MEASURING TRACKING PERFORMANCE OF VIDEO-BASED AUGMENTED REALITY SYSTEMS FOR DESIGN AND CONSTRUCTION

Xiangyu Wang

Key Centre of Design Computing and Cognition, Faculty of Architecture, Design & Planning, University of Sydney, Australia

ABSTRACT: Recent advances in information and communication technology have brought computer vision tools to support construction and design operations. Video-based see-through Augmented Reality (AR) systems can merge live video streams with computer-generated information. Since video-based AR systems have a digitized image of the real environment, it is possible to detect features in the construction site and use those to enforce registration of computer-generated information onto the user's real world view of construction site. Factors of sensing devices (e.g., video camera) including color tracking, zooming capabilities, focusing modes, and motion tracking could impose inherent delays on the user's real world view. This paper presents a study conducted to investigate those factors and augmentation capability of video camera with reference to real-time streaming. The results assess the effectiveness and efficiency of video camera with motion recognition capabilities, and also the ability to augment the reality.

KEYWORDS: virtual reality, augmented reality, computer vision, robotics, construction site.

1 INTRODUCTION

Video-based see-through Augmented Reality (AR) systems can merge live video streams with computer-generated information and display the resulting images onto the screen. The augmentation may be placing virtual geometric objects into the real environment, or displaying non-geometric information about real objects. By exploiting human's visual and spatial skills, AR can introduce digital information into the user's real world instead of constraining the user into the totally computer-generated virtual world. This paradigm for human-computer interaction provides a promising new technology for many applications. Currently, there exists several application domains for this cutting-edge technology including: medical imaging (Bajura et al. 1992; Lorensen et al. 1993; State et al. 1996; Grimson et al. 1994), robot (Milgram et al. 1993), manufacturing (Caudell and Mizell 1992; Rose et al. 1995), etc. Augmented Reality has the capability for passive or active viewing and manipulation of digital information (e.g., 3D design models), but also for "augmented" control interfaces to conventional machines or robotic mechanical systems. For instance, operator sitting in a backhoe cabin equipped with wearable AR control interfaces that can add digitally based information that are sensed and modeled by computer such as subsurface data, in the form of graphics, to be displayed in the operator's view of the real working area can visually locate the accurate as-built position and depth of buried infrastructure with enhanced decision-making capabilities.

Video-based AR systems rely on computer vision approaches for accurate registration. This process is carried out by analyzing images acquired from cameras at different locations, in order to recover the camera location with respect to the scene and to use these parameters to estimate the scene structure. Such 3D spatial information is used to position and orient the video scenes according to the tracked viewpoint of users. Since video-based AR systems have a digitized image of the real environment, it is possible to detect features in the construction site and use those to enforce registration of computer-generated information onto the user's real world view of construction site. Factors including color tracking, zooming capabilities, focusing modes, and motion tracking could impose inherent delays on the user's real world view. The focus of this paper presents a study conducted to investigate those factors and augmentation capability of video camera which is used as the video input device in video-based AR systems with reference to real-time streaming. The study attempts to assess the effectiveness and efficiency of video camera with motion recognition capabilities, and also the ability to augment the reality.

This paper also presents an experimental method to assess the effectiveness of video camera. There is no illusion that comparing an accessible commercial product to the thousand dollar tracking cameras, held for example by digital air, will have almost no relation, however this experiment attempts to establish the limitations for special effects that can be produced from a commercial product. AR researchers for construction applications would find this experiment relevant to find actual values of the capabilities and take these into consideration when recording video streams for their AR systems in the future. Lighting variables and distance may become a real consideration

next time when reality is augmented by placing special effects upon the display.

2 TRACKING IN VIDEO-BASED AUGMENTED REALITY SYSTEMS

Tracking technology is used widely today in a variety of industries including, but not limited to, medicine, building, archaeology, engineering, entertainment, government and defense, aviation, computing and construction. Specific examples include interactive graphics, Virtual Reality games, vehicle and flying simulators, conferencing, Virtual Reality walk-through, augmented and immersive realities, surgical imaging and production techniques for television and movies. As defined by Welch and Foxlin (2002), motion tracking has five main purposes in a computer system:

- *View Control* – controls the 'virtual' camera pose and position – which can be manipulated and merged with 'real' camera position i.e. superimposed.
- *Navigate* – Enables movement through a virtual world
- *Object Selection* – Allows manual manipulation of virtual objects (e.g. a tracked 'glove')
- *Instrument Tracking* – Allows the integrated manipulation of virtual objects with real world objects e.g. computer-aided surgery
- *Avatar animation* – The creation of animated characters through full body motion capture (*MoCap*) of human actors, animals and objects.

In general, tracking system breaks down an image by frames and looks for relationships of changing components in a frame; a sensor samples movement, computation applied and an estimate is given. The rate of estimation is dependant on computation but also the rate of transfer – between hardware, software and back to hardware. This is particularly applicable in real time applications. A motion tracking system may include one or a combination of mechanical, inertial, acoustic, magnetic, optical or radio frequency sensors, this is commonly entitled hybrid tracker. According to Harrington (2001), the major limiting factors in motion tracking is accuracy, jitter, latency and latency jitter. Without accurate tracking and registration, the virtual objects will not appear in the correct location at the correct time and be unbelievable to the viewer. Jitter is the noise or shaking of a stationary virtual object and is caused by calculation variations. Latency refers to the lag time and is quantified in milliseconds (approx. between 16 and 120 ms). Latency jitter refers to frame by frame lag.

3 EXPERIMENT

The objective of the experiment is to observe and gauge the effectiveness and capabilities of the sensing device (video camera) typically used in video-based see-through video-based Augmented Reality, with respect to facial tracking, color recognition, and quality of capture. These results are recorded, coded and interpreted in an extrinsic analysis to derive meaningful results, both visual and contextual. A commercial video camera (Logitech QuickCam MP Sphere) was selected for testing in order to demonstrate the entire experimentation method. This study helps to gain insight into limitations in motion tracking and capture, and how they could possibly be rectified. The results from the study could also help to define the requirements for a reliable sensing device used in video-based AR system, where an acceptable accuracy of merging computer generated information into the real construction working space could be realized in a fast and efficient manner, and construction site personnel will be able to interact with knowledge and information.

3.1 *Materials and procedure*

The experimental devices include overhead lights, lamp, Logitech Sphere Webcam, Sony T-5 digital camera, tripod, laptop computer as shown in Figure 1. There are totally four groups recruited for the experiment: participants went through a number of experimental proceedings in a controlled room with the sensing camera whereby the user was asked to perform an action while appropriate data was recorded computationally (see Figure 1 and 2 for experimental setup). This was complimented by the video stream recording via the digital camera and the sensing camera itself. The data was then interpreted and transcribed in a thorough analysis to derive meaningful results. The room was set up so that every action the user undertakes was recorded visually on the digital camera regardless of data, even though it might have no relevance to the experimental proceedings. This ensures that an extrinsic analysis was undertaken post-experiment. The overhead lights and lamps allowed control over the lighting of the subject thus to control a number of variables that the camera reacted with, these include color, saturation, brightness and hue. Anchor cards symbolized how far the subjects had to step in order to record a reaction from the camera. Experiments upon the 2 axis of x (left/right) and z (front/back) were conducted and recorded by transcription and later inputted into the spreadsheet set up on the computer. The two experiment coordinators timed and recorded the video stream and camera reactions manually and later organize the raw data into tables.

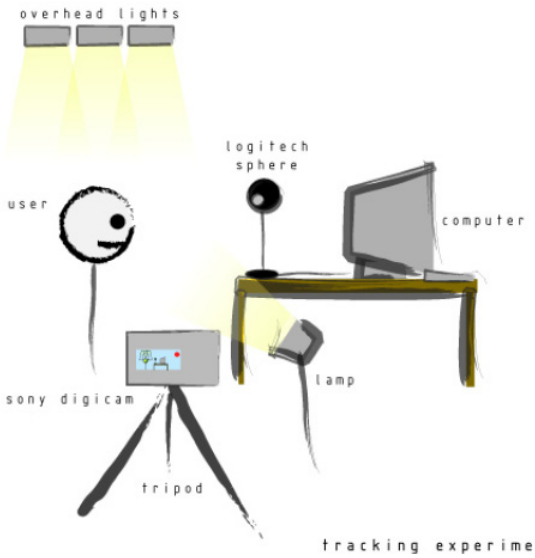

Figure 1. Graphical layout of devices in the tracking experiment.

The participant's actions have direct influence upon the camera's reaction. This creates a unique relationship between the two, that both nodes will affect each other, this will form the basis for the raw data that will be captured. The capture devices are defined in the form of video, pen/paper transcriptions and computational recordings such as camera data streams. After these recordings were captured, an analysis was performed upon this raw data to derive meaningful information and results. The analysis held two stages that must be completed sequentially, beginning with the transcription method. This filtered the raw data into segments, ensuring a smooth encoding (stage 2) was completed.

Figure 2. Actual tracking experiment Setup.

3.2 *Variables*

There are a number of controlled variables (e.g., color tracking, zooming capabilities, focusing modes, and motion tracking) that are manipulated during the experimental proceedings to ensure all extremes of the camera capabilities are met and identified. As mentioned earlier the lighting can affect variables concerned with data recording and camera reactions, thus a lamp was used to provide a flood light upon the subject to drown out unwanted colors. This is necessary for the tracking and color procedures so that irrelevant colors are not picked up and contort data recording. The overhead lights have also added an extra control to the lighting environment to compliment the lamp. The distance of the subject to the camera will also have direct influence upon web camera reactions and tracking capabilities.

Variables concerned with computational activities such as processing speed and pixel resolution will not directly affect the recording proceedings thus they are only defined but not controlled. The webcam interface itself held a number of controlling variables such as color boost, light boost, auto zooming, single face tracking and multiple face tracking. Color boost and light boost will have its greatest influence upon the reaction time latency. The better the lighting conditions, the easier the camera can pick up specific colors and track it.

4 RESULTS AND ANALYSIS

Figure 3 shows each participant moving forward on the z-axis towards the camera. The reaction time is specifically timing how fast the camera reacted to the movement and began zooming. Here we can see the trend that as the participants moved closer towards the camera it took longer for the final movement of the camera to complete, thus increasing the reaction time. The subject appears larger and hence the persons face is magnified; the camera has trouble detecting this as it is unable to zoom out to maintain facial tracking. Thus it takes a its toll on the reaction time as the camera must refocus on the subject, identify itself with a face and then proceed to zoom in. Furthermore, after close speculation we derived that the effectiveness of the camera refocusing and correctly finding a face was inconsistent. It appeared to us during the experiment that the camera became hasty and vaguely tracked on a subjects face. However, it is important to consider the person's complexion and hair style, as these are benefactors to this issue.

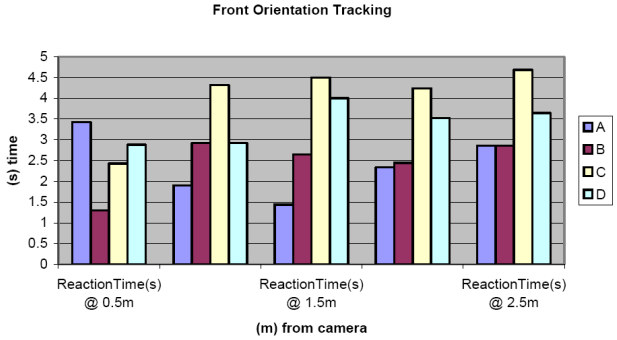

Figure 3. Front Orientation Tracking Reaction Time.

Figure 4 shows that the zoom percentiles are quite consistent as the distance increases significantly. However, at 0.5m and 1.0m participant A and B were zoomed on more accurately than the full 300%, indicating a more precise zoom value tracking for those participants. After observations during the experiment we derived that the auto-zoom selection made the camera zoom into a persons face as much as possible.

Once again after detailed speculation and analysis of the videos we decided that this is the case as the face is the main focus and hence makes the webcam more accurate and responsive in following the subjects movements. However, these are only speculations as after testing movement whilst at 300% the camera was not able to track sudden movements. This is most likely due the maximum zoom selection.

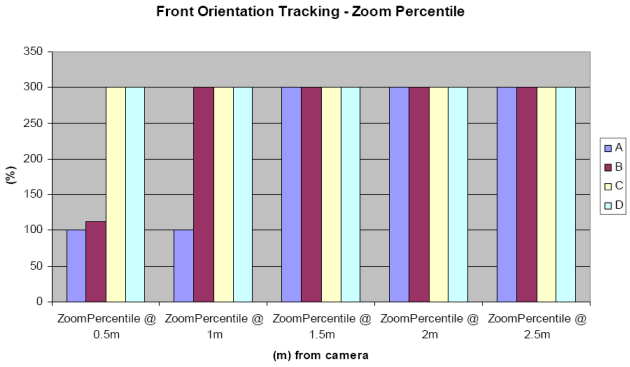

Figure 4. Front Orientation Zoom Percentile.

Here the reaction time was gauged upon the rear movement upon the z-axis in Figure 5. Participants C and D had some unexpected tracking problems at 1.5m and 2.5m. The smallest movement 0.5m sees a consistent tracking reaction time for all participants. However, after this at 1.0m participant C was not tracked at all and progressively for the rest of the increase in distance to the rear. Participant D had the same effects after 1.0m, with no tracking occurring at all from 1.5m to 2.5m. We learned here that once again the tracking of the camera was very inconsistent, however, we have reason to believe that a contributing factor(apart from the factors mentioned earlier) to this is the speed of the subjects movement to their distance from camera.

We have already mentioned that the camera has difficulty in picking up swift movements, and therefore a possibility could have been that they were not tracked. Despite this being a factor, it fails to effectively cover this issue as in some of the cases, movement to the points was slow. An alternative explanation we discussed would be that the object remained in the view of the camera yet no zooming occurred; however, seeing as though the auto-zoom function was enabled, this could not have been the case.

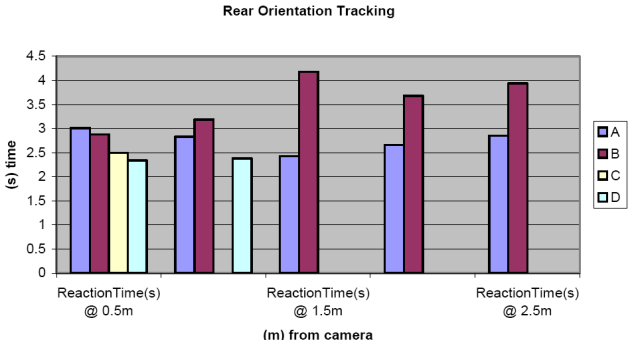

Figure 5. Rear Orientation Tracking Reaction Time.

Again as mentioned above, participant A had precise zoom percentiles resulting in accurate tracking results (as shown in Figure 6). Participant C however failed to track to the rear with no zooming or reaction occurring after the 0.5m mark. As mentioned above this could be as a result of fast movement, however, after studying the results for the subjects, it seems that for this to happen on four consecutive occasions means that another factor is depriving this from occurring. Therefore, we could only speculate on this issue which resulted in somewhat inconclusive results.

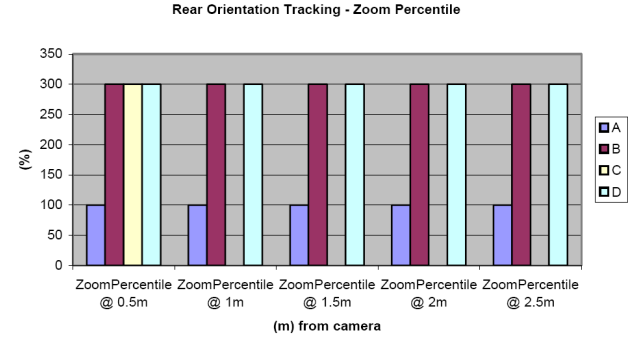

Figure 6. Rear Orientation Zoom Percentile.

The results (see Figure 7) here vary quite a lot with the participants moving along the x-axis to the left of the web camera. At all distances, participant C was tracked most efficiently and quickly with comparison to the participant's times. Participant D had the longest times for the camera to settle with a final destination of movement. We found that the camera was more likely to track a persons face as long as their distance x-axis position from the camera remained the same. As well as this, rather than stepping forward or backwards, the camera, at 100% zoom retains the subject in view and hence facial tracking is more accurate.

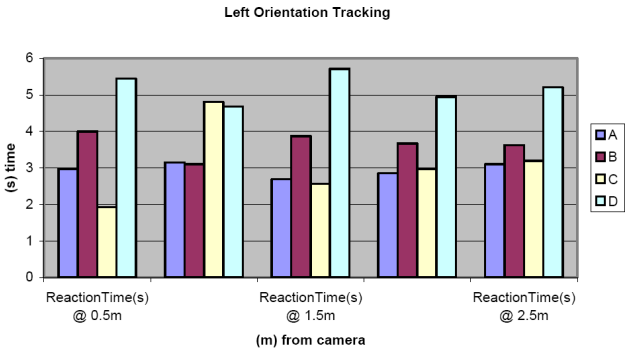

Figure 7. Left Orientation Tracking Reaction Times.

The only variation from the maximum zoom percentage of 300 occurred with participant C whereby they were tracked and zoomed upon more accurately for the increasing distances (see Figure 8). As mentioned above seeing as though the subject remained on the same x-axis this contributed to more accurate face tracking. The movements to the left are more effectively noticed and reacted upon compared to movements to the back or front.

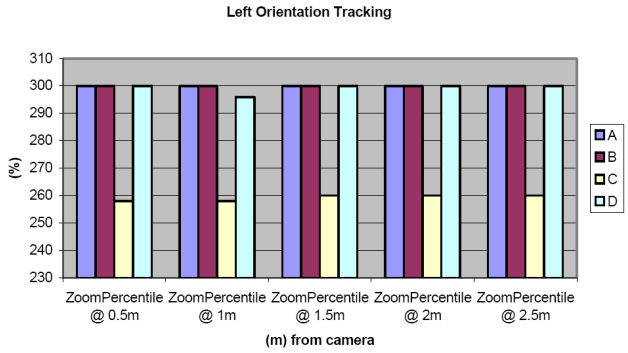

Figure 8. Left Orientation ZoomPercentile.

The results in Figure 9 from the right directional movement of participants along the x-axis graph a consistent reaction time increase with the increase in distance. At 0.5m it would be expected that the shortest reaction time would occur, which it does. At 2.5m each of the participants reaction times have peaked to their largest range. The results and conclusions derived from what we learned in the left orientation tracking applied here too.

Unfortunately no significant information can be derived from the graph in Figure 10, because all the zoom percentiles stayed the same as the distance increased. It would be expected that as the distance increases, the reaction time to settle on a zoom percentile and position would increase, however this did not occur.

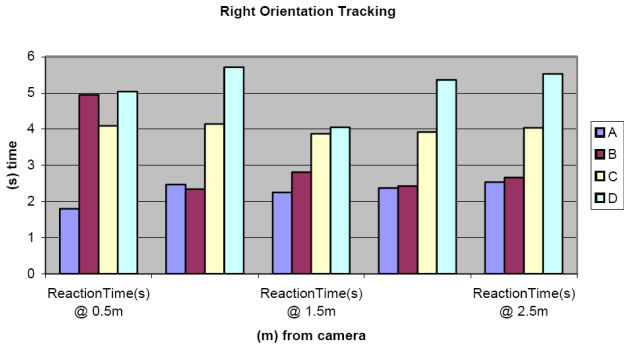

Figure 9. Right Orientation Tracking Reaction Time.

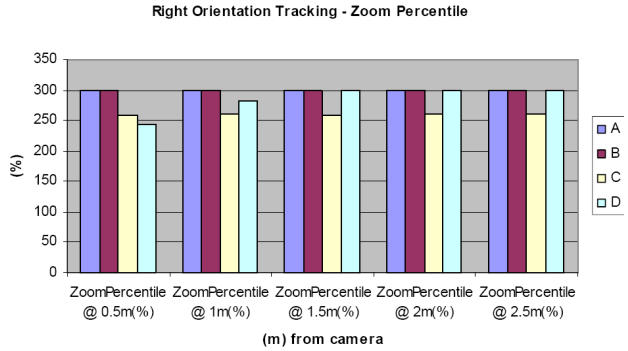

Figure 10. Right Orientation Tracking Zoom Percentile.

5 GENERAL DISCUSSION AND SUMMARY

Although there is significant video stream recording, pen/paper transcriptions and tabulated data spreadsheets, the reaction from the camera itself was not outputting interesting values. Most of the time it would zoom to its full capacity 300% blurring the subject out of total focus, while the reaction time for moving along the x-axis was sometimes lost, perhaps due to the participants hair covering certain pixel mapped areas of the face, whereby the camera could no longer track. These speculations are only that and the results cannot be fully interpreted with the hardware setup that this experiment conducted. However, we do see some valuable information as to how far the camera tracked up until – 3.5m along the x-axis and z-axis proved to be the final limitation on the distance the subject could move until no more tracking occurred. The values themselves in the table are too close together to derive any significant differences between distances or subjects themselves. However, hair style appeared to make a significant difference whether or not the subject was focused upon.

This paper conducted a study to investigate the augmentation capability of video camera with reference to real-time streaming. The results of the study assessed the effectiveness and efficiency of video camera with motion recognition capabilities, and also the ability to augment the reality.

REFERENCES

Bajura M., Fuchs, H., Ohbuchi, R. (1992). Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. Edwin E. Catmull (ed.): *Computer Graphics (SIGGRAPH '92 Proceedings)* 26(2), July 1992, pp. 203–210.

Caudell, T. and Mizell, D. (1992). Augmented Reality: An Application of Heads-Up Display Technology to Manual Manufacturing Processes. *Proceedings of Hawaii International Conference on System Sciences*, January 1992, pp. 659–669.

Grimson, W., Lozano-Perez, T., Wells, W., Ettinger, G., White, S. (1994). An Automatic Registration Method for Frameless Stereotaxy, Image, Guided Surgery and Enhanced Reality Visualization. *IEEE Conf. Computer Vision and Pattern Recognition*, Seattle, WA, June 19-23, 1994, pp. 430–436.

Harrington, M. (2001) *Controlling Motion-Tracking Devices: Navigating simulated worlds* Article published on Dr. Dobb's Portal A website posting software information and related interest articles Pg 1. url: http://www.ddj.com/184410885

Lorensen, W., Cline, H., Nafis, C., Kikinis, R., Altobelli, D., Gleason, L. (1993). Enhancing Reality in the Operating Room. *Visualization '93 Conference Proceedings*, IEEE Computer Society Press, Los Alamitos, CA, October 1993, pp. 410–415.

Milgram, P., Zhai, S., Drascic, D., Grodski, J.J. (1993). Applications of Augmented Reality for Human-Robot Communication. *Proceedings of IROS '93: International Conference on Intelligent Robots and Systems*, Yokohama, Japan, July 1993, pp. 1467–1472.

State, A., Livingston, M., Garrett, W., Hirota, G., Whitton, M., Pisano, E., Fuchs, H. (1996). Technologies for Augmented Reality Systems: Realizing Ultrasound-Guided Needle Biopsies. *Computer Graphics Proceedings, Annual Conference Series:* SIGGRAPH '96 (New Orleans, LA), ACM SIGGRAPH, New York, August 1996, pp. 439–446.

Rose, E., Breen, D., Ahlers, K., Crampton, C., Tuceryan, M., Whitaker, R., Greer, D. (1995). Annotating Real-World Objects Using Augmented Reality. *Computer Graphics: Developments in Virtual Environments (Proceedings of CG International '95 Conference)*, Leeds, UK, June 1995, pp. 357–370.

Welch, G. and Foxlin, E (2002). "Motion Tracking: No Silver Bullet, but a Respectable Arsenal," IEEE Computer Graphics and Applications, special issue on "Tracking," November/December 2002, 22(6): 24–38.