

Mobile Human-Robot Teaming with Environmental Tolerance

Matthew M. Loper
Computer Science Dpt.
Brown University
115 Waterman St.
Providence, RI 02912
matt@cs.brown.edu

Sonia H. Chernova
Computer Science Dpt.
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
soniac@cs.cmu.edu

Chris V. Jones
Research Group
iRobot Corporation
8 Crosby Dr, Bedford, MA
cjones@irobot.com

Nathan P. Koenig
Computer Science Dpt.
University of Southern California
941 W. 37th Place
Los Angeles, CA 90089
nkoenig@usc.edu

Odest C. Jenkins
Computer Science Dpt.
Brown University
115 Waterman St.
Providence, RI 02912
cjenkins@cs.brown.edu

ABSTRACT

We demonstrate that structured light-based depth sensing with standard perception algorithms can enable mobile peer-to-peer interaction between humans and robots. We posit that the use of recent emerging devices for depth-based imaging can enable robot perception of non-verbal cues in human movement in the face of lighting and minor terrain variations. Toward this end, we have developed an integrated robotic system capable of person following and responding to verbal and non-verbal commands under varying lighting conditions and uneven terrain. The feasibility of our system for peer-to-peer HRI is demonstrated through two trials in indoor and outdoor environments.

Categories and Subject Descriptors

I.2.9 [Artificial Intelligence]: Robotics—*Operator Interfaces*; I.4.8 [Image processing and computer vision]: Scene Analysis—*Range data, Tracking*; I.5.4 [Pattern Recognition]: Applications—*Computer vision*

General Terms

Design, Human Factors

Keywords

Human-robot interaction, person following, gesture recognition

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HRI'09, March 11–13, 2009, La Jolla, California, USA.
Copyright 2009 ACM 978-1-60558-404-1/09/03 ...\$5.00.

1. INTRODUCTION

Mobile robots show great promise in assisting people in a variety of domains, including medical, military, recreational, and industrial applications [22]. However, if robot assistants are to be ubiquitous, teleoperation may not be the only answer to robot control. Teleoperation interfaces may require learning, can be physically encumbering, and are detrimental to users' situation awareness. In this paper, we exhibit an alternative approach to interaction and control.

Specifically, we show the feasibility of active-light based depth sensing for mobile person-following and gesture recognition; such a sensor has strong potential for reducing the perceptual (and computational) burden for tasks involving person following and observation. The most essential aspect of our system is the reliability from which accurate silhouettes can be extracted from active depth imaging. Our system is augmented with voice recognition and a simple state-based behavior system for when the user is out of visual range.

Existing approaches integrate vision, speech recognition, and laser-based sensing to achieve human-robot interaction [20, 14, 6, 8]. Other recent approaches focus specifically on people following [16, 5], gesture-based communication [10, 7, 23], or voice-based operation [11, 15]. However, these systems are either typically designed for use in indoor environments, or do not necessarily incorporate following and gesture recognition. Our approach is intended to further the field with viability in both indoor and outdoor environments, via the use of active sensing, robust perception mechanisms, and a ruggedized platform. One promising approach to pose estimation via range imaging by Knoop et al [9] uses an articulated model and iterative closest point search. However, their focus is on pose tracking (unlike our gesture recognition) as they have additional assumptions about initial pose alignment.

At an abstract level, we strive for *environmental tolerance*: the ability of a system to work in a variety of conditions and locales. Of course, such tolerance can take many forms, and we do not make blanket claims of robustness against all forms of environmental variance. Our methods are meant



Figure 1: Our testbed system: an iRobot PackBot EOD mobile robot, a SwissRanger camera, and a Bluetooth headset.

to contribute to the environmental tolerance of existing approaches to person-following and gesture recognition, with consideration to variations on lighting and uneven terrain.

Enabled by active depth imaging, we present an integrated robot system for peer-to-peer teaming with the following properties in mind:

- **Proximity maintainance:** ability of the robot to stay within proximity of a moving human user
- **Perception of verbal and nonverbal cues:** ability to recognize both gestural and spoken commands
- **Minimization of instrumentation:** ability to interact with people in their natural environment, with minimal reliance on markers or fiducials
- **Preservation of situation awareness:** minimize interruptions to the user that are caused by monitoring or correcting of the robot’s behavior; for example, the user should not have to constantly look back to see if the robot is following

2. TASK DESCRIPTION

Broadly speaking, our robot is designed to (1) automatically accompany a human on mostly flat but uneven terrain, and (2) allow hands-free supervision by a human user for the completion of subtasks.

A more specific task description is outlined as follows. To initiate following, a pre-parameterized gesture must be performed by a person. This user should then be followed at a certain distance (always maintaining this distance, in case the person approaches the robot). The next execution of the same gesture toggles the following behavior to stop the robot. A second parameterized gesture is used to command the robot to perform a specialized task: a “door breach” in our case. Although we only use two gestures, more gestures can be learned from human demonstration and used for online recognition as in [7]. Voice commands are used to summon the robot back to the user when out of visual range.

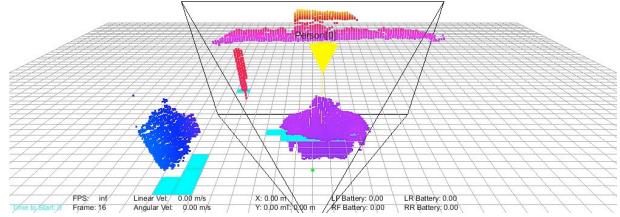


Figure 2: Sample data returned from SwissRanger camera.

Our test environments must consist of mostly flat terrain in two locations: an indoor office space and an paved asphalt parking lot. It is assumed that the user is not occluded from the robot’s view and not directly touching any other objects, although other people can move behind the user from the robot’s view. Although the user height can vary, it is assumed their physical proportions will roughly conform to a medium build. The following sections describe our system in terms of each of its component parts, including the robot platform, perception, and behavior.

3. ROBOT PLATFORM

Our platform consists of an iRobot PackBot, equipped with a 2.0Ghz onboard laptop and a CSEM SwissRanger depth camera as its primary sensor. A Bluetooth headset (used for voice recognition) is our secondary sensor. Details on each of these components will now be described.

The PackBot base is a ruggedized platform with all-weather and all-terrain mobility at speeds of up to 5.8mph. This robot is well suited for tracking and maintaining close proximity to people in both indoor and outdoor environments.

Next we turn to the topic of our primary sensor. We have three basic goals for this sensor: it should be insensitive to global illumination changes; it should not require recalibration for new environments; and it should provide a rich source of data, suitable for detecting and making inferences about a person. A color camera provides a rich data source, but modeling color with strong illumination changes can be difficult, and color calibration is generally required. Stereo can provide a rich source of depth data, but requires strong textures to infer depth, and can suffer from specularity problems. Finally, laser rangefinders are unaffected by global illumination, but (in their typical 1-dimensional form) are not rich enough to robustly detect people and their gestures.

We have opted to use a CSEM SwissRanger, which performs reasonably in both the categories of data richness and illumination invariance. By producing its own non-visible light, and reading the phase-shift of the returned light, this sensor can function in total darkness or in bright light. And because this technology is in its infancy, its capabilities are only bound to improve. Technical specifications for this camera are available online [21].

The SwissRanger provides a real-time depth map to our robot in which the intensity of each pixel represents the distance of the camera from the nearest object along that ray. Such a depth map may be viewed as an image, as in Figure 3(a), or as a point cloud, as in Figure 2. Importantly, extracting human silhouettes from these depthmaps is greatly simplified by the use of this sensor.

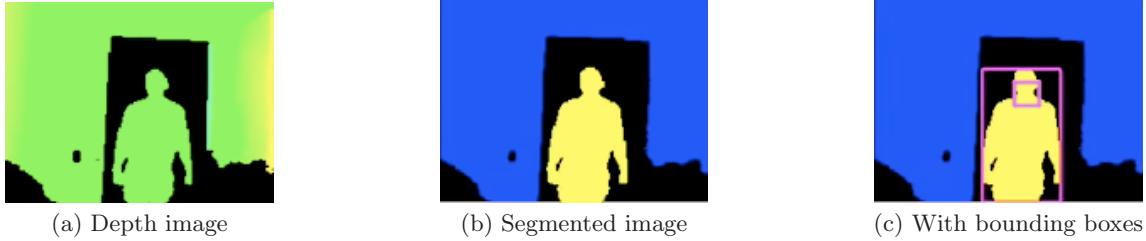


Figure 3: Raw depth image data is segmented into labeled regions, which are then categorized as “person” or “not person.” For “person” regions, bounds for the body and head are then estimated.

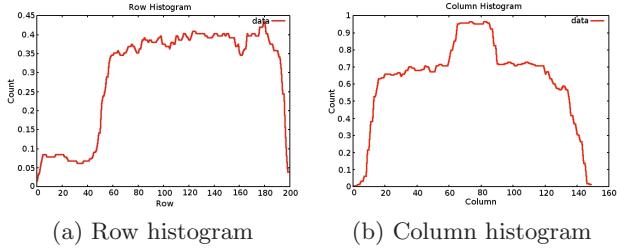


Figure 4: Row and column histogram signature used as a descriptor for the segmented human component in Figure 3(b).

The field of view of the SwissRanger camera is 47.5 x 39.6 degrees, with a maximum range of 7.5 meters. To enable the robot to detect and track humans, the camera is placed at height of roughly 1.5 meters to provide visual coverage of a person’s head, arms and torso. The camera is mounted on a pan-tilt unit to compensate for the small field of view. We found the effective distance range to be between 1 and 5 meters: the camera’s field of view required subjects to be at least 1 meter away, while the camera resolution restricted subjects to under 5 meters.

Our primary sensor is most useful for people detection, following, and gesture-based communications. However, if a person is not in view of the robot, they may still wish to communicate in a natural fashion.

Speech-based communication is a time-tested modality for human-robot interaction. We use the VXI Roadwarrior B150 headset as our secondary sensor. This wireless Bluetooth microphone continually streams audio data to the laptop onboard the robot, providing the user with a means of wireless control over the robot over greater distances.

4. PERCEPTION

The next task is to generate features from our sensor inputs, and use these features to achieve our task goals. These goals include human detection, person tracking and following, and gesture recognition; each will be described in turn.

4.1 Human Detection and Following

In order to detect humans in varying environments, we require algorithms to interpret the depth data obtained by our active lighting system. Our detection algorithm consists of two functions. The first is a pixel-level routine to find collections of pixels in the scene representing contiguous objects. The second function identifies which of the candidate

regions represent humans based on the relative properties of their silhouette.

Tracking and following of a single moving person are then performed using a Kalman filter and PID control, respectively. The robot and the pan-tilt head are separately controlled in this manner, in order to keep tracked person centered in the field of view of our visual sensor.

The first phase of human detection, where all potentially human objects are found, relies on the observation that contiguous objects have slowly varying depth. In other words, a solid object has roughly the same depth, or Z-value in our case, over its visible surface. We have chosen to use a connected components algorithm, based on its speed and robustness, to detect objects. This algorithm groups together pixels in the image based on a distance metric. For our purposes, each pixel is a point in 3D space, and the distance metric is the Euclidean distance along the Z-axis between two points. When the distance between two points is less than a threshold value, the two points are considered to be part of the same object. The output of the algorithm is a set of groups where each group is a disjoint collection of all the points in the image. A simple heuristic of eliminating small connected components, e.g. those with few points, significantly reduces the number of components. The final result is depicted in Figure 3(b).

The second phase of our human detection algorithm identifies which of the remaining connected components represent a human. Motivated by the success of several previous works [4, 12, 17], we use a Support Vector Machine (SVM) trained on the head and shoulders profile to identify the human shape. Our SVM implementation utilizes the libsvm library [3], configured to use C -support vector classification and a radial basis kernel.

Our feature vector consists of the shape of the human in the form of a row-oriented and column-oriented histogram. For a given connected component, the row-oriented histogram is computed by summing the number of points in each row of the connected component. The column-oriented histogram is computed based on data in the columns of the connected component. Figures 4(a) and 4(b) depict the row histogram and column histogram from the connected component found in the center of Figure 3(b). Before computing the histograms, the components are normalized to a constant size of 200x160 pixels. This technique provides a reasonable means to detect a wide range of people in both indoor and outdoor environments, and has shown robustness to variations in person size, clothing, lighting conditions and, to a lesser extent, clutter in the environment.

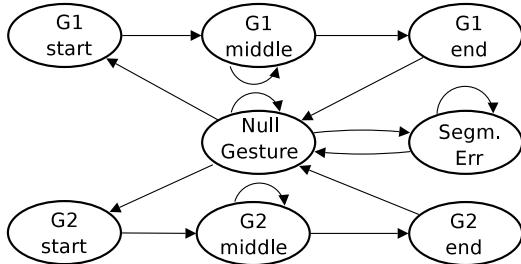


Figure 5: Gesture recognition Markov chain.

4.2 Gesture Recognition

To meet the requirements of environmental tolerance, it is essential that our recognition model be transparent and error tolerant. A Hidden Markov Model [13] is a natural choice for the speed and probabilistic interpretation we desire, and meets our demands for transparency and time-tested reliability. In order to incorporate error tolerance, we include a state to represent segmentation/sensor failure.

The following sections describe our gesture database, states, features, training, and inference.

4.2.1 Gesture Database Construction

Each gesture was recorded offline as a set of observed, ground-truth motions. An actor was asked to perform gestures, and his movements were recorded in a motion capture laboratory with a Vicon motion-capture system. For each gesture, a set of time-varying poses were recovered, stored in 95-dimensional joint angle space.

For the gesture recognition task, it is useful to define gesture progress in terms of the subject's position. Gesture progress is defined as a value in the range $[0, 1]$, such that the boundaries mark the beginning and end of the gesture, respectively.

4.2.2 Gesture State Definition

At any given time, a person is performing one of a set of predefined gestures. We divide each gesture into a beginning, middle, and end. A “null” state identifies when a person is not performing a gesture of interest, and a “segmentation failure” state identifies mis-segmented frames (with unusually high chamfer distance). A Markov chain for these states is shown in Figure 5.

4.2.3 Observation Feature Definition

To recognize gestures, we must infer something about poses over time. We begin with the silhouette and three-dimensional head position introduced in the tracking stage. This information must be converted to our observation feature space, since a silhouette image is too high-dimensional to be useful as a direct observation.

A cylindrical body model is arranged in a pose of interest, and its silhouette rendered. Pose hypotheses are generated from each gesture model in our database, sampled directly from actor-generated gesture poses. A pose hypothesis is then be rendered and compared against a silhouette. Chamfer matching, first proposed in [1] and discussed more recently in [18] is used to compare the similarity of the silhouettes. We opted for a body-model based approach because it has more potential for invariance (ex. rotational invariance), intuitive flexibility (body model adjustments),

and the use of world-space and angle-space error (instead of image-based error).

We then perform a search in the space of each gesture's pose database, finding the best matching pose for each gesture by comparing hypothesized silhouettes with the observed silhouette.

Given n gestures in the training database, we are then left with n “best poses,” each assuming that a particular gesture was performed. We generate our observation feature space by using the gesture progress, the change in gesture progress over time, and the error obtained from the chamfer distance comparison. Thus, given n poses in the gesture database, we are left with $(n \times 3)$ observation variables. We model our observations as being distributed according to a state-specific Gaussian, with a different covariance matrix and mean for each of the states in Figure 5.

4.2.4 Gesture Training and Inference

The HMM was trained on 16 examples total of each gesture, using one female (5'6") and three males (5'10", 5'11", and 6'0") all of medium build. The Viterbi algorithm was run at each frame to recover the most likely gesture history. Because the last few items in this history were not stable, a gesture was only deemed recognized if its “gesture end” state was detected six frames prior to the last frame. This resulted in a recognition delay of 0.5 seconds.

4.3 Speech Recognition and Synthesis

Speech recognition is performed using the free HMM-based *Sphinx-3* recognition system [19]. The central challenge for the speech recognition component is to provide robust and accurate recognition under the noisy conditions commonly encountered in real-world environments. Pre-trained speech recognition systems that are designed for text dictation in office environments perform poorly under these conditions as they are unable to distinguish speech from motor and background noise.

To improve recognition accuracy we train a custom acoustic model using the *SphinxTrain* system. A set of audio speech files containing an abbreviated vocabulary set are recorded using the VXI Roadwarrior B150 noise-canceling headset. Additional audio samples containing common background noises, such as the sound of the robot's tread motors, are used to train the model to differentiate these sounds from speech. The abbreviated vocabulary set limits the word choice to those relevant to the robotic task, improving overall recognition. Our current vocabulary has been selected to include a set of basic operational commands and instructions, such as “stop”, “turn back” and “forward big/little”. In future work we plan to extend this list to include several dozen spoken commands and instructions.

Speech synthesis is performed through the *Cepstral Text-to-Speech* system [2], which enables any written phrase to be spoken in a realistic, clear voice. The Cepstral system allows the robot to verbally report its status, confirm received commands, and communicate with its operator in a natural way. This mode of communication is invaluable as it allows detailed information to be shared quickly, with little distraction to the operator, and without requiring a hand-held device or display unit.



Figure 6: A user’s gesture to stop is followed by a waiting condition, after which the user returns and activates following again.



Figure 7: A user’s gesture to stop is followed by speech-based control, putting the robot into a waiting condition. Speech is then used to retrieve the robot, and a gesture is used to reinitiate following.

5. BEHAVIORS

Because our other modules account for varying environmental factors, our behaviors do not require special-case handling for different physical settings. Input from each of the system components described above is integrated to enable the robot to perform useful functions. The robot’s behaviors consist of time-extended, goal-driven actions which are easy to conceptually understand and use. In this work, we utilize four behaviors, each of which is mapped to a unique command, see Table 1.

The **person-follow** behavior enables the robot to track and follow a user, forward or backward, while attempting to maintain a distance of 2 meters. This behavior is toggled on and off by the gesture of raising the right arm into a right-angle position and then lowering it.

The second behavior, called **door-breath**, is activated by raising both arms to form a T. This behavior looks for a door frame and autonomously navigates across the threshold waiting on the other side for a new command. This maneuver can be particularly useful when exploring structures in dangerous areas.

Behaviors three and four are voice-activated behaviors that can be used to control the robot remotely, even out of view of the operator. The behavior **turn-around** is activated when the person speaks the behavior’s name, and as the name implies the robot rotates 180 degrees in place. The “forward little” command activates a behavior that drives the robot forward for two meters. A finite state machine representing transitions between these behaviors is shown in Figure 8.

An additional behavior, **camera-track**, is used in combi-

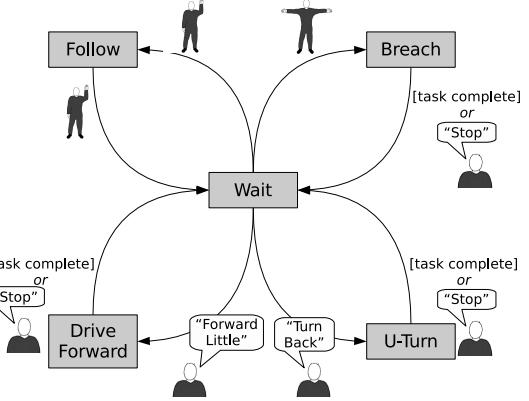


Figure 8: Behavior finite state machine.

nation with above behaviors to control the robot’s camera. The camera tracks the robot’s current target, and resets to a default position when no person is in view.

6. RESULTS

The performance of the system was evaluated within a winding indoor hallway environment, and an open outdoor parking lot environment under cloudy conditions. Both environments were flat in their terrain, and differed principally in lighting and the degree of openness. The attached video demonstrates the robot’s ability to perform the following functions in our test settings:

- Following a person in a closed environment (winding hallway with sharp turns)
- Following a person in an open environment (outdoors)
- Maintaining focus on user in the presence of passers-by
- Accurately responding to gesture commands
- Accurately responding to verbal commands
- Confirming all commands using the speech interface
- Autonomously locating and breaching a narrow doorway
- Interacting with different users

Table 2 presents the average performance of the person and gesture tracking components over multiple runs. Due to reduced sunlight interference, the person detection component accurately detected the user in 91.2% of its sensor frames in the indoor environment, compared to 81.0% accuracy outdoors. The closed indoor environment, which contains many more surfaces and objects detectable by the camera, also resulted in a false positive rate of 1.5%. Both indoor and outdoor accuracy rates were sufficient to perform person following and gesture recognition at the robot’s top speed while maintaining an average distance of 2.8 meters from the user. Gesture recognition performs with high accuracy in both environments, with statistically insignificant differences between the two conditions. Note that the gesture recognition rates are only across frames where successful person recognition took place.

Our system does exhibit certain general limitations. For example, the person in view must face toward or away from the robot (as a side view does not allow gestures to be recognized properly).

Another limitation relates to outdoor operation: although it works well in overcast conditions, bright direct sunlight can be a problem for our chosen sensor. Black clothing also poses an issue, as inadequate light may be returned for depth recovery.

In general, we found that the SwissRanger has advantages and drawbacks that are complementary to those of stereo vision. Indoors, hallway areas are often not densely textured, which can lead to failure for stereo vision algorithms, but which do not impede the use of the SwissRanger. On the other hand, when it works, stereo provides much better resolution than our chosen sensor (which, at 176x144, leaves

Num	Type	Command	Behavior
1	Gesture		person-follow
2	Gesture		breach-door
3	Voice	“Turn Around”	turn-around
4	Voice	“Forward Little”	forward-little

Table 1: Gesture and voice commands, with mapping to behaviors.

	Person Detection	
	% Accuracy (Per Frame)	% False Positives
Indoor	91.2	1.5
Outdoor	81.0	0.0
	Gesture Detection	
	% Accuracy (Per Frame)	% False Positives
Indoor	98.0	0.5
Outdoor	100.0	0.7

Table 2: Person and gesture recognition performance rates in indoor and outdoor environments.

something to be desired in the face of the multi-megapixel cameras of today).

A final limitation relates to our motion generation; although we found it to create adequate following behavior, it is not guaranteed to avoid walls when following around corners, and may back up into a wall (to maintain distance) if a user approaches from the front. Problems such as these could be ameliorated with the introduction of a 360 degree laser sensor.

7. CONCLUSION

In this paper, we presented a robotic system for natural human-robot interaction that shows promise in improving environmental tolerance. We combined a ruggedized physical platform capable of indoor and outdoor navigation with active visual and audio sensing, to achieve person following, gesture recognition, and voice-based behavior. Our choice of ranging sensor and perceptual methods were described in the context of our integrated system for peer-to-peer HRI. Our system demonstrates the feasibility our approach and depth-based imaging as an enabling technology for HRI.

8. ACKNOWLEDGMENTS

This work was funded by DARPA IPTO, contract number W31P4Q-07-C-0096 and the Office of Naval Research, contract number N00014-07-M-0123.

9. REFERENCES

- [1] H. G. Barrow, J. M. Tenenbaum, R. C. Boles, and H. C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, pages 659–663, 1977.
- [2] Cepstral, 2008. <http://www.cepstral.com>.
- [3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, 2006.
- [5] R. Gockley, J. Forlizzi, and R. G. Simmons. Natural person-following behavior for social robots. In *HRI*, pages 17–24, 2007.
- [6] A. Haasch, S. Hohenner, S. Huwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. Fink, J. Fritsch, B. Wrede, and G. Sagerer. BIRON – the bielefeld robot companion. In *Int. Workshop on Advances in Service Robotics*, pages 27–32, Stuttgart, Germany, 2004.

- [7] O. Jenkins, G. Gonzalez, and M. Loper. Interactive human pose and action recognition using dynamical motion primitives. *International Journal of Humanoid Robotics*, 4(2):365–385, Jun 2007.
- [8] W. G. Kennedy, M. Bugajska, M. Marge, W. Adams, P. Fransen, B. R., A. C. D., Schultz, and J. G. Trafton. Spatial representation and reasoning for human-robot collaboration. In *Twenty-second National Conference on Artificial Intelligence (AAAI-07)*, 2007.
- [9] S. Knoop, S. Vacek, and R. Dillmann. Sensor fusion for 3d human body tracking with an articulated 3d body model. In *ICRA 2006: Proceedings 2006 IEEE International Conference on Robotics and Automation*, pages 1686–1691, May 2006.
- [10] N. Kojo, T. Inamura, K. Okada, and M. Inaba. Gesture recognition for humanoids using proto-symbol space. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pages 76–81, 2006.
- [11] M. N. Nicolescu and M. J. Mataric. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 241–248. ACM Press, 2003.
- [12] M. Oren, C. Papageorgiou, P. Sinha, edgar Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Computer Vision and Pattern Recognition*, pages 193–199, June 1997.
- [13] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Readings in speech recognition*, pages 267–296, 1990.
- [14] O. Rogalla, M. Ehrenmann, R. Zollner, R. Becher, and R. Dillmann. Using gesture and speech control for commanding a robot assistant. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, pages 454–459, 2002.
- [15] P. E. Rybski, K. Yoon, J. Stolarz, and M. M. Veloso. Interactive robot task training through dialog and demonstration. In *HRI '07: Proceeding of the ACM/IEEE international conference on Human-robot interaction*, pages 49–56, New York, NY, USA, 2007. ACM Press.
- [16] D. Schulz. A probabilistic exemplar approach to combine laser and vision for person tracking. In *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, August 2006.
- [17] H. Shimizu and T. Piggio. Direction estimation of pedestrian from multiple still images. In *Intelligent Vehicles Symposium*, pages 596–600, June 2004.
- [18] C. Sminchisescu and A. Telea. Human pose estimation from silhouettes - a consistent approach using distance level sets. In *WSCG International Conference on Computer Graphics, Visualization and Computer Vision*, pages 413–420, 2002.
- [19] Sphinx-3, 2008. <http://cmusphinx.sourceforge.net>.
- [20] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, head pose and gestures. In *IEEE/RSJ International Conference Intelligent Robots and Systems*, volume 3, pages 2422–2427, Sendai, Japan, 2004.
- [21] SwissRanger specifications, 2008. <http://www.swissranger.ch/main.php>.
- [22] S. Thrun. Toward a framework for human-robot interaction. *Human Computer Interaction*, 19(1&2):9–24, 2004.
- [23] S. Waldherr, S. Thrun, and R. Romero. A gesture-based interface for human-robot interaction. *Autonomous Robots*, 9(2):151–173, 2000.