Multimodal Interaction with an Autonomous Forklift

Andrew Correa, Matthew R. Walter, Luke Fletcher, Jim Glass, Seth Teller, Randall Davis

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA
{acorrea, mwalter, lukesf, glass, teller, davis}@csail.mit.edu

***Abstract*—We describe a multimodal framework for interacting with an autonomous robotic forklift. A key element enabling effective interaction is a wireless, handheld tablet with which a human supervisor can command the forklift using speech and sketch. Most current sketch interfaces treat the canvas as a blank slate. In contrast, our interface uses live and synthesized camera images from the forklift as a canvas, and augments them with object and obstacle information from the world. This connection enables users to "draw on the world," enabling a simpler set of sketched gestures. Our interface supports commands that include summoning the forklift and directing it to lift, transport, and place loads of palletized cargo. We describe an exploratory evaluation of the system designed to identify areas for detailed study.**

Our framework incorporates external signaling to interact with humans near the vehicle. The robot uses audible and visual annunciation to convey its current state and intended actions. The system also provides seamless autonomy handoff: any human can take control of the robot by entering its cabin, at which point the forklift can be operated manually until the human exits.

***Index Terms*—autonomous; interaction; tablet; forklift; robotic**

I. Introduction

One long-standing goal of research in human-robot interaction is achieving safe and effective command and control mechanisms for mobile robots. This goal becomes increasingly important as robots are deployed into human-occupied environments. We discuss our approach toward this goal in the context of robotic forklifts tasked with autonomously performing warehouse operations in unstructured outdoor environments. Motivated by the need to take military forklift operators out of harm's way, we have developed a system that enables humans to command and interact with a 2700 kg (1350 kg load capacity) forklift that autonomously picks up, transports, and places palletized cargo in response to high-level commands from a human supervisor. For the system to be effective in this domain:

- It must operate in existing facilities that have little or no special preparation.
- It must be usable by current and new warehouse personnel with minimal training.
- It must behave in a predictable way, so that its presence will be acceptable to humans.
- Its interface must enable a single human supervisor to command multiple robots simultaneously.

These requirements motivated the design choices we made in developing our interaction mechanisms. They call for entrusting ever more autonomy to the robot (i.e., not teleoperating it), ensuring that it is subservient to humans, and arranging for it to expose its world knowledge and intent to humans. To that end, we designed novel multimodal interfaces that allow humans to direct and interact with the forklift as it autonomously approaches, lifts, transports and places palletized cargo in an outdoor warehouse. Among these mechanisms is a context-aware tablet interface, shown in Fig. 1, through which humans use speech and stylus gestures to convey task-level commands to the robot.

Fig. 1. The supervisor commands the forklift to lift a pallet from a truck, by circling the pallet on the robot's-eye-view displayed on the tablet interface (inset). The tablet also interprets simple spoken commands to summon the forklift to various task areas within the warehouse.

While the problem of warehouse automation has been explored previously, existing systems [24] target the needs of long-term storage and distribution centers and require carefully prepared indoor environments devoid of people. These systems also impose tight constraints on, and assume perfect knowledge of, any objects to be manipulated. In contrast, our system must operate within the often temporary storage facilities that are typical of the military supply chain and material distribution centers for disaster relief. Such environments are relatively unprepared and unstructured, and are already occupied and used by people accustomed to working in close proximity to human-operated forklifts. Additionally, the cargo handled within these facilities exhibits highly variable geometry and appearance—consisting of anything from a carefully manufactured metal box to several wooden beams tied together by hand. Our goal is to introduce a capable robot into existing environments such that the robot operates safely, and humans quickly learn to command the robot and feel

978-1-4244-4893-7/10/$25.00 © 2010 IEEE

comfortable working near it. We believe that a key element in achieving these goals is the intuitiveness of the robot's interface and its display of its intentions.

Maintaining human safety is particularly challenging within our target environments, as they are populated with pedestrians, trucks, and other forklifts that may be either manned or (in the future) autonomous. These environments are also dynamic, since palletized cargo is constantly being stored, retrieved, and relocated within them. As such, our system must be designed to work safely and seamlessly in close proximity to other moving entities, and with minimal assumptions about environment structure.

Given the extensive uncertainty within these environments, there will be occasions when the vehicle will be unable to complete a task on its own. The system must be able to detect and gracefully recover from such situations. One important element of our recovery strategy is the ability of any human to approach the robot, enter its cabin, and directly operate its controls as though it were an ordinary, manned forklift. This automatic switchover to manual operation makes the robot subservient to people, in the sense that it will never struggle with a human operator for control over the forklift's operation.

The paper offers three contributions to the human-robot interaction literature:

- We demonstrate a system that allows the user to draw on a view of the world as seen by the robot, and that uses the robot's material context – its physical surroundings – to interpret what the user has drawn.
- We build on work in speech understanding to enable spoken commands in noisy, uncontrolled environments, including spontaneous warning utterances that successfully interrupt the robot's operation.
- We describe a set of aural and visual external displays that communicate the robot's intent in a way that seems to be intuitive and transparent. Moreover, the robot is designed to cede autonomy upon any human's approach. Both steps increase the likelihood that the robot will be accepted by people working around it.

II. Related Work

Earlier work in robot control interfaces has given rise to wireless tablet-based devices through which a user can control one or more robots. We describe relevant research in this area in terms of the situational awareness provided to the operator, the level of autonomy given to the robot, and the types of input actions made by the user to command the robot.

As do our efforts, Fong et al. [7] address the problem of designing an interface that can be used to control a ground robot on uneven terrain with minimal user training. Their PdaDriver system uses images from a user-selectable camera for situational awareness, and allows the supervisor to teleoperate the vehicle with stylus gestures using either a two-axis virtual joystick displayed on the screen or by specifying a desired trajectory segment by clicking waypoints on the image. Our interface also makes use of gestures drawn on images as a means of commanding the robot, though for tasks other than navigation. Another similarity lies in the extension to the collaborative control paradigm [6], which emphasizes the importance of interaction between the robot and operator to facilitate situational awareness. A fundamental difference, however, is that our approach explicitly avoids teleoperation in favor of a task-level interface; in principle, this enables a single human supervisor to command multiple robots simultaneously.

Similar to the PdaDriver system, Keskinpala et al. [14] describe a PDA-based touch-screen interface for mobile robot control through which the user teleoperates a ground robot. In addition to providing views from a vehicle-mounted camera, the interface allows the user to view raw LIDAR (laser-range scanner) and sonar returns, either projected on the camera image or on a synthesized overhead view of the robot. The latter view is intended to facilitate teleoperation within cluttered environments, where a forward-facing camera image would provide insufficient situational awareness. Similarly, our interface incorporates the robot's knowledge of its surroundings both as a means of improving the supervisor's spatial awareness, and a means of revealing the robot's interpretation of the supervisor's commands. Our approach is different, in that we render contextual knowledge at object level (e.g., pedestrian detections) as opposed to rendering raw sensor data, which subsequent user studies [13] have shown to add to the user's workload during teleoperation.

Skubic et al. [22] describe a framework in which a user prescribes the path and goal positions for a team of robots within a coarse, user-sketched environment map. Unlike our system, their interface supports only sketch-based interaction and supports only navigation commands. Perzanowski et al. [20] introduce a multimodal interface that, in addition to pen-based gestures, accommodates a limited subset of speech and hand gestures to issue navigation-related commands. Their framework allows the robot to clarify commands with the user to resolve ambiguity, but does not otherwise support the flow of information back to the user.

Sakamoto et al. [21] utilize gestures made on the world to command an indoor robot. Through a limited set of strokes, the user can give simple navigation directives by drawing on a bird's-eye view of the robot's environment displayed on a tablet computer. Originating from downward-facing cameras mounted to the ceiling, the interface requires significant environment preparation; this limits the vehicle's operating region to the cameras' field of view. In contrast, our system uses gestures drawn on a canvas corresponding to either the robot's view or a synthesized top-down view. In both cases, the views are generated based upon on-board sensing and do not limit the robot's operating environment. Other investigators have also shown the utility of enabling a teleoperator to switch between first-person and third-person views of the workspace [5].

Existing research related to multimodal robot interaction [11] uses a combination of vision and speech as input. Our approach is analogous as it combines the supervisor's eyes with speech [8] and sketch [3]. While many approaches to sketch recognition have been explored in previous work across multiple domains [18], [25] and using multiple modalities [1],

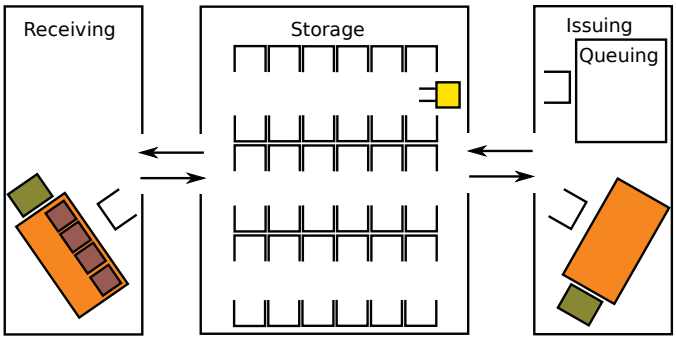


Fig. 2. A typical warehouse consists of a receiving area where loaded trucks arrive, a storage lot, and an issue area where customers take delivery of palletized cargo (occasionally after it is staged in a queueing area).

[12], we chose to design a multimodal system that uses speech and sketch as complementary, rather than merely mutually disambiguating, modes.

One important element of joint human-robot activity is the ability of individual agents to detect each others' cues. Such cues are important aids in interpreting intent or other internal state; for example, eye contact has been shown to play an important role in pedestrian safety [9]. In autonomous vehicles, such cues are often missing by default. Matsumaru et. al. [16] explore several methods for communicating intended future motion of a mobile robot, including the use of synthetic eyes. The eyes indicate future motion, but not perceived objects of interest in the world. Illuminated displays on the robot were shown to provide an effective method of informing people of intended robot speed and direction [15].

Our robot is not anthropomorphic, but we adopt the idea of generating externally evident cues in order to enable people near the robot to understand its current state, world model, and intentions.

III. System Overview

We begin by providing an overview of our system, a brief description of the warehouse environment, and a description of the robotic platform.

A. Conceptual Warehouse Environment

The general structure of the warehouse that we consider consists of three zones: "receiving," "storage," and "issuing," shown in Fig. 2. As the name suggests, the receiving zone corresponds to the region of the warehouse where trucks arrive with deliveries of palletized cargo. The forklift is tasked with unloading pallets from these trucks and transporting them to the storage region where they are deposited on the ground in one of many named storage bays. Customers await delivery of orders in the issuing zone. After the forklift conveys each pallet of cargo to this zone, it places it on the bed of a customer's truck. People periodically walk through and operate manned forklifts within all areas of the warehouse in order to perform tasks such as inventory control and packing or unpacking the contents of individual pallets.

B. The Robotic Platform

The robot [23] is a 2700 kg (1350 kg capacity) commercial forklift that we have modified to be drive-by-wire. The platform is equipped with laser range finders used to detect pallets, obstacles, and pedestrians. Four cameras, facing forward, left, right, and backward provide a view of the robot's surroundings to the interface while four microphones, one under each camera, are used for shout detection (see Fig. 9).

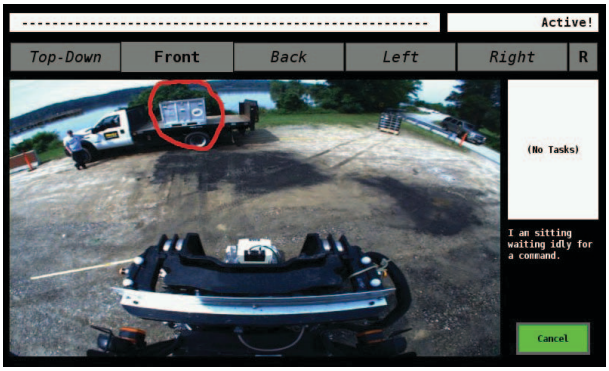
The forklift has three operating modes: autonomous, stand-by, and manual. In autonomous mode, the forklift performs unmanned pallet manipulation and navigation throughout the warehouse. Stand-by mode refers to a pause state in which the robot is stationary, awaiting further instruction from the human supervisor. The robot enters stand-by mode either after completing all assigned tasks, or after determining that it is unable to perform a pending task, for example, due to a perceived obstacle. In this "stuck" case the robot waits for an explicit "resume" signal from the supervisor, which may come after an intervention taking one of three forms. First, the supervisor may choose to modify the environment to make the current task feasible. Second, the supervisor may choose to enter the forklift cabin and operate it manually through the difficult task. Finally, the supervisor may instruct the robot, through the interface, to abandon the current task.

In manual mode, the forklift behaves like an ordinary manned forklift, i.e., it moves only in response to a human operator physically actuating its steering wheel, gas and brake pedals, and other controls while sitting in the cabin. With the understanding that manual intervention will occasionally be necessary, we explicitly designed the drive-by-wire actuation and mode transitions such that operators can seamlessly take control of the vehicle and operate it as they would an ordinary forklift. Whenever the operator exits the cabin, the forklift returns to stand-by mode, from which it can be commanded by the supervisor to resume autonomous operation.

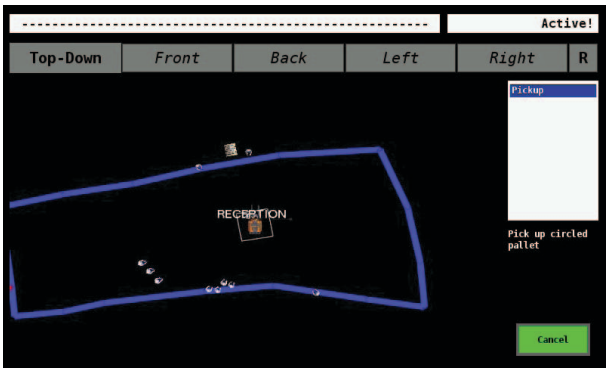
IV. The Tablet Interface

In order for existing warehouse personnel to be able to efficiently direct the robot, its command interface must be easy to use and require minimal training. The design must provide situational awareness to allow a human supervisor to effectively command the system, whether the human is standing nearby and directly observing the robot, or in a remote location. To address these needs, we developed a multimodal command interface enabling the human supervisor to issue high-level directives through a combination of simple spoken utterances and pen-based gestures on a handheld tablet.

To avoid confusion, only one tablet can control a robot at any time, and only the supervisor holding it has authority over the robot's task list. Our interface is based on a Nokia N810 Internet Tablet, which has a built-in microphone, touchscreen and stylus. The user draws on images captured by the robot's cameras (or synthesized by the system) and speaks into the tablet's microphone. Hardware buttons on the exterior of the tablet are used for system and safety-critical functions such as changing the forklift's operating mode and initiating speech



(a) A user's stroke is displayed as it is drawn.



(b) The top-down view indicates the pallet and pedestrian detections.

Fig. 3. The interface allows the user to select between different camera views and a synthesized map. (a) The user circles a pallet on the back of a truck displayed on the front camera image and a pickup task is queued. (b) The map view displays the subsequent LIDAR-based pallet estimate along with the robot's knowledge of pedestrians who were standing behind the vehicle. Note the false positive detection to the right of the pallet. To the right of these views, the interface conveys the robot's current operational state (in this case, "Active") along with a list of queued tasks.

recognition. One button places the robot into autonomous mode; one button places it into stand-by mode; and one button causes audio samples to be captured and analyzed for speech. The pause button constitutes one of the system's safety mechanisms; with it, the supervisor can stop the forklift at any time. These functions were implemented with the hardware buttons, both because hardware buttons are more reliable and because they are easier to trigger than software touchscreen buttons.

Also for safety, the tablet emits a "heartbeat" to the forklift at 10 Hz. This heartbeat lets the forklift know that a tablet is connected, and thus that someone has control over it. If the forklift loses its connection to the tablet (i.e., if it does not hear a heartbeat for 0.3 s), it pauses itself and awaits receipt of an explicit "resume" command from the supervisor (which requires a reconnected tablet for transmission).

A. Situational Awareness and Annunciation

A critical capability of an effective command interface is providing the user with sufficient situational awareness to understand the robot's environment [2], [4], [17]. Doing so is particularly challenging given the resources available

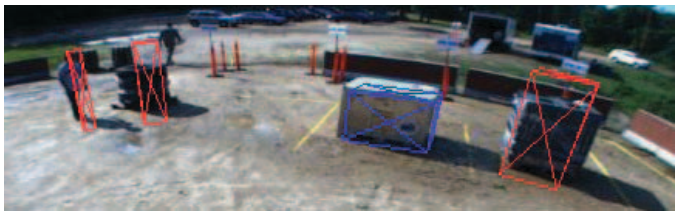


Fig. 4. Front camera view (zoomed and cropped). Objects recognized as pallets are outlined in blue; other objects and people are outlined in red.

on small tablets, and requires careful consideration of what information should be conveyed to the operator. In an effort to give the user sufficient knowledge of the robot's environment, our interface makes use of both live camera images and a synthesized overhead map. Both incorporate a succinct level of information that captures the robot's object-level knowledge of its surroundings.

As one means of providing situational awareness, the tablet allows the operator to view images, refreshed twice a second, from one of four cameras mounted on the robot that, together, form a 360° view around the vehicle. Fig. 3(a) provides a snapshot of the interface, in which the user is issuing a command by drawing on an image from the forklift's forward camera. The array of buttons above the image allows the user to select the camera view displayed in the center of the screen. Each image is augmented with wireframe bounding boxes that indicate objects that the robot detects with its LIDAR sensors, then classifies as people, trucks, or pallets. Fig. 4 shows an image from the interface in which an object that has been recognized as a pallet is outlined in blue, while other detected objects (including people) are outlined in red.

The interface also provides an overhead map view of the robot's local environment that incorporates a topological representation of the warehouse. In similar fashion to the augmented camera images, the map view displays objects that the robot detects in its vicinity, as shown in Fig. 3(b). For both the map view and the augmented imagery, we deliberately chose this level of abstraction of the robot's perception over the option of rendering raw sensor data, in order to minimize the user's cognitive burden in interpreting the data, an important consideration meant to increase the ease of use of the interface.

In addition to displaying the robot's situational awareness, the interface displays updates about the robot's current task that include information about its interpretations of the supervisor's recent commands. Text boxes at the top of the screen inform the supervisor about what the system is doing (e.g., engaging a pallet) and of the interpretation of any recent speech. After an utterance has been recognized, the upper-left text box displays the understood utterance. If the utterance was not understood, it displays "I didn't understand you." The upper-right text box displays the robot's operating mode, which represents whether the robot is operating autonomously and active, being manually driven, or is paused (indicating that it is safe to approach the vehicle). For example, when the forklift is approaching a pallet on the back of a truck during

a pickup task, the left text box says “Approaching truck” and the right text box says “Active: Pickup”.

To the right of the screen is the queue of tasks the forklift is to perform, with the current task at the top. A description of the selected task is displayed; the supervisor can click on any task to select it. The selected task can be canceled by clicking the cancel button. Completed, failed, or canceled tasks are removed from the queue. When a task stalls (e.g., when the forklift cannot find the pallet indicated by a pickup command), the queue remains unchanged until the stall condition abates or the supervisor intervenes.

B. Sketched Commands and Control

The user can command the robot to move, pick up, and place pallets by drawing on the canvas. Gestures have different meanings depending on their context. For example, circling a pallet is an instruction to pick it up, while circling a location on the ground or on the back of a truck is an instruction to put the pallet in the circled place. Drawing an “X” or a dot on the ground is an instruction to go there, while drawing a line is an instruction to follow the path denoted by the line.

C. Drawing on the World

Fig. 5 shows the lexicon of shapes recognized by the system. At the lowest level, all these shapes (except the dot) are a combination of lines and circles, defined by their geometric properties alone. As noted above, however, one shape can have multiple meanings. Depending on context a circle can be a pallet pickup command, a pallet drop-off command, or a circular path. When a shape has been recognized to have a certain meaning, we call it a “gesture.”

Our system recognizes shapes as in traditional sketch recognition systems: it records the timestamped point data that makes up each stroke and uses heuristics to compute a score for each possible shape classification based on stroke geometry. It then classifies the stroke as the highest-scoring shape. We implemented our own sketch recognition system without using more general techniques (e.g., [19], [25]) because our lexicon was so limited, and because the “drawing on the world” paradigm compensates for minor failings on the part of the geometrical sketch recognizer.

To classify shapes as gestures, the system must consider both what was drawn and what it was drawn on. We define the scene (i.e., the “context”) as the collection of labeled 2D boxes that bound the obstacles, people, and pallets visible in the camera view. The incorporation of the scene is what differentiates our approach from ordinary sketch recognition. Fig. 6 shows a high-level view of the sketch interpretation process under this model.

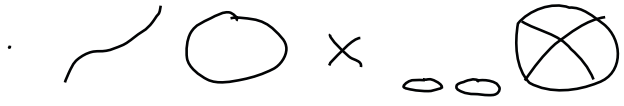


Fig. 5. Recognized shapes, from left to right, consist of: a dot, a line, a circle, an “X”, a double-circle, and a crossed circle.

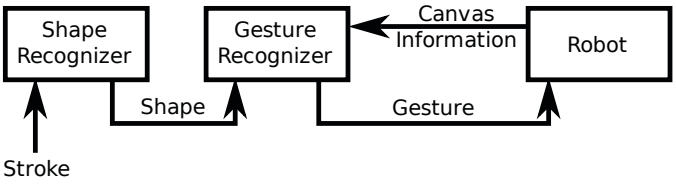


Fig. 6. A high-level view of the sketch recognition system.

D. Sketched Ambiguity & Contextual Corrections

Fig. 7 shows an example of the use of context to disambiguate a stroke. In this example, the stroke (Fig. 7(a)) could be either a circular path gesture that avoids objects (Fig. 7(b)), a pallet pickup command (Fig. 7(c)), or a pallet placement command (not shown). Which it is depends upon the context within which the stroke was drawn. Specifically, when the circle is around a pallet, it is labeled as a pickup gesture, but when it is not around a pallet and is too large to be a drop-off gesture, it is labeled as a circular path. Further, the geometrical shape recognizer could recognize the stroke as either a curved line or a circle. Incorporating scene context into the classification process removes the need to rely solely on stroke geometry for interpretation, making the entire process less sensitive to the accuracy of the geometrical recognizer.

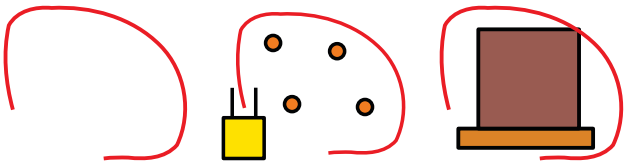


Fig. 7. (a) A circular stroke alone is ambiguous: it could be either (b) a circular path or (c) a pallet manipulation command. Context determines the gesture’s meaning.

This ability to disambiguate shapes into different gestures allows us to use fewer distinct shapes. As a result, the geometrical sketch recognition task is simplified, leading to higher gesture classification accuracy and robustness. The smaller lexicon of simple gestures also allows the user to interact with our system more easily.

Currently, only the pallet manipulation gestures are implemented in our system, using a rule-based engine that takes a shape and a scene as input and produces a gesture as governed by the rules. We are working on adding additional shapes with multiple meanings, such as defining an “X” gesture to correct a pallet recognition (i.e., to indicate “that is not a pallet”) while still maintaining the ability to draw an “X” on the ground to denote a desired destination.

E. Speech Commands and Control

Our interface runs the SUMMIT speech recognition library [8] locally on the tablet [10] to classify each utterance as one of several phrases. The SUMMIT recognizer uses a pre-existing set of telephone-based acoustic models. Allowable phrases are specified with a context-free grammar. Recognized

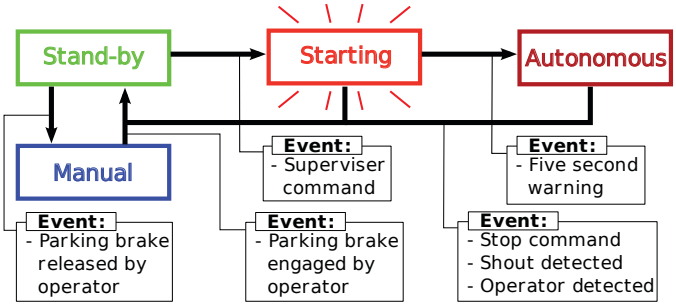


Fig. 8. States and state transition events for the autonomous forklift.

utterances include summoning commands, such as "Come to receiving." In order to simplify the recognition process, a push-to-talk strategy is used to record utterances.

V. Robot Interactions

Working with a 2700 kg forklift requires that particular consideration be given to safety. To that end, we equipped the forklift with annunciation features to make its behavior more transparent, and we made it cautious around and subservient to humans. To do this, we attempted to model robot interactions after warehouse practices that are already in place. Where no such practices could be built upon, we added annunciation features and sensors to interact smoothly with a dynamic, unconstrained environment.

A. Integration of Current Practice

Since it is a required safety practice in existing warehouses (under OSHA and military rules) to engage the parking brake when exiting the forklift cabin, engagement of the parking brake is used as a manual hand-over cue for the forklift. Fig. 8 shows the system's state transition rules. If the parking brake is released manually, the forklift enters manual mode and can be operated normally by the human forklift driver. When the parking brake is engaged, the vehicle enters stand-by mode. When the operator has vacated the vehicle, after a 5-second grace period, the robot can resume autonomous operation upon the supervisor's explicit command.

B. Situational Awareness

Several laser range-finding sensors mounted around the forklift enable the robot to detect pallets, fixed and moving obstacles, and people [23]. These sensors are used to populate the world model. Information in the world model dictates states of autonomy and helps the system interpret the meaning of sketched commands.

C. Subservience

We trust human judgement over that of the forklift in all cases. Accordingly, the robot cedes complete control to any human in the driver's seat. Anyone who wishes to operate the forklift through a difficult maneuver need only enter the cabin and start controlling the forklift. Moving any of the controls triggers the switch from autonomous mode to manual mode.

Fig. 9. The forklift detects an approaching pedestrian and pauses. Colored lights skirting the vehicle indicate the proximity of close objects (red to green lights map range from close to far). The LED signs display the reason for pausing.

If the robot's sensors detect anyone approaching, it will stop, even if it is in the middle of performing an autonomous task.

D. Annunciators

Humans take a variety of cues from each other. We can, for instance, glean a great deal by noting where a human forklift driver is looking, allowing us to anticipate which way the driver is about to move, or judge whether the driver has seen a pedestrian. We attempted to make our unmanned forklift generate analogous cues by fitting it with strings of LEDs along its sides, back, and mast, and placing LED signs and speakers facing forward, backward, left, and right, to notify nearby pedestrians of autonomous mode shifts.

The LED strings display the current mode, motion intent and proximity to nearby obstacles (conservatively assumed to be pedestrians). A chasing light pattern depicts the direction of the forklift's intended motion: strobing forward when about to move forward, and backward when about to back up. Similarly, the lights on the mast indicate the intended motion of the carriage: strobing up along the mast when about to lift, and down when about to lower. If the forklift detects an approaching pedestrian, it pauses and the LED strings on the chassis are used to show the direction and range of the detected person (shown in Fig. 9). People close to the robot induce red lights which shift to green as the person moves further away.

To annunciate intent to nearby pedestrians, all four LED signs display identical informative English text. By default, they display the level of autonomy of the robot (e.g., "Active!" when autonomous and "Manual" when not). When performing a task, they display the task and sub-task. For example, when going to the storage area, they display "Summon: Storage," and when going through the acquisition phase of pallet engagement, they display "Pickup: Detect." Additionally, when the forklift is becoming active, the speaker plays a school-bell

ring, the LED strings flash red and the LED signs count down the seconds, displaying the text: "Active in … 3, 2, 1."

E. Shout Detection

Since the forklift must operate in close proximity to any number of pedestrians, anyone in the vicinity of the forklift may verbally command it to stop moving at any time. The forklift continuously listens for appropriate shouted commands using the four microphones mounted below the cameras. Each input audio channel is connected to a streaming speech recognizer [10] configured to spot a small number of key phrases (e.g., "Forklift, stop moving!"). Training data was collected from twenty people speaking key phrases and random utterances in a variety of forklift operating conditions (e.g., motor noise, beeping) in order to train acoustic models for this task. During operation, the forklift pauses if a key phrase is detected through any microphone.

VI. Exploratory Tablet Evaluation

We performed an exploratory evaluation of the tablet interface, designed to identify areas for detailed study. Twelve undergraduate engineering students volunteered to perform a summoning task and a pallet manipulation task. All participants were male and had little or no experience in robotics. None had seen the forklift or the interface before the study. The evaluation took place in a mock outdoor warehouse that included pedestrians and moving vehicles. Noise within the environment included background conversation, shouting, and vehicle engines.

Each participant began with the forklift in the storage area and a truck loaded with one pallet in the receiving area. The participant was asked to summon the robot to receiving and direct it to pick up the pallet from the truck. The forklift and warehouse were within view of each participant throughout. All on-robot annunciation features were turned off for the study, so the participants' attention would be on the tablet interface.

After reading a sheet of instructions explaining the tablet's interaction modes, each participant was given a brief 5-minute tour of the interface and asked to begin. While commanding the robot, each participant was asked to speak aloud about what he was doing, what he thought was happening, and why he thought it was happening. All interactions were observed by a member of the team, and the system logged the images, views, and raw and interpreted voice and sketch commands made on the tablet. Each participant was given 20 minutes to complete the two tasks, after which he was asked to describe his experience with the interface.

We identified several directions for a user study, stemming from the following observations:

- When speaking aloud, the participants revealed an ability to correctly gauge the extent of the forklift's autonomy and determine what the robot was doing, based upon the text feedback fields and by watching the robot.
- All but two participants could tell when a pickup command had failed, despite no explicit feedback, based upon the task's disappearance from the queue.
- Only two participants exercised any view other than that of the front camera.
- Several participants suggested additional feedback mechanisms (e.g., an explicit pallet detection failure message).

The outcomes of the evaluation suggest a formal user study that focuses on two areas—UI feedback and remote vs. local control—and uses current warehouse personnel as the test group.

The study would explore the effect different levels and types of feedback have on the operator's understanding of the robot's current state. To do this, we would add explicit annunciation features on the tablet for various events (e.g., pallet detection failure) and determine the usefulness of each of these features by selectively enabling them. Further, other types of annunciation (such as displaying recognized pallets) would be selectively enabled or disabled.

The study would also attempt to gauge the efficiency with which a user can command the robot from a remote location, where the tablet interface provides the only view of the robot and its environment. During our evaluation the participants did not walk along with the robot as it navigated within the warehouse, but it nevertheless remained in sight. It would be interesting to see how the participants use the available camera and synthesized map views to control the robot from a remote location and to characterize how this reduced situational awareness might affect their operating efficiency.

We are in the process of formulating a user study which takes these new areas of focus into account. We will ask current warehouse employees to command the robot to perform a series of summoning, pallet pickup, transport and pallet drop-off tasks from both local and remote locations. The study will include controlled tests that evaluate operator performance with varying levels of feedback from the robot and tablet.

VII. Future Work

A goal of our work is to provide our system with more autonomy in an effort to lessen the supervisor's burden. To that end, we are developing capabilities that allow the operator to direct the robot to perform higher-level tasks spanning longer time horizons. One such example is that of tasking the forklift with unloading and storing a truck's entire cargo of multiple pallets via a single directive. Complex capabilities such as this require more contextually diverse gestures and more powerful feedback mechanisms.

One motivation for developing more robot autonomy is the ability to simultaneously command multiple robots. Our current tablet interface is designed to control a single vehicle, but can easily be extended to allow the operator to switch between several robots. In the long run, we wish to develop command interfaces through which the supervisor can specify high-level goals without having to worry about allocation of individual robots. Achieving this capability will require more sophisticated understanding of utterances and gestures within

material context, as well as more capable perception, planning, control, and annunciation subsystems.

VIII. Conclusions

This paper described novel multimodal interface mechanisms that allow people to interact with an autonomous forklift operating in close proximity to humans in a minimally-prepared environment. Chief among these interaction mechanisms is a tablet interface through which users provide high-level directives to the forklift through a combination of spoken utterances and sketched gestures. The interface incorporates object-level knowledge of the robot's surround with live camera and synthesized map views as a means of providing situational awareness to the supervisor. We described the interface's use of a novel interaction paradigm and sketch recognition mechanism called "drawing on the world," and gave a high-level description of its implementation. We demonstrated that the method aids the disambiguation of geometrically identical shapes and allows the use of a smaller set of shapes to mean a larger number of things.

We additionally described interaction mechanisms that enable the robot to operate among people. These include various means of annunciating the robot's intent and its knowledge of the environment to pedestrians in its vicinity. We presented different means by which people can seamlessly change the robot's level of autonomy from being autonomous to being drivable as a standard forklift, and back again.

Finally, we described an exploratory evaluation of the tablet interface and have described areas on which to focus a more detailed user study.

Acknowledgments

We gratefully acknowledge the support of the U.S. Army Logistics Innovation Agency (LIA) and the U.S. Army Combined Arms Support Command (CASCOM).

This work was sponsored by the Department of the Air Force under Air Force Contract FA8721-05-C-0002. Any opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

References

[1] A. Adler and R. Davis. Speech and sketching: An empirical study of multimodal interaction. In *Proc. Eurographics Workshop on Sketch-based Interfaces and Modeling*, pages 83–90, Riverside, CA, Aug. 2007.

[2] J.L. Burke, R.R. Murphy, M.D. Coovert, and D.L. Riddle. Moonlight in Miami: A field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human-Computer Interaction*, 19(1):85–116, June 2004.

[3] R. Davis. Sketch understanding in design: Overview of work at the MIT AI Lab. In *Proc. AAAI Spring Symposium on Sketch Understanding*, pages 24–31, Stanford, CA, March 2002. AAAI Press.

[4] J.L. Drury, J. Scholtz, and H.A. Yanco. Awareness in human-robot interactions. In *Proc. IEEE Conference on Systems, Man and Cybernetics (SMC)*, volume 1, pages 912–918, Washington, DC, October 2003.

[5] F. Ferland, F. Pomerleau, C.T. Le Dinh, and F. Michaud. Egocentric and exocentric teleoperation interface using real-time, 3D video projection. In *Proc. ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 37–44, March 2009.

[6] T. Fong, C. Thorpe, and C. Baur. Multi-robot remote driving with collaborative control. *IEEE Transactions on Industrial Electronics*, 50(4):699–704, Aug. 2003.

[7] T. Fong, C. Thorpe, and B. Glass. PdaDriver: A handheld system for remote driving. In *Proc. IEEE Int'l Conf. Advanced Robotics*, July 2003.

[8] J.R. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137–152, November 2003.

[9] B. Hamilton-Baillie and P. Jones. Improving traffic behaviour and safety through urban design. *Civil Engineering*, 158(5):39–47, May 2005.

[10] I.L. Hetherington. PocketSUMMIT: Small-footprint continuous speech recognition. In *Proc. Interspeech*, pages 1465–1468, Antwerp, Aug. 2007.

[11] H. Holzapfel, K. Nickel, and R. Stiefelhagen. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures. In *Proc. International Conference on Multimodal Interfaces (ICMI)*, pages 175–182, State College, PN, October 2004.

[12] M. Johnston, P.R. Cohen, D. McGee, S.L. Oviatt, J.A. Pittman, and I. Smith. Unification-based multimodal integration. In *Proc. Eighth conference on European chapter of the Association for Computational Linguistics*, pages 281–288, Morristown, NJ, 1997.

[13] H. Kaymaz-Keskinpala and J.A. Adams. Objective data analysis for PDA-based human-robot interaction. In *Proc. IEEE International Conference on Systems, Man and Cybernetics (SMC)*, volume 3, pages 2809–2814, The Hague, The Netherlands, October 2004.

[14] H. Kaymaz-Keskinpala, J.A. Adams, and K. Kawamura. PDA-based human-robotic interface. In *Proc. IEEE Conference on Systems, Man, and Cybernetics (SMC)*, volume 4, pages 3931–3936, Washington, DC, October 2003.

[15] T. Matsumaru. Mobile robot with preliminary-announcement and indication function of forthcoming operation using flat-panel display. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 1774–1781, 2007.

[16] T. Matsumaru, K. Iwase, K. Akiyama, T. Kusada, and T. Ito. Mobile robot with eyeball expression as the preliminary-announcement and display of the robots following motion. *Autonomous Robots*, 18(2):231–246, 2005.

[17] C.W. Nielsen, M.A. Goodrich, and R.W. Ricks. Ecological interfaces for improving mobile robot teleoperation. *IEEE Transactions on Robotics*, 23(5):927–941, October 2007.

[18] T.Y. Ouyang and R. Davis. Recognition of hand drawn chemical diagrams. In *Proc. AAAI Conference on Artificial Intelligence*, pages 846–851, 2007.

[19] B. Paulson and T. Hammond. Paleosketch: Accurate primitive sketch recognition and beautification. In *Proc. Int'l Conf. on Intelligent User Interfaces (IUI)*, pages 1–10, Gran Canaria, Jan. 2008.

[20] D. Perzanowski, A.C. Schultz, W. Adams, E. Marsh, and M. Bugajska. Building a multimodal human–robot interface. *Intelligent Systems*, 16(1):16–21, Jan.-Feb. 2001.

[21] D. Sakamoto, K. Honda, M. Inami, and T. Igarashi. Sketch and run: A stroke-based interface for home robots. In *Proc. International Conference on Human Factors in Computing Systems (CHI)*, pages 197–200, Boston, MA, April 2009.

[22] M. Skubic, D. Anderson, S. Blisard, D. Perzanowski, and A. Schultz. Using a hand-drawn sketch to control a team of robots. *Autonomous Robots*, 22(4):399–410, May 2007.

[23] S. Teller et al. A voice-commandable robotic forklift working alongside humans in minimally-prepared outdoor environments. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, May 2010 (to appear).

[24] P.R. Wurman, R. D'Andrea, and M. Mountz. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI Magazine*, 29(1):9–19, 2008.

[25] B. Yu and S. Cai. A domain-independent system for sketch recognition. In *GRAPHITE '03: Proc. 1st International Conference on Computer graphics and Interactive Techniques in Australasia and South East Asia*, pages 141–146, New York, NY, 2003. ACM.