SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality System

Alex Olwal Hrvoje Benko Steven Feiner
Department of Computer Science, Columbia University
500 W. 120th St. 450 CS Building, New York, NY
{aolwal,benko,feiner}@cs.columbia.edu

Abstract

We introduce a set of statistical geometric tools designed to identify the objects being manipulated through speech and gesture in a multimodal augmented reality system. SenseShapes are volumetric regions of interest that can be attached to parts of the user's body to provide valuable information about the user's interaction with objects. To assist in object selection, we generate a rich set of statistical data and dynamically choose which data to consider based on the current situation.

1. Introduction

A major problem in developing systems that support multimodal interaction through speech and gesture is to determine the object(s) to which a user is referring, since irrelevant objects will likely fall into the user's gaze and pointing direction. We introduce *SenseShapes,* volumetric regions of interest that can be attached to the user, to provide valuable statistical information about the user's interaction with objects in the environment. We have created a multimodal augmented reality (AR) system in which the statistical data generated by SenseShapes assists the fusion of multimodal input into executable commands.

Our multimodal input devices consist of a modified Essential Reality P5 glove to sense hand gestures, a headset microphone, and two InterSense IS900 six-degree-of-freedom trackers to monitor head and hand position and orientation. IBM ViaVoice 10 is used for speech recognition, and a Sony LDI-D100B optical see-through head-worn display presents the AR overlay. Our system receives the gesture events shown in Figure 1 (currently "point," "grab," and "thumbs up") and speech commands that, with the aid of the SenseShapes, are integrated into valid actions. SenseShapes are also being used in a different multimodal AR system [2] that, unlike the one described here, employs mutual disambiguation of unimodal recognizers, does not track the user's fingers, and uses a static weighted combination of statistics instead of the dynamic integration process described in Section 3.

2. SenseShapes

SenseShapes are volumetric primitives (spheres, cubes, cylinders, and cones) that we attach to the user (e.g., a pointing cone attached to the hand). Previous work in AR and VR has included selection of objects using rays or cones attached to the user's hand [3] and head [4] and computing object intersections with these volumes, or has used the projection of a tracked point on a glove to perform selection on the image plane [5]. In contrast, a SenseShape keeps a history of all objects in the scene that intersect it, and stores statistical data about these objects. Our statistical data currently provides us with five different rankings for an object, which are relative to a specific object's behavior in a certain SenseShape during a time period.

Figure 1. An AR user interacting through gesture, speech and SenseShapes. The tracked glove and three possible gestures are shown at the right.

The *time* ranking (T_{rank}) is derived from the fraction of time (T_{object}) the object spent in a volume over a specific time period (T_{period}). The more time the object spends in the volume, the higher the ranking.

$$T_{\text{rank}} = \frac{T_{\text{object}}}{T_{\text{period}}}, \qquad 1 \geq T_{\text{rank}} > 0.$$

The *distance* ranking (D_{rank}) is based on an object's distance D_{object} from the volume's origin (which can be arbitrarily chosen) compared to other objects, where D_{max} is the distance of the most distant object in the volume from the volume's origin. The closer the object is to the volume, the higher the ranking.

$$D_{\text{rank}} = 1 - \frac{D_{\text{object}}}{D_{\text{max}}}, \qquad 1 \geq D_{\text{rank}} > 0.$$

The *stability* ranking (S_{rank}) measures an object's presence in the volume relative to other objects based on E_{object}, the number of times an object enters and exits the volume, and E_{max}, the most times any object enters and exits the volume. Fewer entries/exits yield a higher ranking (a more stable object).

$$S_{\text{rank}} = \frac{E_{\text{max}} + 1 - E_{\text{object}}}{E_{\text{max}} + 1}, \qquad 1 \geq S_{\text{rank}} > 0.$$

Proceedings of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR '03)
0-7695-2006-5/03 $17.00 © 2003 IEEE

The most stable objects don't leave the volume, and have $E_{\text{object}} = 0$ and $S_{\text{rank}} = 1$.

The *visibility* and *center-proximity* rankings reflect an object's visibility relative to selected SenseShapes. We compute the visibility ranking of a cone by rendering a low-resolution version of the scene, from a center of projection at the cone's apex, into an off-screen object buffer [1] containing the cone's base. Each object is rendered with a unique color, allowing it to be identified in the frame through the pixel color, as shown in Figure 2.

We currently generate two object buffers, one for an eye cone and one for a hand cone. The visibility ranking (V_{rank}) is defined as

$$V_{\text{rank}} = \frac{\sum visiblePixels_{\text{object}}}{\sum pixelsInFrame}, \qquad 1 \geq V_{\text{rank}} \geq 0,$$

where $visiblePixels_{\text{object}}$ are the visible pixels an object has in a frame, and *pixelsInFrame* are all the pixels in the frame.

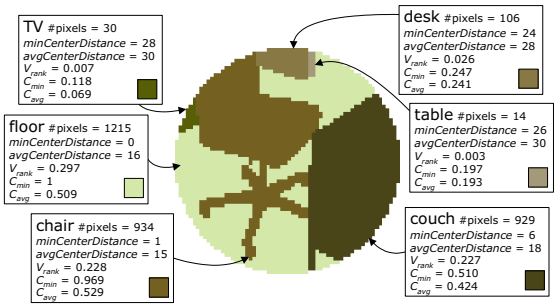

Figure 2. Off-screen object buffer with a 64-pixel diameter. Each object is listed with visibility and center-proximity information and rankings.

The center-proximity rankings (C_{min} and C_{avg}) indicate how close the visible portion of an object is to the center of the shape. The pixel distance is calculated as the Euclidean distance from the center of the object buffer.

$$C_{\text{min}} = 1 - \frac{minDistanceToCenter_{\text{object}}}{maxDistanceToCenter_{\text{frame}}}, \qquad 1 \geq C_{\text{min}} \geq 0,$$

$$C_{\text{avg}} = 1 - \frac{avgDistanceToCenter_{\text{object}}}{maxDistanceToCenter_{\text{frame}}}, \qquad 1 \geq C_{\text{avg}} \geq 0,$$

where $maxDistanceToCenter_{\text{frame}}$ is the object buffer radius, $minDistanceToCenter_{\text{object}}$ is the smallest distance of any pixel of the specific object, and $avgDistanceToCenter_{\text{object}}$ is the average distance for all pixels.

Only the statistics that are applicable to a particular SenseShape are considered during integration of statistics (e.g., visibility ranking might not be relevant for a hand-centered "grabbing sphere").

3. Dynamic Integration

Our preliminary experience with SenseShapes shows that it is useful to have the system perform *dynamic integration*, in which an integration strategy for determining which statistics to use is chosen based on the current gesture, speech and SenseShape. For example, we have used spatial cues in speech (such as "this/these/that/those") to select among a set of alternative rankings:

$$\text{Ranking("make this desk red")} = T_{\text{rank(hand)}} \times V_{\text{rank(head)}} \times V_{\text{rank(hand)}} \times C_{\text{min(head)}} \times C_{\text{min(hand)}} \times D_{\text{rank(hand)}}$$

$$\text{Ranking("make that desk red")} = T_{\text{rank(hand)}} \times V_{\text{rank(head)}} \times V_{\text{rank(hand)}} \times C_{\text{avg(head)}} \times C_{\text{avg(hand)}} \times (1 - D_{\text{rank(hand)}})$$

In this example, closer objects will be weighted higher when "this" is used and "lower when "that" is used. Furthermore, due to the inherent imprecision of pointing at a distance, the average center proximity rank is used for "that," while the minimum center proximity rank is used for "this." When no spatial cues are detected, the system uses both the average and the minimum center proximity rankings, but discards the distance rankings altogether.

4. Conclusions and Future Work

SenseShapes are a set of statistical tools that use instrumented volumes to determine the user's intentions in a multimodal AR system. Our preliminary experiments show that the dynamic integration of SenseShape statistics increases the predictability of selection. We plan to further improve SenseShapes to dynamically adapt their position, orientation, size, and geometry to accommodate different situations. Following that, we plan to conduct a user study to measure the effectiveness and relevance of each of the ranking rules, to allow for more effective dynamic integration.

Acknowledgments

We wish to thank Sajid Sadi and Avanindra Utukuri for their support with the P5 glove. This research was funded in part by Office of Naval Research Contracts N00014-99-1-0394, N00014-99-1-0683, and N00014-99-1-0249, NSF Grant IIS-00-82961, and a gift from Microsoft.

References

[1] Atherton, P. R. 1981. A Method of Interactive Visualization of CAD Surface Models on a Color Video Display. In *Computer Graphics (Proceedings of ACM SIGGRAPH 81)*, 15(3). 279–287.

[2] Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Feiner, S., and Cohen, P. An Architecture for 3D Multimodal Interaction in Augmented and Virtual Reality. To appear in *Proc. ICMI 2003 (Int. Conference on Multimodal Interfaces)*, Vancouver, BC. November 5–7, 2003.

[3] Liang, J., Green, M. JDCAD: A Highly Interactive 3D Modeling System. *Computers and Graphics*, 18(4). 499–506. 1994.

[4] Mine, M. *Exploiting Proprioception in Virtual-Environment Interaction*. Ph.D. Thesis, Department of Computer Science, University of North Carolina at Chapel Hill. 1997.

[5] Piekarski, W. and Thomas, B. H. Using ARToolKit for 3D hand position tracking in mobile outdoor environments. *First IEEE Int. Workshop on Augmented Reality Toolkit*. 2002.