Source of Acquisition
NASA Ames Research Center

Common Metrics for Human-Robot Interaction

Aaron Steinfeld[1]
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA

Terrence Fong
Intelligent Systems Division
NASA Ames Research Center
Moffett Field, CA

David Kaber
Dept. of Industrial Engineering
North Carolina State University
Raleigh, NC

Michael Lewis
Sch. of Information Science
University of Pittsburgh
Pittsburgh, PA

Jean Scholtz
Information Access Div.
NIST
Gaithersburg, MD

Alan Schultz
Intelligent Systems Section
US Naval Research Lab
Washington, DC

Michael Goodrich
Computer Science Dept.
Brigham Young University
Provo, UT

ABSTRACT
This paper describes an effort to identify common metrics for task-oriented human-robot interaction (HRI). We begin by discussing the need for a toolkit of HRI metrics. We then describe the framework of our work and identify important biasing factors that must be taken into consideration. Finally, we present suggested common metrics for standardization and a case study. Preparation of a larger, more detailed toolkit is in progress.

Categories and Subject Descriptors
I.2.9 [**Artificial Intelligence**]: Robotics – *operator interfaces.*

General Terms
Measurement, Performance, Design, Experimentation, Human Factors, Standardization.

Keywords
Human-robot interaction, metrics, unmanned ground vehicles.

1. INTRODUCTION
In the early years of many technical fields, the research community often utilizes a wide range of metrics that are not comparable due to a bias towards application specific measures. Common metrics typically develop as researchers devote more attention to the core questions of the field. This transition allows for greater sharing of knowledge as it becomes possible to compare findings, to benchmark designs, and to draw from an evaluation toolkit.

We believe that *human-robot interaction* (HRI) has reached such a point and, thus, we are working to develop a set of common metrics. Specifically, we have begun identifying methods to assess how much effort human and robot must contribute (independently and jointly) to effectively accomplish a task. Our goal is to provide a foundation upon which to build better HRI and to improve the performance of human-robot teams.

The primary difficulty in defining common metrics is the incredibly diverse range of human-robot applications. Thus, although metrics from other fields (HCI, human factors, etc.) can be applied to satisfy specific needs, identifying metrics that can accommodate the entire application space may not be feasible. As such, it may be necessary to rely on measures that, while not ensuring comparability across applications, provide the benefits afforded by familiar methods and scoring. A good example of this would be the use of subjective ratings scales (e.g., Likert).

Many metrics, such as *time-to-completion*, are highly application or task specific. As such, many fields develop scenario- based reference tasks. The best example of this technique in HRI is the NIST Urban Search and Rescue arena, which is currently used for scoring in a number of robot competitions [23, 24]. Within the arena, the metrics that are used focus on overall human-robot system performance (e.g. number of victims found minus the number of penalties assigned), but do not specifically emphasize how the mission is accomplished (i.e., by the human, the robot, or some combination of the two).

As a means of partitioning HRI, metrics can be organized with respect to system characteristics and interactions [26]. Although there have been several attempts to develop taxonomies [9, 18, 54], the community has yet to develop a consensus for a standard framework. For the purposes of this paper, therefore, we have chosen to analyze HRI in terms of three aspects: human, robot, and system. This enables us to identify and discuss metrics that are useful throughout the application space.

In summary, the goals of our effort are: (1) identify classes of metrics to facilitate comparison of research results; (2) identify common metrics that can be used for evaluations across a wide range of tasks and systems; and (3) provide a measurement toolkit for future studies.

2. SCOPE AND FRAMEWORK
In order to bound the scope of our study, we have chosen to focus our work on task-oriented mobile robots. In particular, we present metrics in terms of five task categories. We selected these tasks because they can be performed with a high-level of human direction (pure teleoperation), a high-level of robot independence (full autonomy), or at any point on the interaction spectrum. By doing so, we believe that: (1) our metrics are broadly applicable to a wide range of applications and (2) we can assess the impact of different levels/types of HRI on performance.

2.1 Navigation
This is a fundamental task for mobile robots: move the robot from A to B [13]. Performing this task requires determining where the robot is (A), where it needs to be (B), how it should get there

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Human Robot Interaction'06, March 2-3, 2006, Salt Lake City, Utah USA.
Copyright 2006 ACM X-XXXXX-XXX-X/XX/XXXX...$X.XX.

[1] Correspondence: 412-268-6346, steinfeld@cmu.edu

(path, resource usage, etc.), and how to deal with environmental factors and contingencies (obstacles, hazards, etc.) encountered on the way.

2.2 Perception

The focus of this task is to perceive and understand the remote environment for applications such as search, surveillance, target identification, etc. This task does not include perception needed for other tasks (e.g., navigation requires localization). Performing this task requires: establishing a context through proprioceptive sensing, interpreting sensor data within this context, seeking/filtering additional sensor data, and deciding what information to give to other agents. Reflecting current practice, we emphasize camera imagery in choosing perception metrics.

2.3 Management

The purpose of this task is to coordinate and manage the actions of humans and robots, acting independently or in groups. Of primary concern is allocating and deploying resources to guarantee appropriate coverage (i.e., having the "right" agent at the "right" place at the "right" time). Performing this task requires assessing availability, understanding capabilities, team coordination, monitoring, recognizing problems, and intervention.

2.4 Manipulation

In this task, the robot interacts with the environment. For our work, we consider manipulation to encompass not only arm-based grasping, but also non-prehensile motions (e.g., pushing) and discrete actions, such as payload drop- off. Applications include ordnance disposal, geology (e.g., rock sampling), construction, and personnel/material delivery. Performing this task requires determining what is to be effected, specifying how it is to be done ("put this there"), executing the process, and verifying the outcome.

2.5 Social

The objective of this task is to perform work that requires significant "social interaction". Applications include tour guiding, health care (mobility assistance, therapy, etc.), entertainment, and testing models of human intelligence. Performing this task requires perceiving and interpreting the world in terms of past experience, recognizing and modeling users, understanding social communication and norms models, and acquiring/exhibiting social competencies [12].

3. BIASING EFFECTS

While this is not meant to be an exhaustive list, there are many factors that may bias or confound HRI effectiveness. Therefore, care should be taken in measuring system effectiveness or attempting to establish benchmarks when such effects are present. An excellent discussion of biasing effects and general issues related to performance measurement can be found in [1].

3.1 Communications

Communications factors, such as delay, jitter, and bandwidth, can have profound effects on human performance. As such, HRI quality may be strongly dependent on the capacity of the communication channel(s) to carry information between human and robot [49].

Delay (aka "latency" or "lag") is the time delay caused by the transmission of information across a communications network. Delay is well known to degrade human performance in motor-sensory tasks with interactive systems as well as planning and performance in teleoperation scenarios [27, 32].

Jitter is the variance in transmission time that measures whether the amount of time between two messages at the receiving end is the same as the when they were sent [19]. In teleoperation, data packets transmitted between a control station and a telerobot may have different inter-arrival times with no data packet loss [16, 47].

Bandwidth describes the data transmission capacity of the communications channel. Bandwidth limitations do not imply loss of information unless techniques are used to promote transmissions speed. For example, video transmission across computer networks (e.g., the Internet) generally requires the use of lossy data compression, which may result in unacceptable loss of visual detail for remote perception.

3.2 Robot Response

Timing factors within the robot may confound time-oriented HRI metrics. This is especially true if these factors are not uniform across examined designs or test conditions. Special care should be taken with prototype and proof-of-concept robots as variable system behavior is likely to occur. Moreover, most conventional robot control architectures are not designed to support interaction at human rates.

Examples include system lag and update rate. System lag is comparable to communication delay, but refers to time spent by the robot processing information. For example, a mobile robot may spend time computing a new collision-free path when given a waypoint request. Update rate (also referred to as "display-system lag") refers to a delay in displaying information (camera images, sensor data, robot status, etc.) to the operator.

3.3 User

Performance shaping factors (PSF) can influence behavior and affect human performance. These include operational factors (tactics, time on station, etc.), equipment factors (physical parameters, workspace layout, etc.), task factors (complexity, repetitiveness, etc.), personnel factors (training, motivation, stress, etc.), and external environmental factors (illumination, visibility, etc). Numerous guidelines for reducing and analyzing the impact of PSF are given in [1].

The human's role may also affect the fluidity and effectiveness of HRI. In Scholtz [42], for example, it is suggested that there five different HRI roles (supervisor, operator, mechanic, peer and bystander) that humans may play, each of which requires different information and awareness. Thus, performance may be dependent on the role required and how well the interface supports it in specific mission situations.

4. TASK METRICS

4.1 Navigation

1) Global navigation: The system needs to have an overall understanding of the locale in which it is working. Some parameters might be adjusted prior to starting a task or mission, such as whether the robot is indoors or outdoors, off-road or on-road, in an urban terrain, wooded terrain, or desert. During task execution, the system needs to know where in this particular area is it. For example, if the robot is moving about inside a building, the system should know on which floor it is located.

2) Local navigation: This is a finer granularity of information that is essential for smoothly moving in an area. The system must

know what potential hazards are close by, such as doorways, stairs, culverts, trees, or pedestrians.

3) Obstacle encounter: Not all navigation is without problems. Obstacles are often encountered and at times, robotic systems may have to extract themselves from ditches or debris. Creating a plan for extraction necessitates knowing characteristics of the obstacle (size, hardness) as well as knowing other potential hazards in the local environment.

Effectiveness measures how well the task is completed. Potential measures include:

- Percentage of navigation tasks successfully completed
- Coverage of area
- Deviation from planned route
- Obstacles that were successfully avoided
- Obstacles that were not avoided, but could be overcome

Critical incidents can be used as an indirect measure of navigation HRI. For example, Scholtz, Young et al. [45] coded critical incidents in an urban search and rescue competition and noted the number of critical incidents that could be attributed to each type of navigation.

Efficiency measures the time needed to complete the task. Efficiency measures include:

- Time to complete the task
- Operator time for the task (includes HRI overhead)
- Average time for obstacle extraction

Amount of non-planned looping in navigating effort, or workload, measures include:

- Number of operator interventions per unit time. Interactions can be planned or unplanned. Unplanned interactions are termed "interventions" [21]. The average number of interventions per unit of time can also be used as a measure of HRI in navigation. The average time needed for the intervention, and the effectiveness of the intervention can also be measured [44].
- Ratio of operator time to robot time. For example, if the operator spends 5 minutes to input a navigation plan that allows the robot to successfully navigate for an hour, we have a 1:12 ratio [55].

4.2 Perception

Perception is the process of making inferences about distal stimuli (objects in the environment) based on proximal stimuli (energy detected by sensors). In HRI, perceptual inference can be performed by the robot (localization, obstacle detection, etc.), by the human (e.g., identifying a victim in a video image), or in combination, such as a robot that directs its operator's attention to an area of interest but leaves inference making to the human.

Veridical perception depends on fusing sensor data about robot state with sensor data about the environment. Inferences about objects viewed in a camera image, for example, depend on whether the image is from an upright robot or a robot that has rolled over and the camera is now pointing to the ceiling [4].

There are two basic tasks involved in perception: interpreting sensed data and seeking new sensor data. HRI metrics for perception can be divided between those addressing *passive perception* (interpretation of received sensor data) and *active perception* (in which multiple sensor readings are obtained to disambiguate or increase confidence for perceptual inference [2]).

1) Passive Perception: Passive perception involves interpreting sensor data: identification, judgment of extent, and judgment of motion. *Identification* measures detection and recognition accuracy for task objects within sensor range.

Potential measures include:

- Detection measures: % detected, signal detection, detection by object orientation, contrasts between detection in cluttered and sparse environments, etc.
- Recognition measures: classification accuracy, confusion matrices, recognition by object orientation

Judgment of extent measures the accuracy of quantitative judgments about the environment. The unaccustomed viewing height and field of view provided by a robot's camera can make these judgments very difficult. Psychophysical data on spatial judgments can, however, provide a normative reference [50].

Potential measures include:

- Absolute judgments of distance, size, or length
- Relative judgments of distance, size, or length
- Platform relative judgments such as "How long would it take the robot to reach the wall?"

Judgment of motion measures the accuracy with which egomotion or movement of objects in the environment is judged.

Potential measures include:

- Absolute estimates of robot velocity
- Estimates involving relative motion such as "Will robot collide with another moving object?"

Other metrics include indirect measures of perceptual performance that reflect the accuracy of the operator's perception. Clear perception of robot attitude, for example, might be inferred from the choice of level paths through uneven terrain [30].

2) Active Perception: Active perception in HRI ranges from relatively passive tasks such as control of pan and tilt of a camera to control of robot movement in search. To differentiate active perception from mobility/navigation tasks we require that active perception involving mobility be initiated by detection of a possible search target.

Active identification measures performance on recognition tasks involving mobility. Potential measures (in addition to recognition measures for identification) include:

- Efficiency: time or effort to confirm identification, improvement in identification over initial detection
- Effort: amount of camera movement [22].

Stationary search measures performance on search tasks that do not involve mobility. Stationary search may involve camera control or data fusion between sensors. Potential measures include:

- Detection accuracy for targets within sensor range
- Efficiency as time to search or non-overlapping coverage
- Coverage as percentage of potential sensor coverage
- Operator confidence in sensor coverage

Active search measures performance on search tasks involving mobility. In this case the initiating stimuli are objects within sensor range that might conceal a target (e.g., [5]). Potential measures (in addition to stationary search measures) include:

- Efficiency: time and effort expended (e.g., for target identification)
- Identification errors: number of incorrect targets, number of targets missed, etc.
- Degree of operator fusion

While humans are effective at synthesizing information, there are known interface characteristics that can hinder this capability. Cataloging how well a system supports the synthesis of information can provide a valuable HRI metric. An example task that can be affected by degree of operator fusion is the ability to utilize information from multiple sensors to develop an accurate awareness of robot state. Essentially, how well does a system support the ability to develop accurate assessments of remote scenarios?

4.3 Management

1) Fan out: Fan out, as defined in Goodrich and Olsen [17], is a measure of how many robots (with similar capabilities) can be effectively controlled by a human. It directly affects the logistical demands related to robot deployment, the difficulty in handling and managing the robot during use, and the total cost-benefit ratio of the robot system in question.

Depending on the value of the robot in question, fan out can be considerably biased in favor of the robot. For example, unmanned aerial vehicles (Predator, Global Hawk, etc.) currently in use by the U.S. military require many humans to operate each vehicle.

This measure is also a good indicator of robot hand-offs between operators and the upper limit of workload for operators. When the number of robots is large and a group of humans are managing them as a team, this begins to emulate the task requirements seen in air traffic control. As such, metrics and methods from this domain may be relevant (e.g., [38]).

2) Intervention response time: Whenever an operator does not devote total attention to a robot, there will be delay between when the robot encounters problems and when the operator intervenes. This is often the case with supervisory control or when multiple robots must be controlled [48]. Operator intervention may be physical (i.e., requiring "mechanic- like" assistance) or cognitive (requiring decision making, perceptual support, etc) [14, 42].

A key metric, therefore, is intervention response time, which can be measured either from when the operator first recognizes the problem or when the robot first requests assistance. Response time can also allow specific details to be examined. For example, response time could be subdivided into: (1) time to deliver the request from the robot, (2) time for the operator to notice the request, (3) situation awareness and planning time, and (4) execution time. The first segment examines system performance while the remaining ones are specific to the interface design and task at hand.

3) Level of autonomy discrepancies: It is becoming increasingly common for autonomous robots to be designed and operated with multiple levels of control and autonomy [49]. In many cases, some levels are more appropriate than others for specific environments, tasks, and events. Anecdotal evidence documented from robot deployments suggests that some robot failures may have been prevented if the robot had either had the ability to enter an appropriate autonomous state or the operator had commanded the robot to do so [51].

In the simplest case, one can measure the ability of the human to accurately and rapidly identify the appropriate level of autonomy. Interfaces that support this process have been highlighted as important in previous research [13, 49]. Related to this process is the performance of the human to subsequently activate autonomy appropriately, e.g., [28]. Greater understanding of how and why autonomy behaves generally leads to more appropriate utilization of the autonomy [36, 49].

This metric encompasses several factors (situation awareness, trust, etc), but serves as a good indicator of system efficiency. It is particularly useful if one knows what the "optimal" autonomy state should be for a given task. Experimenters can then configure test events that require certain states (e.g., impossible to complete without human assistance on object detection) and check if the human-robot system enters the appropriate state.

4.4 Manipulation

1) Degree of mental computation: Certain manipulation activities can be measured by the degree of mental computation performed by the operator. Examples of mental computation tasks are mental rotation, rate tracking, and object-referent association in working memory. For example, because of limited camera views and communication bandwidth, operators may be required to make mental, orthographic projections of 2-D views of an end-effector for control purposes.

For example, Macedo et al. [31] demonstrated that the degree of angular offset of the axes of display rotation relative to hand controller rotation had a significant effect on time-to-control response and accuracy in teletracking tasks. Control-display misalignment increased for non-orthogonal angles and human path tracking performance significantly degraded.

Mental workload is strongly influenced by demands made on short and long-term memory. For example, reliance on working memory for mental labeling of objects (e.g., in a remote work environment) can result in high operator workload. Moreover, the degree of mental computation required for a particular task may depend upon perceptual features of the environment.

2) Contact errors: A key metric in almost all manipulation tasks is contact error. In particular, the number of unintentional (or inadvertent) collisions between a manipulator and the environment (including task objects) is highly indicative of performance (e.g., positional accuracy). Moreover, the type of contact errors (glancing, hard/soft, etc.) is useful for system assessment (e.g., capability for working in cluttered spaces).

Prior research has demonstrated that operator performance and workload are significantly affected by whether joint or world mode (i.e., end-effector position) control is required for task performance [27]. For example, world mode can reduce task completion times, but may also increase the number of contact errors when working in confined spaces in which joints may contact other objects. That is, the operator may have good global situational awareness on the end goal for the manipulator, but may suffer from poor local situational awareness on the position of each manipulator joint, etc.

attentional resources. It is particularly relevant in high workload and time stress situations as a basis for decision-making.

One well-known query-based tool for evaluating SA is the "Situation Awareness Global Assessment Technique" (SAGAT) [11]. SAGAT has been used to assess SA at a various levels of autonomy [25, 44]. In general, the most important aspect of using SAGAT to measure SA is performing a detailed task analysis in order to formulate appropriate operator queries. In Scholtz, Antonishek et al. [44], for example, an analysis of human interventions in autonomous rover off-road driving was used to develop questions for assessing SA at multiple levels.

2) Workload: Multidimensional workload assessment techniques may be useful for relating human perceptions of cognitive load to operator SA, telepresence, performance, and user interface design. For example, the NASA-Task Load IndeX (NASA-TLX) [20], has been widely used to measure human performance and workload in teleoperation scenarios [8, 25, 44]. In general, results have shown that subjective ratings of workload decrease as the level of system autonomy increases and that shorter teleoperation tasks yield lower workload ratings.

At this point in time, there is a need to identify non-intrusive measures of workload that can characterize operator stress in real-time. Such workload information could be used as a basis for dynamically configuring system interfaces to best support operator performance [53]. Substantial research has already been conducted on the use of physiological measures as real-time indicators of cognitive workload (e.g., see [52] for a survey of cardiovascular and respiratory measures).

3) Accuracy of mental models of device operation: Design affordances, operator expectations and stimulus-response compatibility can all impact human performance. The common types of compatibility identified in the literature include conceptual, movement, spatial, and modality compatibility [41]. The benefits of matching interface displays and controls to human "mental" models include reductions in mental transformations of information, faster learning and reduced cognitive load (e.g., Macedo, Kaber et al. [31]).

Numerous studies on user mental model assessment have been published in the human factors literature, primarily for household appliances and desktop computer interfaces [33, 41]. Many of the measures developed in these studies can be directly used for HRI.

5.3 Robot Performance

1) Self-awareness: The degree to which a robot can accurately assess itself will have a direct impact on the ability of the human to efficiently interact with the robot. The less a robot is aware of its capabilities and the less it is able to recognize when it is having trouble, the more human monitoring and intervention is required.

Self-awareness is particularly important when a robot must ascertain if involving the human is useful. For example, if a robot is operating far (in time and/or distance) from the human (e.g., a lunar rover with an Earth-based operator), it must be aware that it cannot ask the human for physical assistance and that obtaining cognitive/perceptual help may take considerable time.

To qualitatively measure self-awareness, we propose assessing the following robot characteristics: (1) understanding of intrinsic limitations (mobility, sensor limitations, etc); (2) capacity for self-monitoring (health, state, task progress) and recognizing deviations from nominal; and (3) effectiveness at detecting, isolating, and recovering from faults (during both planning and execution).

2) Human awareness: A robot can also be scored on the degree to which it is aware of humans. Depending on the application, the robot may need to be sensitive to the human's presence and have knowledge of the human's commands (expectations, constraints, intent) [10]. Clearly, the level of "awareness" depends on the level of autonomy that the robot is expected to achieve and the role(s) played by the human(s) [42]. This capability can be dynamic and may include a user model that helps the robot recognize human behavior and react appropriately [12].

Human awareness implies competency in various skills, the proficiency of which can be assessed independently or collectively. These include: (1) human-oriented perception (human detection and tracking, gesture and speech recognition, etc); (2) user modeling and monitoring (cognitive, attentional, activity); (3) user sensitivity (adapting behavior to user, measuring user feedback, recognizing human state).

A recently proposed metric is the number of "awareness violations" (awareness information that should be provided that is not provided) that occur during task execution [10]. This metric is particularly well-suited to critical incident analysis, in which anomalous situations (operator or robot encounters a problem) are examined post-mortem.

3) Autonomy: The ability of robots to function independently is limited, though continually improving. This is especially true when robots face anomalies, or conditions, that exceed their autonomous capabilities. Though there are many application-specific methods, a useful metric for measuring autonomy in general is "neglect tolerance" [17].

Neglect tolerance directly measures how a robot's effectiveness declines when the human is not attending to the robot. In particular, it measures the amount of the time the robot can be neglected before performance drops below an acceptable level of task performance. Two methods for assessing neglect tolerance are described in [35].

We must note, however, that neglect tolerance encompasses numerous factors: task complexity, robot capability, user interface, and the user. Thus, the metric is only useful for obtaining an overall measure of a robot's autonomy, rather than specific details (e.g., failure modes).

6. USE EXAMPLE

6.1 Laser Range Finder Visualization

Nielsen, Ricks et al. [34] and Ricks, Nielsen et al. [39] tested interfaces that integrated laser information and video from a remote robot to support teleoperation (e.g., Figure 1). The study consisted of two parts: an experiment that used only simulated robots and another experiment that used real robots. Subjects were asked to teleoperate the robot through a series of mazes, following instructions given by visual cues in the world. For the simulation study, subjects were asked to memorize a sequence of five images or words before teleoperating the robot and then recall the sequence after completing the maze. For the real-world study, subjects were asked to remember a sequence of five images or worlds that they would encounter in the world, and were then asked to recall the sequence at the completion of the maze.

The hypothesis being tested was that the integrated display would be easier for subjects to use than the side-by-side display in a

4.5 Social

Some social robots (Cog, Kismet, etc.) are "biologically inspired" and use deep models of human cognition and interaction in order to simulate the social intelligence found in living creatures. This is often the case when the primary function of the robot is to interact socially with people. Other social robots (Nursebot, CERO, etc.) are "functionally designed" and show their social competence only in reaction to human behavior (i.e., they outwardly appear to be socially intelligent, even if the internal design does not have a basis in cognitive science) [12].

This dichotomy is important to understand because the criteria for "good performance" often differs substantially. In particular, "functionally designed" social robots may need only to produce certain experiences for the user, rather than having to withstand deep scrutiny for "life-like" capabilities. The difficulty, of course, is determining which metrics (engineering, psychological, sociological) are most appropriate for evaluating social "effectiveness".

1) Interaction characteristics: One approach is to assess characteristics such as interaction style or social context via observation [6] or conversational analysis [7].

2) Persuasiveness: The robot is used to change the behavior, feelings or attitudes of humans. This is the case when robots mediate human-human interaction, as in autism therapy [7].

3) Trust: Research on trust in automation suggests that this is an important factor to measure. In particular, trust is likely to influence reliance on complex, imperfect automation in dynamic environments that require the human to adapt to unanticipated circumstances [29].

4) Engagement: Social interaction is widely cited as an effective mechanism for engaging users. A key metric, therefore, is to measure the efficacy of various social characteristics (emotion, dialogue, personality, etc.) for capturing attention (acquisition time) and holding interest (duration). See, for example, Bruce, Nourbakhsh et al. [3] and Schulte, Rosenberg et al. [46].

5) Compliance: Social characteristics (appearance, adherence to norms, etc.) can also influence the amount of cooperation a human gives to a robot, which may be critical for tasks in certain domains (e.g., in health care). Thus, measuring compliance can provide significant insight into the effectiveness of the robot design, e.g., Goetz and Kiesler [15].

5. COMMON METRICS

5.1 System Performance

When we assess system performance, we are concerned with measuring how well the human(s) and the robot(s) perform as a team. Although there are many well-known task measures (see ANSI/AIAA [1] for an extensive list), our emphasis is to evaluate the human-robot team and human-robot interactions, rather than task-specific performance.

1) Quantitative performance: Quantitative measures assess the effectiveness and efficiency of the team at performing a task. Since robots are generally designed to operate with some level of autonomy [37], performance measures must consider the autonomy design. Quantitative performance measures include:

- Effectiveness: the percentage of the mission that was accomplished with the designed autonomy. For example, consider a system that is designed to be fully autonomous. If this system successfully performs a task, but a human is required to intervene 20% of the time, then the system is only 80% effective given the design specifications. The number and duration of operator interventions can also be used to compute the effectiveness metric.
- Efficiency: the time required to complete a task. In many cases, a robot may have sufficient competency to perform a task if time constraints are ignored. Thus, efficiency can be calculated for: (1) all tasks completed (regardless of the contributions of the human and the robot); or (2) only for those missions completed with the autonomy design.

2) Subjective ratings: In addition to quantitative measures of performance, subjective ratings can be used to assess the quality of the effort. The effectiveness metric measures the performance of the system (human and robot) but subjective ratings should be compiled from all stakeholders involved, both direct and indirect.

Consider, for example, a search and rescue operation. A human-robot team locates a victim trapped in a collapsed structure. The medical team gets the correct information to provide medical support while the structural engineering team directs rescue operations. Metrics for this mission should assess not just the effectiveness and efficiency of locating the victim but also the quality of the information provided to the medical and structural engineering teams.

3) Appropriate utilization of mixed-initiative: Robots will increasingly possess more self-awareness and more awareness of their operators [14]. One aspect of system performance is the ability of the human-robot team to appropriately regulate who has control initiative. Suggested measures are:

- Percentage of requests for assistance made by robot
- Percentage of requests for assistance made by operator
- Number of interruptions of operator rated as non-critical

Perhaps the main issue in task-oriented HRI, however, is achieving the right mixture of human and robot autonomy. Often it is possible to perform tasks with humans and/or robots, thus it is important to decide and verify which human or robotic assets are most appropriate to use for a given mission.

One method for assessing the performance of human-robot teams is described in Rodriguez and Weisbin [40]. This method focuses on decomposing a work scenario into "functional primitives", allocating these primitives to either human or robot resources, evaluating execution of each primitive, and computing the ratio of performance benefit to resource allocation.

Another method for evaluating the overall effectiveness of human-robot teams is interaction effort, which measures the overall effort required by the human to work with the team [35]. Interaction effort, because it considers the amount of autonomy of each team member, is particularly useful for when the overall mission requires the use of a mix of competencies or sub-groups within the team.

5.2 Operator Performance

1) Situation awareness: Situation awareness (SA) is critical to effective decision-making, operator performance and workload in numerous dynamic control tasks [25, 43]. In general, SA is relevant to human in-the-loop control when there are multiple competing goals and multiple, simultaneous task demands on

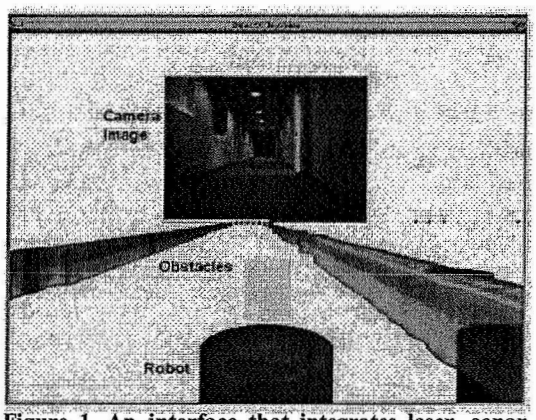

Figure 1. An interface that integrates laser, sonar, and video into a single perspective.

navigation task. This experiment employed the following Navigation and Common metrics:

4.1.3 Obstacle encounter – number of collisions

5.1.1 Efficiency – time-to-complete the maze and average speed

5.1.2 Subjective ratings – reported time to "feel comfortable" with the interface during training, rating scales from one to ten (effort, learnability, and confidence), and interface preference

Each of these metrics allowed effective comparison of the displays tested, thus demonstrating their applicability to HRI.

7. CONCLUSIONS

The continuing work under this effort will expand and refine the material presented here. The eventual plan is to provide a living, comprehensive document that future research and development efforts can utilize as a HRI metric toolkit and reference source.

In closing, we would like to point out the need to select appropriate test populations when applying these metrics. Specifically, as robots are increasingly deployed in applications in which the target user is not an expert roboticist [49], it becomes critical to recruit subjects having a broad range of knowledge, experience, and expertise.

8. ACKNOWLEDGMENTS

We would like to thank Stephen Hughes and Julie Marble for their insightful comments.

9. REFERENCES

[1] ANSI/AIAA. Guide to Human Performance Measurements, AIAA, Washington, DC, 1993.

[2] Bajcsy, R. Active perception. Proc. IEEE, 76 (1988), 966-1005.

[3] Bruce, A., Nourbakhsh, I., and Simmons, R., The role of expressiveness and attention in human-robot interaction. In Proc. AAAI Fall Symposium on Emotional and Intelligent II: The Tangled Knot of Social Cognition, (2001).

[4] Casper, J., and Murphy, R. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. IEEE Transactions on Systems, Man and Cybernetics B, 33 (2003).

[5] Cassimatis, N., Trafton, G., Schultz, A., Bugajska, M., and Adams, W., A task domain for combining and evaluating robotics and cognitive modeling techniques. In Proc. NIST Performance Metrics for Intelligent Systems Workshop, (2002).

[6] Dautenhahn, K., and Werry, I., A quantitative technique for analysing robot-human interactions. In Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), (2002).

[7] Dautenhahn, K., Werry, I., Rae, J., Dickerson, P., Stribling, P., and Odgen, B. Robotic playmates: Analysing interactive competencies of children with autism playing with a mobile robot. In Dautenhahn, K., Bond, A., Canamero, L. and Edmonds, B. eds. Socially Intelligent Agents: Creating Relationships with Computers and Robots, Kluwer Academic Publishers, 2002.

[8] Draper, J., and Blair, L., Workload, flow and telepresence during teleoperation. In Proc. IEEE International Conference on Robotics and Automation (ICRA), (1996), 1030-1035.

[9] Draper, J.V., and Kaber, D.B. Human-Robot Interaction. In Mital, A., Ayoub, M.M., Kumar, S., Wang, M.-J. and Landau, K. eds. Industrial and Occupational Ergonomics: Users' Encyclopedia (Encyclopedia of Ergonomics), 1999.

[10] Drury, J., Scholtz, J., and Yanco, H., Awareness in human-robot interactions. In Proc. IEEE International Conference on Systems, Man and Cybernetics, (2003).

[11] Endsley, M. Measurement of situation awareness in dynamic systems. Human Factors, 37 (1995), 65-84.

[12] Fong, T., Nourbakhsh, I., and Dautenhahn, K. A survey of socially interactive robots. Robotics and Autonomous Systems, 42 (2003).

[13] Fong, T., and Thorpe, C. Vehicle teleoperation interfaces. Autonomous Robots, 11 (2001).

[14] Fong, T., Thorpe, C., and Baur, C. Robot, asker of questions. Robotics and Autonomous Systems, 42 (2003).

[15] Goetz, J., and Kiesler, S., Cooperation with a robotic assistant. In Proc. Computer-Human Interaction, (2002).

[16] Goldberg, K., and Siegwart, R. Beyond Webcams: Introduction to Online Robots. MIT Press, Cambridge, MA, 2002.

[17] Goodrich, M., and Olsen, D., Seven principles of efficient human robot interaction. In Proc. IEEE International Conference on Systems, Man and Cybernetics, (2003), 3943-3948.

[18] Granda, T., Kirkpatrick, M., Julien, T., and Peterson, L., The evolutionary role of humans in the human-robot system. In Proc. Human Factors Society 34th Annual Meeting, (1990), 664-668.

[19] Gutwin, C., Effects of network delay on group work in shared workspaces. In Proc. European Conferences on Computer Supported Work, (2001).

[20] Hart, S., and Staveland, L. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In Hancock, P. and Meshkati, N. eds. Human Mental Workload, North-Holland Elsevier Science, 1988.

[21] Huang, H., and Messina, E., Toward a generic model for the autonomy levels for unmanned systems. In Proc. NIST Performance Metrics for Intelligent Systems Workshop, (2003).

[22] Hughes, S., and Lewis, M., Robotic camera control for remote exploration. In Proc. Conference on Human Factors in Computing Systems (CHI), (Vienna, Austria, 2004), 511-517.

[23] Jacoff, A., Messina, E., and Evans, J., A reference test course for autonomous mobile robots. In Proc. SPIE-AeroSense Conference, (Orlando, FL, 2001).
[24] Jacoff, A., Messina, E., and Evans, J., A standard test course for urban search and rescue robots. In Proc. NIST Performance Metrics for Intelligent Systems Workshop, (2000).
[25] Kaber, D., Onal, E., and Endsley, M. Design of automation for telerobots and the effect on performance, operator situation awareness and subjective workload. Human Factors & Ergonomics in Manufacturing, 10 (2000), 409-430.
[26] Kaber, D.B., and Chow, M.-Y., Human-robot interaction research and an approach to mobile-telerobot interface design. In Proc. XVth Triennial Congress of the International Ergonomics Association (CD-ROM), (Seoul, Korea, 2003).
[27] Kaber, D.B., Riley, J., Zhou, R., and Draper, J.V., Effects of visual interface design, control interface type, and control latency on performance, telepresence, and workload in a teleoperation task. In Proc. XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society, (2000), Human Factors and Ergonomics Society, 503-506.
[28] Lee, J., and Moray, N. Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, 40 (1994), 153-184.
[29] Lee, J., and See, K. Trust in automation: designing for appropriate reliance. Human Factors, 46 (2004), 50-80.
[30] Lewis, M., Wang, J., Manojlovich, J., Hughes, S., and Liu, X., Experiments with attitude: attitude displays for teleoperation. In Proc. IEEE International Conference on Systems, Man, and Cybernetics, (2003).
[31] Macedo, J., Kaber, D., Endsley, M., Powanusorn, P., and Myung, S. The effects of automated compensation for incongruent axes on teleoperator performance. Human Factors, 40 (1999), 541-553.
[32] MacKenzie, S., and Ware, C., Lag as a determinant of Human performance in interactive systems. In Proc. ACM Conference on Human Factors in Computing Systems (INTERCHI), (New York, NY, 1993), ACM SIGCHI, 488-493.
[33] Newman, W., and Lamming, M. Interactive System Design. Addison-Wesley, Boston, 1995.
[34] Nielsen, C.W., Ricks, B., Goodrich, M.A., Bruemmer, D., Few, D., and Walton, M., Snapshots for semantic maps. In Proc. IEEE International Conference on Systems, Man and Cybernetics, (The Hague, The Netherlands, 2004).
[35] Olsen, D., and Goodrich, M., Metrics for evaluating human-robot interactions. In Proc. NIST Performance Metrics for Intelligent Systems Workshop, (2003).
[36] Parasuraman, R. Human use and abuse of automation. In Mouloua, M. and Koonce, J. eds. Human-Automation Interaction, Lawrence Erlbaum Associates, 1997.
[37] Parasuraman, R., Sheridan, T., and Wickens, C. A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man and Cybernetics B, 30 (2000), 286-297.
[38] Rantanen, E., and Nunes, A., Taxonomoies of measures in air traffic control research. In Proc. International Symposium on Aviation Psychology, (2003).
[39] Ricks, B., Nielsen, C.W., and Goodrich, M.A., Ecological displays for robot interaction: A new perspective. In Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), (Sendai, Japan, 2004).
[40] Rodriguez, G., and Weisbin, C. A new method to evaluate human-robot system performance. Autonomous Robots, 14 (2003), 165-178.
[41] Sanders, M., and McCormick, E. Human Factors in Engineering and Design. McGraw-Hill, New York, NY, 1993.
[42] Scholtz, J., Theory and evaluation of human robot interactions. In Proc. Hawaii International Conference on System Science 36, (2003).
[43] Scholtz, J., Antonishek, B., and Young, J., Evaluating human-robot interfaces: development of a situational awareness assessment methodology. In Proc. Hawaii International Conference on System Science 37, (2004).
[44] Scholtz, J., Antonishek, B., and Young, J., Evaluation of operator interventions in autonomous off-road driving. In Proc. NIST Performance Metrics for Intelligent Systems Workshop, (2003).
[45] Scholtz, J., Young, J., Drury, J., and Yanco, H., Evaluation of human-robot interaction awareness in search and rescue. In Proc. IEEE International Conference on Robotics and Automation (ICRA), (2004).
[46] Schulte, J., Rosenberg, C., and Thrun, S., Spontaneous, short-term interaction with mobile robots in public places. In Proc. IEEE International Conference on Robotics and Automation (ICRA), (1999).
[47] Sheiknainar, M., Kaber, D., and Chow, M.-Y. Control gain adaptation in virtual reality mediated human-telerobot interaction. Human Factors & Ergonomics in Manufacturing, 15 (2005), 259-274.
[48] Sheridan, T. Telerobotics, Automation, and Human Supervisory Control. MIT Press, Cambridge, 1992.
[49] Steinfeld, A., Interface lessons for fully and semi-autonomous mobile robots. In Proc. IEEE International Conference on Robotics and Automation (ICRA), (2004).
[50] Tittle, J., Woods, D., Roesler, A., Howard, M., and Phillips, F., The role of 2-D and 3-D task performance in the design and use of visual displays. In Proc. Human Factors and Ergonomics Society's 46th Annual Meeting, (2002).
[51] Verma, V. Anecdotes from Rover Field Operations, Unpublished, 2000.
[52] Wilson, G. Applied use of cardiac and respiration measures: Practical considerations and precautions. Biological Psychology, 34 (1992).
[53] Wilson, G. Real-time adaptive aiding using psychological operator state assessment. In Harris, D. ed. Engineering Psychology and Cognitive Ergonomics, Ashgate, Aldershot, UK, 2001.
[54] Yanco, H., and Drury, J., A taxonomy for human-robot interaction. In Proc. AAAI Fall Symposium on Human-Robot Interaction, (2002), 111-119.
[55] Yanco, H., Drury, J., and Scholtz, J. Beyond usability evaluation: analysis of human-robot interaction at a major robotics competition. Human-Computer Interaction, 19 (2004), 117-149.