Fusing Depth, Color, and Skeleton Data for Enhanced Real-Time Hand Segmentation

Yu-Jen Huang
USC Institute for Creative Technologies
whuang@ict.usc.edu

Mark Bolas
USC Institute for Creative Technologies
bolas@ict.usc.edu

Evan A. Suma
USC Institute for Creative Technologies
suma@ict.usc.edu

Figure 1: Example results of our hand segmentation approach. Green circles represent the original skeleton tracking data, and the rectangles indicate the search area within the depth and color images.

1. INTRODUCTION

As sensing technology has evolved, spatial user interfaces have become increasingly popular platforms for interacting with video games and virtual environments. In particular, recent advances in consumer-level motion tracking devices such as the Microsoft Kinect have sparked a dramatic increase in user interfaces controlled directly by the user's hands and body. However, existing skeleton tracking middleware created for these sensors, such as those developed by Microsoft and OpenNI, tend to focus on coarse full-body motions, and suffers from several well-documented limitations when attempting to track the positions of the user's hands and segment them from the background. In this paper, we present an approach for more robustly handling these failure cases by combining the original skeleton tracking positions with the color and depth information returned from the sensor.

2. METHODS

In our approach, we start by acquiring the skeleton tracking data for the elbows and hands (in our tests we used the Microsoft Kinect for Windows SDK). We informally observed that the elbows are substantially more reliable than the reported hand positions. Therefore, we used the elbows as a reference for defining a rectangular search area where the hands would most likely be located. Although the hand tracking data is often inaccurate, it still provides a likely initial guess that we used to set the length/width of the search area. If the distance between the estimated hand positions is within a predefined threshold, we define one large, single search area instead of two separate ones. This allows our approach to more robustly handle cases where the hands are held together.

The next step in our approach is to more precisely locate the hands using both the color and depth images returned from the sensor. In the color image, we applied an explicit RGB boundaries skin cluster to classify the skin pixels in the search area that were obtained in the previous step [1]. Using the depth image, we make use of the existing Kinect segmentation implementation, which separates the user from the background but does not differentiate between specific parts of the body. We take the intersection of the two pixel sets, thereby eliminating erroneous false positives to provide a more robust overall segmentation. We then apply a connected component labeling algorithm to remove small impossible objects. The results of our hand segmentation algorithm are shown in Figure 1.

3. CONCLUSION

In this paper, we described an approach for enhanced hand segmentation using the Microsoft Kinect. By combining the existing skeleton tracking data with the depth and color image streams, we achieve a more robust segmentation from both the background and other parts of the user's body. In particular, this approach also handles failure cases where the existing skeleton tracking libraries return inaccurate position data, such as when the hands are close together or against the user's body.

4. ACKNOWLEDGEMENT

This work is supported by DARPA under contract (W911NF-04-D-0005) and the U.S. Army Research, Development, and Engineering Command. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

5. REFERENCES

[1] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *Proc. Graphicon*, volume 3, pages 85–92, 2003.

SUI'13, July 20–21, 2013, Los Angeles, California, USA.
ACM 978-1-4503-2141-9/13/07.