Using Vision, Acoustics, and Natural Language for Disambiguation

Benjamin Fransen[1], Vlad Morariu[1,2], Eric Martinson[1,3], Samuel Blisard[1,4], Matthew Marge[1,5], Scott Thomas[1,2], Alan Schultz[1], and Dennis Perzanowski[1]

[1]Naval Research Laboratory
Washington, DC 20375
fransen@aic.nrl.navy.mil
alan.schultz@nrl.navy.mil
dennis.perzanowski@nrl.navy.mil

[2]University of Maryland
College Park, MD 20742
morariu@umd.edu
scthmas@cs.umd.edu

[3]Georgia Institute of Technology
Atlanta, GA 30332
ebeowulf@cc.gatech.edu

[4]University of Missouri-Columbia
Columbia, MO 65211
snbfg8@mizzou.edu

[5]University of Edinburgh
Edinburgh, Scotland EH8 9YL
m.marge@sms.ed.ac.uk

ABSTRACT

Creating a human-robot interface is a daunting experience. Capabilities and functionalities of the interface are dependent on the robustness of many different sensor and input modalities. For example, object recognition poses problems for state-of-the-art vision systems. Speech recognition in noisy environments remains problematic for acoustic systems. Natural language understanding and dialog are often limited to specific domains and baffled by ambiguous or novel utterances. Plans based on domain-specific tasks limit the applicability of dialog managers. The types of sensors used limits spatial knowledge and understanding and constrains cognitive issues, such as perspective-taking.

In this research, we are integrating several modalities, such as vision, audition, and natural language understanding to leverage the existing strengths of each modality and overcome individual weaknesses. We are using visual, acoustic, and linguistic inputs in various combinations to solve such problems as the disambiguation of referents (objects in the environment), localization of human speakers, and determination of the source of utterances and appropriateness of responses when humans and robots interact. For this research, we limit our consideration to the interaction of two humans and one robot in a retrieval scenario. This paper will describe the system and integration of the various modules prior to future testing.

HRI'07, March 10–12, 2007, Arlington, Texas, USA.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing—*Discourse, Language parsing and understanding;* I.2.9 [Artificial Intelligence]: Robotics--*Operator interfaces*, *Sensors*; I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence--*Multiagent systems*; I.4.8 [Image processing and computer vision]: Scene analysis—*Color*, *Motion*, *Object Recognition*, *Range Data*, *Sensor fusion*, *Tracking*; I.5.4 [Pattern Recognition]: Applications—*Computer Vision*, *Signal processing*, *Waveform analysis*; I.5.5 [Implementation]: Interactive systems.

General Terms

Design, Human Factors.

Keywords

Acoustics, Artificial Intelligence, Auditory Perspective-Taking, Dialog, Human-Robot Interaction, Natural Language Understanding, Spatial Reasoning, Vision.

1. INTRODUCTION

To be effective, a human-robot interface should handle not only the sensor information from all of its independently operating modules, such as its vision and auditory components, but it should also handle verbal input, and deal with various types of ambiguity that beset vision, acoustic, and natural language understanding systems. The robot must not only be able to see where it is going and discriminate both objects and people, but hear verbal interactions, know who is talking and to whom, know what is being spoken about, and cooperatively interact with humans for effective and efficient communication. To accomplish these tasks, our independent efforts in robot vision, acoustics, spatial reasoning, and natural language interfacing are brought together to overcome some of the bottlenecks in our separate areas, to achieve more robust human-robot interaction.

To accomplish an adequate level of interaction and the ability to disambiguate sensor and verbal input in successfully completing a task, we are integrating several sensor components on a B21r mobile robot, along with a natural language and gesture interface. Earlier shortcomings of the interface are discussed in the literature [21], but our main concern in integrating the various technologies here is to facilitate the disambiguation of referents in a human-robot dialog, by using the strengths of the various modules in whatever combination seems most appropriate when sensor data or verbal input/output is insufficient to unambiguously identify either humans or objects in the environment. This paper describes the system and integration aspects of our work, as well as algorithms. Therefore, it is not a user study, but we plan to conduct an integration experiment using human subjects in the near future.

2. RELATED WORK

This research is based on our previous work on natural language understanding, gesture recognition, and spatial reasoning [21,24] and can be compared with other research efforts; however, we differ from them in several ways. For example, while Leonardo [7] is designed as a stationary platform, our research focuses on a mobile one, placing restrictions on processing capabilities utilized for the task. We also require algorithms capable of working on a moving platform where it is necessary to re-orientate the sensors constantly while suppressing motion noise. Furthermore, our robot operates in a dynamically changing environment. Also, unlike Leonardo [7], our work does not focus on the mimetic aspects of human-robot interaction.

Like the robot Mel [23], we focus on human-robot interaction in a dialog-based setting. Our work overlaps with theirs; however, we believe the interaction in our retrieval scenario expands the requirements for robot scene analysis, since it integrates dialog and sensor issues affected by mobility.

The integration of sensors and language understanding is not emphasized in other research, such as the Hermes robot of [3]. While Hermes is mobile, as is our robot, Hermes is a service robot, designed to perform actions independently. We, like [7] and [23], focus on the process(es) necessary to disambiguate real world settings through sensors and human interaction, but are doing so in a dynamically changing environment.

Finally, unlike many previous investigations which utilize limited pattern matching, we are committed to robust natural language processing to achieve full language understanding. We do not believe dialog issues can be resolved without full language understanding.

3. SYSTEM OVERVIEW

To achieve a collaborative dialog in which humans and a robot disambiguate objects and locations, we employ a blackboard architecture [9] (Figure 1).

Three components—Vision, Acoustics, and Natural Language—provide input to the Multi-Modal Reasoning Component (MMRC). MMRC determines if the input can be mapped to an appropriate Robot Action. If so, actions, such as movement to a location or identification of an object, occur. If not, MMRC determines what element is missing to produce a Robot Action and either prompts the user for additional verbal, visual or acoustic information, or queries one or more of the three input modalities for additional information. MMRC re-integrates the input with the missing information, or corrects the inappropriate information, for an appropriate Robot Action. This architecture is being implemented on a B21r mobile robot named George (Figure 2).

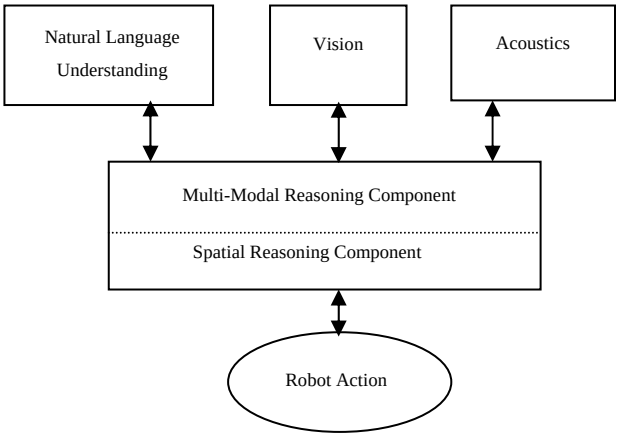

Figure 1. Architecture

In this research we are integrating more sophisticated vision technology, giving us better gesture recognition, as well as human and object detection (Section 4). We also introduce an auditory component for sound localization (Section 5). A new Spatial Reasoning component (Section 6) incorporates the recognition and localization of objects in a 3D model of the world. The natural language understanding system, NAUTILUS, (Section 7) incorporates a dialog manager (Section 8). The basic plan for the dialog scenario is one in which humans converse with each other and with a robot that reacts and interacts with the humans, trying to locate a soda can in a laboratory environment. In Section 9 we present our conclusions and discuss our future plans.

Figure 2. George, a B21r mobile robot

4. VISION

An important aspect of our vision research is the ability to detect and interpret gestures for disambiguating directions and the

locations of objects. To do so, we incorporate a stereo vision system consisting of two cameras mounted on a pan-tilt base (see Figure 2). In this section we discuss gesture recognition input generated by 3D hand and face tracks. We also discuss an additional component of the vision system, an omnidirectional camera, used for human detection and recognition. The latter capability enables the robot to locate human team members with whom it is interacting.

4.1 Gesture Recognition

While previous work in gesture recognition focuses mainly on Hidden Markov Models [11,26], segmental Hidden Markov Models (SHMMs) provide a more tractable system for many types of gestures. SHMMs have been used to recognize speech [19], handwriting [2], and generic waveforms [13]. Our system employs variable length SHMMs for real time recognition of static and dynamic hand gestures, such as "Stop," or "Go over there." The combination of dynamic processor allocation and varying hand speeds produces gestures with both intra- and inter-gesture sampling rate variation. We use SHMMs to accommodate these sampling rate variations and allow sparsely and densely sampled gestures to share common error metrics. By incorporating a priori error models and speed invariant features, training gestures can be generated by a single user action or drawn using a pen-based input.

The novel use of SHMMs for gesture recognition has several advantages; such as, the choice of features facilitates hand speed invariant recognition. Rapid training is made possible by using a priori models. Also, the combination of features and gesture models provides a tractable user interface for rapid training in addition to robust recognition.

4.2 Gesture Model

Gestures are modeled as left to right SHMMs. Each gesture is modeled as a sequence of directed line segments. Each node in the model records a line segment, bounded by a pair of measurements, and a direction:

$$<(x_1,y_1,z_1),(x_2,y_2,z_2),(d_x,d_y,d_z)> \quad (1)$$

where (x_1,y_1,z_1) is the starting location, (x_2,y_2,z_2) is the ending location, and (d_x,d_y,d_z) is the direction. Direction is calculated by normalizing the velocity:

$$d = (V_x / |V|,\ V_y / |V|,\ V_z / |V|) \quad (2)$$

where $V_x=x_2-x_1$, $V_y=y_2-y_1$, $V_z=z_2-z_1$.

Dynamic gestures are defined by a variable length sequence of nodes. For static gestures, only one node is necessary and direction information, d, is not recorded. In this context, static gestures map a specific volume of space to a specific action. To distinguish between static and dynamic gestures during training, a minimum distance for dynamic gestures is defined.

The use of line segments provides invariance to sampling rates and hand speeds. Storing gestures as piece-wise linear motions allows user actions to be compared against a line, not individual measurements. By comparing measurements against line segments, not individual measurements, sparsely sampled gestures that are produced by rapid user actions or slow sampling rates generate models that can share the same error metrics as densely sampled gesture models do.

4.3 Training

Two forms of training are available, permitting users to introduce deictic (pointing) or iconic (symbolic) gestures: a pen-based input for drawing gestures and an online training mode that records user actions. While drawing gestures is efficient for entering static gestures, natural curves associated with human movements are difficult to draw and are best introduced based on user actions. A database of training gestures is maintained for recognition. Adding and viewing gestures is performed using a Gesture Recognition Interface (GRI) (Figure 3).

To add a new gesture, users can press the START button, perform the gesture and then press STOP. Pressing SAVE adds the new gesture to the database. An a priori error model for hand and face tracking measurements makes training via a single training sample possible. Removing the cumbersome process of generating many examples for each new gesture facilitates rapid retraining for new environments. In addition to recording and playback capabilities, the GRI provides the user with configurable error boundaries that facilitate the teaching of gestures to new users.

New users are taught gestures with large error boundaries that are made smaller as a gesture is learned. Eventually, the perceived gesture is mapped to a natural language utterance, so that the gesture can be used in isolation or in conjunction with a spoken utterance. The GRI also provides a mapping between gestures and robotic actions. This mapping capability allows gesture-based robotic control.

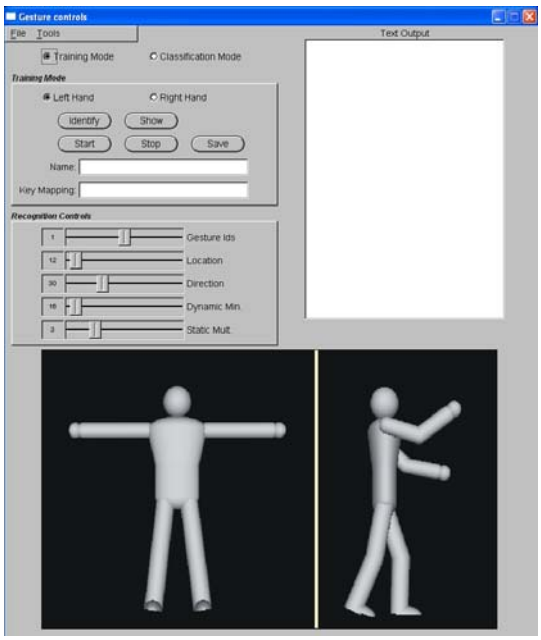

Figure 3. Gesture Recognition Interface

4.4 Recognition

Gesture recognition is performed by solving the Viterbi equation for each training sequence, normalizing the output based on the length of the sequence and comparing the result against a minimum recognition probability. If two gestures exceed the minimum recognition probability, the gesture with greater

probability is returned. Distance traversed by a gesture, rather than time, is used to determine the number of states necessary for two gestures to be compared. A motion record equal to the longest gesture is stored. Each candidate gesture is compared against the most recent user motion of the same length.

4.5 Pointing Gestures

While using gestures to disambiguate speech has a long history, going back to [4], we are focusing on using 3D visual tracking in a dynamic environment and investigating user-originated spatial references. In this environment, pointing gestures are recognized based on the orientation of a user's right arm. A 3D vector is formed between the elbow and forearm to determine pointing directions as depicted in Figure 4.

To transfer the point of reference from the user to the robot, the system translates the origin of the pointing gesture from the user's elbow to the robot. Pointing gestures are recorded based on speech or in combination with other gestures.

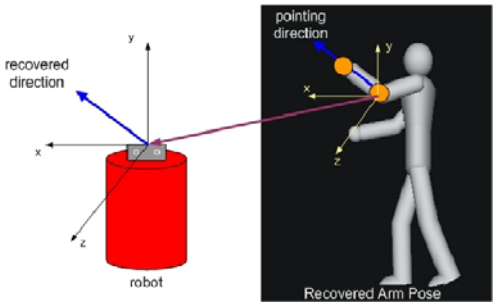

Figure 4. Pointing gesture

4.6 Stereo Tracking and Identification

The stereo tracking system is responsible for identifying and tracking the user's face and hands, extracting 3D coordinates for gesture recognition. Although some others, e.g. Pacquin and Cohen [20], utilize 2D motion tracking to provide motion control for mobile robots, we use 3D tracking. For us, depth information from the stereo camera provides accurate foreground segmentation and facilitates robustness in the presence of multiple people. In addition to detecting and tracking people, the stereo vision system is used for object identification.

4.7 Person Detection

In the person-detection phase, the system looks for frontal views of faces in the scene. Once a face is found, skin color is learned by creating a color histogram of the ellipse enclosed in the detected face region. We use the HSV color model for the histogram, keeping only hue and saturation and ignore the brightness component [25]. For face detection, we use an algorithm based on the boosted cascade object detection algorithm [15], implemented in the OpenCV open source computer vision library [5].

To prevent false positives due to background distractions, we include depth information to segment foreground pixels from background pixels. Thus, only pixels that are within a certain distance from the camera are used as input to the face detector.

4.8 Tracking

Once the system is initialized with a face location and skin color histogram, it tracks the skin-colored regions corresponding to the face and hands across frames. We assume that the three regions of interest are all of the same skin color, and we track those regions, extending the work of [1].

The algorithm we use represents hypotheses by ellipses and attempts to match hypotheses to observed skin-colored regions in each frame. When a skin-colored region is not explained, a new hypothesis is created. Similarly, when a hypothesis is no longer explained by any observed points, it is removed from the list of tracked hypotheses. The algorithm deals with cases in which one region explains two hypotheses, such as when two skin-colored arms touch or intersect. However, it does not handle the case in which two skin-colored regions explain one hypothesis, such as when an arm's skin-colored region is split by a watch into two skin-colored regions. We extended the algorithm by allowing blobs to be merged if they satisfy some criterion (for example, if they are close enough to each other), and by accepting blob pixels for the tracker only if they are within the sphere of a certain radius from the face of the currently tracked person.

When a face is first detected, it is assigned a hypothesis. At each frame, if any of the hands have not been found or were lost in previous frames, newly detected skin-colored regions are classified as either left or right arms using simple heuristics based on the relative position of the blobs to the face and to each other. When the head track is lost, the system reenters the detection phase.

In cases where the whole forearm is visible, such as when a person is wearing a short-sleeved shirt, the ends of the ellipse representing a tracked arm are classified as either 'hand' or 'elbow'. For this classification, we use simple heuristics based on the distance of the endpoints from the head.

4.9 Object Identification

To provide the system with the capability of identifying objects, such as a soda can in our scenario, we use an appearance- and scale-based object identification algorithm. Object identification is performed based on stereo camera input where both distance and color information is available. The process of performing object identification requires two distinct phases: training and recognition. In the training phase, scene regions are selected and used for generating models. In the recognition phase, modeled objects are located in the environment.

To generate a model, users perform online training. Users select an image region for later recognition by dragging a mouse around a region of interest. Upon selection, a hue histogram and size are recorded and stored.

Object recognition is based on the color and size of a region. Scenes are segmented, and regions with similar hue to the target object are selected. The selected regions are then utilized to seed a CAMSHIFT algorithm [6] that adaptively selects regions based on color properties. Scale information is utilized to evaluate the region size compared to the training sample. To scale each region, the candidate region's depth is estimated using depth information from the stereo camera. If more than one region exhibits similar colors and size to the training region, the user is queried for additional information to disambiguate.

4.10 Omnidirectional Person Tracking

To detect and track persons in its nearby surroundings, the robot uses an omnidirectional camera mounted on top of the robot (see

Figure 2). At every frame, the detector searches for persons and initializes a new track when a person is detected whose bounding box does not overlap with existing tracks. A particle filter [12] then locates each person in translation and scale in subsequent frames using color-based features.

4.11 Initialization

To initialize each track, we use the cascade of boosted Haar-like features described in [15] and implemented in OpenCV [5]. First, we use the full body detector of [14] to find regions that resemble upright humans. To reduce the occurrence of false positives, we search in the upper portion of the detected body region for faces using a detector trained on profile views of faces. Once both body and face are detected, the tracker learns the person's appearance model and instantiates a new track.

A person's appearance model is learned by calculating the discriminative color features described in [8]. To obtain the most discriminative feature, we obtain histograms of the person and non-person regions in the detected window and surrounding neighborhood, and calculate the likelihood of a color belonging to the foreground using the two histograms. To avoid including background pixels in the person region, we do not use all pixels in the person's bounding box. Instead, we use a figure-ground expectation sampling (ES) technique [28] to segment the person from the background region (Figure 5), and calculate the most discriminative features between the person and background. To encode spatial information, we subdivide the person region into a rectangular grid, and obtain the color distribution of each sub-region using the most discriminative color features.

Figure 5. (top) 360° panoramic view obtained from omnidirectional camera. Small interior box around person's face shows face detector result. Next larger, interior box encloses segmented person, and outer box shows person detector result. (bottom) Person template is segmented into background and foreground regions.

4.12 Tracking

Once we obtain the appearance model of a person, we use it to track the person across frames. Exhaustively searching for the best match in translation and scale is computationally prohibitive, even for small neighborhoods. Thus, we use a particle filtering framework to obtain the optimum scale and translation. Particle filtering takes into account previous positions by assuming some dynamics that provide a transition probability and evaluates positions in scale and translation using the observation likelihood. The observation likelihood can be estimated by summing the probability that each pixel in the target region does not belong to the model and by using the exponential function, as in [27], to obtain a probability estimate. The observation likelihood is computed once for each of the samples, so tracking becomes much more computationally feasible. Although our current implementation uses only color for representing a person's appearance (which can sometimes fail in the presence of similarly colored regions), the algorithm can be easily extended by incorporating other cues in the evaluation of a target region. One of many examples is the use of color and edge information in the evaluation of the observation likelihood [27].

5. ACOUSTICS

For a robot to know it is being spoken to, who is speaking to it, and what the relative locations of speaker(s) and robot are, the interface requires auditory capabilities beyond speech recognition. With this goal in mind, we have added an auditory component to provide the robot with certain auditory skills useful in disambiguating information from other components.

5.1 Robot Audition

In order to gather acoustic data, we employ an array of 4 AT831b lavalier microphones mounted on top of the robot (see Figure 2). These microphones are each connected to battery-powered preamps mounted inside the robot body and then to an 8-Channel PCMCIA data acquisition board.

Using this audio equipment, the robot has two auditory tasks to perform, speech detection and sound localization: (1) to detect the presence of speech sounds in the environment, and (2) to localize short speech utterances in the vicinity of the robot. We turn now to a discussion of these tasks.

5.2 Detecting Speech

Before a robot can localize speech sounds, it first needs to detect that they are present in the environment. For this task, we calculate the first two mel-cepstrum coefficients [22] for each microphone in the array. Each coefficient is averaged across all microphones, and then compared to an environment-dependent threshold. Although relatively simple and prone to errors when classifying a single sound sample, the speech detection system works well over time to augment vision sensors tracking humans in the environment.

5.3 Localizing a Human Speaker

Once speech has been identified in the environment, the robot may also need to know the location or origin of the speech sounds. If there is only one person in the room, then a vision system can help identify the most likely source, but if there are multiple people, or if there is no vision system available, then audition can help the robot identify who is speaking. The algorithm we use for this speech localization task is spatial likelihoods [18].

Spatial likelihoods are based on the principle of time difference on arrival. As the speed of sound is finite, and the microphones are physically separated in space, the signal received by each microphone due to a single source is offset by some measurable time. If the value of this time difference between the two received signals can be determined, the possible positions of the sound source is constrained to all positions in the room whose

geometrical position relative to the array corresponds to a measured time difference. Spatial likelihoods are calculated using a maximum likelihood method that utilizes time differences to estimate the likelihood associated with every possible location in the room. Figure 6 shows the spatial likelihood output of a sample containing speech, plotted on a contour plot.

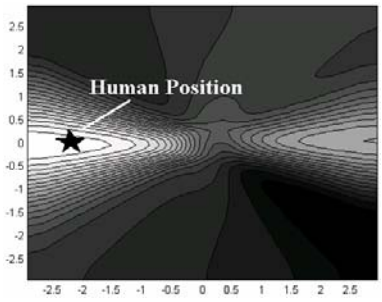

Figure 6. A spatial-likelihoods result for detecting human speech. This result demonstrates the common problem of a strong angular performance, but poor distance estimates.

In theory, given enough microphones in an array, it is possible to localize exactly on the sound source, using the principle of time differences. In practice, however, given the small distances between microphones in the on-robot array, as well as the levels of ambient noise and echoes from the environment, we have observed high amounts of error in the localization from one location. Error tends to be concentrated mostly along the axis stretching from the center of the array out through the sound source location. Thus, cross correlation results are generally better at estimating the angle to the sound source rather than the distance.

6. SPATIAL REASONING

A subcomponent of the Multi-Modal Reasoning Component (MMRC) is the Spatial Reasoning Component (SRC) (see Figure 1) where visual and linguistic data combine to provide spatial descriptions. We are in the process of extending our previous work [24] and integrating it into MMRC.

6.1 Spatial Reasoning

Since the scenario (Section 8) involves locative information and spatial interactions, we have extended our existing work in spatial reasoning to incorporate a 3D model of the world and to interact in more complex ways to disambiguate locative information. Previous versions of SRC computed robot goal locations to the LEFT, RIGHT, FRONT, BEHIND, and BETWEEN objects that were within one standard deviation of a human's location, or given the same instruction, of where the human user wished the robot to move [17]. Also, the previous SRC described the current environment using spatial referencing terms. Evidence Grid (EG) maps of range sensor data generated representations of objects in the robot's environment. While this was sufficient for robot navigation to various spatial localities, we wanted to create a more useful dialog, based on spatial references, for our robots.

To do so, we first employed Lowe's SIFT [16] algorithm with the stereo camera system to generate 3D SIFT point-cloud models [17]. SIFT uses these models to recognize and place objects in the environment through an affine transform. An interesting added benefit to this approach is that once we have the models created, only one camera is necessary to recognize and place the object in 3D space.

3D point clouds are generated (Figure 7 top) and placed into the robot's environment (Figure 7 bottom), projected onto virtual horizontal and vertical planes. These separate planes are fed individually into the SRC to determine the following: (1) the strongest relationship between the objects, (2) the selection of linguistic descriptions from a predetermined language dealing with locations on the horizontal and vertical planes.

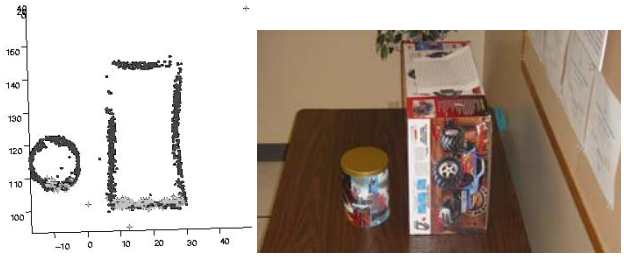

Figure 7. (top) The SIFT point-cloud model demonstrates recognition of SIFT keypoints (in light gray) and placement of the whole model in the robot's environment (in dark gray). (bottom) The scene consists of a large box and can as presented to the vision system. After projection onto vertical and horizontal planes, the language generated is "The box is to the right of the can and extends rearward."

For each of the generated descriptions, a primary description (e.g. FRONT, BEHIND, etc.) is associated with the fuzzy numbers *m1* and *degree1*, and a secondary description is associated with the fuzzy numbers *m2* and *degree2*. These numbers are generated for each of the planes, and we take the max(min(m*, degree*)) to decide whether to use primary or secondary linguistic descriptions from the horizontal or vertical planes.

This information is useful as it allows the robot to describe its environment in three dimensions, thereby enhancing the human-robot dialog. Further, it is a useful tie-in for related work in perspective-taking, allowing the robot to tell its collaborators what and where it sees objects in relationship to each other, to itself, and to the users, or from different perspectives if desired.

7. NATURAL LANGUAGE UNDERSTANDING

As we argued elsewhere [21], natural language provides an intuitively appropriate mode for interaction. We, therefore, incorporate speech to provide human users a natural way of interacting with the robot. Along with the other modules, natural language assists in disambiguating locations and objects. For example, users can verbally provide explanations and clarifications for sensed objects, coupled with the visual and auditory information available in the other components of the system.

7.1 Natural Language Interactions

To permit natural language interactions between humans and the robot, ViaVoice™ maps spoken utterances, such as "Go to the pillar over there," or "Over there", into strings. NAUTILUS, an

in-house natural language understanding system [21], robustly parses and regularizes the utterances for further processing. These representations are combined with gesture information from the vision component to form context predicates. Combined with the semantic information of the utterance, context predicates are used to determine if a gesture is needed, and if one is obtained, that it is appropriate. For example, if the human user tells the robot, "Go over there" but does not gesture to an appropriate location, the robot requests additional information: "Where?" If a gesture is not appropriate, the natural language system provides corrections, and once ambiguities are resolved, either verbally or gesturally, the utterance is mapped to robot commands. We also endeavor to distinguish extraneous gestures, such as scratching one's nose while speaking or indicating beats during a conversation, from truly disambiguating deictic or iconic gestures.

Interactions with the other sensor inputs and with SRC permit humans to describe the locations of objects, as well as to name them for the robot. But difficulties can arise. For example, upon receiving a deictic gesture and an object name (e.g. "That's a chair" +<gesture>), multiple internal representations may result in the SRC. Should this condition arise, the system will ask "Which one?" whereby the human can say "The object nearest to/furthest from you." The robot uses its perspective to determine which of the objects the human is referring to.

8. THE SCENARIO

We have developed a prototype scenario involving two humans and a B21r mobile robot to test the integration of our system. The scenario is a typical retrieval scenario in which two humans give directions to the robot to find a particular object, perhaps in a location deemed unsafe for humans. Further, the humans talk to each other; therefore, the robot needs a natural language understanding system that, with the help of the other perceptual systems discussed here, is robust enough to deal with the human exchange and perhaps even glean useful information from it. The scenario demonstrates the robot's ability to disambiguate information, not only for determining if it is being spoken to, but which objects to retrieve, and to discern which human is directing the robot in a human-robot team.

To assist the robot in identifying its target, a team member holds a replica of the target in front of the robot's video camera. The person states, "George, this (soda can) *[the object is held in front of the robot]* is your target." To disambiguate who instigated the utterance to the robot, we implement a speech disambiguation algorithm similar to [10]. The robot triangulates to focus on the sound source, the position of the speaker, and moves toward it. An omnidirectional camera assists in identifying the whereabouts of the suspect speaker. The robot is now in adequate range to track the speaker's gestures should any be used in subsequent interactions.

If the speaker did not mention the target when initiating interaction with the robot, the robot prompts for the target. As the speaker describes the target, the robot checks if the speaker made any relevant gestures that may assist in finding the goal location.

NAUTILUS, the Natural Language Understanding component, and the MMRC (see Figure 1) determine if help is needed during several stages of the interaction: (1) if enough information is provided in the spatial language (e.g. "The target is over there") to determine a goal; (2) if a gesture is provided toward an approachable location; and (3) if more information is needed to proceed (e.g., "Now turn left of the pillar").

If the linguistic and reasoning components identify a mismatch between a gesture and an utterance, such as an unintended hand motion, the robot requests additional assistance, shaping a question toward the ambiguity (e.g., "Can you point me in the right direction?"). Once the robot has received sufficient information from the speaker, it scans the environment for the target.

The human user(s) follows the robot, remaining in view of the omnidirectional camera (within a one meter radius), suggesting adjustments to the robot's current path at any time. The robot's gesture-tracking camera determines if it has reached the goal by comparing a viewed object to the target's representation stored in memory. The robot achieves the goal when it arrives at the target's location and reports to the speaker that it has found the target.

9. CONCLUSION

Using the scenario involving two humans and a robot that must find an object, we can evaluate how people interact with a robot capable of handling multiple simultaneous feedback from the environment. By combining the various modalities and integrating them in a human-robot interface, we are attempting to leverage the existing strengths of the various modules, and to overcome ambiguities that might arise from their outputs. For example, information from our acoustic module to recognize and further localize human speakers can offset limitations in our vision system. Also, ambiguities that might arise in visual and acoustic signals, when object detection is not robust enough to disambiguate various objects, can be offset by using the natural language component to clarify them. Furthermore, the dialog component directs the discourse along appropriate channels of interaction, noting visual and auditory cues for determining who the correct speaker is and to whom the utterance is intended. By integrating the various vision, auditory, and linguistic modules, we hope to develop a human-robot interface that will allow natural interaction between collaborating humans and robots to achieve a task.

In this report, we have discussed visual, auditory, and natural language components to be integrated in a human-robot interface. We have discussed the components of the interface, noting their individual functionalities as free-standing modules. However, in performing a cooperative human-robot task, ambiguities are bound to arise. Knowing the locations of speakers, who is being spoken to, what topic is being addressed, or what object is being referred to are all problems for completing a task and pose problems for each of the various modules discussed here. The scenario which we have chosen, retrieving a soda can in a laboratory environment, is filled with these and similar problems. In future, we plan to integrate the various components discussed in this report and to perform a user study using the scenario outlined to test how well this integration overcomes the various types of disambiguation mentioned here.

10. ACKNOWLEDGEMENTS

Support for this work was provided by the Office of Naval Research, work unit numbers N0001406WX20001 and N0001406WR20156. Registered trademarks are the property of their respective holders.

11. REFERENCES

[1] Argyros, A. A., and Lourakis, M.I.A. Real-time tracking of multiple skin-colored objects with a possibly moving camera. In *European Conf. on Computer Vision (ECCV 2004)*, vol. 3, 2004, 368-379.

[2] Artières, T., Marchand, J.-M., Gallinari, P., and Dorizzi, B. Multimodal segmental models for on-line handwriting recognition. In *Intl. Conf. On Pattern Recognition*, 2000, 2247–2250.

[3] Bischoff, R. and Graefe, V. Design principles for dependable robotic assistants. In *Intl. Journal on Humanoid Robotics*, vol. 1, no. 1, (March 2004), 95-125.

[4] Bolt, R. A. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th Annual Conf. on Computer Graphics and Interactive Techniques*, (July 1980), 262-270.

[5] Bradski, G., Kaehler, A., and Pisarevsky, V. Learning-based computer vision with Intel's Open Source computer vision library. In *Intel Technology Journal*, (May 2005).

[6] Bradski, G.R. *Computer Vision Face Tracking for Use in a Perceptual User Interface*, Intel Technology Journal, 1998.

[7] Brooks, A.G., and Breazeal, C. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 2006 ACM Conference on Human-Robot Interaction*, (March 2006), 297-304.

[8] Collins, R.T., Liu, Y., and Leordeanu, M. Online selection of discriminative tracking features. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, (October 2005), 1631-1643.

[9] Erman, L.D., Hayes-Roth, F., Lesser, V.R., and Reddy, D. The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. In *Computing Surveys*, vol. 12, no. 2, (June 1980), 213-253.

[10] Harris, T.K., Banerjee, S., Rudnicky, A., Sison, J., Bodine, K., and Black, A. A research platform for multi-agent dialogue dynamics. In *Proceedings of the IEEE Intl. Workshop on Robotics and Human Interactive Communications*, (September) 2004, 497-502.

[11] Iba, S., Weghe, M.V., Paredis, C., and Khosla, P. An architecture for gesture based control of mobile robots. In *Proceedings of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS'99)*, vol. 2, (October 1999), 851–857.

[12] Isard, M., and Blake, A. Condensation—conditional density propagation for visual tracking. In *Intl. Journal of Computer Vision*, (August 1998), vol. 29, no. 1, 5–28.

[13] Kim, S., Smyth, P., and Luther, S. Modeling waveform shapes with random effects segmental Hidden Markov models. In *AUAI '04: Proceedings of the 20th Conf. on Uncertainty in Artificial Intelligence*, 2004, 309–316.

[14] Kruppa, H., Castrillon-Santana, M., and Schiele, B. Fast and robust face finding via local context. In *Joint IEEE Intl. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, (October 2003), 157-164.

[15] Lienhart, L. and Maydt, J. An extended set of Haar-like features for rapid object detection. In *Intl. Conf. on Image Processing*, vol.1, 2002, 900-903.

[16] Lowe, D.G. Object recognition from local scale invariant features. In *Proceedings of the Seventh Intl. Conf. On Computer Vision (ICCV'99)*, (September 1999), 1150-1157.

[17] Luke, R.H., Blisard, S.N., Keller, J.M., Skubic, M. Linguistic spatial relations of three dimensional scenes using SIFT keypoints. In *IEEE Intl. Workshop on Robot and Human Interactive Communication: RO-MAN 2005*, (August 2005), 704-709.

[18] Mungamuru, B., and Aarabi, P. Enhanced sound localization. In *IEEE Trans. on Systems, Man, and Cybernetics Part B*, vol. 34, no. 3, (June 2004), 1526-1540.

[19] Ostendorf, M., Digalakis, V., and Kimball, O. From HMMs to segment models: a unified view of stochastic modeling for speech recognition. *In IEEE Trans. on Acoustics*, vol. 4, 1996, 360–378.

[20] Paquin, V., and Cohen, P. A vision-based gestural guidance interface for mobile robotic platforms. In *Computer Vision in Human-Computer Interaction*: *ECCV Workshop in HCI Proceedings*, 2004, 39-47.

[21] Perzanowski, D., Schultz, A., Adams, W., Bugajska, M., Marsh, E., Trafton, G., Brock, D., Skubic, M., and Abramson, M. Communicating with teams of cooperative robots. In *Multi-Robot Systems: From Swarms to Intelligent Automata*. Kluwer: The Netherlands, 2002, 185-193.

[22] Quatieri, T. *Discrete Time Speech Signal Processing*, Pearson Education, Inc.: Dehli, India, 2002.

[23] Sidner, C.L., Kidd, C.D., Lee, C.H., and Lesh, N., Where to look: A study of human robot engagement. In *ACM International Conference on Intelligent User Interfaces*, (January 1994), 78-84.

[24] Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams W., Bugajska, M. Spatial language for human-robot dialogs. In *IEEE Trans. on Systems, Man, and Cybernetics, Special Issue on Human-Robot Interaction*, (May 2004), vol. 34, no. 2, 154-167.

[25] Sobottka, K., and Pitas, I. Face localization and facial feature extraction based on shape and color information. In *Proc. of the Intl.Conf. on Image Processing*, (September 1996), 483-486.

[26] Starner, T, and Pentland, A. Visual recognition of American Sign Language using Hidden Markov Models. In *Intl. Workshop on Automatic Face and Gesture Recognition*, 1995, 189–194.

[27] Yang, C., Duraiswami, R., and Davis, L.S. Fast multiple object tracking via a hierarchical particle filter. In *IEEE Intl. Conf. On Computer Vision*, vol. 1, 2005, 212-219.

[28] Zhao, L., and Davis, L.S. Iterative figure-ground discrimination. In *Intl. Conf. on Pattern Recognition (ICPR)*, 2005, 67-70.