Joint Navigation in Commander/Robot Teams: Dialog & Task Performance When Vision is Bandwidth-Limited

Douglas Summers-Stay
Army Research Laboratory
douglas.a.summers-stay.civ

Taylor Cassidy
IBM Research
Army Research Laboratory
taylor.cassidy.ctr@mail.mil

Clare R. Voss
Army Research Laboratory
clare.r.voss.civ@mail.mil

Abstract

The prospect of human commanders teaming with mobile robots "smart enough" to undertake joint exploratory tasks—especially tasks that neither commander nor robot could perform alone—requires novel methods of preparing and testing human-robot teams for these ventures prior to real-time operations. In this paper, we report work-in-progress that maintains face validity of selected configurations of resources and people, as would be available in emergency circumstances. More specifically, from an off-site post, we ask human commanders (C) to perform an exploratory task in collaboration with a remotely located human robot-navigator (Rn) who controls the navigation of, but cannot see the physical robot (R). We impose network bandwidth restrictions in two mission scenarios comparable to real circumstances by varying the availability of sensor, image, and video signals to Rn, in effect limiting the human Rn to function as an automation stand-in. To better understand the capabilities and language required in such configurations, we constructed multi-modal corpora of time-synced dialog, video, and LIDAR files recorded during task sessions. We can now examine commander/robot dialogs while replaying what C and Rn saw, to assess their task performance under these varied conditions.

1 Introduction

Our research addresses a paradoxical situation in developing a robot capable of teaming with humans. To know what capabilities such a robot needs, we seek to determine how a human commander would interact — choice of vocabulary and sentence types, expected capabilities and world knowledge, resources used to accomplish tasks efficiently, etc. But without such a robot to interact with, we cannot know how a commander would behave. The prospect of human commanders teaming with mobile robots that are "smart enough" to undertake joint exploratory tasks requires novel methods of preparing and testing actual human-robot teams for these ventures, in advance of actual real-time operations. Furthermore, given the need for human/robot teams during emergencies (such as Japan's tsunami/Fukishima disaster), we are interested in particular in the feasibility of commander/robot shared tasks that include NL communication specifically for network contexts when bandwidth is limited by emergencies. Here we ask, how can multimodal data, as collected and processed by robots, and the robots themselves contribute real-time alerts and responses to human commanders over geographically-distributed networks?

The first phase of our approach is to introduce a human stand-in who navigates the robot, posing as an intelligent control system. At this stage, following our prior work (Voss et al., 2014), we seek to determine how the commander communicates to accomplish different tasks with the robot, while we limit the information made available in passing from the robot's sensors and camera to the commander by way of the stand-in. In future phases, we will progressively automate away this actor's role, replacing the audio that the stand-in hears with what is "understood" by automatic natural language semantic interpretation within a dialog manager, and replacing the joystick that it uses to navigate as the robot

Proceedings of the 25th International Conference on Computational Linguistics, pages 9–16,
Dublin, Ireland, August 23-29 2014.

with "actions" as automatically generated from micro-controller commands produced by transformation of semantic commands.

In this paper, we report *work-in-progress* that maintains face validity of selected configurations of resources and people, as would be available in emergency circumstances. From an off-site post, we ask human commanders (C) to perform an exploratory task in collaboration with a remotely located human robot-navigator (Rn) who actually controls the navigation of, but cannot see, the physical robot (R). We restrict the information Rn receives from R by imposing network bandwidth restrictions comparable to real circumstances which limit what Rn is able to communicate to C. We then examine the commander/robot dialogs and task performance under these varied conditions.

To better understand the capabilities and language required in such configurations, we constructed multi-modal corpora of time-synced dialog, video, and LIDAR files recorded during task sessions. We can now examine commander/robot dialogs while replaying what C and Rn saw, to identify the impact of varying the shared visual information on discourse, and to assess task performance under these varied conditions. We hypothesized that more explicit, mututally available information (visual or verbal) between participants would yield better understanding with more common ground, leading to more task success. We also hypothesized that exploration in a more complex physical environment would lead both to more dialog, as needed in resolving references to more locations, and also then on occasion, to less overall task success. We have found in preliminary analyses that, with more explicit *visual* information, some Cs reduce their level of communication, with fewer requests for images from Rn. In one such case, this led to the Rn getting lost. We also noticed that some Cs increased their level of *verbal* communication, requesting far more still images from the robot when Rn could not itself see the robot's images (as opposed to when Rn had access to sent images). Taken together, these observations suggest—contrary to our hypothesis that more information is better, especially in a complex environment—that there may be a "teeter totter" effect in the communication between C and Rn as visual information varies. When Rn has access to more of the robot's visual information, C communicates less with Rn, possibly assuming more shared information than is correct. Whereas when Rn is able to see less, C communicates more with Rn, possibly compensating for the lack of certainty Rn expresses.

2 Related Work

For human-robot communication in joint exploration tasks, we wish to understand two issues. The first is "scene to text": when exploring new locations, how do people talk about what they see, and how does that inform how they want robot team members to communicate about what they "see" while exploring? The second is "text to scene": given natural language instructions, how do people move about in new locations, and how does that impact their expectations of robot navigation? These issues span both generation and understanding of spatial language. There exists a large literature on spatial language, starting several decades ago (Talmy, 1983; Anderson et al., 1991; Gurney et al., 1996; Bloom et al., 1996; Olivier and Gapp, 1998) *inter alia*. This work yielded linguistic insights into the underlying structure of spatial expressions, that has led more recently to annotation efforts like SpaceML (Morarescu, 2006) and spatial role labeling (Kordjamshidi et al., 2010). These results, theoretical and computational, have been incorporated into NLP research, such as spoken dialog systems (Meena et al., 2014).

For "scene to text" processing, starting from a robot's perception of the scene or environment, exploiting even known dependencies among objects (spatial relations, relative motion, etc.) is a central problem in computer vision research. In the current state of robotics, the perceived world (a.k.a. semantic perception) derived from data collected by the robot is limited by what is available within its immediate sensor and video reach (Hebert et al., 2012). Within computational linguistic research, (Feng and Lapata, 2013) have tackled going from news images to text, leveraging the news story content as contextual knowledge, and automatically generating captions describing the image content as relevant for the story. For "text to scene" processing, a robot "understanding" a commander's language entails going beyond linguistic semantic interpretation down to the the robot controller level, as in, for example, Kress-Gazit et al. (2008). Within computational linguistics, Srihari and Burhans (1994) tackled going from text to images, exploiting the conventions and spatial language in news caption to identify people

by their relative positions in accompanying images. More recently Coyne et al. (2011) presented work for text-to-graphics generation, grounding conceptual knowledge in relational semantic encoding of lexical meanings from FrameNet. These one-way, directional approaches provide strong evidence that text and image modalities can each inform the processing of the other, and that, with concurrent audio and video streaming data, the alignment of time-stamped files across the two data modalities should also yield additional benefits in shared structural analyses and disambiguating references.[1]

3 Approach

In previous work, we had teams search a series of buildings, where all information from the Rn to C was strictly limited to text (Voss et al., 2014). While verbal descriptions of scenery were successfully elicited during exploratory missions, the communication was painfully slow and this scenario yielded unrealistic results from our stand-in: we would not expect a robot to generate the complex verbal descriptions we collected. Furthermore we also learned that our equipment could be adjusted for transmission of LIDAR map data and video stream from the robot to Rn and then to C. In this second study, we allowed individual map and image updates to be sent to C, but only on request. This work provides more explicitly shared knowledge between C and Rn, with its form and quantity more realistically varied and dynamic.

Equipment: We used an iRobot PackBot equipped with a forward-facing Kinect camera and a Hokuyo LIDAR sensor.[2] We use GPS and inertial sensors for Simultaneous Localization and Mapping (SLAM). Each participant had their own laptop with speakers and separate push-to-talk microphones. For navigating the robot, the Rn pushed a joystick on an X-box controller that was held. Additionally for transmitting visual information available from the robot during the missions, the Rn pushed separate buttons on the same controller to transfer image and map data to C, but only at C's request.

Pre-pilot Design: We conducted training sessions at one location and test sessions at a second location. A top down view of these sites is provided in Figure 1. We asked participants to perform distinct missions (task conditions) in the training and test sessions, with different levels of visual information available to Rn (vision conditions). Due to wireless networking timeouts and hardware integration difficulties, a number of sessions ended prematurely. Descriptive statistics for the sessions are in Table 1.

	Vision Condition		
Task Condition - quality of dataset	LIDAR only	LIDAR + Image last-sent	Video + LIDAR + Image last-sent
Mission 1 - complete	–	–	6 sessions (77 min)
Mission 1 - partial	–	–	1 session (1 min)
Mission 2 - complete	4 sessions (57 min)	2 sessions (28 min)	2 sessions (18 min)
Mission 2 - partial	11 sessions (15 min)	3 sessions (3 min)	–

Table 1: Total #sessions attempted by configuration (different task & vision conditions)

Vision Conditions: The Rn always saw (i) a continuously updated LIDAR map built up progressively from the robot's sensors as the Rn navigated the robot using the joystick on an X-box controller. On the map during training, the Rn could also see (ii) an avatar shape for the robot's location based on GPS and (iii) an arrow for the robot's facing direction generated by its internal components (updated intermittently by GPS). However the GPS signal was also sporadic during these sessions, causing confusion for Rn navigating the robot. As a result, during test sessions, we turned off the GPS to avoid this source of confusion, mirroring what actual operators do in this scenario. During test sessions, the Rn only saw (iii) the arrow, again within (i) the streamed LIDAR map. Beyond these Rn screen specifics, we ran three conditions controlling for the visual information that the C and Rn could see. During mission 1 (training), Rn was given "full" view of the streaming video, any specific images sent to C at C's request, and the map with arrow and avatar. During mission 2 (test) in one "partially blinded" condition, the Rn

[1] We are also eager to learn more from recent research examining streaming multimodal data for how and where the composition of natural language and the composition of visual scenes can inform one another (Barbu et al., 2012) and (Barbu et al., 2013).

[2] iRobot, PackBot, Kinect, and Hokuyo are all trademarks or registered trademarks.

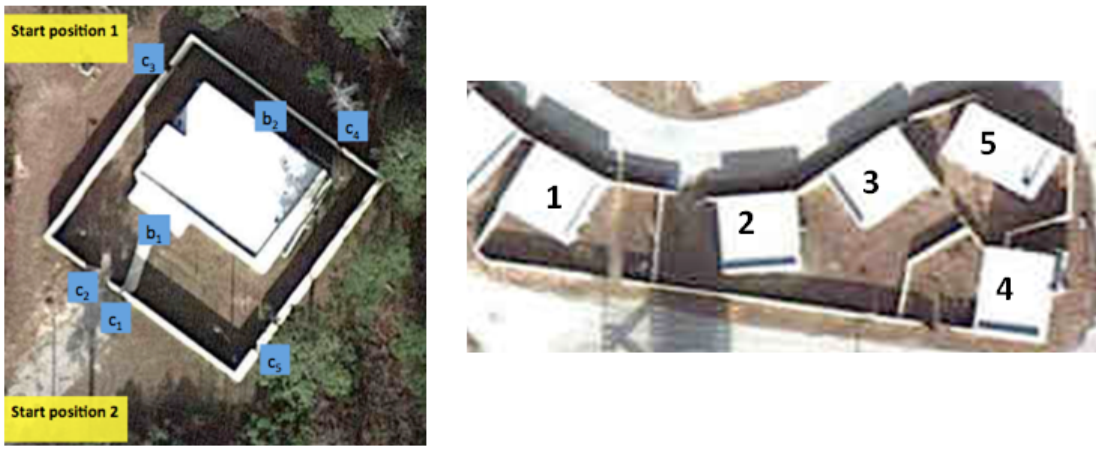

Figure 1: On left side: view of Mission 1 courtyard and building, with doorways marked. On right side: view of Mission 2 courtyards and buildings.

saw no video, but could see the specific images he sent to C as well as the map with arrow, and in the other even "more blinded" condition, Rn saw only the map with arrow. By contrast, the C only ever saw what the Rn sent (by pushing buttons) as snapshots at C's request. During all conditions — independent of what was presented to Rn ("full' view in mission 1, partially blinded or more-blinded in mission 2) — C could always request an updated snapshot image from the video feed or an updated snapshot map from the LIDAR feed or both. As a result, Rn's view was "pushed" and current from the robot's streaming data, whereas the C's view had to be "pulled," requiring C to ask for more snapshots. Note that in Rn's more-blinded condition, images were passed to C with Rn's button push, but Rn could not see the images.

Mission 1: Enter courtyard and building via safe doorways. We hypothesized a robot with the ability to carry on limited conversation regarding simple navigation and exploration, but without sufficient vision capabilities to analyze more subtle clues about whether a doorway was safe to enter. We designed the task to simulate a low-bandwidth condition where constant transmission of the map and video information is impossible. The robot was placed in one of two undisclosed positions outside the courtyard surrounding a building. All sessions adopted the L+I+V vision condition. The site for this mission was a

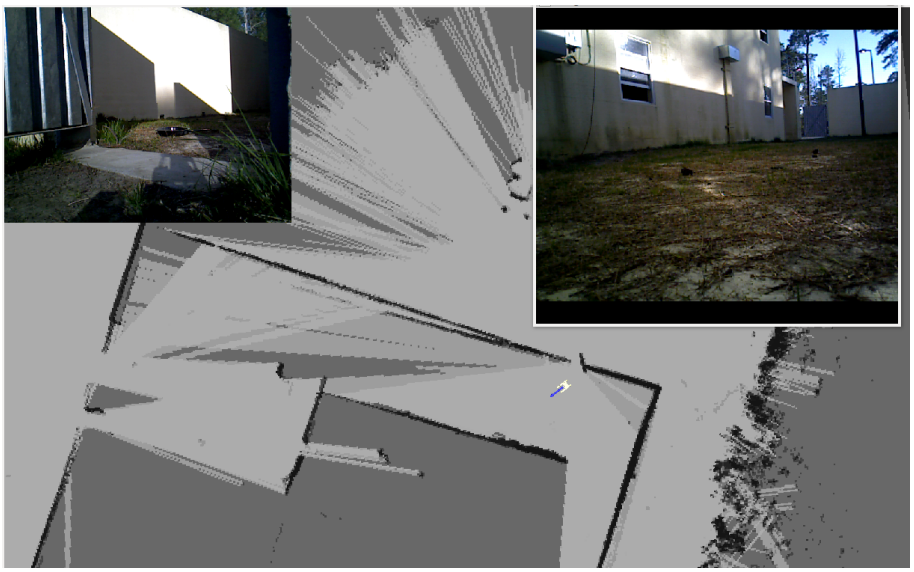

Figure 2: Robot-navigator's screen during Mission 1: upper left is static Image (clip from video, most recently sent to Commander), upper right is video window, gray-scale background is LIDAR map

single rectangular building enclosed by a single rectangular courtyard. The site for mission 2 was more complex, consisting of 5 buildings in a complex series of interconnected courtyards (see Figure 1). There are five doorways into the courtyard and two doorways into the building. These doorways are marked as safe or unsafe in a way that C can recognize but Rn cannot (C is given a key to the meaning of objects placed just beyond open doorways as symbols). The participants are not informed about doorway location or safety status. Figure 2 shows Rn's screen during a mission 1 session. The grey-scale background is an overhead, 2D view of a 3D map being built on the fly by combining various sensor data, which contains a white robot avatar and blue arrow indicating its current pose. C's view is similar, but without video. Success on this task was gauged by whether the robot stayed safe in gaining entry to the house.

Mission 2: Find and classify all building doorways within a compound.

As noted above and shown in Figure 1, the location in this mission had a more complex layout. The robot's location within the compound was not disclosed to C nor Rn (no clues were provided), so that the C and Rn team would need to work hard to place the robot on the map. The team was tasked with thoroughly exploring the compound to capture images of each building doorway. In the LIDAR-only (L) condition, Rn sees only the grey-scale map, whereas in the LIDAR and image condition (L+I) Rn sees the most recently sent image as well as the grey-scale map (same screen layout as in Figure 2 but without video window in upper right). Success on this mission was gauged both by the number of doors (open or closed) that were identified and photographed and by whether the participants were lost at some stage in the exploration.

4 Observations and Preliminary Results

We recorded rich, multi-modal datasets including: dialogue between C and Rn, video, LIDAR 3D point clouds, scene classification output on video frames, and robot pose. The data is used to build up a 3D model of the scene, and automatically align RGB images to the model by mapping pixels to 3D regions. Examples of scene classification performance can be seen in Figure 3. The data for each run consists of a ROS bag file (Quigley et al., 2009) and two audio files.[3]

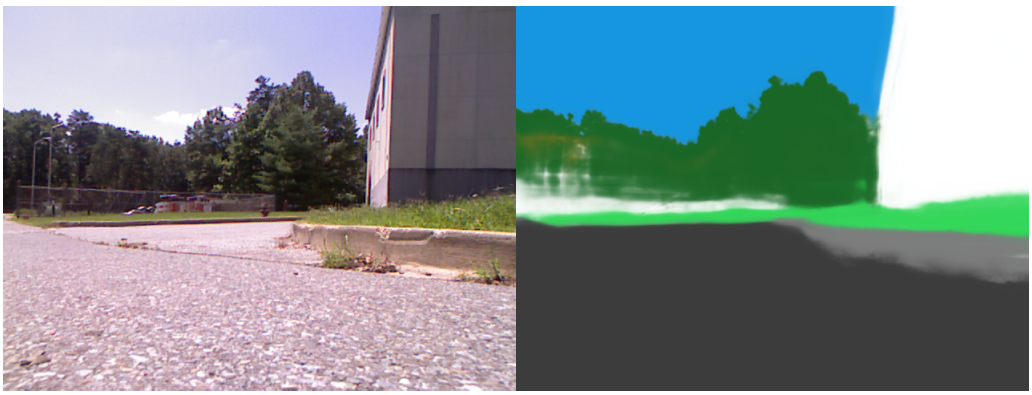

Figure 3: left: view from robot camera. right: automated scene classification. Mix of colors indicates probability of belonging to a particular class. Classes found in this scene include sky, foliage, building, grass, concrete, and asphalt. Performance degrades in lighting conditions unattested in training data.

4.1 Results from Session Path Analysis

Figure 4 shows an overhead 2D view of the final 3D map built using the SLAM module. An orange line depicts the robot's path from mission start to finish, with ordinal numbers indicating the robot's high level trajectory (the robot traveled from "start" to "1", then to the location marked by "2", etc., finally ending on the location marked by "15"). Doorways that were successfully captured in images sent to C are highlighted with a green solid-lined circle, whereas doorways that were passed by are indicated with

[3] A bag file stores nano-second accuracy timestamped, discrete data *messages*, such as an individual video frame, the fact that a joystick button was pressed, or the robot's current velocity.

Mission 1 Sessions (duration)	Vision Condition	Total # Images sent	# Images sent with (any) door	# Images sent with safe door	Task Success: Stayed **S**afe? Gained **E**ntry?
1 (21 min.)	L + I + V	0	0	0	S, E
2 (5 min.)	L + I + V	0	0	0	N, E
3 (17 min.)	L + I + V	3	3	2	S, E
4 (15 min.)	L + I + V	8	7	2	S, E
5 (13 min.)	L + I + V	12	7	4	S, E

Table 2: Mission 1 sessions: These training sessions provided the robot-navigators (Rn) with "full" real-time vision, i.e., their screens displayed all sensed data, as collected by the physical robot (R)

Mission 2 Sessions (duration)	Vision Conditions (**L**IDAR) (**I**mage)	Total #Images (sent w/o map)	Total #Maps (sent w/o img)	Total # Im & Map (sent one, then other)	# deictic refs by C, Rn	# refs to past by C, Rn	Task Success: # **D**oors id? Got **L**ost? **R**ecovered?
A (21 min.)	L map	27	7	5	13, 2	6, 3	9, n/a , n/a
B (20 min.)	L map + I	7	9	7	7, 2	7, 2	7, L, R

Table 3: Mission 2 per-session events: request and reference types, task success.

a dotted line. There is a point in the run depicted where Rn states that he is "lost", which is marked in the figure by a green dot at step 10.

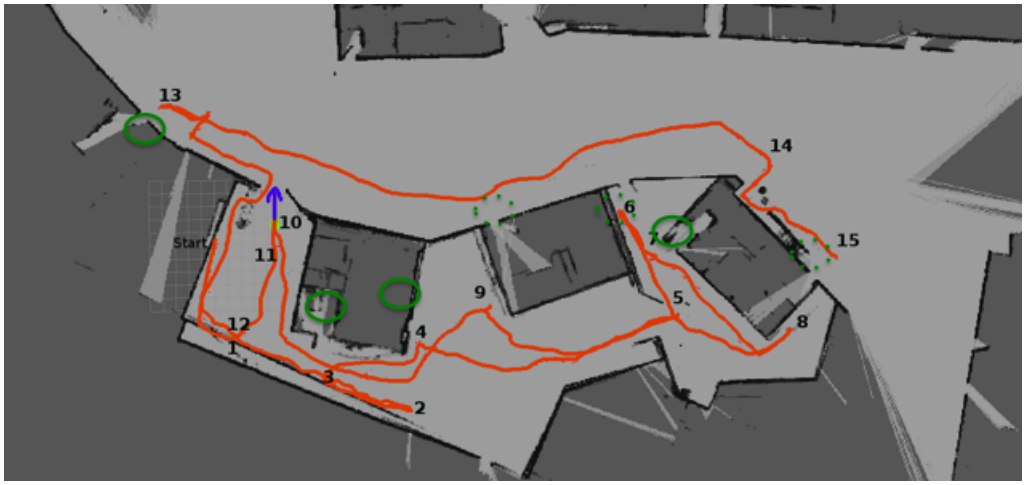

Figure 4: Robot path during Mission 2 session, doorways marked

4.2 Language Phenomena in Dialogs

Referring Expressions: There were few named environment features, necessitating the use of referring expressions. Participants often used pronouns ('behind it'), deictic expressions ('that wall'), and both definite and indefinite noun phrase descriptors ('a wall directly in front of you'). The frequency of referring expressions other than proper names highlights the need for a dialog manager to robustly handle human-robot dialog in our setting. In six mission 2 dialogs consisting of 6,593 words total, we annotated 1,593 referring expressions - 1,213 definite and 380 indefinite. The most common were first and second person singular pronouns (287 and 245), definite expressions of the form *the x* (265) and indefinite expressions of the form *a(n) x* (256). Most references are to things, either in the physical ('face the doorway') or software ('update your map') environment, though there are references to events as well ('do that again').

Lexical Ambiguity: The same objects were sometimes referred to as 'doors' or 'doorways,' although by a dictionary definition, those refer to somewhat different things. Based on context, the robot would need to be able to understand which sense was intended.

Spatial Relations: Since these were navigation and observation tasks, much of the discussion involved spatial language pertaining to object configurations and robot paths. There were references to distances and angles, both specific ('turn 15 degrees to your right') and vague ('turn around.') The robot was asked to 'follow the wall', 'go north', and to travel 'around,' 'behind,' and 'near' various objects.
Clarifications and Suggestions in Dialogs: When uncertain about the meaning of commands, Rn sometimes asked for clarification. At other times, Rn reminded C of its capabilities when appropriate: 'Would you like me to send you an updated map?'

4.3 The Role of Shared Visual Information

Participants were generally able to use both image and map data in conjunction with dialog to gain enough common ground to communicate about the environment and accomplish the tasks at hand. For example, after discussing environment features against the backdrop of an updated 2D map, we were often surprised at the extent to which C apparently kept track of R's location using dialog alone without further map updates, as evidenced by C's ability to correctly use Rn's egocentric frame of reference in verbal descriptions (recall that the robot avatar remained static on C's map between updates). In such cases C and R took advantage of mutually accessible visual information - their 2D maps were identical during discussion. The role of mutually accessible information for achieiving common ground is further supported by the fact that C requested significantly more images in the LIDAR-only condition, when Rn could not see those sent images (see Table 3). Although shared visual knowledge proved useful for resolving referring expressions, C and Rn rarely mentioned the media explicitly ('the building' vs 'the building in the image you sent me'). In this way, the transfer of visual information served to introduce entities into their discourse, but was taken for granted and not called out per se.

5 Ongoing Work

We have found in preliminary analyses that, with more explicit *visual* information, some Cs reduce their level of communication, with fewer requests for images from Rn. In one such case, this led to the Rn getting lost. We also noticed that some Cs increased their level of *verbal* communication, requesting far more still images from the robot when Rn could not itself see the robot's images (as opposed to when Rn had access to sent images). Taken together, these observations suggest—contrary to our hypotheses that more information is better, especially in a complex environment—that there may be a "teeter totter" effect in the communication between C and Rn as visual information varies. When Rn "sees as the robot" with access to more transmitted visual information, C communicates less with Rn, possibly assuming more shared information than is correct. Whereas when Rn "sees" less, C communicates more with Rn, possibly compensating for the lack of certainty Rn expresses. We plan to extend our analysis of how C and Rn communicate uncertainty, and look at how this topic is addressed in first aid and military manuals (US Dept. of the Army, 1993).

We are currently developing a framework to automate many of the tasks currently performed by Rn. Our studies and data collections so far are best understood in the context of the capabilities and limitations of the overall system we are in the process of building. A crucial gap to address is associating referring expressions with corresponding concrete spatial structures in the 3D map. Consider one sentence spoken by the commander in one of the dialogues: "When you get to the wall, turn left and drive along the wall until you reach either a corner or what you believe to be a door." To interpret this correctly, the robot must understand an entire set of points as a single object or part of an object, so it can recognize doors, walls, and corners in the combined vision and point-cloud. Moreover, it needs to plan a path that obeys the constraint "along the wall" and stops at some point which may be a door or a corner, that has not yet been observed. Thus, objects need to be represented independent of the observed world map.[4] At present, scene parsing techniques can analyze images and assign each pixel a probability of belonging to a particular object class (wall, stucco, road, etc.) allowing us to propagate these labels to corresponding points in the 3D model of the scene. In the future, we will use the 3D model to resolve visual ambiguities and attach labels to particular objects that persist from one video frame to the next.

[4]Resolving references to unvisited locations is a largely unexplored problem (Williams et al., 2013; Duvallet et al., 2013).

Acknowledgements

We thank members of the Asset Control and Behavior Branch at ARL for participation in our study and for continuing to provide the technical support that makes our work possible. The work of Taylor Cassidy was funded by IBM under the International Technology Alliance in Network & Information Sciences.

References

A. Anderson, M. Bader, E. Bard, E. Boyd, G.M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, C. Sotillo, H.S. Thompson, and R. Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366.

A. Barbu, A. Bridge, D. Coroian, S. J. Dickinson, S. Mussman, S. Narayanaswamy, D.l Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. W. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. 2012. Large-scale automatic labeling of video events with verbs based on event-participant interaction. *CoRR*, abs/1204.3616.

A. Barbu, S. Narayanaswamy, and J. Siskind. 2013. Saying what you're looking for: Linguistics meets video search. *CoRR*, abs/1309.5174.

P. Bloom, M. Peterson, L. Madel, and M. F. Garrett, editors. 1996. *Language and Space*. The MIT Press.

B. Coyne, D. Bauer, and O. Rambow. 2011. Vignet: Grounding language in graphics using frame semantics. In *ACL Workshop on Relational Models of Semantics (RELMS 2011)*.

F. Duvallet, T. Kollar, and A. Stentz. 2013. Imitation learning for natural language direction following through unknown environments. In *IEEE Intl. Conference on Robotics and Automation (ICRA)*, pages 1047–1053.

Y. Feng and M. Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:4:797–812.

J. Gurney, E. Klipple, and C. Voss. 1996. Talking about what we think we see: natural language processing for a real-time virtual environment. *IEEE International Joint Symposia on Intelligence and Systems*.

M. Hebert, J. A. Bagnell, M. Bajracharya, K. Daniilidis, L. H. Matthies, L. Mianzo, L. Navarro-Serment, J. Shi, and M. Wellfare. 2012. Semantic perception for ground robotics. In R. E. Karlsen; D. W. Gage; C. M. Shoemaker; G. R. Gerhart, editor, *SPIE Proceedings Vol. 8387: Unmanned Systems Technology XIV*.

P. Kordjamshidi, M. Van Otterlo, and Marie-Francine Moens. 2010. Spatial Role Labeling: Task Definition and Annotation Scheme. In *Proceedings of Language Resources and Evaluation Conference*.

H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas. 2008. Translating Structured English to Robot Controllers. *Advanced Robotics Special Issue on Selected Papers from IROS*, Vol. 22, No. 12:1343–1359.

R. Meena, J. Boye, G. Skantze, and J. Gustafson. 2014. Crowdsourcing street-level geographic information using a spoken dialogue system. In *Proceedings of SIGDIAL*. Association for Computational Linguistics.

P. C. Morarescu. 2006. Principles for annotating and reasoning with spatial information. In *LREC*.

P. Olivier and K-P. Gapp, editors. 1998. *Representation and Processing of Spatial Expressions*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

M. Quigley, K. Conley, B. Gerkey, J. Faust, T. B. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. 2009. ROS: an open-source robot operating system. In *ICRA Workshop on Open Source Software*.

R. K. Srihari and D. T. Burhans. 1994. Visual semantics: Extracting visual information from text accompanying pictures. In *Proc. Of Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pages 793–798.

L. Talmy. 1983. How Language Structures Space. In Jr. H. L. Pick and L. P. Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*, pages 225–282. Plenum Press, London.

US Dept. of the Army. 1993. *Physical fitness training: Field manual 3-25.26*. Washington, D.C.

C.R. Voss, T. Cassidy, and D. Summers-Stay. 2014. Collaborative Exploration in Human-Robot Teams: What's in Their Corpora of Dialog, Video, & LIDAR Messages? In *Proceedings of EACL Dialog in Motion Workshop*.

T. E. Williams, R. Cantrell, G. Briggs, P. W. Schermerhorn, and M. Scheutz. 2013. Grounding natural language references to unvisited and hypothetical locations. In *AAAI*.