Virtual Reality Meets Computer Vision

a Dissertation Proposal by

Mike Bajura

May 17, 1993

Abstract:

This proposal explores the merging of computer vision and interactive computer graphics. Computer vision techniques seek to construct models from sequences of images. Computer graphics techniques create sequences of images from models. Combining these two disciplines yields a powerful, interactive, image-oriented system. A prototype system is described which senses a head-mounted display user's environment from the head-mounted display see-through video stream and correctly composites real-world and synthetic head-mounted display images. Such a system is fesable and will provide improved perception for see-through virtual environment applications.

Introduction:

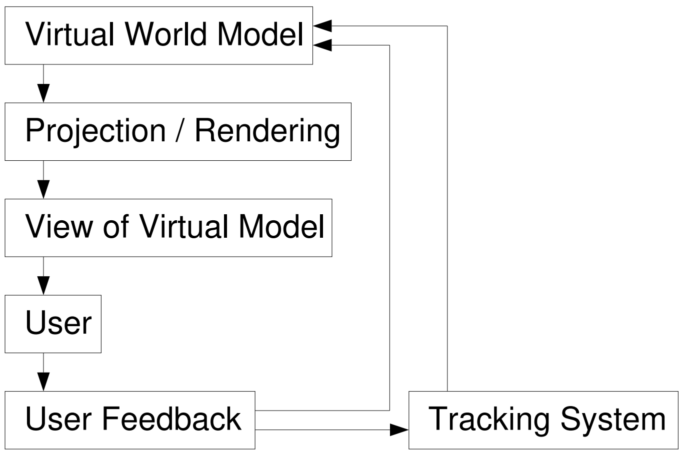

Figure 1: Current Interactive Computer Graphics Systems

Current interactive computer graphics systems, i.e. head-mounted displays, are represented schematically by figure 1. A view of a virtual world model is presented to the user. The user interacts with the virtual world model by changing model parameters either directly, or by means of a tracking system which reports positional information. In most common head-mounted display systems, a tracking system is used to report the position of both a user's head-mounted display and hand. Additional feedback buttons are provided which can directly modify the virtual world model. The user is part of an interactive feedback loop where he responds to images which are altered by his movements. The drawback to this system is that the user can only interact with virtual environments modeled with a virtual world model.

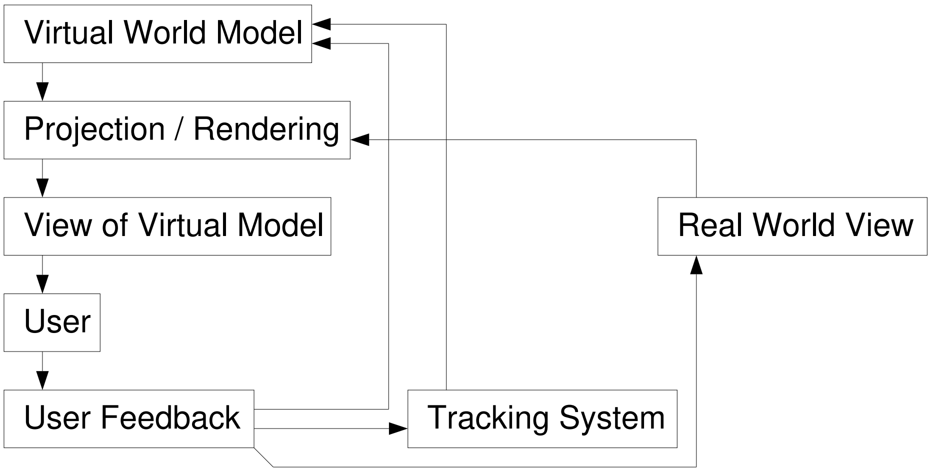

Figure 2: Augmented Interactive Computer Graphics system(Siggraph 1992)

A see-though head-mounted display is diagrammed in figure 2. It allows the head-mounted display user to "see through" his head-mounted display and view both his natural and virtual environments. A see-through head-mounted display is essentially the same system as in figure 1 with the addition of a feedback loop from the user back into the image rendering process. [Bajura, Fuchs, Ohbuchi] used a head-mounted television camera and video mixing equipment to insert a real-world view into a computer-generated virtual-world view. Other systems have used half-silvered mirrors to mix real and virtual images. The drawback with using half-silvered mirrors, however, is the inability to completely mask real, particularly bright, objects in the user's field of view with virtual ones.

The see-through head-mounted display(figure 2) improves over the basic head-mounted display(figure 1) by allowing a user to interact with both his natural and the virtual world environments. However there are problems with this approach. One problem arises as soon as objects in the user's virtual environment need to interact with real-world objects. One would like real-world objects to obscure virtual-world objects and vice-versa when they are either in front of or behind one another. This is difficult to model because it involves knowing the location of objects in the real-world view relative to virtual-world objects. The location of real-world objects for a static environment can be pre-mapped into the virtual world model and used for virtual object and real object interaction, but this is neither a long-term nor a dynamic solution. Other problems with the see-through head-mounted display include poor registration and system lag. Poor registration is apparent when virtual objects and real-world objects do not appear to be in place relative to one another. The registration problem is significantly worse in dynamic situations when there is real-world movement which is not reflected in the virtual-world model for several image frames. This happens because the real-world view feedback loop is much shorter than the virtual world model feedback loop. The problem of registration and system lag is also present, but not as apparent, in non see-through head-mounted display systems.

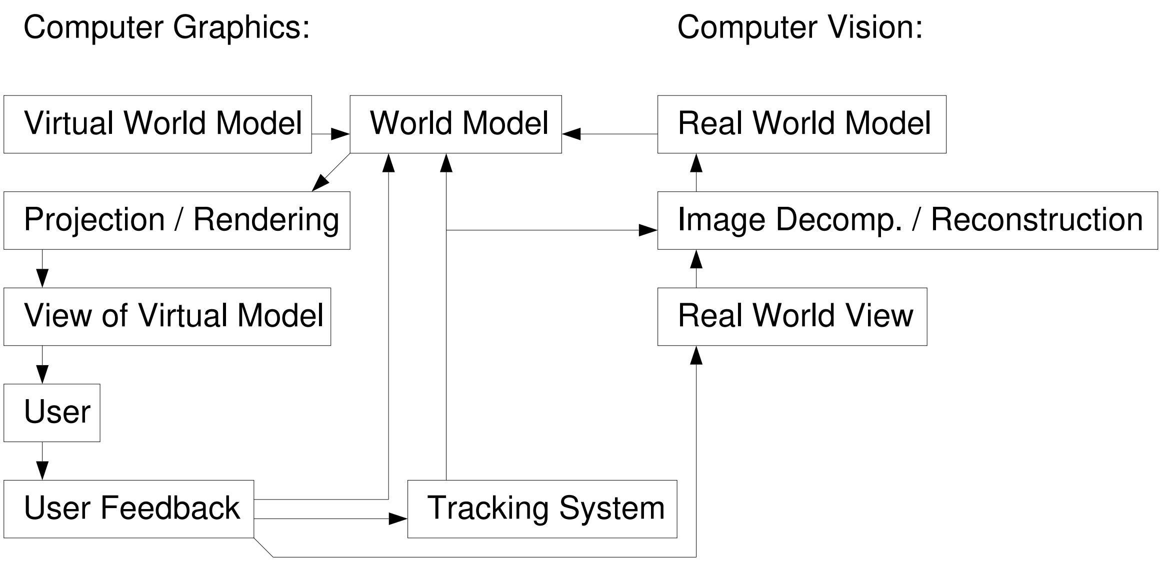

Figure 3: Future Interactive Computer Graphics Systems

An interactive computer graphics system of the future is shown in figure 3. Computer vision techniques are applied to the user's view of his surroundings to abstract a model of the user's real-world surroundings. This model is combined with the virtual-world model to produce a complete model of the user's environment. The user is then presented images of his complete environment. This is the general solution to the problems with the see-through head-mounted display of fig. 2. Virtual object and real-world object interactions can be modelled correctly because there is now a model of them on which to operate. Image registration can be done correctly by locating features in the real-world view and using them to line up the real-world view and virtual world model. The swimming problem associated with system lag is avoided because the correct real-world view is rendered with the correct virtual-world model. However, the problem with absolute system lag still remains.

Several desirable capabilities naturally fall out of this "system of the future":

1) Image composition and real-world and virtual-world interactions are done correctly.
2) Automatic generation of computer models for real-world objects. Parts of the abstracted real-world model can be saved as objects for later use.
3) Tracking in the user's natural environment, i.e. replace the expensive and range-limited tracking systems currently in use for head-mounted display applications. Once a real-world model is created of the user's environment, feature points in the user's environment can be used as tracking beacons [Bishop], much like the optical ceiling at UNC [Ward, Azuma, Bennett, Gottschalk, Fuchs].
4) Optical recognition of primitive hand gestures (no 3D "mouse" needed).

How it can work:

Recent advances in image generation (Pixel-Planes 5 [Fuchs +]), tracking systems [Ward +], and parallel computation (Pixel-Planes 5 + real-time frame grabber and HIPPI interconnect to MasPar and Cray Y-MP) at UNC, and computer vision in general, now make it possible to consider combining computer vision with interactive computer graphics.

The Computer Vision Question: How?:

Due to the computationally intractable problem of matching features in multiple stereo views, the only algorithms left to consider are ones based on monocular views which may vary over time. There are two ways depth can be practically recovered from a monocular view. One way is a "depth from focus" approach where multiple images are taken along the same image axis with different image apertures. The difference in focus for different features in the image is then correlated with distance to the features. [Pentland +] reports using off-the-shelf hardware with this method to compute range over a 1 meter cubed volume to an accuracy of 4.5 bits at 4 Hz. This range is precisely the range most used in head-mounted display applications at UNC.

The second approach is to track image features as they change over time [Broida, Chellappa] [Skifstad]. Eventually this results in the same feature being imaged from multiple directions allowing a triangulation to occur. If the imaging camera's position and orientation are recorded by a tracking system, the 3D position of feature points can be solved by a linear least-squares system. If the 3D position of the camera and the 3D position of the feature points are not known, this amounts to the non-linear structure from motion problem. There are two ways to consider solving the structure from motion problem. One way is a non-interactive start-up procedure which solves for the 3D position of feature points by solving the appropriate system of non-linear equations for some start-up image sequence. Once the 3D position of feature points is known, the camera projection matrix is a linear function of the known feature points. The camera position and orientation can be computed from the camera projection matrix in closed form. The other solution to structure from motion is an iterative solution strategy which dynamically converges to a solution. In either structure from motion solution there is scale factor which cannot be recovered which must be manually adjusted. There is a linear solution to the structure from motion problem [Tomasi, Kanade] but only under the assumption of orthographic projection, which is not a valid assumption in head-mounted display applications.

It is likely that a working system will require a hybrid implementation of both of these computer vision techniques. Feature point detection is needed to accurately measure feature points which are then fit with polygonal approximation based on Delaunay triangulation [Wang, Bowman, Brady, Harris]. Depth-from-focus is needed to help detect range between feature points which will assist the tiling process and will detect spurious features like hand movement in the user's field of view.

Other approaches to recovering depth are based on shape from shading [Oliensis, Dupuis], and detection of image flow [Bergen+]. Neither of these methods has progressed far beyond theory and largely qualitative results.

Computer vision algorithms and image reconstruction using real images must also account for the geometric distortion of a real imaging system. The largest geometric distortion is generally positive radial, or "pincushion", distortion. It is common to correct for distortion by fitting a polynomial approximation to a test pattern imaged through a distorted system [Weng, Cohen, Herniou]. This procedure can be done off-line and need not be interactive.

The Computer Vision Question: Can it be done in interactively?

QUITE LIKELY YES, AND IN THE FUTURE DEFINITELY YES. Current see-through head-mounted display systems are able to correctly model the interactions of a static real environment with a virtual one [Bajura tape 1992]. With the addition of a real-time frame grabber to Pixel-Planes 5, Pixel-Planes 5's i860 processing power, and optional video bandwidth connections to both MasPar and Cray Y-MP computers, a real-time depth-from-focus implementation is definitely possible at UNC. Real-time feature tracking in the Pixel-Planes 5 system is possible by dividing incremental feature windows among i860 processing nodes.

The remaining question is: can the linear least-squares problem (tracked camera) or the non-linear structure from motion problem(non-tracked camera) be solved either completely or incrementally, and can this information be incrementally added to a world model? The linear least-squares problem definitely can be solved in real-time with one or two i860 processors. It is not known how long it would take to solve the structure from motion problem either non-linearly with the downhill simplex method (this is an incremental method which converges quickly from close starting points) or with a least-squares type method which alternately solves for both camera position and feature position which are linear with respect to each other. [Taylor, Kriegman, Anadan] explores a least-squares approach which is claimed to be well-suited to parallel architectures. [Thomas, Oliensis] explores a recursive multi-frame strategy which converges as more images are incorporated into the solution. Another approach could be to solve for structure at a large (blurred) scale, and use that information to constrain a smaller scale solution. The important point is that so long as the solution is "reasonably fast" to update the model of the real world it is acceptable. In the case where both the camera and feature points aren't known, once an estimate for the feature points is known, the camera projection matrix can be solved by a linear least-squares fit to the feature positions, which can be done interactively.

The Dissertation Proposal:

The Thesis:

A video see-through Head Mounted Display (HMD) user's visual comprehension can be enhanced by applying real-time computer vision techniques to the HMD user's see-through video stream to detect surfaces and objects in the user's physical surroundings. Such a system is feasible and practical to build and will provide improved perception for see-through virtual environment applications.

The Experiment:

To design and build a prototype system to sense the HMD user's environment from the HMD see-through video stream and use that information to correctly composite real-world and synthetic HMD images. This will involve:

1) Identifying existing computer vision algorithms which are applicable to this problem.

2) Adapting them to run on the Pixel-Planes 5, a graphics multicomputer with a real-time video digitizer, 5 Gbit/sec. communication network, and 64 40 MHz i860 processor nodes [Fuchs+]. This system has sufficient processing power to execute depth recovery algorithms on a real-time video stream, generate synthetic images, and composite final HMD images from real-time video and synthetic images.

3) Demonstrating an application which uses the computed depth information to perform correct real object and virtual object occlusion.

Contributions of the research:

The contribution of this research will be the details of what is learned by constructing a prototype system. There are some of the questions which will be answered:

- How to use results in computer vision to reconstruct a dynamic real-world model.

- How the reconstructed real-world model interacts with the virtual-world model, particularly in the presence of uncertainty about the real-world model. Should the real-world model be semi-transparent in the regions of uncertainty?

- How to correctly render images of the combined real and synthetic environments.

- Insight into the sources of error (noise) and their effect on the prototype implementation.

Progress Timeline:

Summer 93:	Completion of real-time video digitizer hardware for real-time video input to the Pixel-Planes 5 graphics multicomputer, library research, design of additional optical hardware, initial experimentation with algorithms, simulation. Write draft of introductory chapter of dissertation.
Fall 93:	Conversion of algorithms to run on Pixel-Planes 5 multicomputer, initial experiments with prototype system. Write drafts of middle chapters of dissertation.
Spring 93:	Refinement of hardware and software. Development of prototype application. Refine drafts.
Summer 94:	Finish lab work. Complete drafts, write final chapter of dissertation. Graduate.

This is somewhat aggressive. Satisfying additional requirements and allowing for things to take longer than they optimistically should puts graduation somewhere in spring 1995.

References:

Bajura tape 1992: Test video showing real-world and virtual-world interaction.

Bajura, Fuchs, Ohbuchi: "Merging Virtual Objects with the Real World: Seeing Ultrasound Imagery within the Patient", Computer Graphics Vol. 26 #2 (Proceedings SIGGRAPH 1992), pp. 203-210.

Bergen, Anandan, Hanna, Hingorani: "Hierarchical Model-Based Motion Estimation," ECCV 1992, pp. 237-252.

Bishop: "Self-Tracker: A Smart Optical Sensor on Silicon," PhD Thesis 1984, UNC.

Broida, Chellappa: "Estimating the Kinematics and Structure of a Rigid Object from a Sequence of Monocular Images," IEEE PAMI Vol. 13 #6, June 1991, pp. 497-512.

Fuchs, Poulton, Eyles, Greer, Goldfeather, Ellsworth, Molnar, Turk, Tebbs, Israel: "Pixel-Planes 5: A Heterogeneous Multiprocessor Graphics System Using Processor-Enhanced Memories", Computer Graphics Vol. 23 #3 (Proceedings SIGGRAPH 1989), pp. 79-88.

Oliensis, Dupuis: "Direct Method for Reconstructing Shape From Shading," ?Darpa Image Understanding Workshop, Jan 1992.?

Pentland, Darrell, Turk, Huang: "A Simple Real-Time Range Camera," IEEE CH2752-4/89/0000/0256.

Skifstad: "High-Speed Range Estimation Based on Intensity Gradient Analysis," PhD Dissertation Univ. of Michigan.

Taylor, Kriegman, Anadan: "Structure and Motion in Two Dimensions From Multiple Images: A Least Squares Approach," IEEE Workshop on Visual Motion, Oct. 1991.

Thomas, Oliensis: "Recursive Multi-Frame Structure from Motion Incorporating Motion Error," Darpa Image Understanding Workshop, Jan. 1992.

Tomasi, Kanade: "Shape and Motion from Image Streams: a Factorization Method - Parts 2, 8, 10 Full Report on the Orthographic Case," CMU Technical report

CMU-CS-92-104 , 1992.

Ward, Azuma, Bennett, Gottschalk, Fuchs: "A Demonstrated Optical Tracker with Scalable Work Area for Head-Mounted Display Systems", Proceedings 1992 Symposium on Interactive 3D Graphics (ACM SIGGRAPH).

Weng, Cohen, Herniou: "Camera Calibration with Distortion Models and Accuracy Evaluation.", IEEE PAMI Vol. 14 #10 Oct 1992, pp. 965-980.