Poster: Real Time Hand Pose Recognition with Depth Sensors for Mixed Reality Interfaces

Byungkyu Kang[1]* Mathieu Rodrigue[1]† Tobias Höllerer[1]‡ Hwasup Lim[2]§

University of California Santa Barbara[1]
Korea Institute of Science and Technology[2]

ABSTRACT

We present a method for predicting articulated hand poses in real-time with a single depth camera, such as the Kinect or Xtion Pro, for the purpose of interaction in a Mixed Reality environment and for studying the effects of realistic and non-realistic articulated hand models in a Mixed Reality simulator. We demonstrate that employing a randomized decision forest for hand recognition benefits real-time applications without the typical tracking pitfalls such as reinitialization. This object recognition approach to predict hand poses results in relatively low computation, high prediction accuracy and sets the groundwork needed to utilize articulated hand movements for 3D tasks in Mixed Reality workspaces.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/methodology, Input devices and strategies, Theory and methods

1 INTRODUCTION

Articulated hand movements can potentially play a big role in Mixed Reality interfaces, specifically for object recognition in 3D augmented environments. Without direct hands-on interaction interfaces based on, e.g. haptic, sound or conventional desktop user interfaces (WIMP: windows, icons, menus, and pointing [6]) can be insufficient and downgrade user experience in AR applications. For instance, an online user survey on AR applications [3] claimed that 2D user interfaces in AR applications are obstacles of perception when introduced to the 3D real world. Therefore, we initiated our research from two research questions:

1. *Question A*: What is the best mode of interaction for Mixed Reality applications?
2. *Question B*: How can we improve user experience in terms of robustness and multifariousness of interaction?

For the first research question above, we posit that a hand pose based user interface can be considered as one of the most convenient methods for many Mixed Reality applications. In most real world scenarios involving 3D workspaces, humans interact with physical objects using their fingertips and we can enhance user experience in MR by enabling users to control applications and 3D data with hand interactions.

However, in order for computers to understand a sophisticated and large number of hand poses in real-time, the robustness and computational complexity of an articulated hand system needs to be considered.

To tackle this problem, we propose a hand pose based user interface for Mixed Reality applications, using a composite method, that combines an object recognition approach proposed by [5], an Iterative Closest Point (ICP) [4] and Inverse Kinematics algorithm (IK). Depth sensors simplify the stereo problem as well as the complexity of input data caused by variance in hand, skin color, texture and lighting conditions. Depth sensors are now also widely disseminated among consumers, which makes this system a practical solution. Most importantly, the random decision forest algorithm we employ is fast, and also easy to compute in parallel on a GPU [5] using a parallel computing platform such as Compute Unified Device Architecture(CUDA[1]).

Figure 1: Classification result using 100 training images with manually annotated labels

1.1 Mixed Reality Simulator

The proposed system is aimed towards studying the effect of realistic and non-realistic hand models for use with Mixed Reality simulation [2]. In this particular user study, a static hand model was used in testing the validity of Mixed Reality simulation, replicating an earlier study by Ellis et al. [1]. However, there is uncertainty in understanding the effect of different hand representations and tracking accuracies ranging from non-realistic to realistic. Using the proposed real-time hand pose recognition system, we aim to shed light on this problem by allowing a fully articulated hand to be implemented within our MR simulator platform.

2 HAND POSE RECOGNITION

The method we introduce in this poster consists of two different steps in order to recognize hand poses. This section presents our approach in detail.

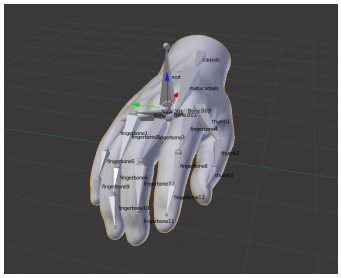
2.1 Finding Joints in Multiple Parts

The first step is based on the algorithm proposed by [5], which provides effective real-time human body pose recognition using depth images by computing simple, depth-invariant features for each pixel. Additionally, the computational speed can be boosted if the features are computed in parallel using a GPU implementation of the randomized decision forest algorithm.

Applying a similar approach to articulated hand recognition, we note that the magnitude of training data is a critical factor in order for classification to perform well for general hand poses. Our preliminary experiment shows that an insufficient number of training images results in more misclassified pixels. Figure 1 shows an example of a classification result with 100 training images. This

*e-mail: bkang@cs.ucsb.edu
†e-mail: mathieu@cs.ucsb.edu
‡e-mail: holl@cs.ucsb.edu
§e-mail: hslim@imrc.kist.re.kr

[1]http://www.nvidia.com/object/cuda_home_new.html



(a) A rigged hand model used for generating synthetic training data set

(b) Color-labeled image for ground truth

(c) Synthetic depth image

Figure 2: 3D hand model and synthetic image example

Table 1: Hand part labels and corresponding parts

Label	Part in Hand	Label	Part in Hand
1	Pinky tip	10	Index tip
2	Pinky middle	11	Index middle
3	Pinky trunk	12	Index trunk
4	Ring tip	13	Thumb tip
5	Ring middle	14	Thumb middle
6	Ring trunk	15	Thumb trunk
7	Middle tip	16	Palm
8	Middle middle	17	Wrist
9	Middle trunk	18	Other

classification result is derived from our own implementation based on the algorithm in [5] with 40 random depth features and 7 finger labels. However, the training images used for Figure 1 were manually labeled with only 7 ground truth labels: five fingers, the wrist and the remaining parts of a hand. We needed to have more labeled parts with significantly larger amounts of training images to generalize all possible hand poses.

To overcome this difficulty, we use a randomly generated set of synthetic depth images. A rigged 3D hand model is used to generate a variety of different hand poses. To facilitate the process of generating synthetic images, we used Blender[2]. Each joint of a hand model is assigned a range of angular rotation in order to generate random poses. A script was implemented in order to automatically generate any number of sample depth images needed along with their corresponding label images to use as a training dataset.

Figure 2 shows the hand model used for generating training samples, a label image and a synthetic depth image, respectively. To implement real hand articulation with our user interface, we divided a hand into 18 part labels for classification. Each label is represented by a unique color and those colors are translated into nominal labels in our classification algorithm. Hand part labels and their corresponding location can be seen in Table 1.

In the training process, we computed 2000 random features per pixel. These features are trained by the random forest classifier and used to classify real-time depth images from a Kinect sensor. Once the classification result is obtained, finger tips and joints in each frame can be estimated by applying the mean-shift algorithm.

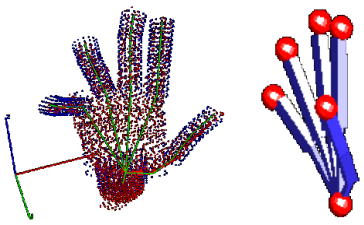


Figure 3: Articulation matching algorithms (Left: Iterative Closest Point [4], Right: Inverse Kinematics)

[2]Blender is an open source 3D content generator

2.2 Interpreting Hand Articulation

Unlike previous approaches for human body pose estimation, hand pose recognition can be more difficult since most depth sensors work best at distances more aligned with full-body tracking. Furthermore, [5] also reported some defects in recognition caused by limb occlusions or overlaps. In order to overcome this problem and obtain better accuracy, we are conducting a follow-up experiment based on a performance comparison across multiple algorithms.

After we compute the joint of each part, we can obtain a hand articulation as a skeletal structure. As a further step, ICP, IK and a combination of both algorithms are used to correct hand articulation. We will compare these three results in future evaluations to the original articulation in both estimation of accuracy and speed. This evaluation helps us to choose the best approach in order to optimize our user interface. Both ICP and IK algorithms are implemented and their examples can be seen in Figure 3.

3 Conclusion and Future Work

In this paper we proposed a system for allowing the use of fully articulated hand movements in a Mixed Reality environment. Future work involves evaluating the ICP and Inverse Kinematics algorithms to retrieve robust joint locations. In follow up work we will employ this system to understand the effects of realistic and nonrealistic hand models within a Mixed Reality simulator.

Acknowledgements

This work was supported in part by NSF CAREER grant IIS-0747520 and ONR grant N00014-09-1-1113.

References

[1] S. Ellis, F. Breant, B. Manges, R. Jacoby, and B. Adelstein. Factors influencing operator interaction with virtual objects viewed via head-mounted see-through displays: viewing conditions and rendering latency. In *Virtual Reality Annual International Symposium, 1997., IEEE 1997*, pages 138 –145, mar 1997.

[2] C. Lee, S. Bonebrake, T. Hollerer, and D. Bowman. A replication study testing the validity of ar simulation in vr for controlled experiments. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 203 –204, oct. 2009.

[3] T. Olsson and M. Salo. Online user survey on current mobile augmented reality applications. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 75 –84, oct. 2011.

[4] S. Pellegrini, K. Schindler, and D. Nardi. A generalization of the icp algorithm for articulated bodies. In M. Everingham and C. Needham, editors, *British Machine Vision Conference (BMVC'08)*, September 2008.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297 –1304, june 2011.

[6] D. Van Krevelen and R. Poelman. A survey of augmented reality technologies, applications and limitations. *International Journal of Virtual Reality*, 9(2):1, 2010.