2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Performance evaluation of monocular predictive display

Adam Rachmielowski, Neil Birkbeck, and Martin Jägersand
University of Alberta

Abstract— **In teleoperation systems, operator performance is negatively affected by time-delayed visual feedback. Predictive display (PD) compensates for delays by providing synthesized visual feedback. While most existing PD methods rely on a priori models (e.g., from laser range finding or stereo vision), recent work on monocular SLAM and SFM makes it possible to acquire PD models in single camera applications. In this work, we evaluate operator performance of PD visual feedback based on a coarse 3D model. We report the experimental results of 12 human tele-operators each performing 96 visual alignment tasks with a 300ms delay. Four operating modes are considered: delayed video (no PD), video-based PD using a stabilizing plane (homography), 3D model-based PD, and no delay (ground truth). The results indicate that vision-based PD (both plane and 3D model-based) is significantly better than delayed video. It reduced task completion time 40% and is nearly as good as the no delay condition. PD based on a sparse a 3D model was somewhat better than the simpler plane based method.**

I. Introduction

Long-distance operation of robots (tele-operation) has applications in space and deep-sea robotics, robot-assisted surgery, or any task where an operator must interact with a remote environment via a robot. A major issue in tele-operation is performing tasks in the presence of communication delay [1]. Early studies considering delayed visual feedback have shown that "sensory-motor adaptation is essentially impossible for delays as small as $0.3s$", leading to inefficient "move and wait strategies" for accomplishing tasks such as placing pegs into holes [2]. Predictive display (PD) attempts to compensate for visual-feedback time delays by predicting and displaying the appearance of the remote environment before the remote signal has actually arrived. Prediction is usually based on the operator's motion control signals to the robot. Since these are available at the master control station, they can drive a simulation that returns predicted visual feedback before the actual video feedback arrives from the remote site.

Early PDs provide simple graphical markup, such as a point indicating the predicted position of a target, or a vector model of the remote robot overlayed on delayed video to indicate a predicted pose change [3]. "Model-based" methods rely on a priori models of the environment and/or robot to provide PD [4] (both predicted visual feedback and predicted force feedback).

Recent PD methods leverage computer vision and graphics to produce high-fidelity predicted visual feedback without requiring a priori models. Burkert, et al. incrementally estimate a geometric model and a set of textures from stereo image pairs [5]. Cobzas, et al. use a pan-tilt camera and laser-rangefinder to estimate an accurate panorama-plus-depth graphics model offline [6]. Both methods provide high-fidelity graphics models that can be rendered from a novel view at frame-rate, but require significant time to build the model or to incorporate new images. Moreover, they are not designed for monocular vision, which may be the only available sensing modality due to cost/payload constraints in mobile robotics or sensor damage in robust systems.

Reconstructing 3D graphics models from a single, moving, monocular camera is a mature computer vision problem [7]. Modern methods [8], [9] provide estimates of camera pose, sparse 3D structure, and keyframe images at frame-rate. With such methods it is possible to concurrently acquire images from a camera, reconstruct a coarse graphics model, and visualize that model, all at frame-rate [10]. Previously, we have used this concurrent method to efficiently collect data for 3D modeling from monocular images [11]. In this work, it provides predicted visual feedback for the vision-based PD experiment.

In vision-based PD the model is built online from the robot camera. This is based on (visual) camera pose tracking, online building of a 3D geometry, and saving keyframe images for use in texturing. When starting the robot in a new place, there is no 3D model of that environment. Similarly, when the robot viewpoint changes to an unexplored area, there is no model. Yet the operator has to be supplied with visual feedback at all times. This suggests a hierarchical approach to PD, with three operating modes:

1) As a baseline, delayed video is available for the operator.
2) Once a reasonably accurate camera pose is available from visual tracking, basic PD can be generated by forward warping the delayed video using a planar homography.[1]
3) After some time when sufficient 3D geometry of the scene has been computed, this geometry is used to provide a richer PD by texturing from keyframes and rendering from the current operator viewpoint.

Two additional advantages of the PD model are that a wider field of view than that of the robot camera can be rendered by texturing from several keyframes, and that it decouples what the operator sees from the current robot pose. The latter is useful in robots where the camera is mounted to the robot and has no separate articulation. Here

[1] While for some robots, e.g. a calibrated manipulator, the robot camera pose would be available, here we consider uncalibrated systems where vision is also used to establish the robot motion

978-1-4244-5040-4/10/$26.00 ©2010 IEEE

a motion/pose trajectory for the robot task at hand may not give good views to the operator, but using a PD model the operator can visualize different viewpoints.

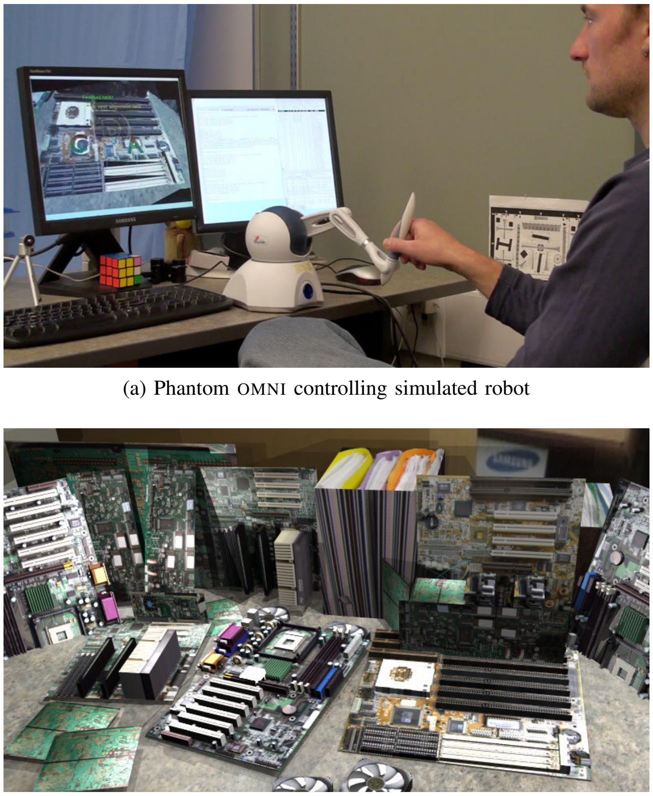
It is worth noting that while conventional computer graphics require dense 3D models to render well from any viewpoint, in this PD application keyframes from relatively close viewpoints are transferred via the geometry to the operator view. The geometry acts as a proxy and even very sparse (10-100 points) models give sufficient results. The simpler planar homography mode is equivalent to a four point model. The visual quality of sparse models can be improved through view-dependent “dynamic” (time and pose varying) texturing for both a planar [12] or 3D mode [13].

We evaluate the three described operating modes (and a ground truth mode) for participants performing a visual alignment task. The task is to remotely move a robot-controlled camera to a specific pose in a static environment (e.g. in order to perform an inspection task, or initiate a manipulation task). We compare time-to-completion and trajectories for both novice and expert participants with no delay (ground truth), delayed video, stabilizing plane, and 3D model-based PD. Experimentally, we determine if a graphics model computed at frame-rate from a single camera is useful as a PD for overcoming communication delay.

II. Experiment Design

In the experiment, participants operated a virtual robot with a Phantom OMNI controller (Fig. 1a). The pose of the OMNI pen-tip mapped directly to the desired pose of the camera mounted on the end effector of the robot (eye-in-hand). The user is provided with visual feedback from the camera, which made this mapping natural. The participants performed 6DOF alignment tasks, which required them to move the robot to a specific pose in its environment. The task environment consisted of circuit boards placed in a workspace approximately $1m \times 50cm \times 50cm$ (Fig. 1b). Each task consisted of aligning a target letter (A, B, C, or D) in the scene with the silhouette of the target letter, which was overlayed on the display (Fig. 2). To complete a task the robot would have to remain within a small threshold of the target pose for $500ms$. Simplified robot dynamics were modeled by limiting the maximum velocity and angular velocity of the robot to $30cm/s$ and $45°/s$, respectively. Lastly, a communication delay of $300ms$ was applied to the visual feedback (except in the ground truth case). This delay was chosen since it is cited as the minimum time at which simple task completion is impeded [2], and it also is representative of typical networking latency over the Internet or other commonly available networks due to a combination of distance and network switches.

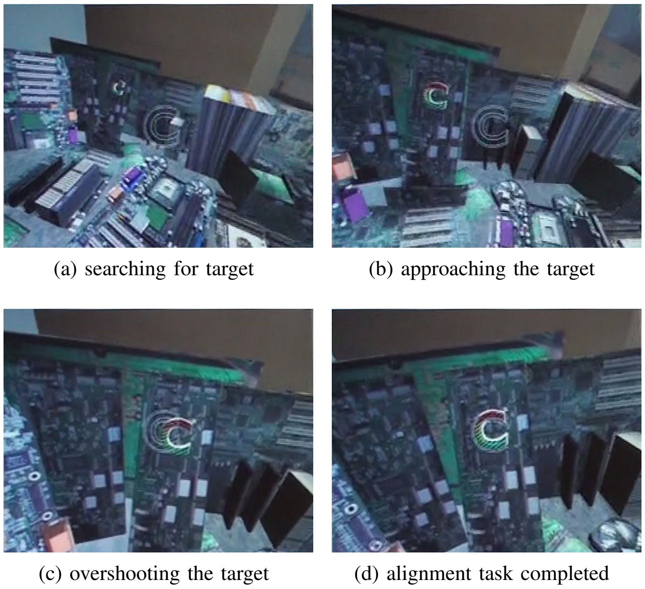
We compared the effect of four possible operating modes on task completion time: delayed image (no PD), stabilizing plane PD, model-based PD, and no communication delay (ground truth). The delayed image (no PD) mode simply presented the user with the image from the robot simulation delayed by 300ms (Fig. 3a). In the stabilizing plane PD mode, the delayed video frame was projected onto a plane at mean depth to the scene (Fig. 3b). The plane was then rendered from the operators current (non-delayed) viewpoint. The stabilizing plane mode represents a simpler view prediction that is possible without sparse geometry. (A version of this is used in consumer camcorders for stabilization of shaky hand-held imaging). This mode is evaluated to quantify whether saving key-frames and triangulating a surface is valuable in this application, or if the simpler approach is adequate.


(a) Phantom OMNI controlling simulated robot

(b) Simulated task environment

Fig. 1: Predictive display experiment setup. Participants move the simulated robot, via the Phantom OMNI haptic device, to specified poses using four different operating modes.

In the model-based PD mode, a sparse structure was pre-estimated (∼500 points) and a view-dependent surface model was estimated at frame-rate [11]. This pre-estimated 3D structure consisted of a set of sparse 3D surface points that were generated by back projecting feature points extracted from renderings of the synthetic model. This sparse 3D point structure is sufficiently different from the underlying ground truth model, and it is intended to simulate the output from a real-time vision subsystem (e.g., [8], [9]). The corresponding surface model used in this PD mode is based on a view-dependent Delaunay triangulation of the surface points visible from the operator’s desired viewpoint. In each trial, key-frame images were stored as the scene was explored, and the surface model was textured using the closest of the key-frames and the most recent video frame (Fig. 3c).


(a) searching for target

(b) approaching the target

(c) overshooting the target

(d) alignment task completed

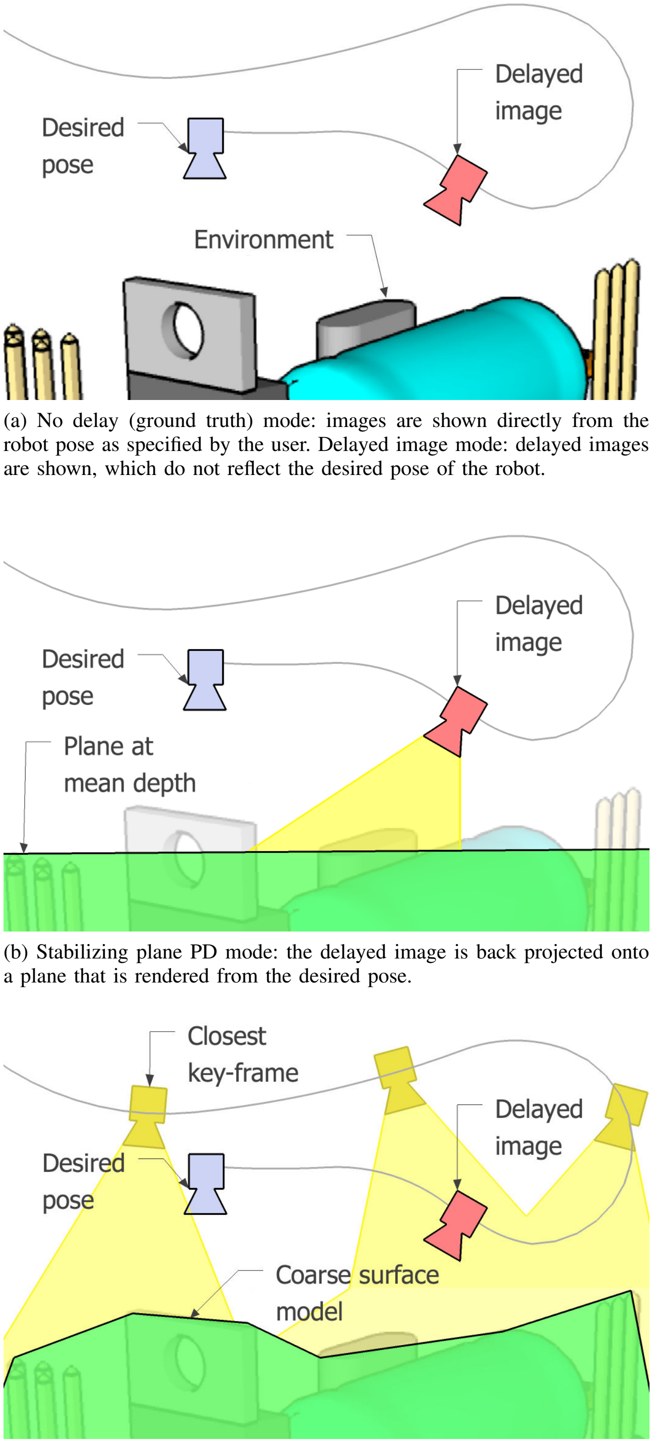
Fig. 2: Visual feedback as the robot is moved to align the C silhouette with the C target in the environment.

At the beginning of the experiment, participants were instructed on the use of the OMNI controller and shown a demonstration of the tasks they would be performing. They were given 2 minutes to practice using the controller to direct a simulated camera (no dynamics constraints) and were asked to perform certain motions (e.g., look to the right or move close to the yellow circuit board) to ensure a sufficient level of competence with the controller. Next, to minimize the effects of learning during the experiment, participants were given 10 minutes to practice performing alignments with each of the operating modes.

Tasks were organized into sequences of four alignments: A, then B, then C, and then D. During the experiment, participants performed 24 tasks with each of the four operating modes, totaling 96 alignments. These were organized into six batches of tasks, where in each batch the participant performed four sequences of four tasks, each with a different mode. To minimize the effects of ordering in the experiment, mode orders within each batch were selected from a randomized list of the possible mode-order permutations, while satisfying the constraint that all sequences are performed with all modes by the end of the experiment.

III. Results and Discussion

Of the 12 participants, six had previously used the Phantom OMNI controller in other contexts or had experience with robotic alignment tasks. They had more consistent performance, and are considered better representatives of real tele-operators. We classify these participants as *experts*. The resulting times to perform alignments during the experiment are shown for all participants (Fig. 4), for experts (Fig. 5), and for individual targets by experts (Fig. 6). The data was first normalized, then a multiple comparison test was performed with analysis of variance (ANOVA). For each



(a) No delay (ground truth) mode: images are shown directly from the robot pose as specified by the user. Delayed image mode: delayed images are shown, which do not reflect the desired pose of the robot.

(b) Stabilizing plane PD mode: the delayed image is back projected onto a plane that is rendered from the desired pose.

(c) Model-based PD mode: the closest key-frames are back projected onto the coarse surface model and then rendered from the desired pose.

Fig. 3: Diagram of simulation and operating modes.

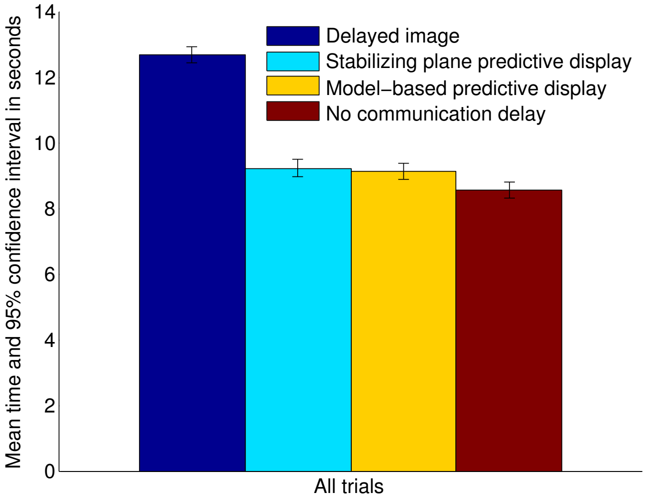


Fig. 4: Time to perform alignment for all subjects. Modes with non-overlapping confidence intervals have significantly different mean times. ∼48% longer to complete with delay than no delay. ∼7% longer to complete with PD modes than no delay.

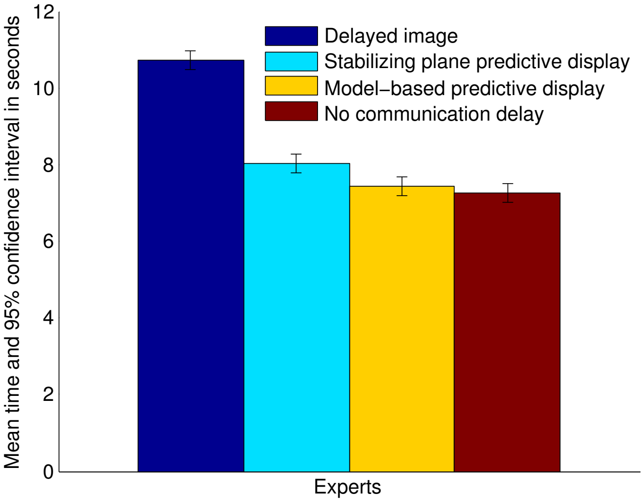


Fig. 5: Time to perform alignment for six expert subjects, who had previous experience using the Phantom OMNI. Modes with non-overlapping confidence intervals have significantly different mean times. Model-based PD is better than stabilizing plane PD.

participant, all times were normalized by the participant's mean time (i.e. each participant contributed to the results equally). Similarly, for each task, all times were normalized by the mean time of the task (i.e. each task contributed to the results equally).

The results indicate that performing these alignment tasks is impeded by communication delay (∼48% longer to complete with delay than no delay), and predictive display almost entirely compensates for the delay (only ∼7% longer to complete with the PD modes than with no delay). For the entire sample population, there is no significant difference between model-based PD and the simpler stabilizing plane PD (Fig. 4). For the expert sub-population, however, model-based PD is better (less time to complete task) than stabilizing plane PD, and in fact there is no significant difference between model-based PD and no delay (Fig. 5). One reason for the similarity between times for PD modes is that the local geometry of the alignment targets is nearly planar (i.e. circuit boards) and hence is well approximated by the stabilizing plane.

The results for aligning individual targets within each sequence (Fig. 6) indicate that while there is no significant difference between the PD modes for the first two targets (A, B), model-based PD is better than stabilizing plane PD for later targets (C, D). This can be explained by considering the alignment task as two separate actions. First, the participant searches for the target and performs a coarse alignment, and then she performs a fine alignment. For tasks C and D, the scene has already been somewhat explored (during the alignment of A and B). The model-based PD then provides the advantage of enabling the search and coarse alignment using the model and key-frames. In contrast, the plane-based PD mode uses only the most recent delayed image and hence is still affected by communication delay. This is evident in the comparison of residuals for a task representing target C alignments (Fig. 7). During the coarse alignment (up to ∼$20cm$ of the target) the stabilizing plane PD is similar to the delayed image, while the model-based PD is similar to no delay; during the fine-scale alignment both PD modes are similar to no delay. The benefit of model-based PD increases for longer tasks with several movements in sequence since the model is incrementally acquired and becomes more detailed over time.

Residuals for a task representing all the trials (Fig. 8) show the move-and-wait strategy (plateaus in the residual) and overshooting (sinusoidal residual) expected in the delayed image mode. Stabilizing plane PD shows some overshooting because the stabilized image is still based on the most recent delayed image. Model-based PD converges relatively smoothly to the target. Some overshooting evident in the no delay residual can be explained by the robot dynamics. The maximum velocity of the robot has some of the same effect as communication delay on task performance. Moving the OMNI to a specific pose is not immediately reflected in the displayed video resulting in overshooting. This also explains why in some specific trials (especially for targets C and D, as explained above), the model-based PD times are better (less time to complete task) than times for no delay.

IV. Conclusion & future work

In tele-robotics, visual feedback is delayed, degrading operator performance; PD compensates for this by immediately rendering the operator's desired viewpoint using a graphics model. In this work, a monocular predictive display system has been evaluated in the context of simulated robotic

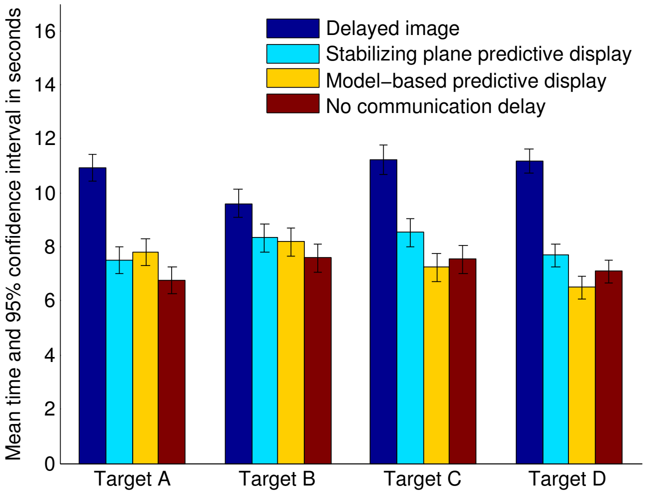


Fig. 6: Time to perform alignment for experts for the first (A) through last (D) alignment target in all six four-target sequences. Within each group, modes with non-overlapping confidence intervals have significantly different mean times.

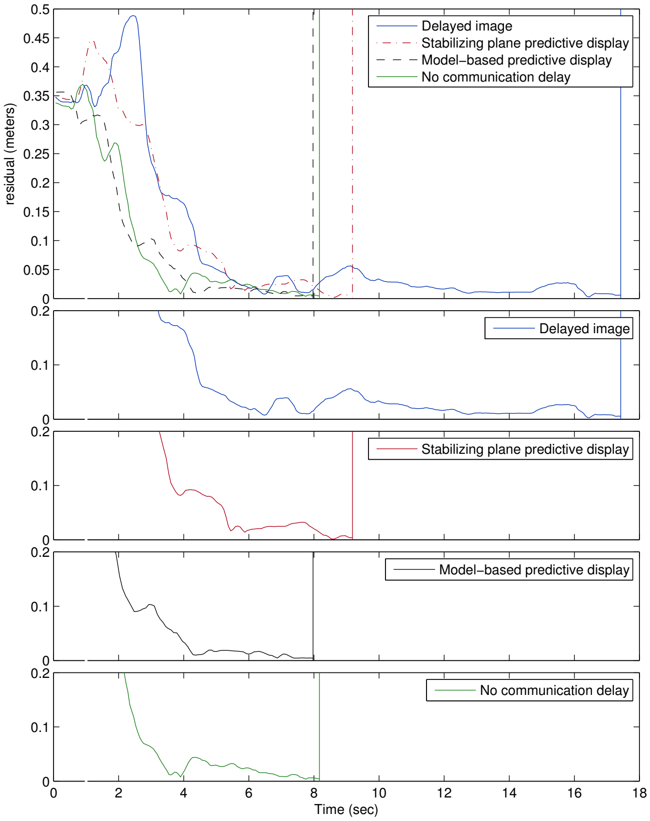


Fig. 7: Position residual for "C target" representative. The task shown is for a single participant and single alignment with ratio of mode times that is closest to the mean ratio for the target C group. The top graph shows all four modes; The bottom four graphs show individual modes within 20cm of the target; Vertical bars indicate when the task was finished for each mode.

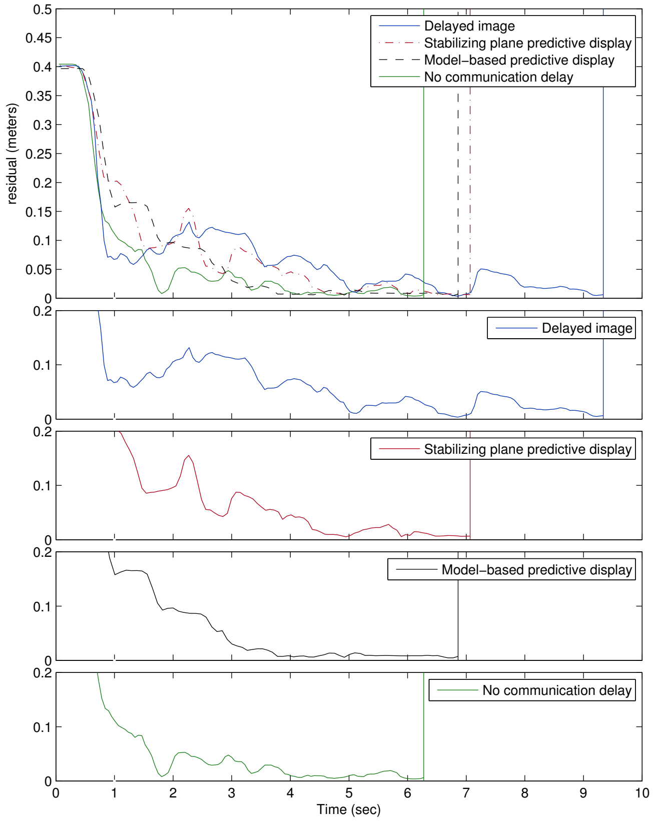


Fig. 8: Position residual for "all trials" representative. The task shown is for a single participant and single alignment with ratio of mode times that is closest to the mean ratio for the sample population. The top graph shows all four modes; The bottom four graphs show individual modes within 20cm of the target; Vertical bars indicate when the task was finished for each mode.

alignment tasks with delayed visual feedback. A simulated environment was chosen for practical reasons and to make the experiments consistent over the many subjects and conditions tested. Yet the simulated model was built to be similar in density and quality to what can be obtained in a monocular computer vision system. The results support the hypothesis that predictive display using a coarse graphics model, of a quality obtainable from computer vision and computed at frame-rate, is useful in compensating for delay. The PD reduced task completion times by 40% compared to delayed feedback, and PD tele-operation performance approached performance with no delay. The model-based approach was shown to be somewhat better than the simpler stabilizing plane approach, particularly for experts. Furthermore, as more tasks were performed in the same visual workspace, and hence a richer PD model was acquired, the advantages of the model-based approach became increasingly significant.

In this work, we pre-estimate sparse structure, use a fixed 300ms delay, provide only visual feedback, and use a simulated static environment. In future work, it would be meaningful to perform trials while concurrently estimating structure (although this would require significantly longer trials, or many more participants). We would also like to investigate the effect of differing amounts of delay and non-uniform delay, as is common in network communication. Initial investigation suggests that PD is almost mandatory for these alignment tasks when delays exceed one second; It took the authors ten times longer to complete tasks with such delays. Participants were only provided with visual feedback in this experiment; Would the proxy geometry estimated by the model-based PD also be useful for providing haptic feedback via the OMNI controller? Moreover, a static environment is considered in this experiment; Would the model-based PD be made suitable for dynamic environments by purging old key-frames and acquiring new frames whenever the camera revisits specific poses? Finally, we have completed preliminary tests of this OMNI and PD configuration with a real camera mounted on a Whole Arm Manipulator robot; In the future, we hope to overcome technical, cost, and safety concerns and to perform trials with this robot.

REFERENCES

[1] P. Hokayem and M. Spong, "Bilateral teleoperation: An historical survey," in *Automatica*, vol. 49, no. 12, December 2006.

[2] T. Sheridan, "Space teleoperation through time delay: review and prognosis," *Robotics and Automation*, vol. 9, no. 5, pp. 592–606, 1993.

[3] A. Bejczy, W. Kim, and S. Venema, "The phantom robot: predictive displays for teleoperation with timedelay," in *Robotics and Automation, 1990*, vol. 1, May 1990, pp. 546–551.

[4] T. Kotoku, "A predictive display with force feedback and its application to remote manipulation system with transmission time delay," in *Intelligent Robots and Systems, 1992*, vol. 1, July 1992, pp. 239–246.

[5] T. Burkert, J. Leupold, and G. Passig, "A photorealistic predictive display," *Presence: Teleoperators and Virtual Environments*, vol. 13, no. 1, pp. 22–43, 2004.

[6] D. Cobzas, M. Jägersand, and H. Zhang, "A panoramic model for remote robot environment mapping and predictive display," *International Journal of Robotics and Automation*, vol. 20, no. 1, pp. 25–34, 2005.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[8] A. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc. International Conference on Computer Vision*, Oct. 2003.

[9] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[10] A. Rachmielowski, "Concurrent Acquisition, Reconstruction, and Visualization," Master's thesis, University of Alberta, Edmonton, Alberta, Canada, April 2009.

[11] A. Rachmielowski, N. Birkbeck, D. Cobzas, and M. Jägersand, "Realtime visualization of monocular data for 3D reconstruction," in *Canadian Conference on Computer and Robot Vision*, 2008, pp. 196–202.

[12] M. Jägersand, "Image-based predictive display for high d.o.f. uncalibrated tele-manipulation using affine and intensity subspace models," *RSJ Journal of Advanced Robotics*, vol. 14/8, pp. 683–703, 2001.

[13] K. Yerex, D. Cobzas, and M. Jägersand, "Predictive display models for tele-manipulation from uncalibrated camera-capture of scene geometry and appearance," in *Proc. of IEEE International Conference on Robotics and Automation*, 2003.