

# Precise Hand Segmentation from a Single Depth Image: Supplementary Material

Minglei Li<sup>1,2,\*</sup>, Lei sun<sup>2</sup> and Qiang Huo<sup>2</sup>

<sup>1</sup>University of Science and Technology of China, Hefei, China

<sup>2</sup>Microsoft Research, Beijing, China

Email: {v-mingll, lsun, qianghuo}@microsoft.com

This supplementary material accompanies the paper "Precise Hand Segmentation from a Single Depth Image". It provides more details on hand RoI extraction and data normalization. Besides, further experimental results and examples are provided. Please also see the supplementary video.

## I. DATA

### A. Hand RoI Extraction

Given a depth image  $I$  and an approximate hand joint location  $\hat{x}$ , a hand RoI is extracted in two steps. First, a square area  $R$  centered at  $\hat{x}$  is set to cover an appropriate area for hand RoI extraction. Then, a hand RoI is extracted in this area using flood fill method with  $\hat{x}$  as a seed.

$R$ 's size is set adaptively. As a camera can be modeled by a usual pinhole, an object's size  $s'$  in a captured image is inversely proportional to its sensing distance  $d$  to the camera.

$$s' \propto \frac{1}{d} \quad (1)$$

To ensure that hand RoIs are extracted from areas with the same size in real world coordinate, we set  $R$ 's size  $w_{\hat{x}}$  as a value inversely proportional to  $\hat{x}$ 's depth, as in equation 2. Based on our experience, we set  $d_{ref} = 1500mm$  and  $w_{ref} = 120$ .

$$w_{\hat{x}} = \frac{d_{ref}}{d_{\hat{x}}} \times w_{ref} \quad (2)$$

Hand RoI is extracted in  $R$ . Seeded by  $\hat{x}$ , a connected region is extracted. A point  $i$  in the connected region belongs to a hand RoI if it is within the square area  $R$  and near to  $\hat{x}$  in depth, as described in equation 3. Point  $i$  is denoted as  $(u_i, v_i, d_i)$ , where  $(u_i, v_i)$  denotes a 2D coordinate in  $I$  and  $d_i$  denotes a depth value.

$$\begin{cases} u_{\hat{x}} - \frac{w_{\hat{x}}}{2} < u_i < u_{\hat{x}} + \frac{w_{\hat{x}}}{2} \\ v_{\hat{x}} - \frac{w_{\hat{x}}}{2} < v_i < v_{\hat{x}} + \frac{w_{\hat{x}}}{2} \\ d_{\hat{x}} - d_{t1} < d_i < d_{\hat{x}} + d_{t2} \end{cases} \quad (3)$$

By Reserving all the points that are in the connected region area and meet the above conditions, hand RoI is constructed as illustrated in Fig. 1.

\*Minglei Li contributed to this work when he worked as an intern with the Speech Group, Microsoft Research Asia.

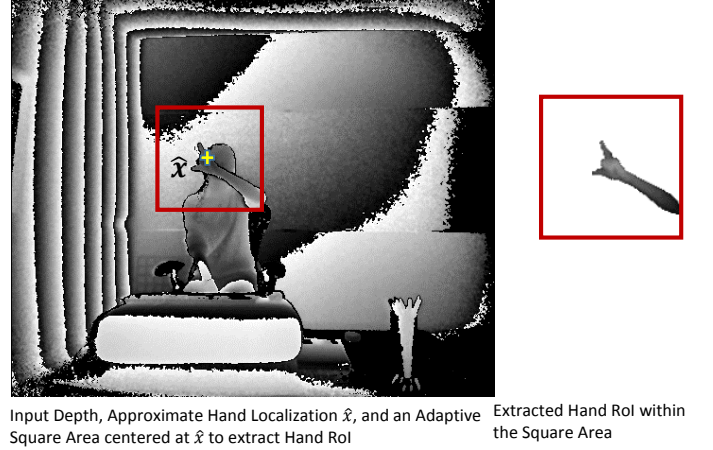


Fig. 1. Hand RoI Extraction

### B. Data Normalization

Translation and scale normalization of hand RoIs in depth maps benefit the segmentation performance and can be easily implemented.

By setting the gravity pixel of the hand RoI or the hand joint location as the center point of a captured hand RoI image, translation invariance is achieved. Since hand RoI is extracted in a square area centered at hand joint location, it is automatically translation invariant.

As a hand RoI's image size is set inversely proportional to its hand joint  $\hat{x}$ 's depth as in equation 2, all hand RoIs' scales can be mapped to the same scale level through being multiplied a scale factor  $\frac{d_{\hat{x}}}{d_{ref}}$ . Here, it is equal to resize all hand RoIs' size to  $w_{ref}$ .

According to the depth range in a hand RoI, foreground points' intensity values are normalized through linearly mapping the depth range to a fixed intensity value range from  $t_1$  to  $t_2$ . In this paper,  $t_1$  to  $t_2$  are set 0 and 150, respectively.

## II. FURTHER EXPERIMENTAL RESULTS

### A. Comparison Method: RDF

The comparison RDF consists of 4 trees with the maximum height of 25. At each node, 10,000 weak learners is sampled.

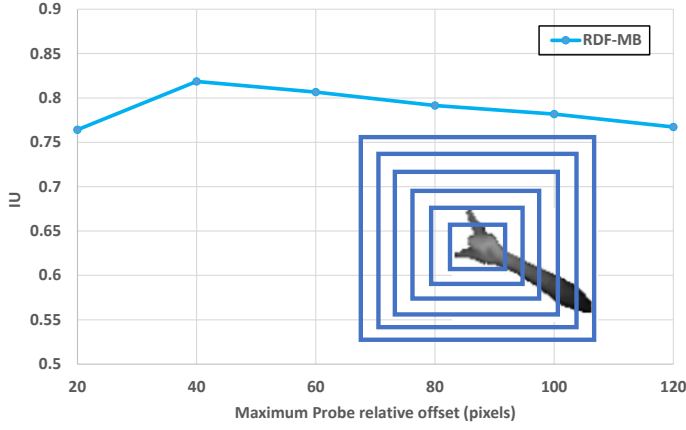


Fig. 2. Range for offset candidates

At a given pixel  $(u, v)$  on depth image  $I$ , its original decision function at each node is

$$I\left(u + \frac{\delta u}{I(u, v)}, v + \frac{\delta v}{I(u, v)}\right) - I(u, v) \geq d_t. \quad (4)$$

where  $I(u, v)$  represents a depth value at  $(u, v)$ ,  $\delta u$  and  $\delta v$  are probe offsets with unit as pixel meters and  $d_t$  is a depth threshold.

Since our hand RoIs are normalized to the same scale level through being multiplied a scale factor  $\frac{d_{\hat{x}}}{d_{ref}}$ , they are equivalent to be those hand RoI images captured at a distance of  $d_{ref}$ . Thus, the decision function at each node on pixel  $(u, v)$  would be degenerated as

$$I(u + \delta u, v + \delta v) - I(u, v) \geq d_t. \quad (5)$$

where  $I(u, v)$  represents an intensity value at  $(u, v)$ ,  $\delta u$  and  $\delta v$  are offsets with unit as pixels and  $d_t$  is an intensity threshold.

Since the range for offset candidates largely affects the performance of RDF, we try to explore an appropriate setup for the range parameters through experiments.

1) *Range of Offset Candidates*: A discrete set of offset range values are used to train the RDF models. With post processing of median filtering and largest blob detection, results on test dataset are evaluated through pixel intersection over union(IU) and drawn in Fig.2. The concentric boxes on the right show the 6 tested offset ranges for a hand point. As offset range is increased, RDF is able to use more spatial context to make decisions. Performance increases with the offset range, but levels off around 40 pixels. In this paper, we used 40 pixels as the offset range.

2) *Range of Thresholds*: Since intensity values of foreground points in hand RoIs are normalized to the range from  $t_1$  to  $t_2$  according to depth, we use the intensity value for threshold  $d_t$  and set its range from  $t_1$  to  $t_2$ .

### III. MORE SEGMENTATION EXAMPLES

More examples of hand segmentation are listed as follows.

