Harnessing the human visual system for image based modeling: an interaction system

Y. Morvan and C. O'Sullivan

Interaction, Simulation and Graphics lab, Trinity College, Dublin, Ireland

Abstract

We present a novel interaction system for modeling 3D geometry from photographs. Given a set of stereo pairs, along with their camera parameters, the interface tracks the movements of the users and allows them to intuitively position the modeling tools in 3D and to browse the stereo pairs using head movements. The geometry being modeled is overlaid on the photographs using a stereo display. The use of the known camera parameters for displaying the geometry ensures proper registration with each photograph and thus results in correct 3D models.

Categories and Subject Descriptors (according to ACM CCS): I.3.0 [Computer Graphics]: General

1. Introduction

Many industries now have to satisfy the great demand for digital visual content. Authoring such content often involves labour intensive tasks, one of them being 3D modeling. While the creation of entirely new shapes does not lend itself to automatization, the reproduction of existing structures holds more promise. In spite of a huge body of research, the techniques that have been developed to pursue that goal still suffer from significant drawbacks. Except for laser scanning, the main idea behind them is to extract geometric information from photographs. These approaches mostly fall into two categories.

Multiple views matching [KS00], photometric stereo [HS05] and coded structured light techniques [ZCS02] reconstruct a dense depth-map of the captured scene. They typically have difficulties handling scenes that exhibit large surfaces of uniform irradiance or highly reflective materials. They are also very data-intensive, but their main drawback is the amount of noise present in the reconstruction. Feature extraction techniques identify corners, edges and/or silhouettes and assemble them into a mesh. They tend to make topological assumptions (concavity [MBR*00], type of scene [WZ02]) and often require human intervention to guide the reconstruction process.

Both categories suffer from the fact that we do not know yet how to make computers "understand" shapes as the human brain does. The motivation of this paper is therefore to leverage efficiently the proficiency of the human visual system at extracting shape from pictures.

The image based modeling system we propose places the user in a situation analogous to copying features of a photograph using a transparent sheet, but in three dimensions. This is achieved through a novel combination of stereo display, geometry projection consistent with the camera poses used to take the photographs, and motion tracking.

2. Related work

Most 3D modeling applications let the user display a background image in workspace viewports. If camera poses are available, it is possible to import them and keyframe them along with the appropriate background photograph so that they match. One can then obtain a viewport where the edited geometry is registered with the photographs of the scene by displaying camera views. The main inconvenience with this image-based modeling setup is that geometry placed in a viewport that appears to coincide with features of the background photograph can in fact occupy an infinity of positions along the rays of the camera projection for that view. Only by changing viewpoints can the user refine the initial placement. If the different viewpoints used do not have convenient alignment properties (*e.g.* orthogonality), this position refining process converges slowly. This is because changes made using one viewpoint will modify the projections in other viewpoints in a non-intuitive manner. This difficulty is compounded by the fact that changing the viewpoint has to be

done using the time slider. No interface that we have tested allows simultaneous browsing of the time slider while moving geometry, leading to a painstaking back and forth between interaction modes. Moreover, the time slider is one dimensional and therefore not an ideal tool to browse through viewpoint positions.

Software that offers image-base modeling features, like Image Modeler or PFTrack take a constraint driven approach to guide geometry editing. They rely on accurate initial features reconstructed by the program. Placing new landmarks can be inconvenient, particularly in regions where no views exhibit strongly contrasted patterns. Their 2D displays can also make the set of initial features (which consists of a vertex cloud) confusing to make sense of.

Many user interfaces have been proposed that combine stereo displays and 3D positioning devices. The work of Fiorentino *et al.* [FdMS02] makes use of motion tracked props and is a good starting point on the topic of 3D modelling interfaces. To the best of our knowledge, we are the first to propose an interface designed for image-based modeling based on these techniques.

3. System description

We describe the physical setup of our system. We then provide some implementation details. Finally, design choices for the user interface are discussed.

3.1. Physical setup

The modeling application is displayed on a conventional "GeoWall" passive stereo screen [SDW02]. The display installation thus consists of a non-depolarizing back projected screen (dimensions $200 \times 150cm$), two 2100 lumens DLP projectors mounted on an accurately adjustable stacker and two circular polarizing filters.

The motion tracking is performed by an optical system: reflective markers are detected by cameras on which infrared LED arrays are mounted. Marker positions are reconstructed by projective geometry using pre-computed calibration information for the set of cameras. Our installation makes use of a Vicon system, and we found that 6 tracking cameras are sufficient for our purpose. They cover a volume large enough to let the user navigate the space in front of the screen for comfortable viewing positions while letting him extend his arms fully. Motion is tracked at a rate of 100 Hz, which falls within the range of 60 to 125 hertz that mouse users are used to.

We decided to track props rather than the user's anatomy. The main reason is convenience: it is much easier to pick up a prop than to attach tracking markers. Additionally, it allows us to extend the user interface in the future by letting the user pick-up different props corresponding to different modeling tools. The props that we currently identify and track are the pair of passive stereo glasses and some wooden pincers.

A wireless mouse is placed in the user's non-dominant hand and its motion is not tracked. Figure 1 shows the setup of our system.

Figure 1: *System setup. Two of the Vicon cameras can be seen in the background. The bright glares are the reflective markers.*

3.2. Software

The application consists of two processes running concurrently:

- A proprietary Vicon server which broadcasts the position of tracked objects (in our case, props) in real time to client programs that connect to it.
- Our modeling program, built on top of OpenGL and GLUT, which connects to the Vicon server through a TCP socket.

The modeling program takes as input a set of photographic stereo pairs as well as the parameters of the camera for each photograph (pose and intrinsic parameters) expressed in cartesian coordinates. Optionally, a set of scene features, which are produced by most camera pose estimation programs, can be provided as initial vertices. The stereo pairs are assumed to be binocular photographs of the scene that the user wishes to model as seen from different viewing angles. The program behaves mostly like a polygonal mesh editing tool based on triangular faces, with a few key differences:

- The user is not free to navigate arbitrarily around the scene being modeled, but browses instead among the input (binocular) viewpoints.
- The stereo photographs corresponding to the current viewpoint are displayed as background images.
- The geometry being modeled is projected on the viewport corresponding to each eye using the camera parameters that were used to take the corresponding photograph.

As a result of these features, the 3D cursor manipulated by the user, as well as any geometry created, will appear to coincide with a feature of the modeled scene in both the left and right viewports if and only if their 3D coordinates match (in the metric space the camera poses are expressed in). When this happens, the binocular disparity cue [CWE94] correctly gives the user the sensation that the cursor or geometry is situated at the same depth as the feature. Figure 2 illustrates this property.

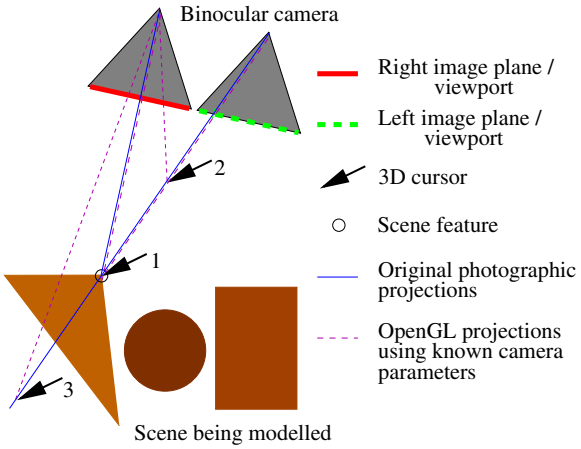

Figure 2: *Top view of a scene being modeled with one binocular viewpoint. Position* 1 *of the 3D cursor coincides with that of a feature and thus gets projected to the same point as that feature in both viewports. Positions* 2 *and* 3 *differ from that of the feature: even though their projections coincide with that of the feature in the left viewport, they do not in the right viewport. Thus the user correctly perceives them as misaligned.*

3.3. User interface

The interface for our modeling application does not use any on-screen menus. We experimented with them but found that the juxtaposition of a 3D workspace and 2D menu was visually confusing for the user. Since we limit ourselves to polygon editing, the combinations of prop states provide enough flexibility to cover all necessary features.

The pincers are held in the dominant hand. Their position as well as orientation is tracked, providing a 6 DOF cursor. The interface can detect whether the user is pinching or not. It also takes into consideration whether the cursor lies within the frustum of the current binocular camera. The wireless mouse is held in the other hand. Its buttons are used in the same way as the augmenter keys (shift, control, alt) are in a keyboard based user interface. The stereo glasses are only tracked for orientation. Tracking their position would indeed be counter productive in this context for several reasons. The first is that the stereo effect works best from specific head positions with respect to the screen. The second is that non depolarizing screens tend to have a reduced half-gain interval, this is the range of viewing angles outside of which projected light is reproduced below half intensity. Lastly, translating one's head often requires moving the whole body, the user could therefore not sit, and long editing sessions could become taxing.

We chose to implement an interface close to Blender's vertex edit mode. One difference is the use of a helper to constrain transformations to desired axes or planes. It is displayed at the barycenter of the set of currently selected vertices. Table 1 summarizes the available operations.

4. Practice

We applied our system to some test scenes. A few practical issues arose.

Some were related to the acquisition of properly registered camera information for each binocular view. Camera tracking software is the most convenient way of recovering camera poses that are consistent with each other within a metric space. However, such programs expect continuous video sequences to function robustly. This implies that, assuming a binocular stereo camera is available, the footage corresponding to one eye has to be tracked independently from that of the other, which leads to difficulties registering the two sets of poses with each other. Moreover, cameras capable of filming continuous stereo footage are uncommon and expensive, and synchronization issues make building one somewhat difficult. We chose to take a single video sequence with a standard monocular camera, and extract satisfactory stereo pairs from the set of tracked poses. This was done by implementing a "stereo suitability" function: given a pair of poses, it computes a grade based on how close they are to being separated in the lateral direction by the typical distance between human eyes ($\approx$ 7.4 cm), and how similar their orientations are. This approach places constraints on the video sequence used, which must consist of lateral translations that are slow enough to contain pairs of poses whose separation is close to 7.4 cm.

In practice, we attached the camera to a bicycle that we slowly pushed around the scene. Pose recovery was done using PFTrack, a commercially available camera tracking application. Its scene scaling feature allowed us to recover real world 3D coordinates for the camera poses from a single distance measurement between two identified features in each scene.

The interface uses small rotating movements of the user's head to let him intuitively browse the available stereo views. Left-right as well as up-down rotations are tracked. However, the set of input binocular viewpoints needs to be pre-ordered in order to make the mapping between head movements and browsing direction intuitive. A simple 2D indexing of stereo views based on their viewpoint's scene coordinates proved satisfactory.

5. Discussion and future work

Our system showed good potential to facilitate the modeling of existing scenes. The main gain is that it makes positioning vertices that match scene features much easier, thanks to the depth cue and the possibility to easily change viewing angle while placing a vertex. Displaying in stereo also makes it easier to edit faces, because vertices that would appear in a confusing cloud on a 2D display can be discriminated by depth.

In its current state, the system suffers from some limitations. One is that the editing options provided are very limited. We are working on adding primitive creation and manipulation through the use of additional props. In the longer term, curve and surface tools could be added. To extend the

Operation	Means
Select*	Place the cursor on a vertex and pinch.
3D Box select*	Place the cursor at the first corner of the box, pinch, drag while still pinching to the opposite corner, then stop pinching.
Move	Place the cursor on a vertex of the current selection, or on an axis of the transformation helper, pinch, drag while still pinching, stop pinching when at destination.
Duplicate	Same as **Move** but with the left mouse button pressed down.
Delete	Same as **Move** but stop pinching when the cursor is out of the frustum.
Undo	Start pinching on an empty space, drag the cursor until is is out of the the frustum, then stop pinching.
Redo	Move the cursor out of the frustum, start pinching, drag the cursor until it is back inside the frustum, then stop pinching.
Browse viewpoints	Hold the right mouse button down and rotate the head slightly in the desired browsing direction, release when the desired viewpoint is reached. The 3D cursor's position is frozen for the duration of viewpoint browsing. It can therefore be used to update the cursor's relationship to the user's hand position, letting him reach other parts of the scene more comfortably. It also helps the user acertain very quickly whether the cursor is at the desired location by observing how it moves with respect to the scene when viewed from a different angle
Create face(s)	Press the middle and right mouse buttons simultaneously. The current selection needs to contain at least three vertices. If more than three vertices are selected, a triangle strip is created based on their order of selection.

Table 1: *List of available operations and the means by which they are invoked. * Selection operations are modified by the mouse buttons: if the left button is down, the new selection is added to the current one. If the middle button is down, it is removed.*

palette of actions that the interface can handle, we plan to implement a menu mechanism similar to Maya's hotbox: a temporary menu would be displayed in 3D within the workspace at the cursor's position when a button is held down. The user could then pick an action with the cursor. Another limitation is that the only depth cues [CWE94] that the system provides are binocular disparity, motion parallax and foreshortening. Improperly conveyed vergence, accommodation, and to a much greater extent occlusion cues can confuse the user when they blatantly contradict the sensation given by binocular disparity. The display device being a flat surface, there is no hope of improving vergence and accommodation cues. However, the use of range information (even inaccurate, provided it is conservative) combined with a depth buffer could greatly improve positioning ease. From a prospective standpoint, one could argue that a system such as ours represents a significant investment. The problem does not lie with the stereo display, as these are already very affordable, but with the motion tracking device. We believe that optical motion tracking is maturing very quickly and could soon be deployed for user interface purposes at a fraction of the cost of the devices currently used by the animation industry.

We are currently working on a formal user study to compare the system we propose to both the traditional workflow of using a desktop geometry modeler with background images and that of dedicated image-based modeling software. We plan to compare execution times, reconstruction quality and user satisfaction.

References

[CWE94] CORENS., WARD L. M., ENNS J. T.: *Sensation and Perception, 4th ed.* Harcourt Brace, New York, 1994.

[FdMS02] FIORENTINO M., DE AMICIS R., MONNO G., STORK A.: Spacedesign: a mixed reality workspace for aesthetic industrial design. In *Proceedings of the International Symposium on Mixed and Augmented Reality* (2002), pp. 86–318.

[HS05] HERTZMANN A., SEITZ S. M.: Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 8 (2005), 1254–1264. Member-Steven M. Seitz.

[KS00] KUTULAKOS K. N., SEITZ S. M.: A theory of shape by space carving. *International Journal of Computer Vision 38*, 3 (July 2000), 199–218.

[MBR*00] MATUSIK W., BUEHLER C., RASKAR R., GORTLER S., MCMILLAN L.: Image-based visual hulls. In *369-374* (July 2000).

[SDW02] STEINWAND D., DAVIS B., WEEKS N.: Geowall: Investigations into low-cost stereo display systems. In *USGS Open File Report* (2002).

[WZ02] WERNER T., ZISSERMAN A.: New techniques for automated architectural reconstruction from photographs. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part II* (London, UK, 2002), Springer-Verlag, pp. 541–555.

[ZCS02] ZHANG L., CURLESS B., SEITZ S. M.: Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *Int. Symposium on 3D Data Processing Visualization and Transmission, Padova, Italy* (June 2002).