Attaining Situational Awareness for Sliding Autonomy

Brennan P. Sellner
Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA USA
bsellner@andrew.cmu.edu

Laura M. Hiatt
Computer Science
Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA USA
lahiatt@cs.cmu.edu

Reid Simmons,
Sanjiv Singh
Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA USA
reids@cs.cmu.edu,
ssingh@cs.cmu.edu

ABSTRACT

We are interested in the problems of a human operator who is responsible for rapidly and accurately responding to requests for help from an autonomous robotic construction team. A difficult aspect of this problem is gaining an awareness of the requesting robot's situation quickly enough to avoid slowing the whole team down. One approach to speeding the initial acquisition of situational awareness is to maintain a buffer of data, and play it back for the human when their help is needed. We report here on an experiment to determine how the composition and length of this buffer affect the human's speed and accuracy in our multi-robot construction domain. The experiments show that, for our scenario, 5 - 10 seconds of one raw video feed led to the fastest operator attainment of situational awareness, while accuracy was maximized by viewing 10 seconds of three video feeds. These results are necessarily specific to our scenario, but we feel that they indicate general trends which may be of use in other situations. We discuss the interacting effects of buffer composition and length on operator speed and accuracy, and draw several conclusions from this experiment which may generalize to other scenarios.

Categories and Subject Descriptors: I.2.9 [Artificial Intelligence]: Robotics – *Operator interfaces*; H.1.2 [Models And Principles]: User/Machine Systems – *Human factors*

General Terms: Experimentation, Human Factors

Keywords: Situational Awareness, User Study, Sliding Autonomy, Case Study

1. INTRODUCTION

It is impossible to create a robotic system that can operate in an open, dynamic environment with zero failures. There are always unexpected error conditions, which the robot's programmers were unable to anticipate. In such cases, a purely autonomous system is left with little recourse: it can attempt a few generalized recovery strategies, but there will always be corner cases that prevent task completion.

HRI'06, March 2–4, 2006, Salt Lake City, Utah, USA.

Rather than attempting to account for every possible error condition, and accept a terminal failure if we encounter an unexpected case, we believe that autonomous robots or teams of robots should instead be designed to recover from common failures and merely recognize, at least in a general sense, rare failures. Most failures can be detected at some level, such as exceeding a limit, but sufficient information for automated recovery is often not available.

Simply detecting errors is of little use to the autonomous system without a way to translate the error into a recovery method. Since it's impractical to do so in code for all conceivable sets of recovery methods and errors, we solve this dilemma by involving one or more humans in the team. When a robot believes it is in trouble, but has no way of recovering on its own, it may request help from a human team member. This takes advantage of the strengths of the autonomous system, such as its ability to quickly and accurately perform routine, repetitive, or long-duration tasks, while using the flexibility of the human partner on an as-needed basis to fill in the gaps in the autonomous system's capabilities.

Our domain is multi-robot construction teams, which is rife with opportunities for coordination and collaboration between robots and humans. The current scenario involves three heterogeneous robots working together to assemble a square structure out of four beams and four connecting nodes. The nodes, mounted on wheeled bases, must be braced before a beam may be connected. This results in three natural roles within the team: a bracing robot, a robot to perform the fine manipulation necessary to connect a beam to a node, and a third to provide a mobile sensor platform.

Within our construction domain, a human teleoperator is available to help the system as needed. We have previously [4] compared teleoperation, autonomy, and two mixed (Sliding Autonomy) approaches. The two mixed approaches varied as to whether the human was allowed to constantly monitor the system's progress and proactively assume control. We investigate here the case in which the human is occasionally asked to provide assistance while they are performing other unrelated tasks. The challenge for the human in such a scenario is to swiftly attain situational awareness when they are called upon to intervene. By situational awareness, we mean an understanding of: the robotic team's and workspace's current state, how far the team has progressed in the assembly process, what caused the team to ask for help, and what is an appropriate next action. There are clearly degrees of situational awareness: knowing that the robot is in the assembly area is clearly different than knowing its precise position relative to the current beam. For the purposes of this paper, we define situational awareness as sufficient understanding of the robot's (or robots') state to formulate a short-term plan of action. This will vary somewhat on a case-

by-case basis, but in general will only involve comprehension of the spatial relationships between the small set of objects and robots directly involved in the current task.

Attaining situational awareness in this scenario is nontrivial, since the operator doesn't continually monitor the team's progress while fulfilling their other responsibilities. It is often quite difficult to determine the problem which triggered the request for help by simply observing the team's current (static) state. For instance, in a high-clutter environment the cause of the request for help may be ambiguous due to limited camera angles and a lack of depth perception. While moving the robots could help to remove this ambiguity, this may not be safe if they are in close proximity to obstacles. One approach to helping the human safely attain situational awareness is to provide a playback of data for some amount of time preceding the request for help - in other words, maintain a buffer of the system's recent activity and display it to the human when help is needed. This helps remove some of the scene's ambiguity, and also provides information about the team's recent actions, which may further help in determining the current problem.

We have conducted an experiment in order to investigate how the length of this buffer and different combinations of camera angles and synthesized views affect the human's acquisition of situational awareness within our scenario. The experiment showed that for our scenario 5 - 10 seconds of one raw video feed yields the fastest response from users, while 10 seconds of three video feeds results in the most accurate responses. While these specific results are necessarily tightly tied to our scenario, we feel that they indicate trends which may be of use in other Sliding Autonomy scenarios.

2. RELATED WORK

There has been a significant amount of work over the years on helping humans maintain situational awareness in a number of different scenarios. Much initial work focused on maintaining the situational awareness of pilots [3] [9], while more recent research has investigated the maintenance of situational awareness during teleoperation of search and rescue robots [12]. The primary focus of the existing situational awareness literature is on maintaining the awareness of an operator who is in continual control (or at least is continually monitoring) a robot or robots. This is in contrast to our domain, in which we are interested in helping the operator repeatedly attain situational awareness without monitoring the system between interaction episodes.

One model of situational awareness applicable to our domain is that proposed by Endsley [1] [2]. That model defines three levels of the situational awareness: Level 1 (perception of environmental elements) consists of basic perception of cues: an operator who has achieved Level 1 situational awareness has successfully comprehended the bits and pieces of information available to them. Level 2 (comprehension of current situation) integrates the data perceived in Level 1 and, once achieved, allows the human to derive task-relevant meaning from the raw data perceived by Level 1. The final stage of situational awareness, Level 3 (projection of future states), involves the projection of the future state of the system. Operators who have achieved Level 3 situational awareness are able to predict future system behavior, including the system's likely reaction to their input, from current events and dynamics perceived in Levels 1 and 2. In our experiment, we attempt to determine how long it takes subjects to attain Level 2 situational awareness when provided with differing sources and amounts of historical information.

Teleoperation systems enhance operator situational awareness in different ways. Many systems provide multiple viewpoints through both external views and views from cameras on the robot(s) themselves. Wang and Milgram [11], however, have tried to limit the mental workload required to reconcile those two views by creating a new type of viewpoint. Their view, called a "tether" view, is a display that is neither external nor robot-oriented; instead, it combines the two by simulating how the scene would look from a kite flying behind the robot in the workspace.

In addition to viewpoint, situational awareness is also improved by studying how spatial information of the workspace should be presented to the operator. Lasswell and Wickens [7] investigated ways to improve information displays for pilots in order to improve taxi-way safety and traffic flow. They showed that a 3-dimensional, perspective-view of the workspace reduced lateral tracking errors, but that a 2-dimensional, plan-view of the workspace supported greater taxi speeds. This suggested that by giving the 2D display a wider field of view, operators would be able to get a better feel for their situation as well as benefit from the advantages of a plan-view of the workspace. Presumably, techniques such as this that are used to provide a greater degree of situational awareness could also be applied to the situations we are looking at, in which the operator is attempting to gain, not maintain, awareness of the workspace.

A different area of science pertinent to our work is that of cognitive psychology. When presented with multiple visual displays, the question arises of how much information operators can process and remember at a time. The cognitive psychology community has done relevant work studying the limitations of human visual working (short-term) memory [10]. Studies have shown that working memory has severe limitations, and can only hold a few pieces of information at a time. What exactly limits how much information working memory can hold is still under investigation; possibilities include the number of objects attended to, the number of features those objects have, and so forth. Regardless, this research could be very helpful in the design of operator interfaces, as it is important to provide operators with enough, but not too much, information.

3. SITUATIONAL AWARENESS AND SLIDING AUTONOMY

We now provide an overview of the form of Sliding Autonomy we investigate in this paper and discuss our approach to helping the human attain situational awareness. As part of the Sliding Autonomy discussion, we examine how the autonomous component decides when to request help from the human component and how an understanding of a human's ability to quickly attain situational awareness can affect this decision. In order to arrive at an interface which will help the human swiftly achieve situational awareness, we look at the limitations of the current approach and some of the questions that arise when designing an interface for this purpose.

Our current research deals primarily with Sliding Autonomy, which addresses the question of how best to meld the complementary talents of human teleoperators and autonomous control systems via various combinations of relative autonomy. On one end of the autonomy spectrum lies pure teleoperation, in which the human teleoperator is in complete control of every aspect of all robots. In general, teleoperation is reliable, but is slow and imposes significant workload on the human operator. The other extreme is pure autonomy, in which the robotic team acts on its own with no human involvement whatsoever. Although pure autonomy is often much faster than teleoperation, it is significantly less robust, especially in dynamic domains where all failure modes cannot be determined *a priori* [4].

Many robotic systems that allow human involvement are limited to these two opposing modes - teleoperation and pure autonomy - with few, if any, options in between. Our goal is a system which melds the respective benefits of teleoperation and pure autonomy

while avoiding their associated shortfalls. In this experiment, we examine a Sliding Autonomy mode in which the autonomous system requests help as needed from a human performing other, unrelated, tasks.

When operating in this mode, the autonomous system maintains models of both its own past performance and any available human teleoperators' skills. These models allow the autonomous system to decide whether to request help from the human due to its belief that the human will be more efficient. In addition, a robot may request help if it believes it will be unable to recover from a detected failure. The goal of this mode of operation is to allow the autonomous system to increase its robustness by taking advantage of the human's skills and flexibility at need while not unduly loading them. The net effect is a mode that is more reliable than pure autonomy and faster than pure teleoperation. The human may also attend to other tasks while the autonomous system is operating, since they do not need to constantly monitor the team.

Here, we investigate how the composition and amount of data presented to the operator affect their ability to swiftly and efficiently acquire situational awareness when transitioning from an unrelated task. Addressing this problem yields two benefits: a more efficient user interface, which allows the human to attain situational awareness more quickly; and a more accurate model of how long this acquisition will take, which allows the autonomous system to make more informed decisions about whether to request assistance with a particular tasks. These benefits combine to yield a more efficient human-robot team.

The standard operator interface in this situation is identical to that of a full-time teleoperator: it reflects the current state of the robots, and provide information about the workspace in a variety of forms. However, such an approach suffers from two major problems when the operator has not been monitoring the team prior to the request for help. The first is its stationary nature: a limited, static view of the world results in ambiguity about the spatial relationships between objects and robots, especially in a high clutter environment with limited camera angles. The canonical solution is to shift a robot's viewpoint. However, if the robot is near obstacles which the human is unable to accurately localize, such uninformed motion could prove dangerous. A static view of the world also hampers the acquisition of situational awareness. As discussed in [2], time and the perception of temporal dynamics greatly influences the acquisition of Level 2 and 3 situational awareness.

A second limitation of a "traditional" teleoperation interface in this scenario is its lack of history and its lack of support for determining the intentions of the autonomous agents. One aspect of attaining situational awareness is determining what action the robots were attempting to perform when they asked the human for help. This is often not obvious from a static view of the world: for instance, if a manipulator is in contact with an object, is it having trouble picking it up, or accurately placing it?

One approach to easing the acquisition of situational awareness is to maintain a buffer of information, which can be played back to the human when operator assistance is required. This allows the human to "get up to speed" by allowing them to view information about the system as it approached the state that triggered the request for help. This makes it possible to both infer the robot(s) intentions and avoid needless motion through potentially hazardous terrain. The two obvious questions for implementing such a buffer are (1) what data is most relevant to attaining situational awareness, and (2) how much data it should buffer.

One may be tempted to include all available data in the interface on the theory that more data results in greater information intake by the human. However, the human brain's ability to winnow information from chaff is distinctly finite, and the operator will quickly become overwhelmed [2]. Not only are the human's resources finite, but the ability of the robots to store and transmit large amounts of data is also restricted, especially when the robots are not collocated with the human. Bandwidth is always limited, and serves as a firm upper cap on the amount of data which may be presented to the human.

Alternatively, the human could be given their choice of display elements. However, since the human cannot know which are the most useful data streams *a priori* and the system cannot know which streams the human will choose, performance would suffer greatly. The human would often not have the proper data available, and the system would be unable to perform effective caching due to bandwidth limits, resulting in large delays before the human would be able to intervene.

The question of how much data to buffer is also complex. In some scenarios, "key events" will exist for some or all error conditions, which, when observed in their entirety, will allow the operator to identify the error. In such situations, it is important that the data buffer contain the entirety of these events, and the length of the buffer will not be correlated to accuracy.

However, it is often the case that the key events are not observable. In such situations, the operator must observe a playback of the evolving system to determine which, if any, error has occurred, as the signs are often subtle and ongoing. If observable key events do not exist in a scenario, the length of data buffer is of vital importance, and the best length is a function of the scenario itself, the speed of the robots, and the human's ability to attain situational awareness. The vast majority of the examples used in this experiment did not contain observable key events.

While it may appear that a longer buffer will result in a greater degree of situational awareness, and thus a faster response time, this is not necessarily the case. We hypothesize that there is a point at which longer playbacks provide no more useful information about the problem at hand. In addition, when considering efficiency, one must take into account not only how long it takes the human to react after viewing the playback, but also how long is spent watching the playback. There may be a tradeoff between playback time and time exclusively spent thinking about the situation, and it is possible that the optimal overall reaction time is not necessarily the case resulting in the minimum time devoted exclusively to cognition.

Although the specific answers to the questions of data relevance and buffer length are in part task- and domain-dependent, there are some principles that apply to a range of similar domains. In order to investigate these principles, we evaluate them using our own construction domain and robot team, and discuss our specific results and how they may apply to other similar scenarios.

4. EXPERIMENTAL DESIGN AND METHODOLOGY

We conducted an experiment to assess how much and what types of information shown to a human operator correlate with how quickly the operator is able to gain situational awareness. In the experiment, the subjects were asked to observe a set of prerecorded data streams and then determine both why the autonomous system requested assistance and identify an appropriate action. Between trials, we varied which data streams were available to the subject, as well as the streams' length. When not responding to a simulated request for help, the subject performed a concentration-intensive distractor task to simulate a multitasking operator.

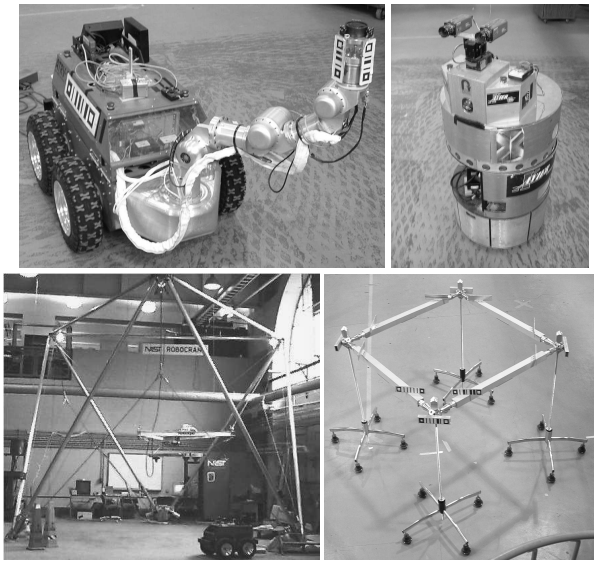

Figure 1: The Mobile Manipulator (top left), Roving Eye (top right), RoboCrane (bottom left), and the completed structure (bottom right).

4.1 Scenario and Robots

The assembly scenario used in this experiment involves four beams and four planarly compliant nodes that are assembled into a square structure (Figure 1). In order to weakly simulate conditions in space, the nodes are supported by casters that roll easily along the floor. Thus, bracing of the nodes is required before the end of a beam may be inserted into the node.

We have decomposed this scenario into tasks that can be completed by agents fulfilling three different roles: an agent that provides information about the state of the world (the Roving Eye; Figure 1), an agent that braces the nodes during docking (the Crane; Figure 1), and an agent that does the actual manipulation and insertion of the beams into the nodes (the Mobile Manipulator; Figure 1). Neither the Crane nor the Mobile Manipulator possess any extrinsic sensors and must rely on positional data transmitted to them by the Roving Eye, which is equipped with stereo cameras.

4.2 Interface

The information streams available to the human include three video feeds and one synthesized "technical drawing"-style visualizer (Figure 2). The video feeds are from one of the Roving Eye's cameras, a fisheye camera placed in the Crane looking down onto the workspace, and a stationary external camera placed outside the workspace looking towards the structure. The Roving Eye's stereo pair is also used to estimate the relative positions of the various objects in its field of view [5]. This information is in turn used by the visualizer to display the relative positions of the beam and node from above and in front of the beam ((4) in Figure 2). This provides data that is at times not otherwise available to the user, due to the lack of depth perception from single cameras. However, as a result of the data's noisiness and the autonomous system's reliance on it, the visualizer is never used alone, in order to give the operator an opportunity to recover from errors resulting from data corruption. Neither the camera on the Crane nor the external camera are utilized by the autonomous system.

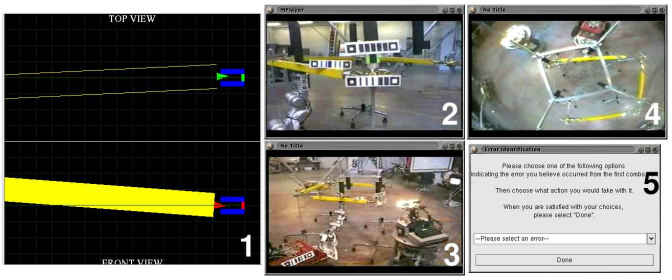

Figure 2: The subject interface, including three video streams (the Roving Eye's cameras (2), an external camera (3), and a Crane-mounted camera (4)), a synthesized view of the beam and node (the visualizer) (1), and the error categorization dialog (5). The "minimal video" conditions incorporate (2), while the "maximal video" conditions incorporate (2), (3), and (4).

4.3 Experimental Design

Our two experimental variables are the composition and length of the data feed that is presented to the subject. We chose to investigate four lengths (0, 5, 10, and 20 seconds) and four different combinations of available data (see Figure 2):

1. Minimal video: Roving Eye video only (Min vid)
2. Minimal video + visualizer: Roving Eye video and the visualizer (Min vid + viz)
3. Maximal video: Roving Eye, Crane overhead, and external videos (Max vid)
4. Maximal video + visualizer: Roving Eye, Crane overhead, and external videos, as well as the visualizer (Max vid + viz)

This yields a total of 16 different test conditions. Since all possible combinations of data feeds could not feasibly be investigated, these combinations were chosen to allow the comparison of minimal data (1, above) against maximal data (4, above), as well as several points in between. They are then evaluated in conjunction with playback time of varying lengths, ranging from static feedback (the 0 second playback condition) to 20 seconds of feed.

The example requests for assistance used in the experiment were drawn exclusively from the task of docking one end of a beam with a node. This is a precise manipulation task performed by the Mobile Manipulator and is rich in potential errors. This provided a large variety of situations in which the robots could request help, lowering the probability of the subject randomly guessing the correct answer. In addition, this placed all the examples within a subset of the domain, which made training subjects much more tractable.

During each trial, the subject was asked to identify why the robot requested help during the observed docking. The errors fell into four broad categories: false alarms, obscurements, interference by non-target objects, and interference by the target node. False alarms occurred when a successful docking had been incorrectly labeled an error by the system. An obscurement error occurred when the Roving Eye lost sight of the beam for any reason. Interference by non-target objects consisted of the beam hitting either another (non-target) node or the Crane's end effector. Finally, the beam could have become stuck in an undocked position on the target node, as a result of an error in the Roving Eye's data or the Mobile Manipulator approaching the node along an erroneous vector.

4.4 Experimental Procedure

The experimental procedure was a combination of training and testing. The subject's training began with reading a written overview of the task and hardware at hand, with the experimentor answering any questions.[1] The subject was then shown one example of each of the seven types of errors via the graphical interface (Figure 2), using the maximal video + visualizer and 20-second playback condition. The experimentor discussed each example with the subject in order to ensure the subject understood each error's characteristics. These training examples were not used during the test phase. If the subject did not feel fully trained by this point, the same training examples were repeated until the subject and experimentor felt the subject grasped the problem.

After training (which typically took 20-30 minutes), the following test procedure was used. The subject first played a Tetris-like game requiring significant concentration [8] for a time chosen from a normal distribution centered at one minute, with a standard deviation of 5 seconds. After this time had elapsed, the subject's display was switched to the interface (Figure 2), with the current condition's data streams visible. The length of playback associated with the current condition was immediately shown, with all displayed data streams synchronized. Once playback was complete, all interface elements were frozen on the final frame of the playback buffer and the classification entry dialog was displayed (note that this prevented the subject from choosing an error prior to the completion of playback). As soon as the subject selected one of the seven error classifications, they were returned to the distractor task. The actual time elapsed during playback and the time taken to classify the error. The time needed for playback plus error classification is a fairly tight upper bound on the time needed to attain Endsley's Level 2 situational awareness (comprehension) [2].

Each user was tested on four of the 16 conditions, with six example errors chosen per condition. Each set of six examples was chosen randomly without replacement from a pool of 29. In order to ensure that no error type occurred more than once per test condition, when an example was picked and removed from the pool, all other examples of the same type were marked. Marked examples were removed from the pool until the six examples for the current test condition were selected, after which they were unmarked and returned to the pool. The entire experiment, including training and testing phases, consumed an average of 1.5 hours per subject.

To account for ordering effects, we applied Latin squares to both effects and ran the combined conditions. A Latin square is a statistical technique which allows experimentors to test effects while controlling for two other known sources of variation (here, inter-subject variability and ordering effects). Since each subject was evaluated under four test conditions, 16 subjects were required to cover all the possible orderings. We evaluated 32 subjects in all. Our subjects were students at Carnegie Mellon. None had prior experience with the task, and their backgrounds spanned the Carnegie Mellon student population.

5. RESULTS

During the experiments, we recorded the time it took subjects to classify each example, both including and not including the time it took them to watch the data feed, as well as the accuracy of their classifications. We define "response time" as the time between when playback of the data stream finished and when the users classified the current error via the dialog box. We now analyze this data in the context of the 16 test conditions (4 available data feeds by 4 lengths of playback) using a univariate ANOVA test.

[1]Since a significant portion of the training consists of interactions between the experimentor and the subject, a single experimentor conducted all the experiments, in order to avoid training bias.

Data feed composition	Response time		Playback + response time		Classification accuracy
	μ	σ	μ	σ	
Min vid	19.2	17.7	28.6	17.6	44.8%
Max vid	26.9	30.0	36.4	28.8	62.0%
Min vid + viz	20.4	16.5	29.8	16.3	53.6%
Max vid + viz	27.2	23.3	36.6	23.3	55.7%

Table 1: Error classification time (seconds) and accuracy (probability of correct response) as a function of data stream composition. Each row of this table contains data from all data feed length conditions and comprises the entirety of the corresponding data stream rows in Figures 3–5.

5.1 Data Feed Composition

The results show that the "minimal video" data feed condition resulted in the shortest response and playback + response times (Table 1) [2]. However, there was not a statistically significant difference (at a 95% confidence level) between the minimal video and minimal video + visualizer conditions in either case, according to the Bonferroni [3] post hoc test (Table 2). The presence of additional video feeds appeared to be an important factor, as classification times under both conditions incorporating maximal video were significantly slower than either minimal video condition (Table 2).

While fewer data feeds appears to be an advantage when it comes to eliciting a rapid response, this is not the case if one is concerned with accuracy. As can be seen from Table 1, accuracy peaks at a 62% likelihood of a correct classification for the "max vid" data feed composition. Note, however, that the only statistically significant difference is between the max vid and min vid conditions (Table 5). We may only draw the inference that additional video increases accuracy. We hypothesize that the visualizer may make up some of the lack of the min vid condition, while overloading the operator in the max vid + viz case. However, these hypotheses are only supported by trends, *not* statistically significant differences in the data. We attribute the relatively low accuracies achieved by our subjects to their unfamiliarity with the scenario and the low quality of the video feeds. Note that at no time does accuracy decrease below random choice (14.2%).

5.2 Data Feed Length

The results also confirmed our hypothesis that a longer video playback time results in a shorter user response time (Table 3) (recall that we define "response time" as the time elapsed between the end of the playback and the classification of the error). This trend was true for each condition, and is illustrated in Figure 3. However, using a Bonferroni post hoc test, significance was not found between the 5 and 10 or the 10 and 20 second conditions (Table 4). This is not unexpected, as the data feed lengths are closely spaced.

These trends differ when we examine the playback + response time data. Here, the 5 and 10 second playback conditions were the fastest, with the 0 and 20 second playbacks taking significantly longer (Table 3). We can see from Table 4 that the only insignificant differences are those between 0 and 20 seconds and 5 and

[2]In this and the following tables, we use μ to indicate the mean and σ to represent the standard deviation of the presented sample.

[3]This is one type of post-hoc test, used to examine whether there are significant differences between individual categories, such as minimal and maximal video.

Significance		Max vid + viz	Min vid + viz	Max vid
Response time	Min vid	**Y**	n	**Y**
	Max vid	n	**Y**	–
	Min vid + viz	**Y**	–	–
Playback + response time	Min vid	**Y**	n	**Y**
	Max vid	n	**Y**	–
	Min vid + viz	**Y**	–	–

Table 2: Significance between data stream conditions from a pair-wise Bonferroni post-hoc test at a 95% confidence level. Significant differences are denoted by a bold 'Y'. Note that the only two differences which are not significant are between the two maximal video conditions and the two minimal video conditions.

Data feed length	Response time		Playback + response time		Classification accuracy
	μ	σ	μ	σ	
still frame	35.0	26.5	37.4	26.5	42.7%
5 seconds	23.3	19.6	28.4	19.6	51.6&
10 seconds	18.7	21.3	28.8	21.3	58.3%
20 seconds	16.7	18.5	36.8	18.5	63.6&

Table 3: Error classification time (seconds) and accuracy (probability of correct response) as a function of data feed length. Each row of this table contains data from all data stream composition conditions and comprises the entirety of the corresponding data feed length rows in Figures 3–5.

10 seconds. This is unsurprising, as they represent the trough and peaks of the data feed length x playback + response time curve.

Data feed length has a clearly salutary effect on accuracy, as can been seen in Table 3. Unsurprisingly, increasing length results in greater accuracy, with only the neighboring length categories showing statistically insignificant differences (Table 5). This, when considered alongside the playback + response time trends, indicates that a complex nonlinear tradeoff may be made between overall response time and classification accuracy by the system designer or the autonomous system.

5.3 Interaction Effects

The ANOVA test also revealed trends in the interaction effects between the two experimental variables' effects on response time, with a significance of .083 and .084 in the response time and playback + response time cases, respectively. [4] Figures 3 and 4 illustrate these effects. The graph of response time alone (Figure 3) suggests that the maximal video and maximal video + visualizer conditions are more affected by video playback time than the other data stream conditions. It also shows that the data feed conditions have much less of an effect on users' mean response time when the playback video length is 20 seconds than when it is only a still frame. The graph that incorporates playback time (Figure 4) confirms these effects. These interactions are discussed further in the next section.

[4]This means that there is approximately an 8% chance that data feed length and composition independently affect response time. This is on the borderline of being statistically significant, as 5% is the commonly accepted upper bound for a level of significance.

Significance		20 seconds	10 seconds	5 seconds
Response time	still frame	**Y**	**Y**	**Y**
	5 seconds	**Y**	n	–
	10 seconds	n	–	–
Playback + response time	still frame	n	**Y**	**Y**
	5 seconds	**Y**	n	–
	10 seconds	**Y**	–	–

Table 4: Significance between data feed length conditions from a pair-wise Bonferroni post-hoc test at a 95% confidence level. Significant differences are denoted by a bold 'Y'

Significance: Error Classification Accuracy				
		Max vid + viz	Min vid + viz	Max vid
Data stream	Min vid	n	n	**Y**
	Max vid	n	n	–
	Min vid + viz	n	–	–
		20 seconds	10 seconds	5 seconds
Data feed length	still frame	**Y**	**Y**	n
	5 seconds	**Y**	n	–
	10 seconds	n	–	–

Table 5: Significance for error classification accuracy between the various conditions from a pair-wise Bonferroni post-hoc test. Significant differences are denoted by a bold 'Y'.

While data feed composition and length may interact with respect to response time, there are no interactions with respect to accuracy, as the ANOVA test on the accuracy data yielded a 0.462 interaction significance. This indicates there is a 46% chance the observed data could occur if data feed composition and length independently affect accuracy. This implies that if the designer's sole goal is accuracy, data feed composition and length may be "dialed in" independently of one another. This may be intuited to an extent from Figure 5.

6. DISCUSSION

The measured response times are an implicit measurement of situational awareness, as opposed to subjective (self-rating) and explicit (questionnaires administered during a suspension of the task) measurements [6]. Most established methods for subjectively or explicitly measuring situational awareness, such as SAGAT [3] and SART [9], are designed to measure ongoing situational awareness during a long-term task, and are thus not immediately applicable to the domain we are investigating. However, we believe our implicit measurements to be a good measure of the ease or difficulty

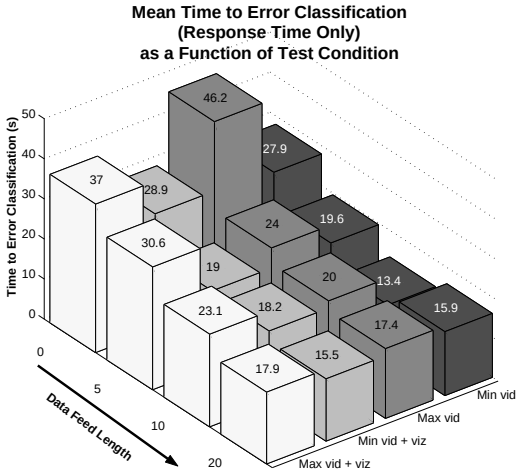

Figure 3: Average subject response time, by test condition

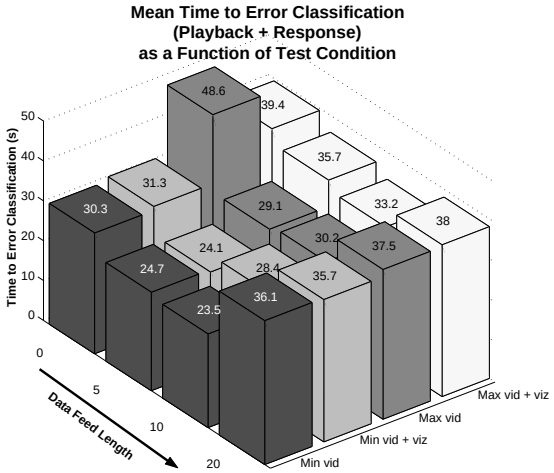

Figure 4: Sum of average subject playback + response times, by test condition

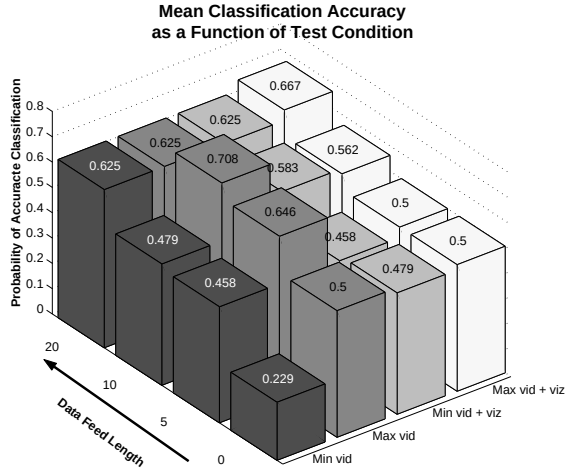

Figure 5: Mean error classification accuracy, by test condition. Please note that the ordering of the data feed length axis is reversed relative to Figures 3 and 4 for purposes of readability.

with which the subject attains situational awareness. By examining the time taken to classify the current error, we are able to directly measure how long it takes to attain Level 2 (comprehension) situational awareness, using Endsley's model [2]. Since the subject has been performing a high-concentration distractor task for longer than a trial's length prior to each trial, we are confident that we are capturing the entire process of attaining situational awareness. In addition, because the subject's sole task once a trial begins is to determine the error, we also believe that we are not overestimating the response time, except by the relatively constant factor of the time required to manipulate the classification interface. Thus, we believe that we are in fact able to capture a reasonably tight upper bound on the time taken to achieve Level 2 situational awareness in our domain.

The data from the minimal video and maximal video data feed composition conditions suggest that a simpler display (here, just the Roving Eye video) leads to the shortest response time, although this incurs a significant accuracy penalty. We believe that since subjects have limited visual information to consider while making their choice, it takes them less time to decide how to respond given the available information. When presented with more information, however, subjects took significantly longer to respond, suggesting that the extra videos add a significant processing overhead. While response time was significantly longer in the max vid case, accuracy also improved significantly (Table 5), indicating that a tradeoff must be made between response time and accuracy.

However, when the visualizer information is added to each of these conditions, user response time does not significantly increase. One possible explanation is that information presented by the visualizer is easier to process than information that must be extracted from the videos. This is especially true since the visualizer presents 3D information in a natural way, requiring little extra mental processing, whereas subjects are required to fuse the multiple video streams in order to extract 3D information from them. Another possible explanation is that subjects made little use of the visualizer, and instead concentrated their attention on the videos. In order to directly measure this, attention tracking data (such as that from a gaze tracking system) is needed.

While the addition of the visualizer has little impact on response time, it improves accuracy when added to the minimal video condition, but degrades it when added to maximal video (Table 1). Although neither of these effects is statistically significant, they may indicate an interesting trend. The visualizer provides information not available in the minimal video case, but may overlap significantly with the maximal video's content. This overlap results in information overload, with the operator less able to glean the relevant data. This may indicate that providing the same information in more than one form is in fact detrimental.

When considering the data feed length results for response time alone, we can at first glance see that the longer the data feed the operator is presented with, the better - there is a clear inverse relationship between the feed length and the time subsequently taken to select an error condition. While this is clearly not the metric of use to the autonomous system or system designer (both of which will instead use playback + response time), it serves to illustrate that the operator is still accumulating and processing additional useful information during the playback process. If we were to further extend the data feed length, we would expect the operator's response time to eventually plateau to the time needed to manipulate the classification interface. This hypothesis is supported by the nonlinear form of the data feed length x response time curve (see Figure 3).

Upon examining the more relevant playback + response time data (Figure 4 and Tables 3 and 4), we see that after a certain point increasing the length of the data feed does not result in improved classification time. This is because it takes longer to both watch the

display and respond than it does to watch a shorter display buffer and take a bit longer to respond. Thus, the data suggest that a playback in the range of 5 to 10 seconds will result in the quickest responses for this configuration. However, accuracy steadily improves as the length of playback increases. No one choice of playback length optimizes both response time and accuracy: this is a tradeoff that must be made by the system designer, and is a decision which can be guided by our results. If accuracy is imperative, a cost in additional time will need to be paid. While the specifics will vary between systems, it seems likely that similar playback length vs response time and playback length vs accuracy curves will exist in other teleoperation systems.

The interaction effects with respect to response time between the two experimental variables provided further insight into this situation. The increased sensitivity of the two data feed conditions that include all three video playbacks to data feed length supports the earlier conjecture that processing information in the form of raw video output takes longer than processing information from a simpler component, such as the visualizer. Also supported is our earlier theory about the human user's performance plateauing after viewing a certain length of playback. Because the difference between data stream conditions decreases as the playback time increases, it can be seen that users are becoming saturated with information, and that more information will most likely not decrease their response time any further. In addition, this indicates that it is possible to trade off playback time against the bandwidth required for the playback while maintaining a given level of performance.

7. FUTURE WORK AND CONCLUSIONS

Due to lack of hardware, we were unable to log one obvious aspect of how situational awareness is attained: which data streams the subject was actually attending to at any given moment. This means that we are unable to distinguish between whether the subject was attending to a video that was of little utility for the current error or the subject was attending to a useful video, but did not comprehend the error. One approach that would allow us to separate these factors is to add a gaze tracker to the system.

As mentioned above, the response time of users under increasing data feed lengths probably plateaus after a certain point. It would be interesting to study further what exactly determines where this point is, and whether it is affected by other factors, such as the selected data streams and the speed of the robots in autonomous mode. It may be that a human assisting a robot team which moves quickly requires less absolute time to achieve situational awareness. If this is the case, a slow-moving system may improve the human's response time by playing back the buffer at a faster rate.

While our specific results are unlikely to be of direct use in other scenarios, some relevant trends will be useful:

(1) There is unlikely to be a single configuration which optimizes both accuracy and response time. The system designer must decide whether he wants to optimize one or the other, or whether a compromise is in order. This is true when adjusting both data feed length and composition.

(2) While increasing data feed length improves accuracy, it will eventually stop decreasing response time and begin regressing. This is one of the tradeoffs that must be made. Similarly, increasing the amount of information available to the operator improves accuracy for a time. However, the operator eventually becomes inundated with data, at which point accuracy begins to fall off.

(3) Data feed length and composition cannot be controlled independently. In general, increasing playback length decreases the effect of data feed composition, especially with respect to accuracy. This allows the designer to trade off instantaneous bandwidth against the duration of bandwidth usage while maintaining a target accuracy level.

(4) Although duplicating data in different forms (e.g. video and visualizer) may intuitively seem to be helpful, it instead contributes to information overload. Care must be taken when selecting teleoperation interface elements. They should be considered with respect to each other as well as their stand-alone usefulness to the task in order to avoid needless duplication.

In this paper we have discussed the importance of quickly gaining situational awareness in systems involving sliding autonomy. In order for a human operator to be an effective team member in a system asking for help, they must be able to switch tasks, gain situational awareness, and diagnose the error as quickly as possible. We introduced various factors that affect a human operator's ability to quickly gain situational awareness of their workspace when shown a brief, partial history of the robots' movements. We varied which data feeds were shown and how long the feeds were, in an attempt to determine the optimal interface for our scenario and team. Human subject experiments have shown that 5-10 seconds one video feed results in the quickest response, while 20 seconds of three video feeds results in the most accurate operator response. This further illustrates the maxim that everything has a cost: the system designer must make tradeoffs between speed, accuracy, and bandwidth to build a system suited to the scenario at hand.

8. ACKNOWLEDGMENTS

This work has been supported by NASA grant NNA04CK90A.

9. REFERENCES

[1] M. R. Endsley. A methodology for the objective measurement of situation awareness. *Situational Awareness in Aerospace Operations (AGARD-CP-478)*, pages 1/1 – 1/9, 1989.

[2] M. R. Endsley. *Situation Awareness Analysis and Measurement*, chapter Theoretical Underpinnings of Situation Awareness: A Critical Review. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[3] M. R. Endsley, S. J. Selcon, T. D. Hardiman, and D. G. Croft. A Comparative Analysis of SAGAT and SART for Evaluations of Situation Awareness. In *42nd annual meeting of the Human Factors and Ergonomics Society*, Chicago, October 1998.

[4] F. W. Heger, L. M. Hiatt, B. Sellner, R. Simmons, and S. Singh. Results in Sliding Autonomy for Multi-Robot Spatial Assembly. In *8th International Symposium on Artificial Intelligence, Robotics and Automation in Space (iSAIRAS)*, Munich, Germany, September 5-8, 2005.

[5] D. Hershberger, R. Burridge, D. Kortenkamp, and R. Simmons. Distributed Visual Servoing with a Roving Eye. In *Proceedings of the Conference on Intelligent Robots and Systems (IROS)*, Takamatsu Japan, October 2000.

[6] A. T. Hjelmfelt and M. A. Pokrant. Coherent Tactical Picture. Technical Report CNA RM 97-129, Center for Naval Analyses, Alexandria, Virginia, 1998.

[7] J. W. Lasswell and C. D. Wickens. The Effects of Display Location and Dimensionality on Taxi-Way Navigation. Technical report, Aviation Research Laboratory, Institute of Aviation, University of Illinois at Urbana-Champaign, May 1995. Prepared for NASA.

[8] D. R. Nelson and A. Sayman. Crack Attack! Manual. Game's home page: http://www.nongnu.org/crack-attack/, http://aluminumangel.org.

[9] R. M. Taylor. Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design. *Situational Awareness in Aerospace Operations (AGARD-CP-478)*, pages pp. 3/1 – 3/17, 1990.

[10] E. K. Vogel, G. G. Woodman, and S. J. Luck. Storage of Features, Conjunctions, and Objects in Visual Working Memory. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1):92–114, 2001.

[11] W. Wang and P. Milgram. Viewpoint Optimization for Virtual Environment Navigation Using Dynamic Tethering - A Study of Tether Rigidity. In *47th Annual Meeting of the Human Factors and Ergonomics Society*, Denver, Colorado, 2003.

[12] H. A. Yanco and J. Drury. Where Am I? Acquiring Situation Awareness Using a Remote Robot Platform. In *IEEE Conference on Systems, Man and Cybernetics*, October 2004.