Photogeometric Sensing for Mobile Robot Control and Visualisation Tasks

Alonzo Kelly[1], **Dean Anderson**[1], **Erin Capstick**[2], **Herman Herman**[1], **Pete Rander**[1]

Abstract. Photogeometric sensing is a relatively new sensor modality that tightly integrates geometry and appearance sensing into a single package. Such a sensor produces imagery that encodes the appearance and the range to every sensed point in the scene. This new type of sensor enables much higher fidelity virtualized reality displays that can be produced in real time from the data gathered by a moving robot. Such displays exhibit several ideal characteristics for human robot interaction tasks that enable new approaches to supervisory control and remote visualization. Photogeometric sensors suitable for HRI applications cannot yet be purchased but they can be constructed by co-locating ranging and appearance sensors and combining the data at the pixel level. This paper outlines our approach to the construction of such sensors as well as their successful use in several applications.

1 INTRODUCTION

We will use the term *appearance* to refer to sensing modalities which are sensitive to the intensity of incident radiation including visible color, visible intensity, and infrared modalities. Conversely, *geometry* will be used to refer to modalities that register any of depth, range, shape, disparity, parallax, etc. The term *photogeometric* (PG) sensor will refer to a sensing device that produces both kinds of data in a deeply integrated manner. For our purpose in this paper, the data is deeply integrated if the spatial correspondences of the data are known. Ideally, as shown in Figure 1, the resolutions are matched as well so that a one-to-one mapping exists between geometry and appearance .

Figure 1: Photogeometric Data Set. Every color pixel in the left image has an associated range pixel in the right image.

The deep integration of appearance and geometry data can be a powerful technique for enabling effective human-robot interaction. In many applications, robots must understand the geometry of the environment in order to move around competently while avoiding collision. In such applications, geometry sensing is often the preferred modality of the robot designers. Conversely, humans process appearance data more readily and we can assimilate geometry perceptually only when it is converted to appearance data. For example, two stereo views or the parallax evident in a moving cloud of points on a computer screen will enable humans to perceive depth.

When images of both modalities are available – and their correspondence is known – it becomes possible to convert between the modalities relatively seamlessly. For supervisory control, such conversion makes it possible to extract accurate 3D coordinates when a pixel in a video stream is designated. For visualization, such conversion makes it possible to render synthetic views of the scene from arbitrary perspectives which may never have been the site of any real sensor.

The paper is organized as follows. Section 2 provides a broad overview of related work. Section 3 explains our technique for producing a photogeometric sensor. Section 4 describes an experiment using such sensing for mobile manipulator teleoperation. Section 5 describes an experiment using such sensing for outdoor mobile robot teleoperation. Section 6 provides a brief summary and outlook.

2 RELATED WORK

The notion of aiding an operator by displaying the perception data produced by a remotely controlled robotic system must have occurred to the first designers of such systems. Numerous techniques for supervisory control and teleoperation of manipulators, and even telepresence were clearly outlined as early as the mid 1980s [16]. The same concepts were considered early for legged vehicles [13] and wheeled mars rovers [3].

In broad terms, although perception data is nominally a view of the state of the environment, it is more properly described as a view of the robot's model of that environment. Hence, such data is equally a view into the internal state of the robot. It is natural for engineering displays to use such data during system development but it also quickly becomes clear that a good way to understand robot behavior is to know what it "thinks" it perceives in its immediate surroundings.

Of course, this mode of tele-operation depends on the use of adequate sensing. Military and consumer markets have driven the development of guidance systems and TV cameras that are relevant to mobile robots today. While laser ranging sensors are now commercially produced for factory robots, systems designed specifically for outdoor mobile robots are either single axis, immature products, or of inadequate performance for our purposes. For these reasons, our work continues a long tradition [11] of custom sensor development for lack of any alternative

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, 15213
[2] DCS Corporation. 1330 Braddock Place, Alexandria, VA, USA, 22314

which continues in robotics labs around the world up to the present time [12].

Given the sensor data needed, the earliest approaches simply displayed the raw sensor data or showed the robot in a 2D overhead view in the context of its surrounding perceived objects. Applications like space exploration generated a strong impetus to develop more realistic virtual displays as early as 1991 [8]. These systems were tested terrestrially [6] and derivatives were ultimately used on the Mars pathfinder mission. Contemporary developments include more emphasis on sensor fusion [5] as well as efforts which display both forms of data in a less integrated but more useable way [15].

Our work presented in this paper is consistent with trends to provide more realistic views for the purpose of real-time control. However, our work emphasizes the construction of a novel sensing device which performs sensor fusion at the pixel level. This device has been designed to be well suited to solving the problem of virtualizing a real environment in real-time.

3 IMPLEMENTING PG SENSING

At some point in the future, flash lidar devices may be available which share apertures with color cameras in order to produce photogeometric data in hardware. Until that day comes, we find the value of PG data to be worth expending effort to produce it in whatever manner we can today.

Our implementation approach centers on the goal of producing an integrated data set of appearance and geometry data from two different sensors. The data may be organized arbitrarily but our two most common formats are camera-derived color data augmented with range, *rangified color* (RC), and lidar-derived range data augmented with color, which we call *colorized range* (CR) data.

Computational stereo vision is a natural RC modality because range is produced for every pixel in the reference appearance image. However, its utility in applications can be limited due to the relatively poor quality of the range data. This is often the case in our applications. Flash lidar sensors also continue to advance [1] but none yet meet our requirements for operation in outdoor environments. Conversely scanning lidar devices have been our preferred geometric sensing modality for two decades. Nevertheless, we will discuss PG sensing where the range data is provided by a scanning or a flash lidar.

In general, every appearance modality can potentially be paired with every geometry modality. Ideally, each sensor of a pair would image the same region of the scene as the other at the same resolution and frame rate from the same position. In practice, numerous technical issues arise due to the different attributes of the two sensors including:

- **Projective Geometry**. Lidar is often spherical polar, whereas cameras (and flash lidars) provide a perspective projection.
- **Resolution**. Scanning lidar typically produces 1% of the angular resolution (solid angle) of a camera so there can be up to 100 camera pixels for each lidar measurement.
- **Field of View**. Standard camera lenses, spherical mirrors, and lidar scanning mechanisms rarely provide the same field of view.
- **Location**. Displacement of one sensor center of projection or emission relative to another leads to parts of one view missing from the other – even if all other parameters match.
- **Frame Capture and Beam Scanning**. In cases where data is gathered on the move, each point of lidar data is captured from a different sensor position whereas all pixels in a camera frame come from a single position.

3.1 Establishing Pixel Correspondences

A basic property of cameras is their projective geometry which projects a 3D scene onto a 2D photosensitive sensor array. While the azimuth and elevation coordinates in the image are related to the equivalent directions in the scene, information about the depth of objects is lost when a camera image is formed.

Hence, the most valuable attribute of PG imagery is its recovery of the depth dimension which is lost when a real scene is imaged with a camera. This information is recovered by:

- establishing an association of lidar range points with camera pixels
- geometric transformations to convert lidar data to camera coordinates

For RC data, the color data is augmented with depth so that the result is an augmented image. For CR data the range data is colorized and the result is an augmented range image or point cloud. In either case, the mechanism to establish correspondences is the same. Consider Figure 2 which expresses the essence of the problem when both sensors are viewed from overhead.

For now, suppose that both sensors are stationary with respect to the scene and let us define a lidar "image" to mean the data produced by one sweep over the scene of the lidar scanning mechanism.

While it is not clear how to directly map color pixels onto a lidar data set, the reverse operation is conceptually straightforward. Hence both RC and CR datasets rely on a common procedure to establish correspondences. Let the letter L designate a coordinate frame attached to the lidar center of emission and let the letter C designate one at the camera center of projection. The homogeneous transform matrix that converts coordinates of a point from frame L to frame C is denoted T_L^C. Let the letter I designate row and column coordinates in the camera image plane. The projective transformation matrix that provides the image coordinates of a 3D point will be designated P_C^I. The homogeneous dimension will be omitted from vectors unless the matrices are written out. Under this notation, the camera image coordinates $\underline{r}^I = \begin{bmatrix} x & y \end{bmatrix}^T$ of the point imaged by a lidar point $\underline{r}^L = \begin{bmatrix} x & y & z \end{bmatrix}^T$ are:

$$\underline{r}^I = P_C^I T_L^C \underline{r}^L \qquad (1)$$

If the scene has sufficient 3D (non-planar) structure, the spatial separation of the sensors introduces characteristic problems of triangulation:

- **Missing parts**. Even with perfect field of view overlap, surfaces oriented perpendicularly (and invisibly) to the viewing direction of one camera may be visible to the other.
- **Depth ambiguity**. It is possible for the lidar to have ranged to a point on a background object that is behind a foreground object which was imaged by the camera.

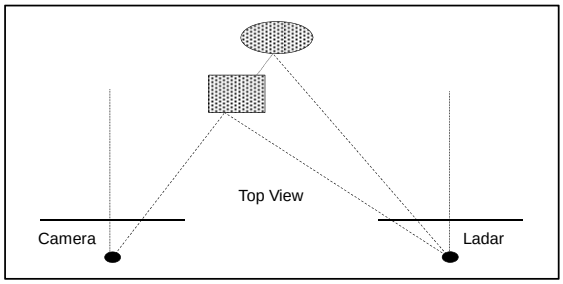

Figure 2: Multi Sensor Geometry and Depth Ambiguity. The camera measures the angle to objects whereas as the lidar measures angle and range. Due to the baseline separating the sensors, a lidar may image more than one object along the line of sight of a camera pixel.

While the first problem has no solution, the second can be solved by forming a depth buffer of all of the lidar data as viewed from the perspective of the camera image. All lidar data can be projected into bins which are sorted by depth or the processing may simply retain only the smallest range value in each bin. In either case, when two or more lidar pixels fall on the line through a given camera pixel, only the closest lidar point should be associated with the color pixel. All others are occluded and invisible to the camera so their color is unknown. While these triangulation issues cannot be eliminated entirely, they can be mitigated significantly by placing the two sensors very close together relative to the depths being imaged. However when the lidar is mounted on a moving vehicle, its continuous scanning process places limits on what can be achieved.

3.2 Forming Photogeometric Datasets

Given the correspondences between elements in each data set, either CR or RC data may be formed. The production of CR data using lidar is easiest to illustrate. In this case, the sensor intrinsic data format is a temporally ordered set of 3D points expressed in Cartesian or polar coordinates relative to the sensor center of emission. Each lidar point is simply augmented by the color of its associated camera pixel, if any. The color information might be the color of the closest camera pixel, the average over a region, or a block of pixels forming a small texture map.

In the case of RC data, the goal is to produce range data for every color pixel in a color image. Typical camera angular resolutions are 1 millirad whereas lidar is typically 10 millirad. Hence, once the lidar correspondences are computed, only 1% of the camera pixels can be expected to have associated lidar points. In other words, there will inevitably be holes in the coverage of the image by the range data. Small holes will be due to the reduced angular resolution of the lidar and larger ones due to missing parts or nonoverlapping fields of view.

When dense range data is desired, interpolation can be justified on the basis that the lidar is really providing the average range of the region of the scene that is spanned by a large number of camera pixels. The range data can be interpolated using the dilation operation of computer vision to fill small holes. The dilation radius can be related to the expected angular lidar footprint in the camera image. When both sensors are close together, the effect of surface orientation is minimal.

3.3 Sensor Configuration

Due to many considerations including the numerous robotic platforms that we construct annually and the desire to standardize solutions across programs, we have been continuously refining our photogeometric sensor concept for many years.

Two recent sensor designs are shown in Figure 3. For scanning lidars, we typically purchase an off the shelf scanning lidar which scans in one degree of freedom (called the *fast axis*) and then we actuate the housing in a second degree of freedom (called the *slow axis*) in order to produce a scanning pattern that spans a large angle in both azimuth and elevation. For flash lidars or stereo ranging systems, the interfaces to these devices are equal or similar to those of cameras so the process is more straightforward.

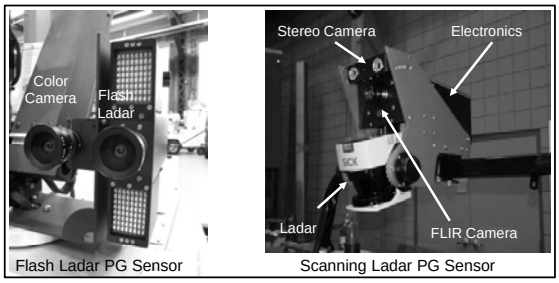

Figure 3: Two Custom Photogeometric Sensors. The device on the right fuses data from a commercial scanning lidar by SICK, stereo cameras, and a forward looking infrared (FLIR) camera. The device on the left fuses a PMD-Tec flash lidar with a color camera.

The interface to the composite device is a combination of fast Ethernet (used for high bandwidth data) and CAN Bus (used for low latency for control). One design goal is to render the composite device interface standard, high level, and easy to use.

The lidar pointing control system provides precisely timed feedback on the angle of rotation. This data stream is merged with the range and angle data coming from the lidar to form a 2D scanning lidar data stream. This stream is then optionally merged with any camera data and transmitted to the host computer system.

4 REMOTE MOBILE MANIPULATION

Mobile manipulation is a task for which human-robot interaction is often needed due to the difficulty of dexterous manipulation and the higher stakes associated with forceful interaction with the environment. While robots can often competently control their gross position, the final operations of the end-effector tooling may need to be performed with a human in the loop.

We recently conducted an effort to construct a mobile manipulation system that is analogous to commercial platforms and to endow it with a photogeometric sensor in order to study the benefits achievable when an operator designates a target to be manipulated on a video display [2]. Given the historical lack of range data, standard solutions to this problem include implementing a visual servo or using the robot navigation system to drive in the general direction designated until the operator issues a stop command.

However, in complex environments or in cases where the operator needs to direct attention elsewhere, it is more effective to have the robot decide when to stop. Furthermore, if autonomous obstacle avoidance is used, the robot can perform much more intelligently when it knows the precise 3D target for the manipulator end-effector. If manipulation and mobility are to be automatically coordinated, it again is necessary to know the 3D coordinates of the target. Hence, this is a case where it is valuable to have the human look at video while geometry (derived from the video) is communicated to the robot.

4.1 Platform Design

The robot used for these experiments was a modified LAGR mobile robot [9], fitted with a custom three degree-of-freedom manipulator arm and a gripper end-effector (Figure 4). The base vehicle has proved to be a very flexible research platform: in addition to over 40 standard models deployed at various universities, custom versions with LIDAR, metal detectors and omnidirectional cameras have been built.

The Photogeometric sensor consisted of two color video cameras, and the flash LIDAR unit of Figure 3 provided by PMDtec [14]. One camera is mounted on the manipulator arm near the wrist, for use during manipulation. The second camera and flash LIDAR unit is mounted to the shoulder yaw joint for driving and target acquisition.

The PMDtec sensor had a 30 Hz frame rate, a 64x48 range pixel array, a maximum range of 7.5 m, and was capable of operation in indoor and outdoor environments. The field of view was adjustable using standard C-mount optical lenses. For our experiments, a lens with a focal length of 4.2mm was used. This provided a 60 degrees horizontal and vertical field of view. By using a projective lens model, each pixel's range can be converted a 3D point in the workspace.

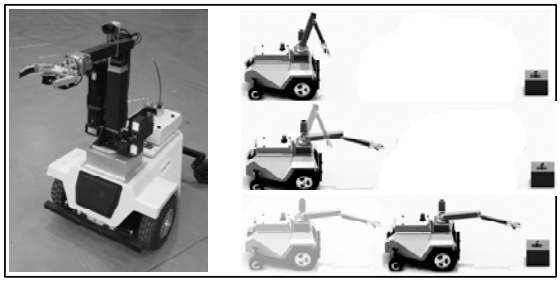

Figure 4: Test Platform for Click and Grasp. The LAGR robot platform was retrofitted with a manipulator and a custom photogeometric sensor.

4.2 Algorithm Design

We implemented a "Click and Grasp" function which allows an operator to click on a color image to designate a target, and have the system then a) recover the location in 3D space, b) navigate to within manipulation range, and then c) either grasp the target or put the end-effector as near as possible. In addition to our PG sensor, key aspects of the solution included control algorithms that coordinated the motion of the platform and the manipulator.

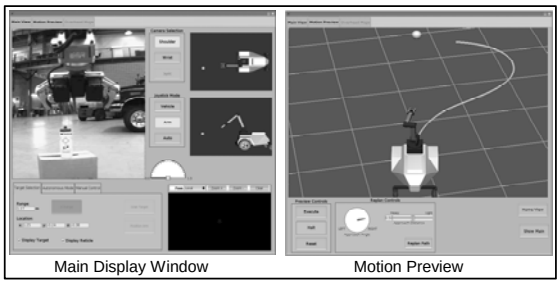

Figure 5: Operator Control Unit Display. The operator selects a target position in the left display. The system then plans and displays the entire motion and optionally asks for confirmation before execution.

4.3 Results

A custom graphical user interface was developed for the test (Figure 5). Two sets of experiments were performed to evaluate the effectiveness of the system in terms of increasing efficiency and usability.

In the first test, an object to be grasped was placed within reach of the manipulator and time to achieve the grasp was measured in various control modes.

Operator	Target 1 (secs. / errs)	Target 2 (secs. / errs)	Target 3 (secs. / errs)
Auto	28.27 / 0	22.52 / 1	24.20 / 0
Expert (JS)	41.92 / 0	40.84 / 0	45.73 / 0
Expert (WS)	47.73 / 0	32.75 / 0	45.37 / 0
Novice (JS)	38.43 / 0	36.95 / 0	32.70 / 0
Novice (WS)	29.21 / 1	22.85 / 0	33.36 / 0

Table 1: Pick Up of Object Within Reach of Manipulator. The operators were allowed to operate the manipulator in both joint-space (JS) and end-effector workspace (WS). "Auto" corresponds to using the automatic "click and grasp" system.

The automatic system was able to accomplish the task significantly faster than both (expert and novice) operators using manual teleoperation, and workspace controls significantly improved the completion time of the novice operator. On average, the automatic click and grab system was able to perform the static manipulation task 13% faster than both users.

Operator errors were noted whenever the test target was knocked over by the manipulator. The test was then reset and the operator permitted to retry. If a failure was due to insufficient grasping force at the end effector, it was not counted as an operator error. Our intent was to focus on characterizing the utility of the system as an aid to the precise positioning of the end effector.

In the second test, the object was outside the manipulator workspace so the platform had to be moved in either an automated or manual fashion.

Operator	Target 1 (secs./errs)	Target 2 (secs./errs)	Target 3 (secs.)	Target 4 (secs.)
Auto	37.08 / 0	68.43 / 2	54.61 / 0	56.21 / 1
Expert	48.37 / 0	40.77 / 0	54.75 / 0	43.00 / 0
Novice 1	45.42 / 0	48.98 / 0	40.52 / 0	32.42 / 0
Novice 2	49.52 / 1	49.43 / 0	43.17 / 0	35.48 / 0

Table 2: Pick Up of Object Outside Reach of Manipulator. The automatic "Click and Grab" system performed comparably to the human operators.

Work-space controls reduced both the time required to complete the task as well as the number of errors made. On average across four trials, operators reduced their number of errors from three to one, and reduced their completion time by 11%. The results demonstrate that the autonomy and perceptive capabilities of our system eases the workload on the operator while increasing task efficiency.

Results from the automatic system were potentially limited by the accuracy of the flash lidar range and co-registration. At short ranges, accuracy was sufficient to reliably grasp an object. However, at longer ranges, errors were large enough to cause manipulation errors. Instead, "click and grab" at long range required several operator interventions to re-designate the target once the base had positioned itself within range.

4 MOBILE ROBOT TELEOPERATION

Effective operation of any mobile platform without direct line-of-sight is intrinsically difficult to achieve. In video-based teleoperation, the loss of peripheral vision caused by viewing the world through the "soda straw" of a video camera reduces driving performance and increases the operator's frustration and workload. Wireless communication links are also subject to dropouts and high levels of latency. Their bandwidth limitations typically cause a large reduction in image quality relative to the fidelity of the underlying video cameras.

When the robot undergoes significant or abrupt attitude changes, the operator response may range from disorientation, to induced nausea, to dangerous mistakes. In contexts where the operator is also in danger, the need for high attention levels deprives operators of the capacity to pay attention to their surroundings. Wireless communications issues and difficulty controlling the robot also increases time on task and increases the time required to become a skilled operator.

4.1 PG Sensing for Autonomy

We have been working on improved operator displays for at least a decade on our robot autonomy programs [3]. PG sensing was originally motivated by its capacity to disambiguate natural obstacles and non-obstacles of the same shape (such as a rock and a bush) by examining their color signatures (see Figure 6). Once the data was available for use in autonomy however, we began to produce specialized point cloud displays and quickly recognized the potential of the PG data for human interfaces.

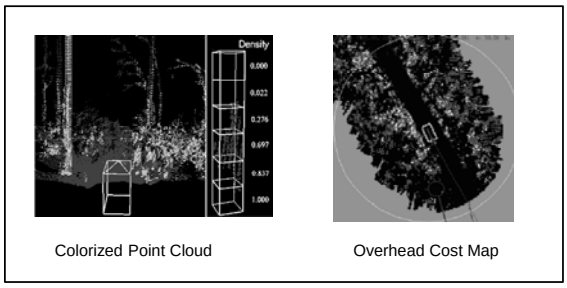

Figure 6: Original Engineering Displays of PG Data on Autonomy Programs. The display of traversability / cost or elevation from an overhead display (right) is traditional in robotics. In recent years, colorized point clouds have also been used. The evolution of the left figure toward photorealism was a natural extension of ongoing efforts.

4.2 3D Video

Photogeometric sensing enables a new capacity to address many of the problems described in the introduction to this section by providing a photorealistic, synthetic, line of sight view to the robot based on the content of geometry-augmented real-time video feeds. The operator experience is equivalent to following the robot in a virtual helicopter that provides arbitrary viewpoints including an overhead viewpoint and the over-the-shoulder view that is popular in video games.

The fusion of video and geometry produces a database whose content is much closer to a computer graphics rendering database than basic video. If the geometry is converted to faceted surfaces and the imagery is converted to textures, the PG sensor data has been converted to a rendering database. If this conversion is performed in real-time, a kind of hybrid *3D Video* is produced which can be viewed from arbitrary perspectives while exhibiting the photorealism and dynamics of live video.

If the sensors are omnidirectional and/or if the system remembers and integrates the rendering primitives over time, the net result is the real-time virtualization of the scene which enables the operation of the robot quite literally as if it were a 3D video game.

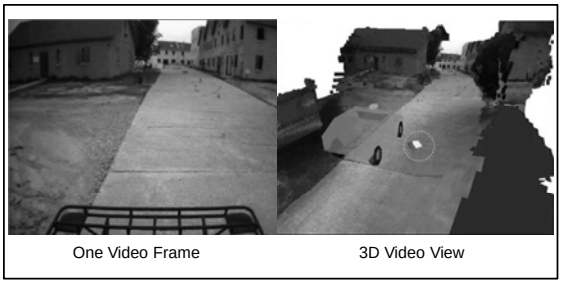

Figure 7: 3D Video View of a Mobile Robot. Left: A video frame produced from a camera on a moving vehicle. **Right**: The 3D Video view produced from all of the video that has been received in the last few seconds by the vehicle. The operator can

look at this database from any angle, at any zoom, while it continues to be updated in real time.

The ability to synthetically generate a viewpoint via computer graphics leads to the following capabilities:

- A natural mechanism to introduce virtual operator aids into the display.
- The capacity to zoom into objects of interest and view them from different perspectives (for example, from above or from the side) from just a few inches away.
- The capacity to have multiple operators cooperate from multiple views, perhaps even using cooperating robots.

5.3 Results

The goal of 3D Video technology is to increase an operator's situational awareness of the vehicle being controlled, thereby reducing operator errors and increasing the speed with which tasks are completed.

We conducted an operator performance assessment over a period of one week involving five operators of different skill levels. The participants averaged 20 years of automobile driving experience. Three subjects had prior experience teleoperating a live vehicle, including one with a 3D Video system. Two of these subjects had participated in one other experiment, while the other had extensive experience, teleoperating a vehicle in many experiments. Three subjects had minimal experience teleoperating a *simulated* vehicle (two of these included in the group with live vehicle experience). Four subjects had been playing driving-based video games for an average of 13 years, with one subject playing as often as a few times per week. One subject had never played a driving based video game.

The test platform was a John Deere eGator vehicle retrofitted for remote control and teleoperation. Participants completed four test conditions, which were counter-balanced across participants to minimize order effects related to course and Operator Control Station (OCS) familiarity:

1. Manually drive from seat
2. Basic Teleoperation with live video
3. Teleoperate with 3D Video - without motion prediction
4. Teleoperate with 3D Video - with motion prediction

Motion prediction refers to a method used to alleviate the effects of video latency. We use the most recent navigation state received from the robot and predict the robot position based on the terrain shape and the history of operator inputs. Due to latency in sending commands to the robot, the instant of time being predicted is not "now" but rather the moment in the future when the commands are predicted to arrive at the vehicle. In principle the display will then respond instantly to operator inputs and it will correspond to a point in time slightly ahead of where the vehicle is now. The availability of lidar data makes it possible to predict robot motion to relatively high accuracy compared to the alternative of ignoring the latency.

The course consisted of a paved roadway with traffic cones set up to guide drivers at particularly ambiguous areas such as intersections. Course features included slaloms, decision gates, and discrete obstacles as a series of loose and tight turns.

Performance metrics included course completion time, course accuracy, average speed and errors as well as subjective input on workload [7], impressions of the system and recommendations for future improvement. Errors were defined as hitting a cone, (having the vehicle emergency-stopped before) hitting an obstacle along the edges of the course (concrete barriers, fences, and hay bales occurred sporadically along the perimeter of the course), or deviating from the defined region of the course (driving off the road). In the end, course accuracy was not measurable due to data collection equipment availability.

Course Completion Time Results: As the figure below indicates, 3D Video enabled operators to complete the course faster than basic teleoperation: 3D Video alone led to completion times approximately 20% lower, while times were 30% lower when 3D Video was combined with motion prediction (MP). As expected, manual driving (in the vehicle) is still far superior, with course completion time approximately 75% lower than basic teleoperation.

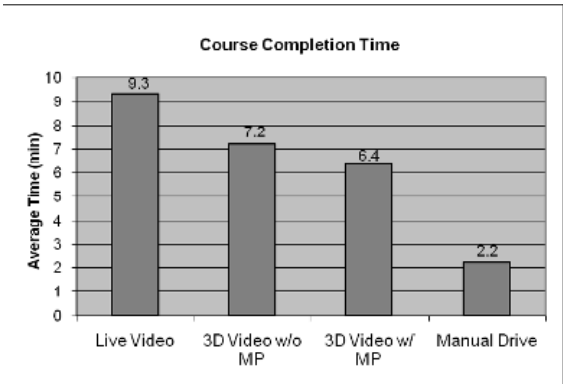

Figure 8: Average Course Completion Times

Speed Results: The benefit of 3D Video follows the same trend as completion time. Basic teleoperation achieved 1.0 m/s average speed, while 3D Video alone led to 30% faster driving, and 3D Video with motion prediction increased speed by 50%. Manual driving was more than three times faster.

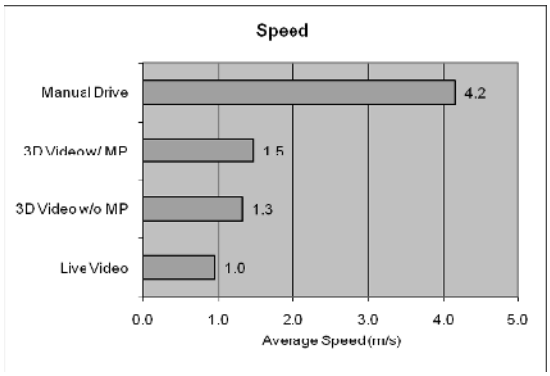

Figure 9: Average Speed

Number of Stops: One interesting repetitive event was operators choosing to stop the vehicle. This was a common response when relevant information was not available due to limited field of view or because latency disoriented the operator. 3D Video configurations reduced the frequency of stopping by 43% when compared to basic teleoperation. No drivers stopped during the manual driving configuration.

Error Rate: Fewer errors were made with the 3D Video than basic teleoperation. With 3D Video alone, the error rate dropped by almost 50%, while the error rated dropped by about 20%

when 3D Video was combined with motion prediction. Manual driving is again the gold standard, with an error rate approximately 75% lower than basic teleoperation. Interestingly, the course was sufficiently complex that drivers did commit errors even with manual driving. The average rate was 2.4 errors per run, and every driver committed at least one error over the course.

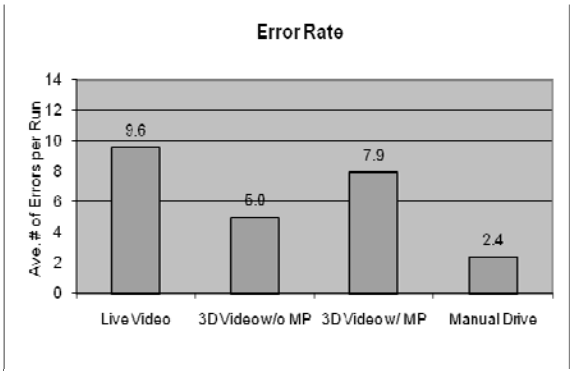

Figure 10: Error Rate

One result is particularly interesting: drivers made markedly fewer errors without motion prediction than with. In other metrics, motion prediction generally shows a small but positive benefit, while here, the contrast is substantial. Analysis of the motion prediction system likely explains why: a combination of i) variable latency invalidating the constant-latency model used in the software, ii) sub-optimal vehicle model parameters and iii) inaccuracies in the pose data, all contributed to errors in motion prediction that were at times substantial (relative to the tolerance of many of the course decision gates, for example).

Workload: The NASA TLX workload questionnaire was administered after each run, allowing operators to rate perceived mental demand, physical demand, temporal demand, own performance, effort and frustration associated with each driving condition. Overall workload scores indicate the least amount of workload was required with the 3D visualization system alone. As expected, the highest workload was achieved with live video, while 3D Video with motion prediction and manual drive were rated similarly. In general, manual driving workload was rated higher than expected. This may be due to the physical effort required to use the eGator steering wheel and a lower than anticipated perceived performance rating.

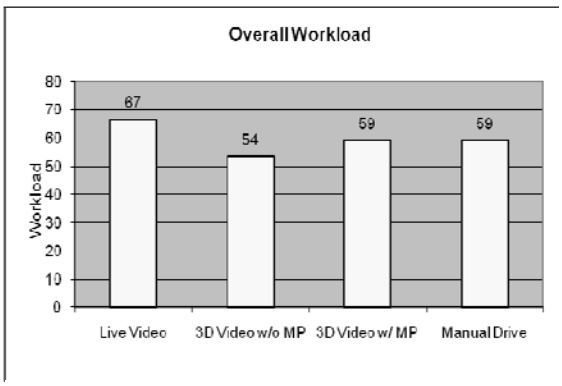

Figure 11: Overall Workload

Looking at the dimension scores associated with overall workload, differences between driving conditions become more apparent. Live video required significantly more mental demand than other driving conditions, as well as higher temporal demand, perceived effort and frustration levels. Temporal demand ratings were very close, which is not surprising given drivers were told to complete the course as quickly as possible, thereby creating time-based workload across all conditions. 3D visualization conditions were rated similarly, but frustration levels were higher without motion prediction. Drivers reported the lowest physical demand with the 3D Video conditions.

Exit Interview: The exit interview was completed with each participant at the conclusion of all runs. The most commonly requested improvements for basic teleoperation include decreased latency, higher video frame rate and more cameras or unique viewpoints. Participants also mentioned better resolution, wider field of view, and an indication of vehicle position in the video frame, which would allow them to drive through tight spaces. In general, operators wanted the ability to judge where the vehicle is positioned in the world by having a direct reference to all objects in the environment.

Participants felt the greatest strength of the 3D Video is the vehicle model presented within the video. The model made it easier to recover from mistakes and allowed operators to judge upcoming course events with respect to the vehicle, thereby allowing them to respond to the environment more accurately. "I could go faster between events and then slow down before an event. I could time the slow down better." 3D Video also provided a wider field of view, latency compensation, and selectable viewpoints. These features provided a "less stressful" environment and reduced the amount of time spent "paying attention to the vehicle," potentially freeing up time for other vehicle control and mission-related tasks.

The following artifacts were present in 3D Video: "video jittering around corners, straight lines in the middle of a road bending, cones disintegrating and appearing flat on the surface, 3D objects smearing as vehicle drove by, and square pixels appearing at the edge of imagery."

3D Video improvement suggestions include reducing artifacts, a higher video frame rate, improvements in latency compensation, and a wider field of view for turns. A higher frame rate was suggested to make driving at a higher velocity easier.

The final portion of the exit interview allowed participants to rank their preferences for driving condition and 3D Video viewpoints. Manual driving was preferred, followed by 3D Video with motion prediction, 3D Video without, and Live video. Three viewpoints were available within the 3D Video: native camera, over-the-shoulder, and overhead (bird's eye view). The overall preference for viewpoints was unanimous: over-the-shoulder, followed by Overhead (Bird's Eye View) and native.

Comments indicate bird's eye view was useful when navigating left or right for a short distance, such as in a slalom, and native location was useful if driving on straight roads for a long distance. Over the shoulder was more or less the "all purpose" preferred viewpoint.

Summary: In perhaps the most significant metric of task completion time, 3D Video showed improvements of approximately 20-30% compared to standard teleoperation.

Other metrics showed improvement as well, with average speed increasing 30-50 % and error rates dropping by 20-50%.

6 CONCLUSIONS & FUTURE WORK

As long as humans use displays of data generated on a remote device in order to control it, the sensors deployed on the device will be used to produce those displays. While single use sensors are common, the dual use of robot perception sensors for autonomy and visualization is already well established.

This paper has proposed a method to expend significant engineering effort in order to produce a virtual sensor with the ideal characteristic of reducing, in real-time, the environment around a remote device to the essence of a computer graphics rendering database. In other words, photogeometric sensing has the capacity to virtualize reality to produce displays with both the photorealism of video and the interactivity of a video game.

PG sensor technology can be transformative for certain human-machine interface tasks, providing solutions to problems for which there has been little hope of significant progress for some time. We have produced many instances of PG sensors in recent years and deployed them in diverse applications. In the two discussed here, user studies have verified substantial gains in the effectiveness of the man-machine system.

PG sensor technology introduces entirely new and highly effective approaches to latency compensation and video compression which have not been elaborated here for reasons of space and focus.

We are reasonably convinced that PG sensing is a sensor modality of choice with unique advantages that enable a new level of shared mental model between a robot and a human. Based on it, communications between the two can become less frequent, more terse, and more precise.

We have produced PG sensing by integrating distinct geometry (lidar) and appearance sensors (camera) into a virtual unit. Over time, the eventual development of high accuracy shared aperture sensors which are integrated in hardware at the pixel level seems inevitable.

7 ACKNOWLEDGEMENTS

The colorized ranging sensor concept described here was based on a concept originally developed under funding from John Deere Corporation. The work in mobile manipulation was funded by the US Navy EOD Technology Division (NavEOD Tech Div). The work in applying 3D Video to robot teleoperation was funded by the US Army Tank Automotive Research, Development, and Engineering Command (TARDEC). The autonomy work described in Figure 6 was funded by DARPA.

REFERENCES

[1] D. Anderson, H. Herman, A. Kelly, “Experimental Characterization of Commercial Flash Lidar Devices”, In Proceedings of International Conference on Sensing Technologies, November 21-23, 2005 Palmerston North, New Zealand.

[2] D. Anderson, T. Howard, D. Apfelbaum, H. Herman and A. Kelly, "Coordinated Control and Range Imaging for Mobile Manipulation" 2008 International Symposium on Experimental Robotics.

[3] R. Chatila, S. Lacroix, T. Simion, M Herrb, “Planetary exploration by a mobile robot: mission teleprogramming and autonomous navigation” Autonomous Robots, 1995.

[4] C. Dima, N. Vandapel, and M. Hebert, “Classifier Fusion for Outdoor Obstacle Detection”, International Conference on Robotics and Automation, April, 2004, pp. 665 - 671.

[5] T. Fong, C. Thorpe, and C. Baur, "Advanced Interfaces for Vehicle Teleoperation: Collaborative Control, Sensor Fusion Displays, and Remote Driving Tools. Autonomous Robots 11, pp 77--85, 2001.

[6] T. Fong, H. Pangels, D. Wettergreen , “Operator Interfaces and Network-Based Participation for Dante II”, SAE 25th International Conference on Environmental Systems, San Diego, CA, July 1995.

[7] S. G. Hart and L. E. Staveland (1987). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

[8] B. Hine, et. al., “The Application of Telepresence and Virtual Reality to Subsea Exploration”, the 2nd Workshop on Mobile Robots for Subsea Environments, Proceedings ROV ‘94, Monterey, CA, May 1994.

[9] Jackel, L., Krotkov, E., Perschbacher, M., Pippine, J., Sullivan, C.: The DARPA LAGR program: Goals, challenges, methodology, and phase I results. Journal of Field Robotics 23, 945–973 (2006)

[10] A. Kelly et al., “Toward Reliable Off-Road Autonomous Vehicles Operating in Challenging Environments”, The International Journal of Robotics Research, Vol. 25, No. 5–6, pp. 449-483, June 2006.

[11] R.A. Lewis and A.R. Johnston, “A Scanning Laser Rangefinder for A Robotic Vehicle” Proceedings of I.ICAI~5, 1977.

[12] J. Ryde, H. Hu, “3D Laser Range Scanner with Hemispherical Field of View for Robot Navigation”, Proceedings of the 2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, July 2 - 5, 2008, Xi'an, China.

[13] D. A. Messuri and C. A. Klein, “Automatic Body Regulation for Maintaining Stability of a Legged Vehicle during Terrain Locomotion”, IEEE Journal of Robotics and Automation, RA-1, pp 132-141, Sept 1985

[14] T. Möller, H. Kraft, J. Frey, M. Albrech, R. Lange.: Robust 3D measurement with pmd sensors. In: Proceedings of the First Range Imaging Research Dat at ETH Zurich (2005). ISBN 3-906467-57-0

[15] B. Ricks, and C. Nielsen and M. Goodrich. "Ecological Displays for Robot Interaction: A New Perspective". Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.

[16] T. Sheridan, “Human Supervisory Control of Robot Systems”, Proceedings of the 1986 IEEE International Conference on Robotics and Automation, Apr 1986.

[17] G. Terrien, T. Fong, C. Thorpe, and C. Baur. "Remote driving with a multisensor user interface". In Proceedings of the SAE ICES, Toulouse, France. 2000.