What the Robot Sees & Understands Facilitates Dialog

John Zelek & Dave Bullock & Sam Bromley & Haisheng Wu
Intelligent Systems Lab, School of Engineering
University of Guelph
Guelph, ON, N1G 2W1, Canada
jzelek@uoguelph.ca

Abstract

The particular applications that interest us include search and rescue robotics, security, elder/disabled care or assistance, or service robotics. In all of these application areas, visual perception and internal representations play a pivotal role in how humans and robots communicate. In general, visual perception plays an important role in communicating information amongst humans. It is typically imprecise and humans usually use their cognitive abilities to interpret the intent. We explore real-time probabilistic visual perception in three different roles: (1) the tracking of human limbs; (2) stereo vision for robot and human navigation; and (3) optical flow for detecting salient events and structure from motion. Our visual perception efforts are expressed in probability distribution functions (i.e., Bayesian). The robot requires to have this uncertainty propagated for any subsequent decision-making task. A related application we are also exploring is using real-time stereo vision to convey depth information to a blind person via tactile feedback. We anticipate that this will provide us some glues for internal representations that can form the basis for human-robot communications. We propose that this representation be based on a minimal spanning basis set of spatial prepositions and show how it can be used as a basis for commands. We assume that uncertainty can be conveyed through the linguistic representation in a fuzzy descriptor.

Introduction

In many applications robots perform tasks in environments that humans occupy and function. The human may be: (1) the user of the robot (e.g., elder care, service); (2) the task of the robot (e.g., search and rescue, security); (3) working alongside the robot (e.g., human-robot co-operation, safety); or an (4) insignificant (bystander) to the task at hand (e.g., vacuum cleaning, safety). In all of these diverse situations, the robot needs to be aware of the humans in the environment as well as the role they play in the current scenario. The types of scenarios that we focus on are unstructured and cluttered environments as opposed to highly structured manufacturing environments.

In order for the robots in the various applications of interest (e.g., search and rescue robotics, security, elder/disabled care or assistance, or service robotics), the robots need to be affordable, low power consuming. Thus they will have to rely on passive sensors such as vision. Vision is imprecise in its description of the world for machines and therefore that has to be represented and conveyed. In addition, the internal representation that the robot uses to describe the world has to have some commonality with human representation for effective communications to exist. We propose that the bond be linguistic in nature.

Visual Perception

The probabilistic framework we have adopted for visual routines is referred to as *Particle filtering*, which is also called the *Condensation algorithm* (Isard & Blake 1998c), is usually used for tracking objects where the posterior probability function is not unimodal or can be modeled by a predefined function such as a Gaussian. The Condensation approach is useful when there are multiple hypothesis and it is necessary to propagate them across time. A Monte Carlo technique of factored sampling is used to propagate a set of samples through state space efficiently. The posterior probability $P(X_t \mid I(x,y,t))$, can be computed by using:

$$P(X_t \mid I(x,y,t)) = \frac{P(I(x,y,t) \mid X_t) P(X_t \mid X_{t-1})}{P(I(x,y,t))} \quad (1)$$

$$= \alpha P(I(x,y,t) \mid X_t) P(X_t \mid X_{t-1}) \quad (2)$$

;where X_t expresses the state at time t. The prior $P(X_t \mid I(x,y,t-1))$ is inferred from predicting $P(X_{t-1} \mid I(x,y,t-1))$ through a temporal model $P(X_t \mid X_{t-1})$ which is used for computing the measurements (observations) $P(I(x,y,t) \mid X_t)$ (i.e., the likelihood), from which the posterior follows. The temporal model typically includes a deterministic drift component and a random diffusion component. It is also a set of samples $S_t = [s_1, s_2, ..., s_N]$ selected from S_{t-1} using a sample-and-replace scheme that are propagated. The posterior is only computed to an unknown scale factor α.

We have used this formalism for the visual perception techniques used by our robot for the basic reason that vision is uncertain and the principle of least commitment should be adhered to as long as possible. This permits a robot to

explain its vision-based actions to a user in a probabilistic form. It also permits the robot to convey this information to a user for the user to use their decision making abilities (cognitive) to make the actual decision.

We have started to use the *particle filtering* framework in three visual routines: (1) tracking; (2) optical flow; and (3) stereo vision. We would also like to explore this framework for face detection. Humans communicate non-verbally via the movements of their arms and hands (e.g., gestures), and therefore we have explored the probabilistic tracking of human limbs. The particle filtering framework permits us to convey our certainty on the locations of the joints in 3D so that a proper judgment can be made on either a gesture or even the intent of the human subject, depending on the role of the human (i.e., user, target of robot).

Gestures from Human Limb Tracking

Automated 3-D tracking of the human body is a necessary prerequisite for human-robot interaction, particularly when there exists the need to direct a robot or visually describe a course of action. Currently, visual tracking technologies use artificial markers and a feature tracking methodology to recover the target user's pose (Aggarwal & Cai 1999r). As well, most tracking systems alter the working environment (Goncalves *et al.* 1995) and/or include multiple camera viewpoints (Gavrila & Davis 1996) in order to solve issues of occlusion and depth ambiguities. Furthermore, many vision-based limb tracking systems rely on computationally intensive procedures which remove the system's ability to perform in real-time (Sidenbladh, Black, & Fleet 2000). We have developed a near real-time 3-D limb tracking system which recursively estimates the joint angles of the user's arms using monocular visual cues in an unconstrained environment. As this is an attempt to reconstruct a 3-D configuration from a 2D data source, the problem is inherently under-determined and ill-posed. This problem is particularly challenging due to the nonlinear dynamics of human limbs, reconstruction ambiguities resulting from the loss of depth information, self-occlusions, and a noisy measurement source due to loose fitting clothing on the target structure. It is a problem which emphasizes inference as much as it does measurement.

A vision-based human-robot interface requires not only target tracking capabilities, but target detection and target initialization as well. The target detection component must be able to differentiate between target users and individuals who are simply passing through the robot's field-of-view. This is an issue often neglected in visual tracking systems, but of great importance to allow for ubiquitous and transparent interaction with robotic devices. We rely on visual recognition of a pre-defined initialization cue to differentiate target users from people who pass by. Once detected, an initial model acquisition stage is used to learn the physical dimensions and appearance model of the target limb during the first few seconds of usage. The target limb is modeled as a set of volumetric cones, connected by joints whose values are tracked probabilistically in state-space. In our experiments with arm tracking, we model the arm as a four degree-of-freedom articulated structure, thus creating a four dimensional state-space. We address the inherent ambiguity of 3-D limb tracking by propagating a multi-modal target posterior which can handle ambiguity in the input data by delaying decisions regarding the target state until less ambiguous data is received Currently the target detection and initialization components are able to perform within real-time limits ($\geq 10fps$). However, the target tracking component is presently performing at sub real-time rates ($\sim 1fps$), though there exists substantial room for optimization via multi-threading and parallelism.

Target Detection and Initialization We have developed a passive initialization scheme in which the robot identifies the target user by visual recognition of the user performing a simple pre-defined initialization cue (e.g. waving an arm several times). This is accomplished through the generation of motion-history images (Davis 1999) which are a means of representing temporal and spatial motion information in an image format. These motion-history images are continually generated during the target detection stage, and are compared to pre-calculated action templates to determine if a user is presently performing the required initialization cue. Our implementation of this event detection method has been quantitatively proved to be subject and location invariant in the recognition of the initialization cue (Bullock & Zelek 2002a).

Once the initialization cue has been recognized, the system learns the physical and appearance models of the target limb in an initial model acquisition stage. A restriction enforced here is that the target limb remain in motion while the model acquisition takes place. This is not a significant constraint as most of the initialization cues we have experimented with involve the user waving or moving the target limb. A motion map is generated using optical flow techniques (Camus 1998), and combined with anthropometric limb information to allow the system to quickly learn the physical dimensions of the limb and the target spatial-chromatic appearance model. This model acquisition stage has been shown to perform well within real-time constraints and is able to accurately describe a target limb by combining spatial and chromatic information into a single target appearance model (Bullock & Zelek 2002a).

Visual Target Tracking Deterministic tracking techniques force the system to make a decision as to the target state (i.e. limb pose) at each time step. In this there is a finite chance of the system making an errant decision, a series of which could lead to permanent loss of the tracked target. Consequently, we track the limb's pose using probabilistic techniques which propagate an entire state-space probability density, rather than a single target state estimate. This offers a mechanism for propagating uncertainty and ambiguity in the measurements. Many visual tracking algorithms use the Kalman or Extended Kalman Filter (Welch & Bishop 2000) for this purpose. However, the Kalman filter is inherently ill-suited to tracking in complex environments since it can only model the target posterior as a uni-modal Gaussian distribution. While this can allow for the representation of uncertainty, it forces the posterior to be modeled as having a single dominant hypothesis. This is often in-

adequate when depth or kinematic ambiguities create input data which tends to support multiple conflicting hypotheses. This motivated us to implement the Condensation particle filtering algorithm (Isard & Blake 1998a) which represents the target posterior not by a Gaussian distribution (a multi-variate mean and variance), but instead by a large set of weighted state-space samples. Each sample, or particle, represents a separate hypothesis as to the true nature of the target, and is weighted according to its calculated likelihood. These particles are made to propagate through state-space according to a motion model ($p(State_t|State_{t-1})$) tuned to the target's behavioral tendencies, and the observed image data. The complete set of particles can combine to form an asymptotically correct estimate of the target state posterior, $p(State_t|Image_t)$. The asymptotic correctness of the tracker output is illustrated in figure reffg:correct. In this figure, the mean positional error of the 3-D hand location estimate is shown to approach zero as the number of samples (and computational resources required) increases. Figure 2(b) shows the estimated limb posterior for the image shown in figure 2(a).

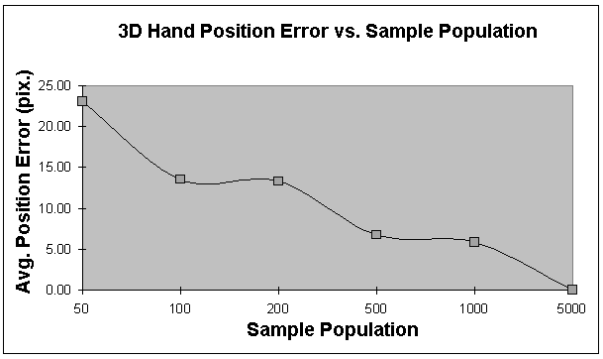
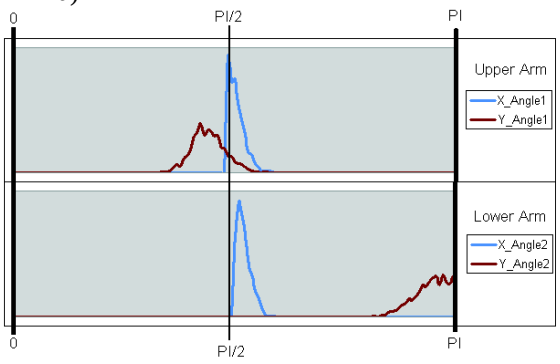


Figure 1: **Accuracy of the Hand Position Estimate:** *The accuracy of the estimated hand position gradually approaches zero as the number of samples (and computational resource required) increases.*

Hypotheses (state-space particles) are weighted according to how well the image data in the region of the hypothesized arm fits the spatial-chromatic appearance model. While this is an adequate tracking cue when the target is clearly visible, during periods of occlusion the state-space particles may drift away from the high-probability regions of state-space and ultimately loss the target. Therefore, a method of focusing the particles into high-probability regions of state-space is required to combat the effects of occlusion. We use the Monte Carlo technique of importance sampling (Isard & Blake 1998b) to redistribute a portion of the particles at each time step using a secondary image cue. We use a novel ridge segment detector which is able to estimate the possible 3-D pose of the arm from the projected contour information. Contours are a natural complement to colored-blobs and thus the two can combine to form a powerful tracking cue in which one excels where the other fails. In our experimentation the system has demonstrated exceptional resilience to target occlusion and temporary target disappear-

a)



b)

Figure 2: **Estimated Arm Pose:** *The estimated 3-D arm pose of the user is shown super-imposed over the original image in (a). In (b) the posterior estimate of the joint angles is plotted for both the upper and lower arm segments.*

ance by employing this search focusing mechanism (Bullock & Zelek 2002b).

The output of the target tracking component is the set of estimated limb joint angles. These can be used to render an image of the estimated arm pose (as is done in figure 2a), or interpreted by a gesture understanding software component for interfacing with the robot. At this state the target tracking component performs at sub-real-time frame rates ($\sim 1fps$), but there exists significant room for optimization by multi-threading and parallelism.

Probabilistic Optical Flow

Optical flow is what results from the recovery of the 2-D motion field (i.e., the projection of the 3D velocity profile onto a 2-D plane; or the resulting apparent motion in an image). Most optical flow techniques assume that uniform illumination is present and that all surfaces are Lambertian. Obviously this does not necessarily hold in the real-world, but we assume that these conditions do hold locally. Optical flow describes the direction and speed of feature motion in the 2D image as a result of relative motion between the viewer and the scene. If the camera is fixed, the motion can be attributed to the moving objects in the scene. Optical flow also encodes useful information about scene structure: e.g., distant objects have much slower apparent motion than close objects. The apparent motion of objects on the image plane provides strong cues for interpreting structure and 3-D motion. Some creatures in nature such as birds are chiefly reliant on motion cues for understanding the world.

Optical flow may be used to compute motion detection,

time-to-collision, focus of expansion as well as object segmentation; however, most optical flow techniques do not produce an accurate flow map necessary for these calculations (Barron, Fleet, & Beauchemin 1995). Most motion techniques make the assumption that image irradiance remains constant during the motion process. The optical flow equation relates temporal (I_t) changes in image intensity ($I(x,y,t)$) to the velocity (i.e., disparity) ((u,v)).

$$I_x u + I_y v + I_t = 0 \tag{3}$$

This equation is not well posed and many approaches (Horn & Schunk 1981) use a smoothness constraint to render the problem well-posed.

$$E^2(x,y) = (I_x u + I_y v + I_t)^2 + \lambda({u_x}^2 + {u_y}^2 + {v_x}^2 + {v_y}^2) \tag{4}$$

Motion field computations are similar to stereo disparity measures albeit for the spatial differences being smaller between temporal images (because of a high sampling rate) and the 3-D displacement between the camera and the scene not necessarily being caused by a single 3D rigid transformation.

A recent hypothesis (Weiss & Fleet 2001) is that early motion analysis is the extraction of local likelihoods which are subsequently combined with the observer's prior assumptions to estimate object motion. Ambiguity is present in the local motion information, either as a result of the *aperture problem* (e.g., the vertical motion component is not attainable from a horizontally moving edge just based on local information) (Wallach 1935) or the *extended blank wall problem* (i.e., both vertical and horizontal gradients are zero and many motion velocities (u,v) fit the brightness constancy equation) (Simoncelli 1999).

The goal in a Bayesian approach to motion analysis is to calculate the posterior probability of a velocity given the image data (Weiss & Fleet 2001). The posterior probability is computing using the spatio-temporal brightness observation (i.e., measurement) $I(x,y,t)$ at location x,y and time t and the 2D motion (u,v) of the object, where α is a normalization constant independent of (u,v):

$$P(u,v \mid I(x,y,t)) = \alpha P(u,v) P(I(x,y,t) \mid u,v) \tag{5}$$

Assuming that the image observations at different positions and times are conditionally independent, given u,v, then:

$$P(I(x,y,t) \mid u,v) = \alpha P(u,v) \prod\nolimits_{i,j} P(I(x_i,y_i,t_j) \mid u,v) \tag{6}$$

where the product is taken over all positions x_i, y_i and times t_j.

The quantity to compute is the likelihood of a velocity $P(I(x_i,y_i,t_j) \mid u,v)$. This also assumes that we are only concerned with a single object which many not necessarily be the case. $P(u,v)$, the prior, has been hypothesized (Weiss & Fleet 2001) that it should favor slow speeds.

For the image velocity likelihood, we have argued that SD (sum difference) can also be expressed as a likelihood (Zelek 2002). Thus making the simplistic optical flow approach proposed by Camus (Camus 1997) a candidate algorithm for a Bayesian approach for real-time optical flow computation. Rather than computing a single likelihood for the scene, we compute a likelihood for each overlapping patch. We also argue that there are really three different likelihood function cases: (1) a well defined symmetric likelihood; (2) an anti-symmetrical likelihood (i.e., aperture problem), and (3) a flat likelihood (i.e., extended blank wall or zero flow). We postulate that the shape of the likelihood (i.e., variance) is an indicator of the reliability of the optical flow value at that location. A tight symmetrical likelihood translates to a good estimator. We also suggest that likelihoods should be propagated spatially in two steps before temporal propagation. Firstly, the *aperture* problem is addressed and secondly the *extended blank wall* problem is solved. We hypothesize that temporal propagation via particle filtering resolves ambiguity.

(a) (b)

Figure 3: **Dense Flow Estimate:** (a) shows where optical flow vectors were detected using the Camus algorithm (Camus 1997), while (b) shows the result of motion detection based on only spatially propagating significant flow vectors.

Stereo Vision: Blind Aid

The work in vision substitution (Meijer 1992) has focused on two main issues: (1) reading and writing; and (2) obstacle detection and avoidance. We are interested in the latter of the two, which is really the problem of spatial orientation. Spatial orientation refers to the ability to establish and maintain an awareness of one's position in space relative to landmarks in the surrounding area and relative to a particular destination (Ross & Blasch 2000). There are approximately 11.4 million visually impaired people in the U.S. In addition, blindness prevalence increases with age and the average population age is gradually increasing. The older population is less likely to be interested in acquiring new skills and may also be subject to loss of hearing, and physical and cognitive function impairment (Ross & Blasch 2000). The need for a personal navigation system has also been discussed by others (S.H. Cheung & J.Patterson 2000).

The role of obstacle avoidance is to present spatial information of the immediate environment for improving orientation, localization and mobility. This is similar to mobile robot navigation and thus is an interesting application problem from which we can reflect on a robot's internal representation. The two oldest aids for this purpose are the *walking cane* and *guide dog*. The walking cane is an effective me-

chanical device which requires certain skills of the person using it to interpret the acoustical reflections that continually result from tapping. The cane's range is only a few feet (limited by the person's reach extent and the length of the cane). Some people find the cane difficult to master or spend significant amounts of time in the learning phase. The guide dog alleviates some of the limitations of the cane but little information regarding orientation and navigation is conveyed to the blind traveler. In addition, dogs require constant care and extensive training for both the dog and person.

Electronic devices have been developed for this purpose and are typically known as *electronic travel aids* (*ETA*'s) or *blind mobility aids*. Early devices relied on an acoustical sensor (i.e., sonar) or a laser light (Meijer 1992). Unfortunately, power consumption is an important consideration and typically laser-based devices require heavy battery packs. Sonar is problematic due to incorrect interpretations when presented with multiple reflections (e.g., corners). Environmental sensing via a camera is attractive due to the wealth of information available through this sense, its closeness in function to the human eye, typical low power consumption (i.e., passive vision) and the low cost that technology has recently presented. In terms of feedback, the typical feedback mechanisms include auditory (Meijer 1992) and tactile (Back-Y-Rita 1995). The Tactile Vision Substitution System (TVSS) (Kaczmarek *et al.* 1985) was developed in the early 70's and displayed map images from a video camera to a vibrating tactile belt worn on the abdomen. Other invasive types of substitution include an approach where an array of electrodes are placed in direct contact with the visual cortex (Dobelle, Mladejovsky, & Girvin 1974),(Hambrecht 1995).

Recently, two devices were developed that evolved from research in mobile robotics, specifically, the NavBelt and the GuideCane (Shoval, Borenstein, & Koren 1998). (Borenstein & Ulrich 1997). The NavBelt provides acoustical feedback from an array of sonar sensors that are mounted on a belt around the abdomen. The array of sensors either provides information in a *guidance mode* (i.e., actively guides the user around obstacles in pursuit of a target) or in *image mode* (i.e., presents the user with an acoustic or tactile image). The GuideCane is a cane attached to a small mobile robot platform with an array of ultrasonic sensors, an odometer, compass and gyroscope sensors. The robot steers around obstacles detected by the sensors. The user receives information via the current orientation of the cane. Unfortunately, the robot is a wheeled platform and therefore restricted to travel along relatively flat surfaces. A problem with the NavBelt is the complexity of learning the patterns of the acoustical feedback and the typical problems of multiple reflection associated with sonar sensors.

A recent development is the further enhancement of a device that converts depth information to an auditory depth representation (Meijer 1992). Rather than using sonar information as input, a camera is used as the input source. The image intensities are converted to sounds where frequency and pitch represent different depths. With a single camera, image intensities do not typically correspond to depth information which is necessary for navigation. Subsequently, an anaglyphic video input has been used (red filter in front of the left camera and a green filter in front of the right camera) where the two video images are superimposed on top of each other. This is analogous to the red-green 3D glasses used for watching 3D movies. Again this anaglyphic image is transferred to an acoustical pattern that the operator interprets. The problem with acoustic feedback is that this ties up the hearing of the person when trying to engage in conversation with other people. In addition, a significant amount of learning is necessary for interpreting the varying beeps and learning the correspondence with the environment. The constant beeping feedback which bears some correlation with the environment can also tend to be annoying.

Ideally, in order to minimize learning and reliance of operator subjectivity, it would be more appropriate for the vision system to only provide a processed depth map to the feedback mechanism. A prototype was constructed, which we plan on continually revising. The prototype is inexpensive and consists of two USB cameras, glove feedback and a wearable computer (currently we use an inexpensive laptop but a small embedded platform is an appropriate substitution). The feedback system is used to relay visual information via tactile feedback through the user's fingers. Simple experimentation has shown the feasibility of this approach and we soon plan on experimenting with the planned demographics for such a device.

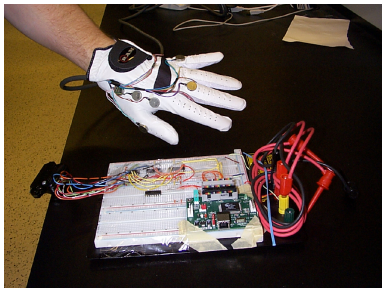


Figure 4: **Second Generation Glove Prototype:** of the tactile feedback unit is shown. There is a major compression of bandwidth in trying to convey a depth map in a tactile form on the hand. It is essential that an appropriate representation ease that transition.

Due to the similarity of trying to solve the correspondence problem in both binocular vision as well as optical flow, we are also trying to cast our stereo vision algorithm into the particle filtering framework. Additional projects that are in their infancy include looking at real-time face detection, which is relevant for finding humans for search and rescue or engagement in dialog.

There is a high bandwidth compression when translating from a depth map to the tactile feedback mechanism. We would like to have an underlying architecture where the stereo vision system can also be used as a sensor for navigating a mobile robot platform. Critical to the tactile conversion of information (e.g., depth map, terrain information, etc.) is some condensed representation of the spatial world (i.e., both obstacles and terrain need to be represented).

We speculate that the glove can also be used as a tac-

tile feedback mechanism when communicating with a robot, playing the role of, lets say, someone tapping you on the shoulder to get your attention. The relevancy in applications such as search and rescue is apparent because the human rescuers will be conducting search concurrently with the robot and the robot only needs to notify the humans when something is of interest for more detailed interaction.

Spatial Representations

One of the most developed theories of spatial representations for large-scale spaces is the *Spatial Semantic Hierarchy* (SSH) (Kuipers 2000). The SSH formalism shows that multiple representations are necessary and can depend on each other. In summary, SSH consists of the following levels:

- the *control level* is used for navigating among locally distinctive states with appropriate control laws (i.e., sensorimotor, feedback control);
- the *causal level* consists of schemas of views and actions (which is an abstraction of the control level), where views can be image-based or descriptors and actions are sequences of control laws;
- the *topological level* are collections of places, paths and regions that are linked by topological relations such as connectivity, order, boundary and containment; and these are created from experienced representations as a sequence of views and actions; and
- the *metrical level* is the global 2D analog (e.g., occupancy grid), with either a single global frame of reference (note: suffers from problems of localization and the space-time cost of mapping algorithms) or a patchwork mapping (loosely coupled collection of local patch maps).

One of the goals is to have a minimally spanning basis set of descriptors for any of the levels. As a start, the abstract levels part of SSH are adequate for any representational systems but the practical implementation requires defining a lexicon for the operators and descriptors. We hypothesize that the descriptors for each of these levels can be obtained from the language we use for describing our world.

Linguistic Representation

Interaction between robots and humans should be at a level which is accessible and natural for human operators. There has been very little research done pertaining to human-machine natural language interaction and communication in the field of autonomous mobile robot navigation (Lueth *et al.* 1994). Natural language permits information to be conveyed in varying degrees of abstraction subject to the application and contextual setting. A major function of language is to enable humans to experience the world by proxy, "*because the world can be envisaged how it is on the basis of a verbal description*" (Johnson-Laird 1989). A minimal spanning language has been used as a control language template onto which recognized speech can be mapped in the SPOTT (Zelek 1996) mobile robot control architecture.

The language lexicon is a minimal spanning subset for human 2D navigational tasks (Landau & Jackendoff 1993; Miller & Johnson-Laird 1976). The task command lexicon consists of a verb, destination, direction and a speed. The destination is a location in the environment defined by a geometric model positioned at a particular spatial location in a globally-referenced Cartesian coordinate space.

A key element of the task command is a minimal spanning subset of prepositions (Landau & Jackendoff 1993), (Zelek 1997) that are used to spatially modify goal descriptions (e.g., near, behind), and to specify trajectory commands (e.g., left, right, north). The spatial relationships used are sparse, primarily including qualitative distinctions of distance and direction. The quantification of the spatial and trajectory prepositions depends on two norms: the definitions for the spatial prepositions *near* and *far* in the current environment and task context. In language design, the descriptors (e.g., spatial prepositions) filter out metric information (i.e., not explicitly encoded), and similarly, such descriptions may be instrumental for providing the structure for a level in a cognitive map. The spatial preposition can also be used for encoding map information in a form that is analogous to the SSH topological level.

Prepositions The preposition is a key element in the linguistic expression of place and path, and there are few of them in comparison to the number of names for objects (Landau & Jackendoff 1993). The totality of prepositional meanings is extremely limited. This fixes the number of prepositions in the English language (Landau & Jackendoff 1993) which makes them ideal for use in a robot task command language. There are two different types of prepositions that are of interest for a robot task command lexicon: (1) one type describes a spatial relationship between objects; and (2) the other describes a trajectory. A preposition in its spatial role is only an ideal. The actual meaning is a deviation from the ideal. It is determined by the context of a specific application. A level of *geometric conceptualization* mediates between *the world as it is* and language (Herskovits 1985).

A locative expression is any spatial expression involving a preposition, its object and whatever the prepositional phrase modifies (noun, clause, etc.):

$$NP_1(preposition)NP_2 \tag{7}$$

where NP is a *noun phrase*. If a noun phrase (NP_i) refers to an object (O_i), then the locative expression can be rewritten as follows:

$$IM(G_1(O_1), G_2(O_2)) \tag{8}$$

where G_i is the geometric description applied to the object (O_i), and IM is the ideal meaning of the preposition. The ideal meaning of a preposition is such that (1) it is manifested in all uses of the preposition, although shifted or distorted in various ways, and (2) it does not apply to the referents of the noun-phrase, but to geometric descriptions associated with these referents (Herskovits 1985). If $[T(IM)]$ is the transformed ideal meaning, then the geometric scene representation can be stated as follows:

$$[T(IM)](G_1(O_1), G_2(O_2)) \tag{9}$$

The different types of categories that typify a geometric description function are as follows (Herskovits 1985):

Spatial Prepositions			
about	above	across	after
against	along	alongside	amid(st)
among(st)	around	at	atop
behind	below	beneath	beside
between	betwixt	beyond	by
down	from	in	inside
into	near	nearby	off
on	onto	opposite	out
outside	over	past	through
throughout	to	toward	under
underneath	up	upon	via
with	within	without	
Compounds			
far from		in back of	
in between		in front of	
in line with		on top of	
to the left of		to the right of	
to the side of			
Intransitive Prepositions			
afterward(s)		apart	away
back		backward	downstairs
downward		east	forward
here		inward	left
N-ward (e.g., homeward)		north	outward
right		sideways	south
there		together	upstairs
upward		west	
Nonspatial Prepositions			
ago	as	because of	before
despite	during	for	like
of	since	until	

Table 1: **The Minimal Spanning Set of English Prepositions**. The possible meanings of all prepositions is extremely limited (Landau & Jackendoff 1993). The above lists the set of all prepositions that minimally span all the prepositional meanings possible in the English language.

1. Functions that map a geometric construct onto *one of its parts*. Examples of parts include a 3D part, edge, or the oriented base of the total outer surface.
2. Functions that map a geometric construct onto *some idealization of the object*. The idealization can be an approximation to a point, line, surface, or strip.
3. Functions that map a geometric construct onto some associated *good form*. A good form is obtained by filling out some irregularities or implementing Gestalt principles of closure or good form.
4. Functions that map a geometric construct onto *some associated volume that the construct partially bounds*. Adjacent volumes include the interior, the volume or area associated with a vertex, and lamina associated with a surface.
5. Functions that map a geometric construct onto *an axis, or a frame of reference*.
6. Functions that map a geometric construct onto *a projection*. The projections can either be a projection on a plane at infinity or a projection onto the ground.

The meaning is transformed due to various contextual factors bearing on the choice and interpretation of a location expression (Herskovits 1985). $T(IM)$ and G functions are frequently fuzzy: however, quantification of a spatial prepositional expression into a fuzzy definition is difficult because context is the key and its quantification is difficult (Herskovits 1988).

Understanding the representations of space requires invoking mental elements corresponding to places and paths, where places are generally understood as regions often occupied by landmarks or reference objects. The spatial preposition is an operator that takes as its arguments, both the figure object and the reference object, and the result of this operation defines a region in which the figure object is located:

$$F_{sp}(O_f, O_r) = R \tag{10}$$

where F_{sp} is the spatial preposition defining a function operating on the figure O_f and reference O_r object models (i.e., O_f is O_1 and O_r is O_2 in Equations 8 and 9) in order to obtain a region of interest R.

The use of spatial prepositions may impose certain constraints on the figure or reference objects, or the resulting region. The restrictions placed on the form of the reference object or figure object by spatial prepositions are not very severe, and only refer to the gross geometry at the coarsest level of representation of the object (Landau & Jackendoff 1993). The object's axial structure plays a crucial role. In a spatial expression defining a location, there are no prepositions which cause the figure or reference object to be analyzed in terms of a particular geon (Biederman 1987). In general, there are no prepositions that insist on analysis of the figure or reference object into its constituent parts. A reference object can be schematized as a point, a container or a surface, as a unit with axial structure, or as a single versus aggregate entity. A figure object can be schematized as most as a single lump or blob (no geometric structure whatsoever), a unit with axial structure that is along, at most, one of its dimensions, or a single versus distributed entity.

Quantifying Prepositional Expressions The spatial relations encode several degrees of distance and several kinds of direction. Many complexities arise from assigning different frames of reference. Some spatial expressions involving axes do not leave this choice of reference system open. The choice of axes can either be defined by the reference object, the speaker (i.e., operator), or the actor (i.e., robot).

The quantification of spatial prepositions depends on the purpose of the use. One possible method is to quantify a spatial expression into regions with a degree of membership, which is done by *fuzzy sets* (Zadeh 1974). This representation is useful as an information source for search strategies that determine the likelihood of locating an object in a particular region, which is encoded by the spatial prepositional expression.

The four levels of distance that are described by English spatial prepositions (Landau & Jackendoff 1993) are:

1. location in the region interior to the reference object (e.g. *in, inside*),
2. location in the region exterior to the reference object but in contact with it (e.g. *on, against*),
3. location in the region proximate to the reference object (e.g. *near*), and
4. location distant from the reference object (e.g. *far, beyond*).

Besides the boundary defined by the reference object, the quantification of a spatial expression only depends on a norm for describing what is meant by *near* and *far*. Although humans are able to represent distance at finer levels for tasks that require finer control, it appears that spatial prepositions do not encode this precision.

Vandeloise (1991), in describing the French equivalents of the English spatial prepositions **near** and **far**, claimed that they are often described in terms of the following factors:

1. The access of the target (i.e., physical such as reachability or seeing such as furthest recognizable object),
2. The dimension of the landmark and to a lesser extent, the size of the target,
3. The size of the speaker, and
4. The speed of the target.

The norm that defines *near* and *far* needs to be defined with respect to some context. Denofsky (1976) claimed that the norm for *far* can be approximately equal to four times the norm for *near*; however, this is clearly not applicable in all situations. For mobile robot navigation, some categories (i.e., or context) that can help quantify the definition are manipulator access, perception access, and the bounds of the environment:

- Manipulator access can define the norm for *near* as being the maximum reach of the manipulator arm from the body surface of the robot.
- Perception access can define the norm for *far* as being equal to the furthest distance that the robot can perceive objects.
- A bounded environment's (e.g., a room) maximum dimensions can define a norm for *far*.

Even though there are only two norms to define (i.e., near and far), it is not clear that such simple definitions for their quantification are adequate across different contextual settings.

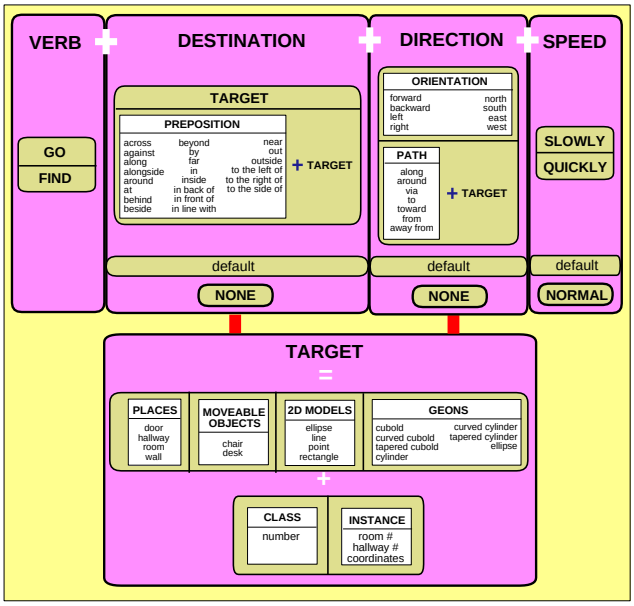


Figure 5: **The Task Command Lexicon**. The basic syntax is given by a verb, destination, direction and speed choice. The destination is a located region that can be modified in a spatial expression. The set of spatial preposition modifiers chosen are 2D because SPOTT is currently only able to perform 2D navigation. A set of trajectory prepositions (i.e., direction) are used to bias the trajectory defined by the local path planner.

A task command as shown in Figure 5 with a basic syntax, given by a verb, destination, direction and speed choice, was formulated for the SPOTT robot control architecture. The targets chosen are based on what the robot (i.e., SPOTT) can perceptually recognize (e.g., door, wall), what the robot knows about (e.g., rooms, hallways in the CAD map), 2D models, and 3D models. Only 2D spatial prepositions are chosen because only 2D mobile robot navigation has been investigated. The role of specifying the direction is to bias the path planning. The lexical set of speed variables is small but a more discriminating set of speeds can be easily accommodated (e.g., very fast, very slow). The lexical basis set can also be used to represent the world in a topological fashion. For the application of navigation, this can define the world and be used for causal schemas for obstacle avoidance (i.e., for both the blind navigator and a mobile robot). Typically, the application of search and resuce will involve terrain that is made up of rubble, uneven terrain that may be difficult to traverse but is traversable under certain operating and control conditions. This is also true for a blind person in urban settings when staircases are encoun-

tered. Some methods (Seraji 2000) have explored classifying terrain with regards to its traversability but none have included control schemas that require initialization under these circumstances. We propose that terrain descriptors be grounded in the SSH formalism causal level, with linguistic indentifiers for states and actions.

Discusssions

We have shown two visual routines (tracking and optical flow) and their probabilistic frameworks. We are currently exploring framing other visual routines such as depth-from-stereo and face detection with a particle filter infrastructure. One of our projects is the exploration of converting depth maps produced from stereo vision into a tactile representation that can be used as a wearable system for blind people. Key to this project is the representation of the environment that facilitates the necessary data reduction. We suggest that the *Spatial Semantic Hierarchy* (SSH) framework be adopted with a linguistic set of operators. We show how spatial prepositions, which are found to be a minimally spanning basis set for the English language, can be used for formulating robot control commands as well as a way of representing topological information about the world for such tasks as obstacle avoidance. Terrain can possibly be represented in the causal SSH level as opposed to the topological level, but is an area for future investigation. By linking internal representations with linguistic representations helps prepare a natural way for communicating with the robot for either issuing commands or for relating world experiences and descriptions.

In this paper we have yet to tie together the Bayesian *pdf*s that result from the various visual routines with the linguistic descriptors. Linguistic terms easily lend themselves to fuzzification and much has been written about this subject matter and we anticipate that the fuzzification requires little attention and will be easily integrated once a complete linguistic descriptor set is available. The description of terrain is an area that requires attention and we feel that the blind urban navigator as well as the search and rescue robot are excellent vehicles for exploring this issue.

References

Aggarwal, J., and Cai, Q. 1999r. Human motion analysis: A review. In *CVIU (73:3)*, 428–440.

Back-Y-Rita, P. 1995. Sensory substitution. *"Nonsynaptic Diffusion Neurotransmission and Late Brain Reorganization" in Demos Publication* 177–203. isbn 0-939957-77-9.

Barron, J.; Fleet, D.; and Beauchemin, S. 1995. Performance of optical flow techniques. *International Journal of Computer Vision* 12(1):43–77.

Biederman, I. 1987. Recognition-by-components: A theory of human image understanding. *Psychological Review* 94(2):115–147.

Borenstein, J., and Ulrich, I. 1997. The guidecane - a computerized travel aid for the active guidance of blind pedestrians. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 1283–1288.

Bullock, D., and Zelek, J. 2002a. Limb target model event initialization for real-time monocular 3-d visual tracking. In *Submitted to CVIU*.

Bullock, D., and Zelek, J. 2002b. Real-time stochastic tracking of human limbs for vision-based interface devices. In *To be submitted to CVIU*.

Camus, T. 1997. Real-time quantized optical flow. *Journal of Real-Time Imaging* 3:71–86.

Camus, T. 1998. Real time quantized optical flow. In *Journal of Real Time Imaging*, 71–86.

Davis, J. 1999. Recognizing movement using motion histograms. Technical Report 487, MIT Media Laboratory, 20 Ames Street, Cambridge, MA.

Denofsky, M. E. 1976. How near is near? Technical Report AI Memo No. 344, MIT AI Lab.

Dobelle, W.; Mladejovsky, M.; and Girvin, J. 1974. Artificial vision for the blind: Electrical stimulation of visual cortex offers hope for a functional prosthesis. *Science* 183:440–444.

Gavrila, D. M., and Davis, L. 1996. 3d modelbased tracking of humans in action: A multiview approach. In *Proc. of IEEE CVPR 1996*, 73–80.

Goncalves, L.; Bernardo, E.; Ursella, E.; and Perona, P. 1995. Monocular tracking of the human arm in 3d. In *ICCV'95*, 764–770.

Hambrecht, F. 1995. Visual prostheses based on direct interfaces with the visual system. *Bailliere's Clinical Neurology* 4.

Herskovits, A. 1985. Semantics and pragmatics of locative expressions. *Cognitive Science* 9:341–378.

Herskovits, A. 1988. Spatial expressions and the plasticity of meaning. In *Topics in Cognitive Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company. 271–297.

Horn, B., and Schunk, B. 1981. Determining optical flow. *Artificial Intelligence* 17:185–204.

Isard, M., and Blake, A. 1998a. Condensation: Conditional density propagation for visual tracking. In *Int. Journal of Computer Vision*, 5–28.

Isard, M., and Blake, A. 1998b. Icondensation: Unifying low level and high level tracking in a stochastic framework. In *Proc. of ECCV*, 893–908.

Isard, M., and Blake, A. 1998c. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision* 29(1):5–28.

Johnson-Laird, P. 1989. Cultural cognition. In *Foundations of Cognitive Science*. MIT Press. 469–499.

Kaczmarek, K.; Back-Y-Rita, P.; Tompkins, W.; and Webster, J. 1985. A tactile vision-substitution system for the blind: computer-controlled partial image sequencing. *IEEE Transactions on Biomedical Engineering* BME-32:602–608.

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

Landau, B., and Jackendoff, R. 1993. What and where in spatial language and spatial cognition. *Behavioral and Brain Sciences* 16:217–265.

Lueth, T.; Laengle, T.; Herzog, G.; Stopp, E.; and Rembold, U. 1994. Kantra: Human-machine interaction for intelligent robots using natural language. In *IEEE International Workshop on Robot and Human Communications.* 106–111.

Meijer, P. 1992. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering* 39(2):112–121. updated info avaiable at http://www.seeingwithsound.com.

Miller, G. A., and Johnson-Laird, P. N. 1976. *Language and Perception*. Harvard University Press.

Ross, D., and Blasch, B. 2000. Evaluation of orientation interfaces for wearable computers. In *IEEE ISWC00*.

Seraji, H. 2000. Fuzzy traversability index: A new concept for terrain-based navigation. *Journal of Robotic Systems* 17(2):75–91.

S.H. Cheung, S.deRidder, K. L., and J.Patterson. 2000. A personal indoor navigational system (pins) for people who are blind. Technical Report Tech. Rep. Technical Report 5051, University of Minnesota, Psychology, Minnesota, USA.

Shoval, S.; Borenstein, J.; and Koren, Y. 1998. Auditory guidance with the navbelt - a computerized travel aid for the blind. *IEEE Transactions on Systems, Man and Cybernetics* 28(3):459–467.

Sidenbladh, H.; Black, M.; and Fleet, D. 2000. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV'00*, 702–718.

Simoncelli, E. P. 1999. Bayesian multi-scale differential optical flow. In *Handbook of COmputer Vision and Applications*. Academic Press. 397–422.

Vandeloise, C. 1991. *Spatial Prepositions: A Case Study from French*. The University of Chicago Press. *Translated by R.K. Bosch.*

Wallach, H. 1935. Ueber visell whargenommene bewegungrichtung. *Psychologische Forschung* 20:325–380.

Weiss, Y., and Fleet, D. J. 2001. Velocity likelihoods in biological and machine vision. In *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press. 81–100.

Welch, G., and Bishop, G. 2000. An introduction to the kalman filter. Technical Report TR95041, University of North Carolina, Dept. of Computer Science, Chapel Hill, NC, USA.

Zadeh, L. 1974. A fuzzy-algorithm approach to the definition of complex or imprecise concepts. Technical Report Memo No. ERL-M474, Electronic Research Laboratory, University of California, Berkeley.

Zelek, J. S. 1996. *SPOTT: A Real-time Distributed and Scalable Architecture for Autonomous Mobile Robot Control*. Ph.D. Dissertation, McGill University, Dept. of Electrical Engineering.

Zelek, J. 1997. Human-robot interaction with a minimal spanning natural language template for autonomous and tele-operated control. In *Proceedings of the Tenth IEEE/RSJ International Conference on Intelligent Robots and Systems*, 299–305.

Zelek, J. 2002. Bayesian real-time optical flow. In *Vision Interface*, 266–273.