Substituting Touch for Vision

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

Abstract

We are currently exploring relaying navigational information (e.g., obstacles, terrain, depth) to a visually impaired person using a tactile glove we have developed. The glove consists of a collection of vibrating motors. The collective patterns of motor activity are used for conveying the navigational information which is mapped from an artificial perception system derived from a wearable camera and computer. The tactile glove has a reduced bandwidth when compared to the visual input stream. Three exploratory routes of tactile mapping include: (1) encoding information in terms of a minimally spanning basis set of spatial prepositions; (2) organizing the hand in terms of functionality (e.g., obstacle motors, terrain motors); and (3) a direct fovea-periphery retinal distinction on the hand. The glove strongly relies on the information provided by the artificial perception system. We have explored a probabilistic framework (e.g., Particle filtering) for modelling dynamical visual processes (e.g., tracking, optical flow, depth from stereo). We suspect that a probabilistic encoding is necessary to model the uncertainty in visual processing. In addition, the integration of temporal stream redundancy helps the reliability of the perceived scene. The internal representations developed for this application will also be useful for mobile robot navigation.

Introduction

Previous depth conveying devices for the blind have relied on active sensors and audible feedback. Active sensors such as sonar are power consuming and have other limitations with the detail and the manner the scene information is relayed back. Passive sensing such as vision overcomes these limitations by relying on the ambient energy in the environment. Audio feedback burdens the one sensory channel that a visually impaired person relies on for communication and safety. Touch feedback is an innovative and new way for conveying scene depth information.

We have proposed a navigational system for the visually impaired that consists of using passive stereo machine vision and a haptic feedback glove capable of conveying complex scene information. The system needs to convey obstacle and terrain information to a blind individual through the haptic channel. The goal is an affordable, low power consuming system that is comfortable and provides safe navigation without the hindrances of more active sensing techniques, nor the disadvantages and interference of auditory navigational aids. Our intent is to allow the user to "feel" their local environment.

Our visual perception routines rely on a probabilistic framework which is modelled on a particle fi lter framework. We have used particle fi lters for human limb tracking and are currently trying to formalize optical flow and stereo vision algorithms into such a framework. We plan on using probablistic stereo vision algorithms and are currently using Konolige's stereo vision algorithm (Konolige 1997). We have started to explore three exploratory routes of tactile mapping including: (1) encoding information in terms of a minimally spanning basis set of spatial prepositions; (2) organizing the hand in terms of functionality (e.g., obstacle motors, terrain motors); and (3) a direct fovea-periphery retinal distinction on the hand.

Probabilistic 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 predefi ned 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 effi ciently. 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.

Visual Target Tracking

Deterministic tracking techniques force the system to make a decision as to the target state (e.g., 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 inadequate 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).

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 vis-

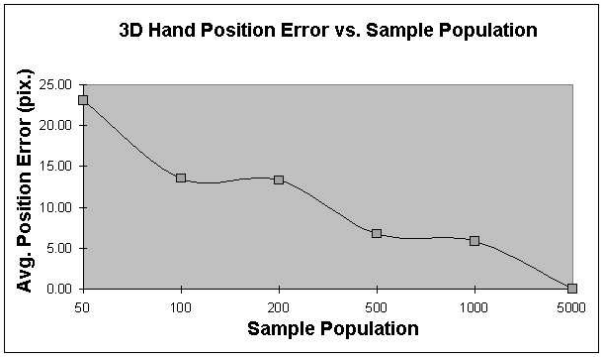

Figure 1: **Accuracy of the Hand Position Estimate:** *The accuracy of the estimated hand position gradually approaches zero as the number of samples (and computational resource required) increases.*

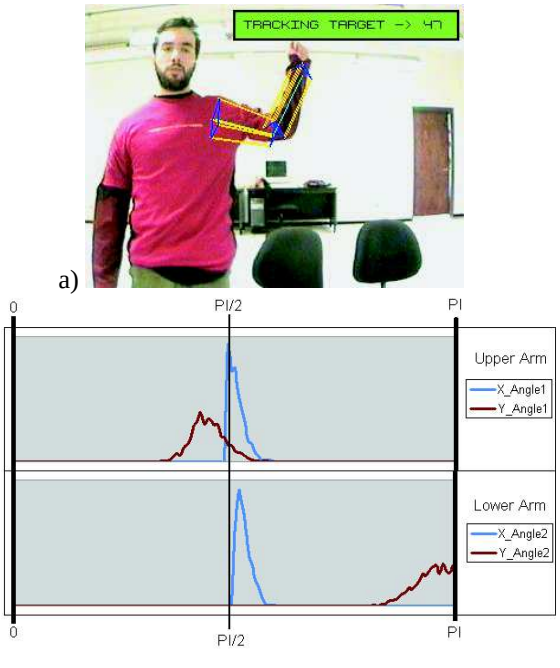

Figure 2: **Estimated Arm Pose:** *The estimated 3-D arm pose of the user is shown super-imposed over the original image in (a). In (b) the posterior estimate of the joint angles is plotted for both the upper and lower arm segments.*

ible, 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 disappearance by employing this search focusing mechanism (Bullock & Zelek 2002).

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

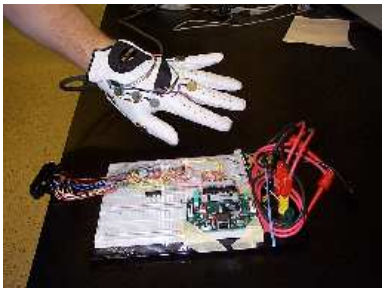

Figure 4: **Second Generation Glove Prototype:** of the tactile feedback unit is shown. There is a major compression of bandwidth in trying to convey a depth map in a tactile form on the hand. It is essential that an appropriate representation ease that transition.

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. 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 tactile 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.

Modeling Touch Perception

A complete model that provides a coherent explanation of touch fibers is a necessary precursor for the design of haptics and sensory substitution devices. In particular, we are interested in a model that will predict the perception of stimulated vibro-tactile patterns superimposed on a tight fitting glove worn by a visually impaired person. We have developed a tactile feedback glove that is to be used in conjunction with a wearable computer and camera vision system. Artificial perception algorithms executing on a wearable computer ingest stereo camera information to produce depth and obstacle information as a navigational aid conveyed via a tactile glove for the blind person. We are currently investigating how to best utilize the glove's bandwidth to convey the necessary navigational information in a timely and intuitive manner. We are about to conduct empirical studies with blind subjects to determine what sensations they experience and which patterns are most intuitive. An accurate model is necessary to predict the responses from the test subjects in order to establish some assurance of safety robustness and predictability for the eventual users.

The haptic community has done very little modelling of the touch system of the human hand. Local properties of mechano-receptors are understood but not their collective interactions. The modelling of a single mechano-receptor (including the mechanics of the skin, end organ, creation of a generator potential, the initiation of the action potential and branching of afferent fibres) has recently been studied for single collections in the fingertips (Pawluk & Howe 1995). This work requires further development into the population responses of neighbourhoods with both excitatory and inhibitatory activity. Another related investigation (Ritter 1992) explored the formation of the somatotopic (projection of the body surface onto the brain cortex) map by a computer simulation of Kohonen's algorithm. The somatotopic map is analogous to other brain sensory processing units (e.g., visual processing pathways). Empirical investigations have provided us with rough estimations on the sizes of the excitatory portion of the receptive fields for touch on the hand. The receptive field distributions are not unlike the fovea-periphery distinction for visual perception where the touch receptors in the fingertips correspond to the fovea. However, we are not interested in analyzing the fingertip touch receptors since these receptors will be left for other sensory activities such as reading (i.e., via Braille). There are approximately 100,000 nerve cells in the hand and 20 different nerve cells, with approximately 12 being of the mechanoreceptor variety that we are interested in. There are approximately 2500 mechanoreceptors per cm. in each fingertip region. The various receptors all have different field sizes (scale), as well as varying dynamic and static properties.

Psychophysical experiments have shown that humans are able to perceive equivalent stimuli via touch that are usually associated with visual perception (e.g., moving bars). We

will develop a set of maps that are analogous to the early visual cortex maps such as maps for edges, bars with various orientations, curvature, and moving bars to name a few. It appears to be obvious that we will probably use touch intensity in a similar fashion as visual brightness but it is a matter of investigation to determine if we can equate vibration frequency with brightness also or perhaps colour.

Linguistic Spatial Representation

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 robot control language template onto which recognized speech can be mapped (Zelek 1997). 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 (*Spatial Semantic Hierarchy*) topological level (Kuipers 2000).

We have also thought of using this topological representation as a form of organizing how the environment is represented on the glove. Another mapping that we are considering is purely function, where different regions of the hand are used for varying roles. For example, the fingers can be used to convey angular (direction) obstacle information (e.g., absence or detection).

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 with a particle filter infrastructure, in particular, as a front-end to our project that is exploring 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. One suggestion is that the *Spatial Semantic Hierarchy* (SSH) framework be adopted with a linguistic set of operators. One other possible representation scheme includes an analogous visual perception stream mapping onto touch perception of the hand (e.g., the fingertips represent the fovea, and the rest of the hand represents periphery regions). Yet another is conveying environmental information (e.g., obstacles, terrain, certainty) using different regions of the hand.

We hope to derive some insight into appropriate environmental representations for mobile robots from our work with the *seeing with touch* project. Staircase navigation for the blind is an area where terrain information representation will be crucial and we hope to use a walking robot in conjunction with blind people for our field trials to test the robustness of our algorithms. This will provide us with some feedback on the appropriateness of our robot-human representational scheme mappings.

We have recently conducted experiments with 10 blind test subjects. The test was to navigate a flat indoor obstacle course with tactile stimulus that indicated directional heading (i.e., three fingers were used to indicate direction: left, in front of, right). With this limited bandwidth, the learning time was minimal (i.e., less than a minute) and all the test subjects were able to navigate the course consisting of three boxes at a normal walking pace. The results of this experiment are preliminary and have yet to be analyzed and will be published shortly. The results indicate a preference for receiving directional information with an appropriate intuitive physical mapping, but this may change when the entire bandwidth of the glove is approached.

Acknowledgments

The authors express thanks to funding from the National Science and Engineering Research Council (NSERC), Canadian National Institute for the Blind (CNIB) and Communications and Information Technology Ontario (CITO).

References

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

Bullock, D., and Zelek, J. 2002. 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.

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.

Konolige, K. 1997. Small vision system: Hardware and implementation. In *Eighth International Symposium on Robotics Research*.

Kuipers, B. 2000. The spatial semantic hierarchy. *Artifi cial 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.

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

Pawluk, D., and Howe, R. 1995. A holistic model of human touch. *Fifth Annual Computational Neuroscience meeting*.

Ritter, H. 1992. Modelling the somatotopic map. In *Neural Computation and Self Organizing Maps*. Academic Press.

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

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.

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.