https://doi.org/10.2352/ISSN.2470-1173.2019.3.SDA-659

Subjective comparison of monocular and stereoscopic vision in teleoperation of a robot arm manipulator

Yuta Miyanishi, Erdem Sahin, Jani Makinen, Ugur Akpinar, Olli Suominen, Atanas Gotchev; Tampere University; Tampere, Finland

Abstract

Remote operation of robot manipulators plays a crucial role in various safety-critical application areas such as forestry, mining, surveillance, etc. The capture and display of visual information of the real operation environment in an information-rich way is one of the key factors in achieving effective and robust teleoperation capability. In this paper, through subjective experiments, we comparatively investigate the roles of monocular, i.e. 2D, and stereo capture and display systems in the context of teleoperation of a robot arm. In particular, the positioning task is considered and, for both monocular and stereoscopic cases, two different modes of capture are implemented, which are static capture setup separately located from the robot arm and dynamic capture setup positioned at the tip of the robot arm. The experiments, conducted on 10 subjects (aged 24-33), show that stereo vision systems enable significant increase in accuracy of positioning compared to conventional 2D capture and display cases. In average, the tasks are completed with highest accuracy in the case of dynamic stereo capture setup.

Introduction

Teleoperation (remote operation) systems plays a crucial role in many safety-critical fields such as forestry, mining, construction, surveillance, etc. [1]. Teleoperation enables an operator to stay in a safe and comfortable place, instead of a dangerous and exhausting site. In addition to the benefits for an operator, remote operation technology can allow concentration of human resources into one place such as a control station, which will give considerable improvement in efficiency in the whole work system.

In teleoperation, an operator mainly relies on the visual information which is critical in creating strong situational awareness, feeling of control and safety, and sense of telepresence or immersion. Thus, relaying rich and reliable visual information from a site to a teleoperation environment, e.g., a remote operation room, is a key factor for teleoperation systems. The success in the relay of visual information can be quantified based on comparison of the conveyed visual information to that information available to the operator in the natural (on-site) viewing.

A human senses the three-dimensional (3D) structure of a space from the images perceived by the two eyes. And the visual sensation of the depth is crucial for performing various operation tasks. On the other hand, a camera images the three-dimensional world onto a two-dimensional sensor. That means, in principle, the dimension of the depth is lost in the captured image. But this does not mean that all depth cues, i.e., information sources of the depth from an image, are lost in the image. A 2D image can still provide several monocular depth cues such as linear or aerial perspective, texture distribution, relative size, occlusion, shade, etc. A stereoscopic capture and display system can add another important depth cue for a human observer, namely binocular disparity cue, by capturing and displaying two images one for each eye. It is well known that binocular disparity is a metric and precise depth cue for human vision and it makes very clear depth perception.

In the review of McIntire et al. [2] on advantage of a stereoscopic display with regard to human task performance, the authors argued that the benefit of a stereoscopic display for a task performance is dependent on the characteristic of the task. That is, a stereoscopic display improves a teleoperation performance when the depth-related task is complex, difficult, and unfamiliar for operators. It is reported that stereoscopic displays improve task completion time or teleoperation accuracy in various teleoperation task scenarios, such as simulated driving and peg-in-hole task, compared to conventional monoscopic (2D) displays [3, 4, 5, 6].

In this study, we comparatively examined the performance of teleoperation for the task of lateral positioning of a robot arm, in monoscopic and stereocopic viewing conditions. The performance measure was the accuracy of the positioning. The task and measure of performance differ from common tasks in previous works like a peg-in-hole task and time to completion. The distinctive feature of the lateral positioning task is absence of performance feedback. In a peg-in-hole task, for instance, there is obvious feedback through the completion of the task because a participant keeps trying until the task is completed. On the other hand, in the lateral positioning task, a participant finishes a trial when he or she feels the robot is reached the correct position. In other words, participants will never know the solution to the task. Therefore we believe the performance of this task reflects the subjective estimation which participants have.

We consider two view scenarios in our experiments, which we define as the operator view and the device view. The operator view corresponds to the view of an on-site operator, e.g., an operator in the cabin of a work machine, where the viewpoint does not change during operation of the robot arm. The advantage of the operator view is that the operator can continuously obtain the overview of the whole scene. The device view is the view from a camera which is attached on the robot arm, and the viewpoint changes as the robot arm moves. Through the device view, the operator gets a close-up visual information about the object of interest. In this study, both views were simulated in the experimental setup, and the effectiveness of a stereoscopic capture and display was evaluated in both scenarios.

Material and methods

A subjective experiment is executed in this study to quantify how accurately naive operators can perform teleoperation of a robot arm in different viewing conditions, in particular, natural, monocular and stereoscopic viewing.

Participants

Ten participants, whose ages were in 24–33, took part in the experiment. They were the university's students and staffs and there was a reward of a confection. All of the participants had normal or corrected-to-normal vision. We verified that there was no participant who had stereo blindness, with Randot stereo test. All of the participants were untrained in operating a robot arm and they were not told the hypothesis and purpose of the experiment. No one reported nausea, dizziness, or visual disorder after the experiment.

Experimental setup

Figure 1 illustrates the experimental setup. The participants sat on the manipulator's seat and had direct view to the experimental field or indirect view through the stereoscopic cameras and display. The surface of the experimental field was black fabric and the targets were white paper with printed black bull's eye. The static stereo camera was fixed in front of a participants so that the view from the static camera was kept to be similar to the direct (natural) view of the participant. The dynamic stereo camera and an L-shaped beam with a white tip were attached to the end of the robot arm. The tip could be seen from both cameras.

The static and dynamic stereo cameras were implemented via the same model of a CMOS camera (Basler acA1920-50gc; Basler AG, Germany) in each pair. The images from the two stereo cameras were processed with a workstation computer with a single NVidia Quadro K2000 video card. The stereoscopic image from the cameras were shown on the stereoscopic display (zSpace 100; zSpace Inc., USA). The display resolution was 1920 × 1080 pixels, and the bit depth was 8 bits. In each monocular viewing condition, only the left image from the stereo camera was shown on the display.

The task of the participants was to manipulate an industrial robot arm (KUKA KR 16 L6-2; KUKA AG, Germany) with a remote controller, to put the white tip above the crosshair of the designated target as close as possible. The participants controlled the arm by pressing one of the four buttons at the same time, which were responsible for moving the arm along the positive and negative directions of the horizontal x and y axes.

During the experiment, the position of the tip of the robot arm and the three targets were recorded by using an infrared optical tracker system (OptiTrack V120:Trio; NaturalPoint, Inc., USA).

Tasks and procedure

A participant manipulated the robot arm in five viewing conditions to position the tip of the arm laterally at one of the three targets, which was designated by an experimenter verbally in every beginning of a trial.

Viewing conditions

The participants wore the polarized glasses in all of the five viewing conditions. In the *natural viewing* condition, the participants saw the field directly from the seat. In the other viewing conditions, the participants observed only the display which was showing a monocular or binocular image from the static or dynamic cameras. The first three conditions, namely the natural

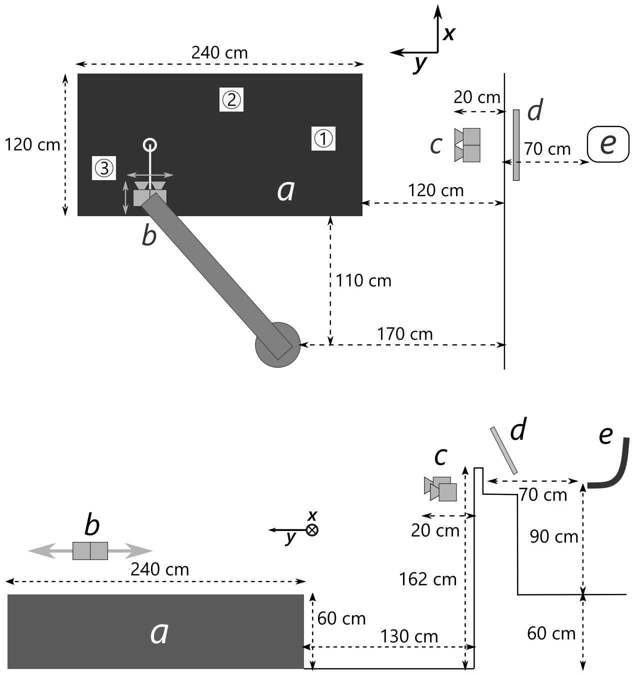


Figure 1. *Experimental setup. The upper and the bottom pictures illustrate the top view and the side view, respectively. **(a)** The experimental field. The tip of the robot arm was allowed to move only above the field. **(b)** The tip of the robot arm and the dynamic stereo cameras. **(c)** The static stereo cameras, which position was fixed in front of the manipulator's seat. **(d)** The stereoscopic display. In the natural viewing condition, the display was removed so that the participants could see the field directly. **(e)** The manipulator's seat on which the participants sat during the experiment.*

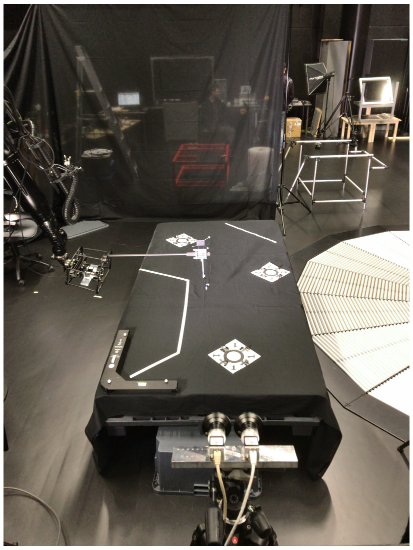


Figure 2. *Picture of the experimental setup. The photo was taken from a point above the manipulator's seat. The stereo camera at the bottom of the picture is the static stereo cameras. At the middle of the image, one can see the dynamic camera and the tip. The tip was at the home position. The white lines on the field are the boundary of the area which the tip can reach. The L-shaped calibration target at the left bottom of the field was removed during the experiment.*

viewing, static-monocular and static-binocular conditions, shared the same direction of view. Because the view was imitation of that from an operator in an actual work machine, the view direction is called as the *operator view* in this study. The last two conditions, which are the dynamic-monocular and dynamic-binocular conditions, shared the other direction of view. It is the view from the cameras at the tip of the robot arm imitating a monitoring camera on a work machine. Therefore, this view is called as the *device view* in this study. Because the participant's seat and the robot arm were on adjacent sides of the rectangular test field, the axes of the operator and device views were orthogonal to each other.

viewing condition	view	depth cues
natural viewing	operator	all
static-monocular	operator	pictorial
static-binocular	operator	pictorial + bin.disp.
dynamic-monocular	device	pictorial
dynamic-binocular	device	pictorial + bin.disp.

Viewing conditions, view, and available depth cues

Procedure

There were 15 conditions (5 viewing conditions × 3 targets) and each condition was repeated 3 times, so one participant performed 45 trials in total. Three trials consisted one block and the three targets were assigned to these trials[1]. Three trials in each block had the same viewing condition, which means that there were 15 blocks (5 viewing conditions × 3 repetitions) for one participant. The order of executing the 15 blocks was randomized. The order of the targets in each block was also randomized.

At the beginning of each block, the tip of the robot arm was at the *home position*, which was in the middle of the three targets (see Figure 2). Then the experimenter told the target number and the participant started manipulation to reach that target. After the participant was satisfied with the positioning, he/she told the experimenter that the task was completed. Then the experimenter recorded the timestamp electronically and proceeded to the second trial. At the beginning of the second trial, the experimenter told the next target number, and the participant started to manipulate the robot arm from the first target to the second target. The third trial was performed in the same manner. After the third trial was completed, the robot arm went back to the home position by pressing another button on the controller.

The experiment took about 40–60 minutes for one participant.

Analysis

In each trial, the finishing location of the tip was identified from the tracking data and the timestamp. Then the error was defined as the lateral vector from the designated target to the recorded position of the tip. Thus, an error vector which has two component of x and y was obtained in every trial (see Fig. 1 for the orientation).

At first, we compared the norm of the errors for different viewing conditions. After obtaining averaged norm for each combination of the targets and viewing conditions, we normalized the errors by dividing them by the averaged error in the natural condition, for both x and y directions. The normalized error indicates the ratio of the error to the average error in the natural condition.

Next, we evaluated the direction of the positioning errors (before the normalization) in the cases of operator and device views. The details are described in the following section.

Results

Norm of the positioning errors

Figure 3 shows the normalized errors for the three targets in the three viewing conditions which have the operator view. In total, the error tended to increase in the order of natural viewing, binocular-static, and monocular-static condition, which corresponds to the poorness of visual information. In the monocular condition, the source of depth information is limited to pictorial cue so it should be the reason for the largest errors. In the binocular condition, binocular disparity cue could be used in addition to pictorial cues; which explains the smaller errors than the monocular condition. The richest information, namely all of the visual information, was available in the natural viewing condition, which explains the smallest errors.

In the natural viewing condition, the normalized error for the target 3 was bigger than the errors for the other two targets. The most possible factor which account for this is the viewing distance. Because the target 3 was the farthest target, the participants couldn't be able to specify the location of the tip at the target 3 as precise as the other nearer targets[2]. In the other two viewing conditions, there was a clear trend in the normalized error that the norm of the error increased by the viewing distance. The difference in viewing distance would also explains this trend. Unlike the natural viewing condition, the visual information was quite limited by using the display in the static-binocular and -monocular viewing conditions; the image the participants saw was limited in resolution, dynamic range, etc. Limited resolution should impact the accuracy in the positioning especially for the farther targets because the apparent size of a target decreased as the viewing distance to the target increases.

Figure 4 shows the normalized errors for the three targets in the two viewing conditions which had the device view. The errors in the monocular viewing condition were obviously bigger than the errors in the binocular viewing condition, which can be explained by the availability of the binocular disparity cue. And there was no clear difference among the errors for the three targets, in both of the binocular and monocular viewing conditions. This is reasonable because the distance from the dynamic cameras to each target were almost the same for all targets at the moment of finishing the positioning, then the image of a target on the display seemed to be quite similar for every target.

Direction of the positioning error

We compared the direction of the error in positioning between the view types, namely the view from the static cameras (the operator view) and the view from the dynamic cameras (the device view). The participants had the operator view in the natural viewing, static-monocular, and static-binocular conditions, where the viewing (depth) axis corresponds to y-axis in the field according to coordinates introduced in Figure 1. The participants had the

[1] In other words, one target never appeared twice in one block.

[2] Note that the target 1 was the nearest target to the participants and also to the static camera in the operator view. The target 3 is the farthest, and the target 2 was between them

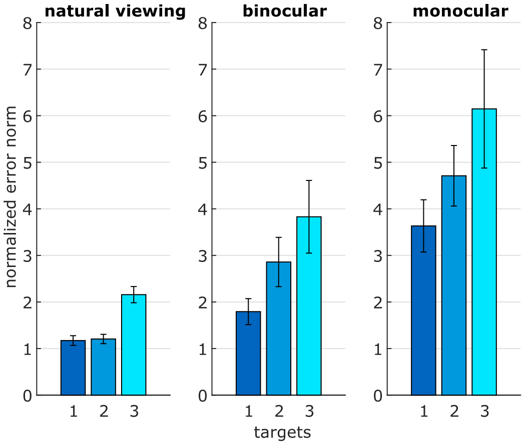


__Figure 3.__ The norm of the positioning errors across the targets and the three viewing conditions in the operator view. Each panel shows the errors in each viewing condition. The errors are normalized by the averaged error in the natural viewing condition. Each bar represents group mean and all error bars are ±1 standard error of the mean.

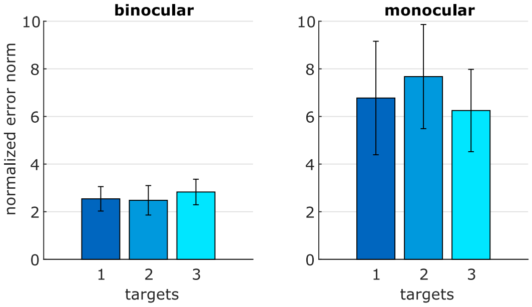


__Figure 4.__ The norm of the positioning errors across the targets and the two viewing conditions in the device view. Each panel shows the errors in each viewing condition. The errors are normalized by the averaged error in the natural viewing condition. Each bar represents group mean and all error bars are ±1 standard error of the mean.

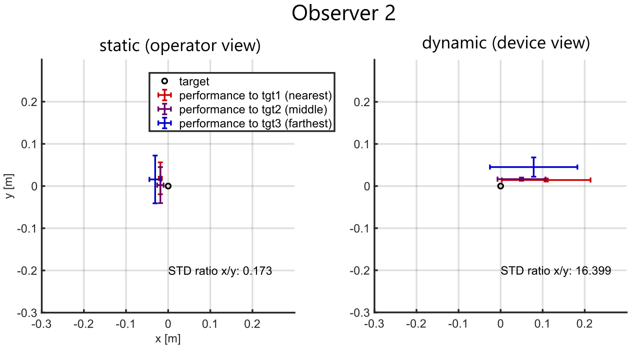


__Figure 5.__ Mean and standard deviation of each x and y component of the positioning errors. Typical result from one participant is shown. In the left panel, the errors were averaged across the three viewing conditions that have the operator view, for each target. In the right panel, the errors were averaged across the two viewing conditions that have the device view, for each target. The STD ratio in each view is shown in the left bottom of the panel.

device view in the dynamic-monocular and dynamic-binocular conditions, where the viewing axis corresponds to x-axis in the field. In each view type, we calculated the standard deviation in the positioning error on x and y axes, for each target. All of the participants demonstrated similar results and Figure 5 shows the typical result from one participant. In the operator view, the standard deviation on y-axis was bigger than that on x-axis, in the positioning to each target. On the other hand, in the device view, the standard deviation on x-axis is bigger than that on y-axis, in the positioning to each target. Then we derived the ratio of the standard deviation on the x-axis to the standard deviation on the y-axis, and averaged them over targets. The resulting ratio is called as the STD ratio in this paper. An STD ratio becomes bigger than 1 when the error on the x-axis is bigger than the error on the y-axis, which means the precision in the positioning performance is worse on the x-axis, and vice versa. An STD ratio which is close to 1 means that there was small difference between the performance on the x and y axes. The STD ratios in the performance of the participant are shown in Figure 5.

The STD ratios in the results were similar for most of the participants, for each view type. In the operator view, the group mean of the STD ratios was 0.15±0.02 (mean±SD)[3]; this indicates that the precision was worse on the y-axis, which is the depth axis of the view, than the x-axis. In the device view, the group mean of the STD ratios was 15.82±9.19; this indicates that the precision was worse on the x-axis, which is also the depth axis in the view, than the y-axis. In both of the operator and device view, the STD ratios obviously differed from one. There was anisotropy in the precision in the positioning between the axes in the view, and the precision was worse on the depth axis than the other lateral-in-image axis. These result implies that the error in human operators' performance tends to be larger in the direction of the depth in the display image.

[3] This value is after outlier removal. One participant was removed in the outlier removal.

Discussion

In this study, we conducted the subjective experiment which evaluated quantitative accuracy of the remote positioning task in the different viewing conditions and the view types. In the operator view, the performance in the natural viewing, monocular viewing, and binocular viewing conditions were compared. In the device view, the performance in the binocular viewing and monocular viewing conditions were compared. In addition to that, we evaluated the difference in the direction of the positioning error between the operator view and the device view.

The task and measure of performance

The benefit of using a stereoscopic display in teleoperation depends on the difficulty of a task. A stereoscopic display improves the teleoperation performance when the task is difficult, complex, and unfamiliar for the operators [2]. In this study, the task is not very difficult, because the test field was illuminated well and the targets had high color contrast (see Figure 2). Still the performance in the binocular viewing condition was clearly better than the monocular viewing condition, in both of the operator view and the device view.

We believe two reasons can explain this. One is the fact that the participants were not trained to manipulate a robot arm; hence the performance had been greatly affected from the presence of additional visual information, i.e., binocular disparity cue. The other reason is the task and measure of the performance we employed. Unlike the common tasks in previous works, there was no feedback in our positioning task; therefore the participants could hardly improve the positioning throughout the experiment. And we evaluated the positioning errors, not time to completion, which directly reflects the errors in participants' perception.

Limitation of a display: comparison to on-site view

As discussed in the literature, the visual information is quite limited in teleoperation by the technical limitation on a display [7]. A display has narrower dynamic range and fewer resolution than natural view, which restricts the visual information especially in a farther or smaller object. Our results confirmed that these limitation impairs the performance in teleoperation compared to on-site operation, since the errors in the binocular and monocular conditions were clearly bigger than the errors in natural viewing condition, in the operator view. The fact that the performance was worse in farther targets in the binocular and monocular viewing conditions strongly suggests that the limitation of a display hinders accurate teleoperation performance to objects in longer viewing distance.

Even in the binocular viewing condition, the performance was worse than the natural viewing condition and affected by the viewing distance. It implies that adding the binocular disparity cue to a display is not enough to replicate the on-site view at a teleoperation control station. As we discussed, some of visual information, such as dynamic range and resolution, is lost by using a display. Therefore, it must be important to identify which class of the missing information is needed, in addition to the binocular disparity cue, to improve a teleoperation display system to be the adequate replication of an on-site view.

References

[1] S. Lichiardopol, "A survey on teleoperation," in *DCT rapporten*, vol. 2007.155, 2007

[2] J. P. McIntire, P. R. Havig, E. E. Geiselman, "What is 3D good for? A review of human performance on stereoscopic 3D displays," in *Proc. SPIE 8383, Head- and Helmet-Mounted Displays XVII; and Display Technologies and Applications for Defense, Security, and Avionics VI*, 83830X (10 May 2012);

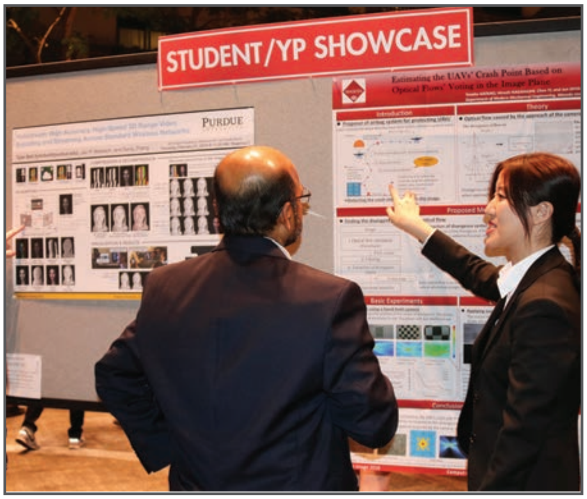
[3] J. Y. C. Chen, R. N. V. Oden, C. Kenny, J. O. Merritt, "Stereoscopic Displays for Robot Teleoperation and Simulated Driving," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, no. 19, pp. 1488-1492, 2010

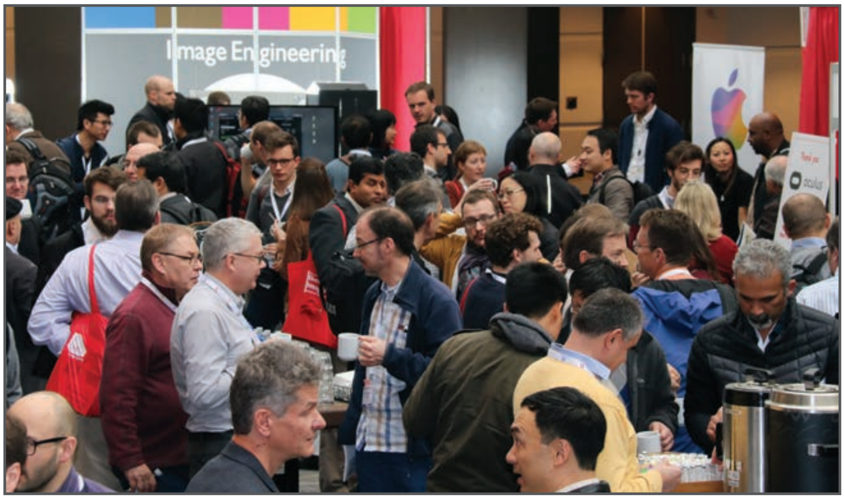
[4] E. H. Spain, "Stereo advantage for a peg-in-hole task using a force-feedback manipulator," in *Proc.SPIE*, vol. 1256, pp. 1256 - 1256 - 11, 1990

[5] E. H. Spain, K. P. Holzhausen, "Stereoscopic versus orthogonal view displays for performance of a remote manipulation task," in *Proc.SPIE*, vol. 1457, pp. 1457 - 1457 - 8, 1991

[6] D. Drascic, J. Grodski, "Defense teleoperation and stereoscopic video," in *Proceedings of SPIE (Stereoscopic Displays and Applications IV)*, pp. 58-69, 1993

[7] J. Y. C. Chen, E. C. Haas, M. J. Barnes, "Human Performance Issues and User Interface Design for Teleoperated Robots," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1231-1245, Nov. 2007.





- AR is a fundamental and transformative interface and platform.
- 20+ years from my initial working demo to today's AR displays
- Optimistic that it won't take another 20 to reach ubiquitous consumer AR!