research Centre
Monash University
Wellington Road
Clayton Vic 3168
+ 61 3 9905 3410

Sutono.Effendi@eng.
monash.edu.au

Zhi Li
Intelligent Robotics
Research Centre
Monash University
Wellington Road
Clayton Vic 3168
+ 61 3 9905 3410

Zh.Li@eng.
monash.edu.au

Abstract

This paper outlines the framework of a complex system to demonstrate multimodal spatial and transactional intelligence in a robot which autonomously supports aged, frail, or otherwise disabled people in a domestic assistive technology context. The intention is that the robot be able to navigate around a known multi-room environment along optimal, collision-free paths in search and retrieval of requested objects such as spectacles, books etc. and must also be capable of tracking and following humans and of reminding them of times for meals, medication etc. and to lead disoriented subjects to their meal place at appropriate times and even dispense medication, if necessary. The modes of communication interchanges with the supported human include spoken speech and gestures (including eye gaze direction) within the context of situational analysis which accommodates recent history, temporal factors and individual user behavioural models. This paper provides an overview of an ambitious research project in its early stages, describes many components developed to date and outlines future work.

Keywords

Intelligent robotics, assistive technology, scene analysis, robot navigation, gesture recognition, human/machine interaction.

1. Introduction

As robots emerge from the structured industrial environments they have habituated for some time into the relatively unstructured spaces of the built and natural world it is clear that they require increasing levels of intelligence, informed by rich sensory sources, to survive, adapt and serve humans, where humans and robots mix freely. In assistive technology environments, where the served humans are perhaps aged, infirm or otherwise disabled, either mentally or physically, the useful interactions between robots and humans need to be particularly sensitive and sophisticated, since no expert knowledge of the technology can be assumed to reside with the served humans. Also, in some cases, even the basis of normal human to human interaction may be partially undermined by physical or mental dysfunction.

Two quite distinct, but functionally linked, types of intelligence need to be mastered by the robot. The first is 'spatial intelligence', which is the understanding of how the working environment is structured in terms of geometry, occupancy and functional designations (eg. kitchen, bedroom etc.) and how to deal with (find, recognise, handle) common objects in it (eg. cups, knives, forks, books, spectacles etc.) in fulfilment of useful duties. This type of intelligence includes the capacity to achieve goal directed path planning and following, obstacle avoidance, collision-free robot arm trajectory planning and object recognition/manipulation. Each of these requirements are serious challenges in themselves. Having them all function together in a smooth operation to achieve some overall mission is at least one order of magnitude higher in complexity, in operational management, context integration, conflict resolution and hierarchical planning terms.

Spatial intelligence, as defined above, has been well researched, in its various components, by the robotics research community over the last two decades or so (Jarvis and Byrne, 1987; Jarvis, 1997; Durrant-Whyte and Guivant, 2000; Jafari and Jarvis, 2005 ; Rawlinson and Jarvis ,2007), with various levels of maturity and reliability being achieved in each.

The second type of intelligence, which must also be mastered to achieve the overall goal of serving humans correctly, safely, and with grace, is 'transactional intelligence'. By 'transactional intelligence' is meant the understanding of how to communicate with humans to be able to correctly interpret and carry out their wishes and to clarify uncertain or detailed aspects of the task at hand. The interaction is a kind of negotiation process which should result in purposeful and correct interpretation and action, with intermediate interchanges sometimes needed to resolve ambiguity whenever it arises. Understanding the intent of the human and how to go about fulfilling it are the essential requirements within the assistive technology context. Accommodating the practical limitations imposed by the physical state of the environment (and the objects within it) and the capabilities of the robot is crucial to these ends. Understanding what a human intends is often fraught with considerable ambiguity. Resolving this ambiguity within the above constraints is the basic goal of 'transactional intelligence' in this context.

Multimodality provides the crucial key, which unlocks how best to resolve ambiguities of interpretation in both the spatial and transactional components of this project. Multiple sensor modes, including stereoscopic and panoramic vision, tactile sensing, laser range finding and range cameras are deployed to resolve issues concerning how to navigate efficiently through unoccupied space (avoiding fixed and dynamic obstacles), searching for and recognising target objects, manipulating objects (and carrying them), and finally putting them before or in the hands of the requester. Likewise, multiple communication modes, including spoken language understanding, gesture recognition, face recognition and gaze direction analysis, are used to achieve 'transactional intelligence' in the context of environmental constraints (e.g. where the robot can go, where objects are, what things can be picked up etc.), individual behavioural modes of the user (once recognised) and maybe even the time of the day (e.g. the approach of meal or medication times).

The following section touches briefly on related work. The next outlines the robot's attributes, its control structure, its sensory capability, and the methods of achieving localisation (i.e. position/pose), environmental modelling, path planning, obstacle avoidance, object recognition, and manipulation. This is followed by a section on the way in which various modes of human/machine communication are to be combined to resolve ambiguity, including the possibility of querying the human to help resolve and/or refine intention. Then follows a section describing a functional subdivision of project tasks to enable various aspects of practical implementation to be put into operation. The next section describes progress so far and the type of experiments being planned for the future. References to papers developed in the author's laboratory have been deliberately given emphasis to indicate the background work supporting this current project. This paper arose out of a previous conference presentation (Jarvis, 2009).

2. Related Work

An increasing amount of research work has recently been focussed on the area of assistive technology with the realisation of an anticipated large number of old/frail people likely to be inhabiting most Western nations over the next twenty years and the corresponding reduction of younger/fit people available to look after them. Many proposed solutions look to modern technology for support. The overall scope of assistive technology is very wide. The more traditional work has been to provide smarter wheelchairs which can provide sensor informed assistance to their users, allowing them to be able to move about purposely with both independence and safety (Jarvis, 2002; Hu et. al., 2007) and to use robotics to provide physical prosthesis (Carrozza et. al.,2001; Pons et. al.,2004)and therapeutic manipulation(Volpe et. al.,2009). More recently there has been a keen interest in tracking (Chakravarty and Jarvis, 2006) old/sick/feeble people to check whether their movement habits are changing, possibly as a harbinger of some serious physical or mental deterioration or to indicate when a fall has occurred (Lee and Mihailidis, 2005) and requires immediate attention. Also, there is an interest in the electronic monitoring of blood pressure, blood sugar and heart rhythms (Gao et. al. 2005) which can be relayed directly to a medical centre for attention if required. Some of this effort has strong associations with security based surveillance (Kanade et. al., 1997), particularly if purely passive methods of observation are used. Combining robotic wheelchair navigation with on-board robot arm manipulation (Prior, 1990) has also attracted research interest; the requirement for robotic hand/eye coordination (Hagar and Chang, 1995) and pattern recognition is clearly evidenced here. In the field of robotic companions (Dautenhahn et. al. 2006), there has also been implications regarding their application for assistive technology as well as entertainment (Tamura et. al. 2004). More generally, there has been considerable work on improving human/machine interaction (BuŸhler et. al., 2002) with a focus on more natural modes of indicating human needs and intentions; clearly, whilst these efforts can enhance the way the general population communicate with computers, they have particular importance in assistive technology applications, especially where the user has limited mental capacities.

In our project, to be outlined in what follow, we have combined the areas of multimodal human/machine interaction with both robotic hand/eye coordination and robot navigation. This holistic approach is fairly unique, as represented in the literature. Whilst mobile robot navigation and robotic hand/eye coordination have long been central to Intelligent Robotics research, until recently, questions relating to human/machine communications have mostly been of interest to the Human-Machine Interaction (HMI) research community. Now that robotics researchers have realised the importance of this topic, they have been enthusiastic in their inclusion of these human-centric elements in their research. Nevertheless, the combination of HMI, navigation and robotic hand/eye coordination in the one project is rare but is the main emphasis of our project which sets it apart from other research efforts.

3. The Robot, Sensors, and 'Spatial Intelligence' Methodologies

The robot [See Figures 1 (a) and (b)] consists of two main parts. The mobile base is simply an adapted electric wheelchair motor/gear/control set, which can carry a human payload for up to six hours between battery charging. It is differentially steered with castor wheels at front and back (for stability) and can turn on the spot. A UMI six degree of freedom robot arm is mounted on the mobile base. This arm is safe to use in the vicinity of humans since it is slow, relatively weak and has plastic coverings. It has an extensive vertical movement axis, which makes it ideal for retrieving objects off various height tables and shelves and has a simple two fingered gripper. The control schematic is shown in Figure 2. An onboard laptop computer drives a serial four port server. One port sends commands to the robot manipulator and a second drives a 32 channel servo motor (hobby type) controller.

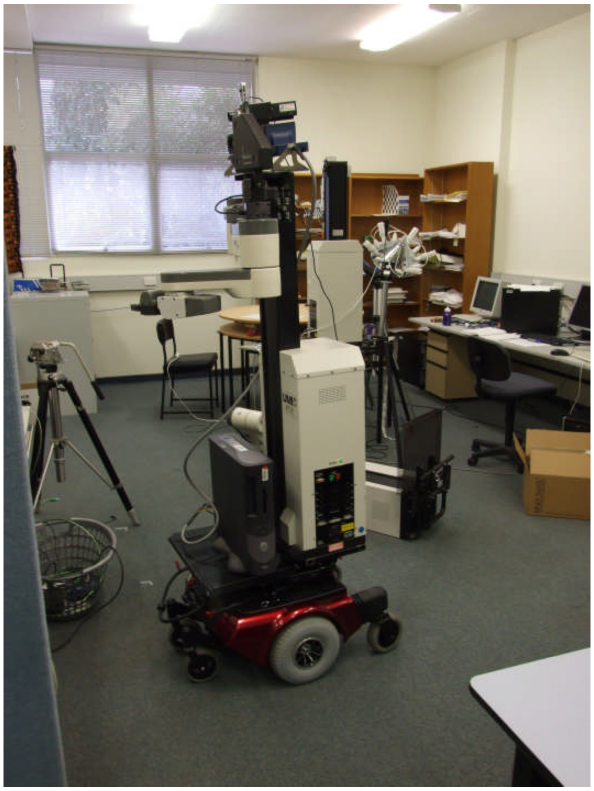


Figure 1(a) Instrumented Robot with Manipulator

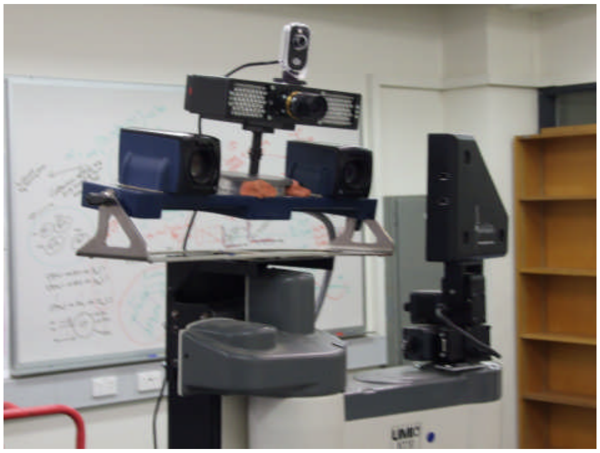


Figure 1(b). Gaze Tracker, Stereo and Range Cameras

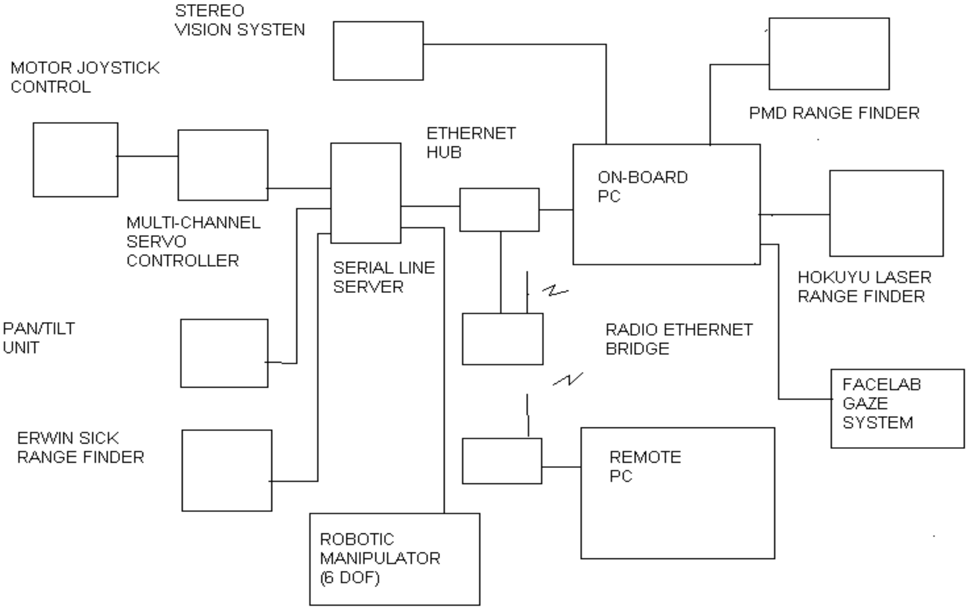


Figure 2. Control/Sensor Data Acquisition Schematic

Two of these channels control the joystick normally used to drive the wheelchair base. A third serial line driver port can control a pan/tilt head for directing a stereo camera/colour camera system towards various targets and the fourth collects range data. Sensors onboard the robot include a Hokuyu line scan laser scanner to be mounted on the robot manipulator hand and an Erwin Sick laser range finder low at the front of the robot, a colour panoramic camera at the very top and a stereo gaze direction analyser (SeeingMachine's Facelab), currently pictured at the base of the robot but to be relocated at head height. A simple localisation scheme will use panoramic vision mapping with pre-scanned laser range/finder camera maps of the working environment acquired by a Riegl LMS Z420i scanner/imager [See Figure 3.].

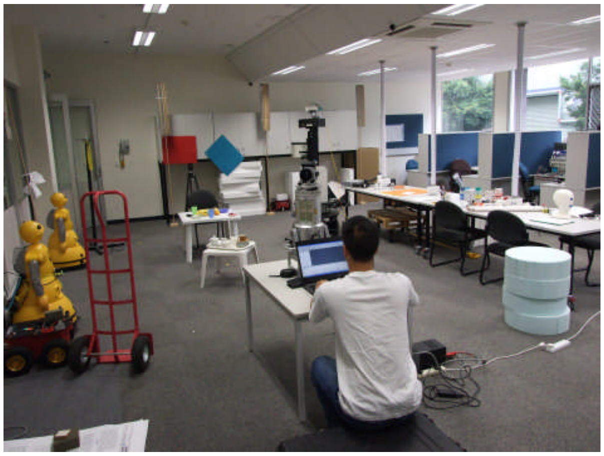


Figure 3(a). Riegl Laser Range Scanner/Camera

The detailed 3D colour map thus acquired will be hand annotated to indicate functional spaces (eg. kitchen, bathroom etc.) and functional large objects (tables, shelves, refrigerator etc.). The overhead camera will be able to note changes of position of chairs and tables for navigation purposes but human intervention will be required for changing functional annotations if such is required over time. Distance Transform (Jarvis, 1985; Jarvis, 1994) path-planning methodology is to be used for global planning with semi/reactive obstacle avoidance adapted from an earlier project (Jarvis, 2000). Further details follow.

Scene analysis of objects on table tops (to determine the existence, location, pose and identity of relevant objects) will be carried out using a combination of laser range finding (Hokuyo laser scanner) and passive stereo vision (Pointgrey's Triclops/Bumblebee). Details follow. Some limited tactile sensing between the robot manipulator's grippers is envisaged to confirm the identity and actual grip of objects to be manipulated and carried. A typical intention such as 'please find my blue mug, which is usually in the kitchen and brings it to me' would be supported by this system. The location of humans would also be tracked so that the robot can approach them and or follow them if they are moving, as the situation may require (Chakravarty and Jarvis, 2006). For example, it makes sense to bring a requested item to the requester, even if he/she has moved since the request.

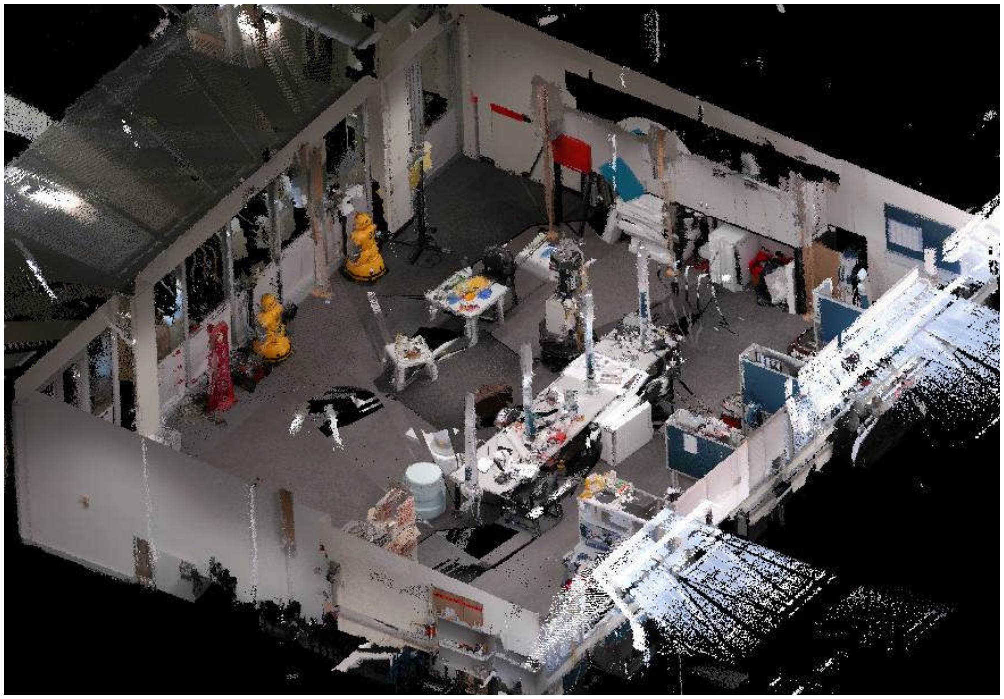


Figure 3(b). Range/Colour Scan of Laboratory

4. Multimodal Human/Robot Transactions

The dominant modes of human/robot communication for this project are spoken language understanding and gesture recognition (including eye gaze). A complex aspect of the language understanding component involves the use of dialogue history and user models as disambiguating knowledge sources and is thus dynamic (anytime) rather than static process. A simple 'first try' version of this combination (speech/gesture) has been published (Harte and Jarvis, 2007). Also included are to be face recognition to establish the identity of the user so that her/his particular habits of communication may be accommodated, and to establish authorisation and 'attachment' for a period sufficient to complete a simple task following instruction. Face recognition can also be used to check whether a visitor is a familiar one or a stranger who perhaps should not be given free entry without further checks by authorised personnel. A gaze direction system (Facelab), previously used to help a disabled user to navigate a wheelchair (Jarvis, 2002), is also to be used to refine gesture interpretation which will mainly be concentrated on arm and hand movement (e.g. say when looking at an object being roughly pointed at). Details follow. The overall schema for the project is shown in Figure 4. However, it is hard to use such a conceptual breakdown as an implementation guide. A simpler way of resolving ambiguities of human intention than is shown in Figure 4. will be used in the first instance. If several modalities (say speech and gesture) are in conflict as to the user’s intention, the confidence weighted probabilities of interpretations for each mode separately will be used as votes, looking for feasible alternatives and the most likely correct one. If there remains high uncertainty after this process, the user can be asked for clarification or to simply repeat the request more carefully and perhaps slowly. Clearly we would like to avoid these kind of clarification requests as they would annoy the users. Hopefully, if we are able to associate certain communication habits for individuals, we can minimise their use.

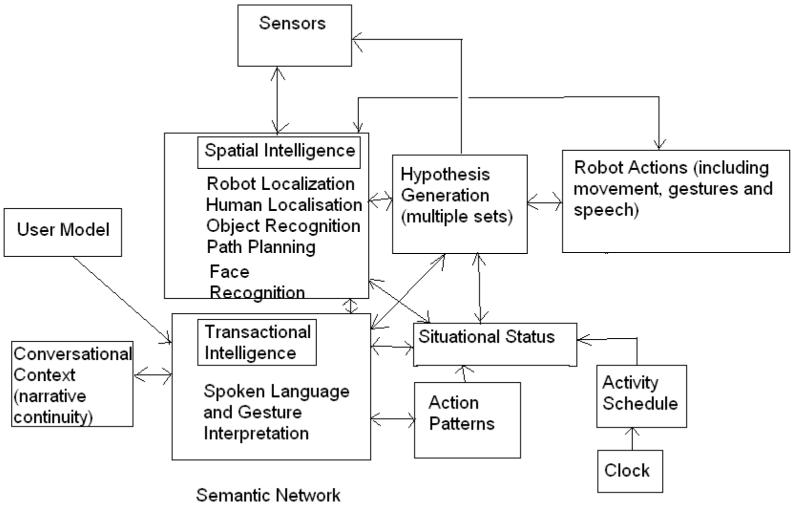


Figure 4. Semantic Network Schematic

5. Task Subdivisions for Project Development

In order to allow two fairly distinct groups of researchers, one predominantly working in robotics, the other on language understanding (within the context of the robot's capability and pertinent human requests for assistance), a functional subdivision of the system has been developed with the intention of being able to integrate the two teams' efforts in a seamless way. Three interlinking nodes can be defined in this task subdivision plan. Firstly, one node will be responsible for the generation of hypotheses of the intent of the user and possible reasonable robot responses and tasks, and the resolution of ambiguity towards discovering a dominant hypothesis. This system will use current dialog, dialog history and user modelling as well the spatial (existence and probable location of objects) and temporal contexts(time of day and scheduled events), together with lists acceptable task possibilities. Some kind of extended negotiation (transaction) will sometimes be needed to obtain some convergence between what the user wants and what is reasonable for the robot to do. Some clarification questions may also have to be posed for this purpose and feasible alternatives may be offered for selection and acceptance.

The second node embodies the capabilities of the robot, its various skills, such as collision-free navigation in dynamic environments (details follow), scene analysis, hand/eye coordination and the means of triggering tasks relevant to the assistive technology domain. Inputs instructions to this node from the hypothesis generation/resolution and task definition node are to be unambiguous and reasonable (but may prove to be unfeasible in a particular instance) and ready for immediate execution. The robot will then attempt the assigned task and report success or failure of completion, with the possibility of specifying one of severable failure modes (such as navigation passage blocked, object can not be found , object found but inaccessible, batteries exhausted etc.).

The third and most vital node concerns the database of spatial geometry (room dimensions, locations, fixed furnishings etc.), object lists (specific as well as generic), estimations of probable location, time stamps of all actions taken which modify the database and the source of the modification (robot/sensor system or hypothesis generator/ambiguity resolver system). Flagging all modifications for validation would be a useful mechanism to support database integrity.

The initial database would be based on dense 3D geometric and surface colour data from the Riegl laser range/colour image scanner which would collect information off-line as a habitat specification , done only once beforehand. This raw scanner data will be hand annotated to label all relevant spaces, furniture, fittings, objects and utilities (stoves, fridges, heaters, toasters, kettles etc.) and extract size, colour and location data for the database. Fixed and movable items will be so classified as would be specific (e.g. a particular book) and generic items (e.g. regular mugs, plates etc.). As object moving actions and/or sensor observations dictate, the database would be modified to reflect the new reality. Clearly all proposed changes will need validation before execution. Uncertainties can be specified probabilistically. Whilst one could consider the robot (plus sensors) being able to construct this database piece by piece as it moves about, the idea of using pre-scanned data is much more practical and almost certainly more accurate. It also permits advancing the project to the transactional intelligence development stages with minimal delay.

We intend to move towards an integrated system where the time of day, the likelihoods of various objects being at various locations (normal expectations plus history of use), the behavioural particulars of a user, the history of language dialogues and gestures, the risk factors associated with making incorrect interpretations and the nuisance value of too many clarification queries can all be taken into account within a working physical system where a real robot carries out useful tasks for an aged, fragile or otherwise impaired human in the familiar surroundings of a home-like environment, where people and robots freely and safely mix.

6. Progress to Date and Plans for the Future

The entire project is a quite ambitious and complex one with many interlinked components. However, many of these have already been addressed in earlier projects and are readily adapted to this one. The Distance Transform methodology (Jarvis, 1985; Jarvis, 1994) for global path planning has been fully investigated and applied successfully in conjunction with barcode scanned localisation on an indoor robot capable of finding its own collision-free way around an obstacle strewn environment (Jarvis, 1997). Localisation using a panoramic vision system and image matching against pre-scanned 3D plus colour detailed maps of the environment has also been demonstrated (Jarvis, Ho and Byrne, 2007), yet the system of overhead panoramic camera localisation is preferred for simplicity. Fusing simple speech recognition with primative gesture and object recognition has also been demonstrated (Harte and Jarvis, 2007). Gaze tracking and dynamic obstacle reactive avoidance in relation to a semi-autonomous wheelchair project has also been employed successfully (Jarvis, 2002) as has human target tracking (Chakravarty and Jarvis, 2006) and face recognition (Axnick and Jarvis, 2005). The challenge is to combine all these previously tested systems into a common framework and to provide semantically related clues and sophisticated ambiguity resolving methodology to meet the requirements of the assistive technology environment targeted.

It is intended (an already commenced PhD. Research project) that a sophisticated gesture capture and recognition system be developed for markerless subjects, using a combination of colour video and range camera sensors (details follow), fusing these two sources of spatial data to extract the dynamic parameters of link movements on a skeletal model (upper torso and head only, initially) of a human and then to train this system to bind sequences of movement to intended communication tokens (for each individual subject). Once an individual is identified using face recognition their raw gesture sequences can be interpreted in a customised way, since different people often have differing ways of specifying intention which may be sometimes culturally dependent as well as individual. Gaze direction vectors and mouth movement detection (for verifying that the person fronting the robot is speaking) will also be extracted.

In what follow, some details of progress to date are reported as extensions of general approach material presented earlier in the paper. These include navigational, scene analysis and gesture recognition work completed so far.

A. Mobile Robot Navigation

The three essential sub-system requirements for autonomous mobile robot navigation are localisation (determining the location and orientation of the robot), environmental modelling (capturing relevant details of the working environment) and path planning (determining the collision-free movements of the robot through the environment from a start point to a nominated goal).

Localisation can be performed using simple odometry (from measuring wheel rotation) with a known starting position/orientation but, due to slippage, imperfect circularity of wheels, undulation of the floor and wheel shape variations under load, errors accumulate incrementally and can, eventually, render the evaluated location/orientation unusable. Beacons at known locations can also be used but this requires careful, possibly tedious, site preparation. Natural landmarks are an alternative, but the computational load is quite high, depending on what on-board sensors are used. In our project we have the advantage of a pre-scanned environment (using a Riegl LMS Z420i laser range scanner/camera) within which particular objects (eg. doors, tables, fridge, book shelves, cupboards etc.) can be annotated and localisation can be determined using an onboard panoramic vision system. However, we have chosen to use a fixed high vantage point panoramic video camera that can recognise the position and orientation of the robot as well as track people and note variations in the obstacle strewn space, both slow and fast changing. Thus the path planning can take into account both static and dynamic obstacles and also accommodate the positions and movements of people, perhaps even using predictions of human movement intention into consideration.

Environmental modelling requires either prior knowledge of the working environment (maps, plans, scans etc.) or a means of incrementally constructing a model using on-board sensors whilst the robot moves around the environment. A considerable body of published work addresses the notion of combining localisation with environmental mapping (Simultaneous Localisation and Mapping – SLAM) (Durrant-Whyte and Guivant, 2000) but, until the map is sufficiently complete, optimal path planning cannot be carried out. In our situation, a complete detailed scan of the environment is taken just once using a Riegl LMS Z420i laser range scanner/camera. The particular advantage of this approach (in contrast to SLAM) is that all functional details of the environment (tables, doors, stoves, cupboards) can be annotated in the data for proving goals for subsequent path planning in accordance to the task required to be carried out by the robot. The overhead panoramic camera mentioned in the previous paragraph supplements this data to include updates and note dynamic aspects, particularly the movement of people.

Path planning has aims, firstly, to arrive at a nominated location without obstacle collision and, secondly to do so efficiently as determined by some optimality criterion. Both static and dynamic obstacles should be avoided, perhaps some busy traffic areas a voided if possible as a courtesy to humans and perhaps even the approach to a human made unobtrusively yet without startling the human from behind included in the path planning strategy. From amongst the large number of path planning strategies available we have chosen the Distance Transform approach since all the aspects mentioned above, such as static and dynamic obstacle avoidance preferred no-go zones human movement predictions, human approach preferences etc., can be easily taken into account and the path plan re calculated frequency when required.

The Distance Transform (DT) path planning algorithm is very simple to construct. Details are provided elsewhere (Jarvis, 1994), but the gist of approach can be easily described:

1. Construct a tessellated floor map with each cell representing a 'floor tile' of appropriate dimensions relative to the robot dimensions (say, a 2x2 set of tiles approximately the 2D size of the robot). Each cell should contain a positive cost representing the cost of entering that cell, obstacles being given an infinite cost. Preferred no-go zones can have high costs. It is even possible to have the cost of entering a cell depend on which neighbour the entrance comes from, thus allowing preferred directions and one way only constructs but this will not be included here.

2. In a separate map same structure as the cost map, described above, goal point cell is set to zero and all free space (space not occupied by obstacles) set to a large number). This is called the DT map.

3. In a forward raster order (left to right, top to bottom), skipping over obstacle cells, replace the constants of each cell by the least of recently visited neighbour's (three above and one to the left) number plus the cost (from the cost map) of entering that all.

4. In a reverse raster order (right to left, bottom to top) repeat the strategy of 3, noting that recently visited neighbours consist of three below and one to the right.

5. Repeat 3 and 4, above, until no change occurs. Now the DT map is the cost weighted Distance Transform.

6. From any point in free space, the steepest descent trajectory in the DT map leads optimally to the goal. The DT map, itself, is starting point independent.

The starting point independence of the DT map has the advantage that, should the robot wander off or deliberately move off the planned path (to avoid a fast mobbing obstacle not yet in the cost map), it can recover by continuing a steepest descent path from its new location. Since the DT provides the distance to the nearest goal for each free cell, the steepest descent trajectory can be followed from any point in fee space to reach the goal in an optimal path.

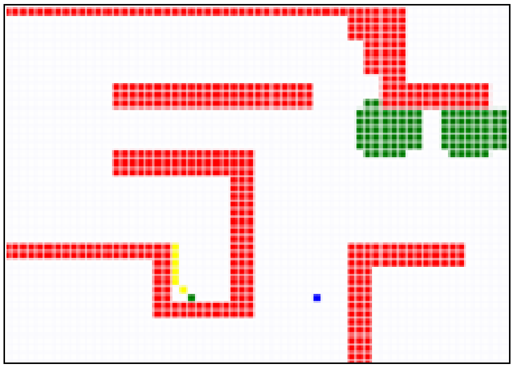
Multiple goals (any one of which when achieved is sufficient) can be included without altering the algorithm (putting zeros in the DT map), the steepest descent trajectory from any point in free-space leading to the goal with the least cost of achieving.

The DT strategy can be extended to any number of dimensions and can also be adapted to cope with the six degree of freedom (or more) movements of robot manipulations using configuration space.

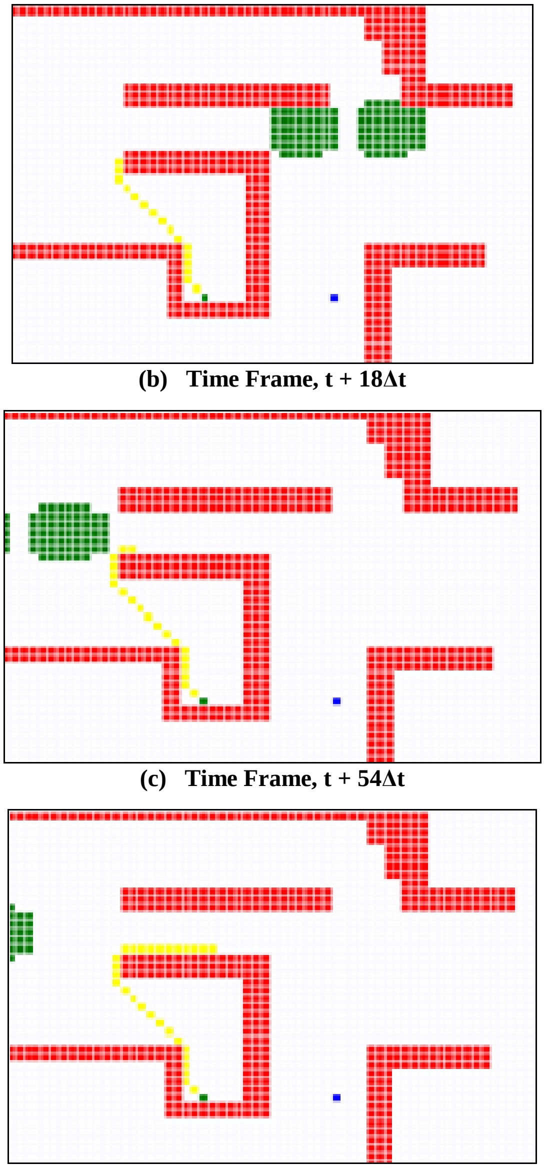
Should one of the dimensions be time a spatio-temporal DT can be generated. The only significant difference when compared to physical dimensions is that, time being irreversible, the construction of the DT only requires one pass on the time dimension.

Suppose a time dimension were added to a two physical dimension tessellated map space. A stack of 2D maps, each representing are point in discreet time, can be piled up, with the top one being at the present and the bottom most one the time count in the future represented the extent of the future factual details of obstacle space or even precautions of such. Cells in the 3D (2D + time) stack marked zero represent goals in time/space (rendezvous). If these are all fixed in position (on a vertical pole through the stack) it means that a nominated position can be reached at any time (of course we still want cost optimality in reaching it). Isolated goals must be reached precisely in the appropriate time interval. A continuous strength of goals through time is like trying to meet a moving goal. Again a corresponding cost may contain non-negative cost values. If only distance and waiting is costed a simple cost structure where waiting costs one unit, front/back and sideways moves costs 2 units and diagonal moves costs 3 units $\left(\frac{3}{2} \approx \sqrt{2}\right)$. Details are given in (Jarvis, 1994). The algorithmic scan moves from the most future layer backwards in time until the top (present) level, replacing costs at each level in parallel (each cells replacement can be calculated independently). Only one pass is necessary. From the present level start point the steepest descent path through time space leads to the least costly achievable goals. The cost replacement rule is to calculate the cost from each neighbour and the cell itself in the next time interval plus the cost of moving into the cell and choosing the least sum. No physical movement costs one unit for waiting cost (like a taxi charge).

Of course, in the simplest case one must have a perfect prediction of the locations of all obstacles in the future until the time interval represented by the bottom map. However, these costs can be probability values calculated from predicted movement observations. As the robot moves physically in time the future horizon can be extended and repopulated with estimated cost with the DT calculated repeated as open as required. Since the space/time DT is a one pass evaluation and essentially a parallelisable algorithm it can be executed quickly. If the obstacle prediction is based on observed movements and simple expectations of straight line trajectories the future costs can be simply determined. Uncertainty can be modelled using Gaussian or other distribution functions and even the distortions of the spread functions caused by not being able to penetrate fixed obstacles can be taken into account. Some example of dynamic obstacle field path planning are shown in Figure 5.



(a) Time Frame, t + 6Δt



(b) Time Frame, t + 18Δt

(c) Time Frame, t + 54Δt

(d) Time Frame, t + 65Δt

Figure 5. A scenario where waiting is preferred by path planning algorithm.

Figure 5 represents a collection of key snapshots from the simulator at different time slices. In the figure, the green shaded region represents the moving obstacle. When the robot encounters the predicted moving obstacles, it waits until the obstacle passes before continuing towards its destination. Although, it may seem that the moving entity is colliding with the red-shaded static obstacles, this is not the case as both the red-shaded static obstacles and the green-shaded moving entities are dilated in order to avoid getting too close to each other and the robot.

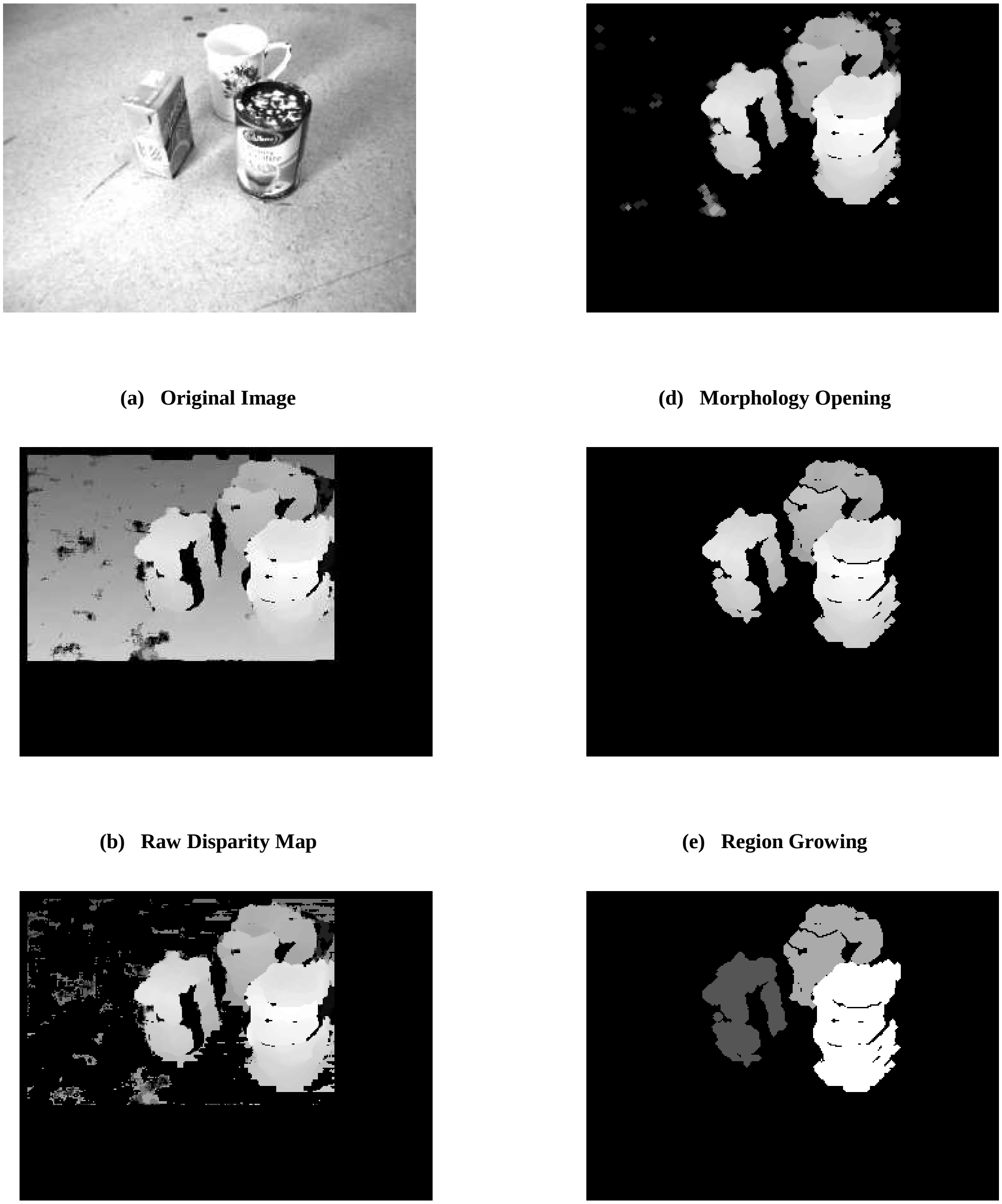
Thus in some scenarios, it can be argued that the optimal route may require a robot to stop and wait approach to avoid moving obstacle.

Unpredictable dynamic changes in the environment can be observed from an overhead panoramic video camera and can be quickly incorporated into the DT planner.

From time to time it may also be required that the annotated map of the environment be adjusted if more permanent changes are to be accommodated. This can be done manually.

B. Robotic Hand/Eye Coordination

When the robot approaches the table or shelf where a target object (eg. cup, fruit, spectacles, pencil etc.) might be found, the local scene needs to be analysed to identify the target object and estimate its position and pose so that it can be grasped and withdrawn from its environment without collision (so that it might be taken to the human requester). Scene analysis concerns describing objects in terms of identity, position/pose and juxtaposition components may have to specify proximity, access, support or containment aspects of the target object in the context of its relevant neighbours. Robotic hand eye coordination's refers to the capability of directing a robot manipulator to manipulated objects in the scene based on it s visual analysis results. In our case, the robot arm is on a mobile robot also carrying the scene analysis sensors. The aim of robot hand/eye coordination in the context of our application is to retrieve the target object without collision. Scene analysis can be particularly difficult if objects in the scene and in a jumble or when some are visually obscuring others. There are two classes of object recognition tasks involved. The first class is where a specific, unique object needs to be found against a database of possibilities. Methodologies based on Scale Invariant Feature Transform (SIFT) (Lowe, 1999) features are ideal for solving this type of pattern recognition problem. A number of localised features are extracted by sight as signatures of the objects they belong to. Finding sufficient matching signatures with respect to a particular object in the database objects (pre calculated) is all that is required to identify a specific object (eg. a particular mug, bottle, pan or spectacles etc.). The object database can have multiple entries of the same object viewed from a variety of positions; this permits recognition even in severely obscured conditions and for varied poses. We use a Triclops (Pointgrey) stereo camera to initially segment the objects in a scene and to identify the supporting plane (eg. table surface). SIFT can then be used to identify the target object if it is present and the stereo disparity (3D data) and segmentation results can be used to construct a virtual bounding box around the identified object to allow grasping to be planned and executed from above (thus reducing the collision problem). A simple example is illustrated in Figure 6.



(a) Original Image

(d) Morphology Opening

(b) Raw Disparity Map

(e) Region Growing

(c) Ground Plane Removal

(f) Object Extraction Result

Figure 6. Stereo Segmentation Analysis Sequence

The second class of recognition problem which needs to be solved in the generic one where the target object is in a class where unique individuals are not represented in the database. Finding an apple or pear or a glass, fork, knife etc. would be examples of the aim for this type of recognition. This is a more difficult problem, which we are still working on, the obscurance problem being a particularly difficult aspect of this task. The simple-minded approach of picking up and separating each object and then attempting to recognise it in isolation is not considered a satisfactory methodology for our application.

C. Gesture Recognition

Gesture recognition is to be one key component of the transactional intelligence aspects of this project. Combining gesture with voice recognition to reduce ambiguity in expressing a task intention to the robot is seen as proving a natural and robust human/machine interface, which could be used by non-technical, possibly frail users. Off-the-shelf face recognition and voice recognition systems are used to support this aim. We adopted a face recognition software package called 'verilook' (version 3.2), which is provided by Neurotechnology. It supports simultaneous multiple face processing in live video and still images. It does not require high resolution images; webcams or other low cost cameras can be used. A quality threshold can be used during face enrolment to ensure that only the best quality face template will be stored into database. VeriLook has certain tolerance to face posture that assures face enrolment convenience: rotation of a head can be up to 10 degrees from frontal in each direction (nodded up/down, rotated left/right, tilted left/right).The voice recognition software we use is called 'Dragon Naturally Speaking' (version 9) provided by Nuance Company. It translates the user's speeches into text. The recognition accuracy can be significantly improved by intensive training. In addition, we can add words or phrases into our own vocabulary set and put emphasis on them in the training stage, so that we can achieve a high recognition rate for the frequently used phrases in our application. By recognising an individual, his/her expected behaviour with respect to voice and gesture communications can be used to simplify and improve the quality of gesture/voice command and dialog recognition.Also, the same person using a fetch command can be the recipient of the fetched object even if they have moved, provided their movements are tracked by the overhead camera. Face recognition can be used to verify that tracking has been carried out correctly. Gestures are useful and a natural way of expressing geometrical and spatial information, particularly pointing and come and go and stop indications.

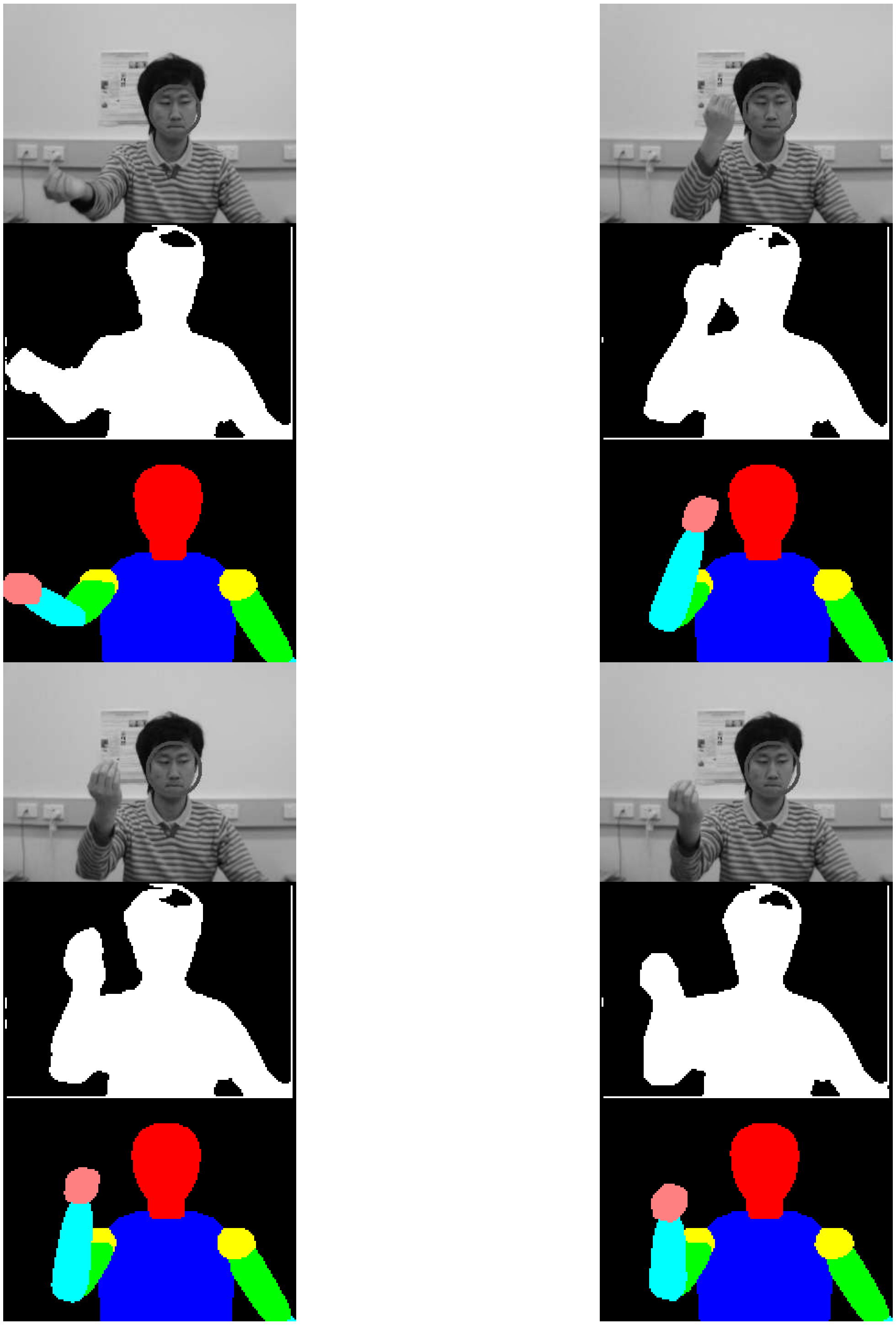
We are using a PMD Technologies range camera which provides frontal surface distances at low resolution (160x120) but high speed. The technology behind this type of camera is infrared laser time-of-flight measurement with each pixel evaluating this measure. A colour web-cam, mounted on the range camera, is also used as it provides higher resolution colour image data. Detecting a face is a first step to determining the position of a person as a preliminary to arm and hand gesture recognition. The resolution of the range camera is too low for reliable face detection so the colour web-cam image is used instead. The algorithm used is based on Hair- like features (Lienhart and Maydt, 2002) and boosted classifiers (Lienhart, Kuranov and Pisarevsky, 2003). Subsequently, the position of the face is found in the range camera data using the SAD (Sum of Absolute Difference) matching method.

We anticipate using a Xuuk's Eyebox 2 camera which detects infrared light reflected through the retinas (exploits the 'red-eye' effect) to count and localise people looking towards the camera to instantly position the vision/range system mounted on the mobile robot in an appropriate position for face and subsequent gesture recognition. We also hope to identify the person who is giving commands to the robot, amongst those facing the robot, by lip movement and audio cues.

When the user's face has been located in the range camera data, the distance between that user and the robot can be estimated. On the assumption that the human body (upper torso, arms and head) can be enclosed in a cubic volume determined by the 3D position, all other objects (background) can be excluded from further analysis. This approach performs well in cluttered and dynamic environments and is clearly insensitive to the colour of clothes worn, even if these colours are similar to background colours.

Upper body gestures are recognised using a model based approach which is less sensitive to noise than model-free alternatives and provides a comprehensive understanding of all major joint angles (finger joints have not been tackled). The degree of freedom of the required model depends on the specific application and a trade-off between accuracy and the number of parameters of the model to be determined (Gavrila, 1999) must be made. Superquadrics are used for our body models.

Whilst other researchers have used multiple cameras from a variety of surrounding viewpoints, this approach does not suit our application where the robot carries the gesture recognition sensors and directs them at selected humans. Thus we are constrained to single viewpoint methods. We use both silhouette and depth information from a single viewpoint to extract upper torso 3D joint angles by finding the best matches between the observed data and body models. Figure 7 shows recognition results when the user's arm moves forwards and backwards. This example shows that the system can deal with self-occlusion problems.



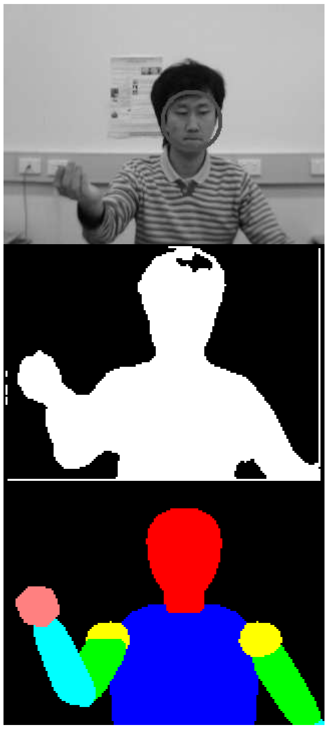


Figure 7. Image, Silhouette and Gesture Model Sequence.

The gaze direction of a person's attention is an important cue regarding nearby objects with which he/she is indicating. People tend to look at the target in the early stages of a hand pointing gesture. Using a FaceLAB (from SeeingMachines) system manufactured by Seeing Machines, which uses a pair of video cameras to stereoscopically determine head pose and eye gaze direction, we can determine a vector of the users gaze direction. The head pose and eye-gaze data can be combined using confidence-weighted factors to estimate the visual attention vector. Intersecting this vector with a hand/arm pointing_gesture and assist in identifying a target object.

So far we have concentrated on static gestures, but will soon extend the systems to understand dynamic gesturers, most likely by combining HMM (Hidden Markov Models) and FSM (Finite State Machines).

Since different particular gestures for a given intention are behavioural aspects for certain races, ethnic groups and individuals, we will customise our gesture recognition system for each registered individual who can be identified using face recognition. For those not registered or identified a default set of general gesture parameters will be used but with a lesser expectation of success.

We hope to eventually, not only combine gesture and speech but also temporal and individual behavioural cues to attempt to correctly interpret the human intention of a robot command. Furthermore we intend to permit the robot to query ambiguous commands to refine its expectations of completing a mission successfully.

Using the high fidelity 3D/colour scan of the working environment (using a Riegl LMS Z420i laser/camera scanner, as mentioned earlier) we will be able to label objects with functional attributes and to build up probabilistic models of where various objects might be found. We will then be able to design optimal search paths for the robot to find the target object with the least expected time of discovery. Behaviour patterns of individuals can also be learned and fed into the system to enhance both the Spatial Intelligence and Transactional Intelligence of the system.

We plan experiments to demonstrate the following kinds of supporting service to aged and/or fragile or otherwise disabled users:

(a) Find and fetch common objects such as spectacles, books, mugs, utensils etc.
(b) Reminders and dispensation of medications
(c) Identify users (and maintain the continuity of a transaction).
(d) Check visitor's identity at the door
(e) Lead a user to a nominated place (eg. dining hall).
(f) Track and follow humans.

The particular behaviour of characteristics of the users, time of day, known special circumstances (eg. anticipating a visitor) would be accommodated by the multi modal transactional intelligence system and appropriate clarification queries prompted by the nature of unresolved ambiguities designed to minimise the nuisance value of such. Eventually it is hoped to have a learning system adaptively refine the whole system to gradually improve the efficiency of the service processes and minimise the annoyance of clarification queries. A lot has still to be done but the framework is now in place and a number of individual components ready for integration.

Conclusions

This paper has briefly outlined the framework of a multi-model spatial and transactional intelligence system directed at having robots help aged, fragile or otherwise disabled people cope with their living environments in an assistive technology context. Much has yet to be done but the way forward is clear though complex. Only actual physical demonstration will eventually prove the system functionally feasible and worthwhile. Such is the overall goal of the project. Much is yet to be done, but the elements for an integrated system are well advanced. However, it is likely that integration will itself be a very difficult process and one which must be planned carefully, each member of the research team being aware of the need to provide clear interface data exchanges between the components in a unambiguous format devised in concert. As the project progresses we would not be surprised to find that the natural ambiguity of human to human communication will have to be resolved to an extent not initially envisioned as a pre-requirement for effective human-robot collaboration.

Acknowledgment

This research is supported by the Australian Research Council through a Discovery Grant (DP0878195).

References

[1] Axnick, K.B.J. and Jarvis, R.A. (2005), Face and pose recognition for robotic surveillance, Proc. 2005 Australasian Conference on Robotics and Automation, 5th to7thDec., Sydney, pp. 1-9.

[2] Bühler, D., Minker, W., Huluser, J and Kruger, S. (2002) Flexible Multimodal Human-machine Interaction in Mobile Environments, 7th International Conference on Spoken Language Processing, Sept. 16-20, Denver, Colorado, USA.

[3] Carrozza, M.C., Massa, B., Micera, S.,Zecca, M. and Dario, P. (2001) A ‘Wearable’ Artificial Hand for Prosthetics and Humanoid Robotics, Proc. 2001 IEEE-RAS International Conference on Humanoid Robots.

[4] Chakravarty, P. and Jarvis, R.A. (2006). Panoramic Vision and Laser Range Finder Fusion for Multiple Person Tracking, accepted for presentation at the International Conference on Intelligent Robots and Systems, Oct. 9-15, Beijing, China.

[5] Dautenhahn, K., Walters, M., Woods, S., Koay, K.L., Nehaniv, C.L., Sisbot, A. and Simeon, T.(2006) How may I serve you?: a robot companion approaching a seated person in a helping context, ACM/IEEE International Conference on Human-Robot Interaction, Proc. 1st ACM/SIGART conference on Human-robot interaction, Salt Lake City, USA, pp. 172-2006.

[6] Durrant-Whyte, H. F. and Guivant, J. (2000) Simultaneous localization and map building using natural features in outdoor environments, Intelligent Autonomous Systems 6. Vol 1, pp 581-588, IAS-6, Italy, 25-28 July.

[7] Gao, T., Greenspan, D. Welsh, M., Juang, R.R. and Alm, (2005)Vital Signs Monitoring and Patient Tracking Over a Wireless Network, Proc. 27th Annual International Conference of the IEEE EMBS, Shanghai, Sept.

[8] Gavrila, D.M. (1999) The Visual Analysis of Human Movement. A Survey, Computer Vision and Image Understanding.

[9] Hager, G.D., Chang, W-C., and Morse, A.S.(1995) Robot Hand-Eye Coordination Based on Stereo Vision, IEEE Control Systems Magazine, Vol.15. pp. 30-39.

[10] Harte, E. and Jarvis, R.A. (2007). Multimodal Human-Robot Interaction in an Assistive Technology Context using Environmental Knowledge, accepted for presentation at the Australasian Conference on Robotics and Automation , 10th to 12th. Dec. Brisbane, Australia.

[11] Hu, H.H.,Jia, P.,Lu, T. and Yuan, K. (2007) Head gesture recognition for hands-free control of an intelligent wheelchair,Industial Automation: An International Jopurnal, Vol. 34, No. !, pp. 60-68.

[12] Jafari, S and Jarvis, R. (2005). Robotic Hand Eye Coordination from Observation to implementation, International Journal of Hybrid Intelligent Systems (IJHIS), 2, pp. 269-293.

[13] Jarvis, R.A.(1985) Collision-Free Trajectory Planning Using Distance Transforms, Proc. National Conference and Exhibition on Robotics, Melbourne, 20-24th Aug. 1984, also in Mechanical Engineering Transactions, Journal of the Institution of Engineers, Vol.ME10, No.3, Sept. pp. 187-191.

[14] Jarvis, R.A. and Byrne, J.C.(1987) An Automated Guided Vehicle with Map Building and Path Finding Capabilities, invited paper, Fourth International Symposium of Robotics Research, University of California at Santa Cruz, 9-14 Aug. pp. 155-162.

[15] Jarvis, R.A. (1994). On Distance Transform Based Collision-Free Path Planning for Robot Navigation in Known, Unknown and Time-Varying Environments, invited chapter for a book entitled 'Advanced Mobile Robots' edited by Professor Yuan F. Zang, World Scientific Publishing Co. Pty. Ltd., pp. 3-31.

[16] Jarvis, R.A. (1997). Etherbot - An Autonomous Mobile Robot on a Local Area Network Radio Tether, Proc. Fifth International Symposium on Experimental Robotics, Barcelona, Catalonia, June 15-18, pp. 151-163.

[17] Jarvis, R.A. (2000). A User Adaptive Semi-Autonomous All-Terrain Robotic Wheelchair, 6th International Conference on Intelligent Autonomous Systems, July 25-27, Venice, Italy, pp. 671-678.

[18] Jarvis, R.A. (2002). A Go Where you Look Tele-Autonomous Rough Terrain Mobile Robot, 8th International Symposium on Experimental Robotics (ISER’) 2, Sant’Angelo d’Ischia, Italy, 8-11 July.

[19] Jarvis, R.A., Ho, Nghia and Byrne, J.B.(2007). Autonomous Robot navigation in Cyber and Real Worlds, CyberWorlds 2007, Hanover, Germany, Oct. 24th to 27th, pp. 66-73.

[20] Jarvis, R. A.(2009) Multimodal Robot/Human Interaction In An Assistive Technology Context, The 2nd International Conference on Advances in Computer-human Interaction,(ACHI 2009), Cancun, Mexico,1st to 6th Feb.

[21] Kanade, T., Collins, R., Lipton, A., Anandan, P., Burt, P. and Wixson, L. (1997) Cooperative Multi-Sensor Video Surveillance, Proc. 1997 DARPA Image Understanding Workshop, May, pp. 2-10.

[22] Lee, T. and Mihailidis, A. (2005) An Intelligent emergency response system; preliminary development and resting of automated fall detection, J. Telemed Telecare, Vol 11, pp. 194-198.

[23] Lienhart, R., Maydt, J. (2002). An extended set of Haar-like features for rapid object detection, Proceedings of 2002 International Conference on Image Processing, Vol. 1, pp 900-903.

[24] Lienhart, R., Kuranov, E. and Pisarevsky, V. (2003) Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection, in DAGM 25th Pattern Recognition Symposium.

[25] Lowe, David G. (1999) Object recognition from local scale-invariant features, *International Conference on Computer Vision*, Corfu, Greece (September), pp. 1150-1157.

[26] Pons, J.J, Rocon, E., Ceres, R., Reynaerts, D., Saro, B., Levin, S. and Van Moorleghem, W. (2004) The MANUS_HAND. Dextrous Robotics Upper Limb Prosthesis: Mechanical and Manipulation Aspects, Autonomous Robots, 16, pp. 143-163.

[27] Prior, S.D. (1990) An Electric Wheelchair Mounted Robotic ARM-A Survey of Potential Users, Journal of Medical Engineering and Technology, Vol. 14, Issue 4, July, pp.143-154.

[28] Rawlinson, D. and Jarvis, R.A. (2007) Topologically-directed navigation, Robotica: International Journal of Information, Education and Research in Robotics and Artificial Intelligence (on-line publication 10th Sept., (hard copy journal version imminent).

[29] Tamura, T., Yonemitsu, S., Itoh, A., Oikawa, D., Kawakami, A., Higashi, Y., Fujimooto, T. and Nakajima, K. (2004) Is an Entertainment Robot Useful in the Care of Elderly People With Severe Dementia?, The Journals of Gerontology Series A: Biological Sciences and Medical Sciencews, 59, M83-M85.

[30] Volpe. B.T., Huerta, P.T., Zipse, J.L., Rykman, A., Edwards, D., Dipietro, L., Hogan, N. and Krebs, H.I. (2009) Robotic Devices as Therapeutic and Diagnostic Toots for Stroke Recovery, Arch Neurol. Vol. 66, No. 9, pp. 1086-1090.