Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth Camera [*]

Zhou Ren Junsong Yuan
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
{renzhou, jsyuan}@ntu.edu.sg

Zhengyou Zhang
Microsoft Research Redmond
Washington 98052-6399, USA
zhang@microsoft.com

ABSTRACT

The recently developed depth sensors, e.g., the Kinect sensor, have provided new opportunities for human-computer interaction (HCI). Although great progress has been made by leveraging the Kinect sensor, e.g. in human body tracking and body gesture recognition, robust hand gesture recognition remains an open problem. Compared to the entire human body, the hand is a smaller object with more complex articulations and more easily affected by segmentation errors. It is thus a very challenging problem to recognize hand gestures. This paper focuses on building a robust hand gesture recognition system using the Kinect sensor. To handle the noisy hand shape obtained from the Kinect sensor, we propose a novel distance metric for hand dissimilarity measure, called Finger-Earth Mover's Distance (FEMD). As it only matches fingers while not the whole hand shape, it can better distinguish hand gestures of slight differences. The extensive experiments demonstrate the accuracy, efficiency, and robustness of our hand gesture recognition system.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; I.4.8 [**Scene Analysis**]: Depth cues, Shape

General Terms

Algorithm

Keywords

Hand Gesture Recognition, Human-Computer Interaction, Kinect Sensor, Finger-Earth Mover's Distance

1. INTRODUCTION

Hand gesture recognition is of great importance for human-computer interaction (HCI), because of its extensive applications in virtual reality, sign language recognition, and computer games [12]. Despite lots of previous work, traditional vision-based hand gesture recognition methods [4, 11] are still far from satisfactory for real-life applications. Because of the limitations of the optical sensors, the quality of the captured images is sensitive to lighting conditions and cluttered backgrounds, thus it is usually not able to detect and track the hands robustly, which largely affects the performance of hand gesture recognition.

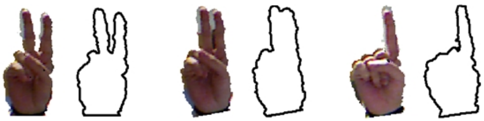


Figure 1: Some challenging cases for hand gesture recognition, using depth cameras: the first and the second hands have the same gesture while the third hand confuses the recognition.

To enable a more robust hand gesture recognition, one effective way is to use other sensors to capture the hand gesture and motion, e.g. through the data glove [8]. Unlike optical sensors, such sensors are usually more reliable and are not affected by lighting conditions or cluttered backgrounds. However, as it requires the user to wear a data glove and sometimes requires calibration, it is inconvenient for the user and may hinder the naturalness of hand gesture. Also, such data gloves are expensive. As a result, it is not a very popular way for hand gesture recognition.

Thanks to the recent development of inexpensive depth cameras, e.g., the Kinect sensor, new opportunities for hand gesture recognition emerge. In spite of many recent successes in applying the Kinect sensor for face recognition [3] and human body tracking [10], it is still an open problem to use Kinect for hand gesture recognition. Due to the low-resolution of the Kinect depth map, typically, of only 640×480, although it works well to track a large object, e.g. the human body, it is difficult to detect and segment a small object from an image with this resolution, e.g., a human hand which occupies a very small portion of the image with more complex articulations. In such a case, the segmentation of the hand is usually inaccurate, thus may significantly affect the recognition step.

To illustrate the above problem, Fig.1 shows some examples. It can be seen that the contours have significant local distortions in addition to pose variations. Due to the low resolution and inaccuracy of the Kinect sensor, the two fingers in the second hand of Fig.1 are indistinguishable as they are close to each other. Unfortunately, classic shape recognition methods, such as shape contexts [2] and skeleton matching [1], cannot robustly recognize the shape contour with severe distortions. Clearly, recognizing noisy shapes is very challenging, especially if there are many gestures to recognize.

[*] Area chair: Alexander Hauptmann

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.

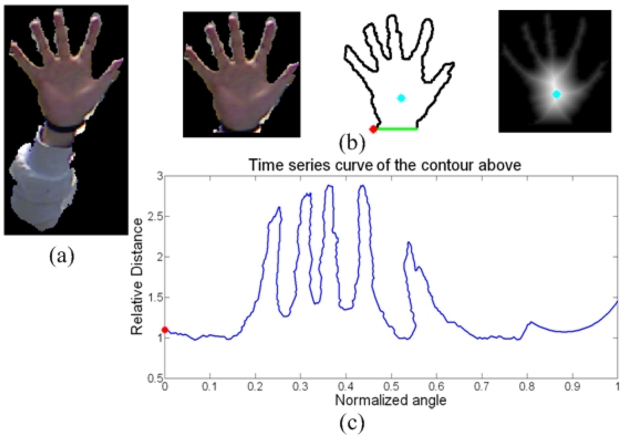


Figure 2: Hand detection. (a) The rough hand segmented by depth thresholding; (b) A more accurate hand detected with black belt (the green line), the initial point (the red point) and the center point (the cyan point); (c) Its time-series curve representation.

In order to address this problem, we propose a novel shape distance metric called Finger-Earth Mover's Distance (FEMD). FEMD is specifically designed for hand shapes. It considers each finger as a cluster and penalizes unmatched fingers. By testing on a 10-gesture dataset, our method is accurate, efficient, and performs robustly to articulations, local distortions, orientation and scale changes. To the best of our knowledge, this is the first attempt in real-life hand gesture recognition using Kinect sensor.

2. HAND DETECTION

We use Kinect sensor as the input device, which captures the color image and the depth map at 640×480 resolution.

2.1 Hand Segmentation

As for hand segmentation, we require the user to cooperate in two aspects (both are reasonable requirements in HCI): first, the user need to make sure that the hand is the frontmost object facing the sensor. Thus, by thresholding from the nearest depth position with a certain gap, a rough hand region can be obtained, as shown in Fig.2(a). Second, the user need to wear a black belt on the gesturing hand's wrist. We use RANSAC to locate the position of the black belt, and thus, a more precise hand shape can be detected, as shown in Fig.2(b). The hand shape is generally of 120×120 resolution, which may cause severe distortions.

2.2 Shape Representation

After detecting the hand shape, we represent it as a *time-series curve*, as shown in Fig.2(c). Such a shape representation has been successfully used for the classification and clustering of shapes [6]. The time-series curve records the relative distance between each contour vertex to a center point. We define the center point as the point with the maximal distance after Distance Transform on the shape (the cyan point), as shown in Fig.2(b); and the initial point (the red point) is defined according to the RANSAC line detected from the black belt (the green line).

In our time-series representation, the horizontal axis denotes the angle between each contour vertex and the initial point relative to the center point, normalized by 360°. The vertical axis denotes the Euclidean distance between the contour vertices and the center point, normalized by the radius of the maximal inscribed circle. As shown in Fig.2, the time-series curve captures nice topological properties of the hand, such as the fingers.

3. HAND GESTURE RECOGNITION

With the hand shape and its time-series representation, we apply template matching for robust recognition, i.e., the input hand is recognized as the class with which it has the minimum dissimilarity distance: $c = \arg\min_{c} \mathrm{FEMD}(H,\ T_c)$, where H is the input hand; T_c is the template of class c; $\mathrm{FEMD}(H,\ T_c)$ denotes the proposed Finger-Earth Mover's Distance between the input hand and each template.

3.1 Finger-Earth Mover's Distance

In [9], Rubner *et al.* presented a general and flexible metric, called Earth Mover's Distance (EMD), to measure the distance between signatures or histograms. EMD is widely used in many problems such as content-based image retrieval and pattern recognition.

EMD is a measure of the distance between two probability distributions. It is named after a physical analogy that is drawn from the process of moving piles of earth spread around one set of locations into another set of holes in the same space. The locations of earth piles and holes denotes the mean of each cluster in the signatures, the size of each earth pile or hole is the weight of cluster, and the ground distance between a pile and a hole is the amount of work needed to move a unit of earth. To use this transportation problem as a distance measure, i.e., a measure of dissimilarity, one seeks the least cost transportation — the movement of earth that requires the least amount of work.

Grauman and Darrell applied EMD to contour matching and contour retrieval [5], which represents the contour by a set of local descriptive features and computes the set of correspondences with minimum EMD costs between the local features. However, the existing EMD-based contour matching algorithms have two deficiencies:

1. Two hand shapes differ mainly in global features while not local features. As shown in Fig.3(a)(b), the fingers (global features) are their major difference. Besides, the large number of local features slows down the speed of contour matching. Therefore, it is better to consider global features in contour matching.

2. EMD allows for partial matching, i.e., a signature and its subset are considered to be the same in EMD measure: as in Fig.3(c)(d), the EMD distance of these two signatures is zero because the signature in Fig.3(d) is a subset of Fig.3(c). However, in many situations partial matching is illogical, such as in the case of Fig.3(a)(b), where the finger in Fig.3(b) is a partial set of the fingers in Fig.3(a). Clearly, they should be considered being very different.

Our Finger-Earth Mover's Distance (FEMD) can address these two deficiencies of EMD-based contour matching methods mentioned above. Different from the EMD-based algorithm [5], which considers each local feature as a cluster, we consider the input hand as a signature with each finger (the global feature) as a cluster. And we add penalty on empty holes to alleviate partial matches on global features.

Formally, let $R=\{(\mathbf{r_1},\ w_{\mathbf{r_1}}),\ ...,\ (\mathbf{r}_m,\ w_{\mathbf{r}_m})\}$ be the first hand signature with m clusters, where $\mathbf{r}_i$ is the cluster representative and $w_{\mathbf{r}_i}$ is the weight of the cluster; $T=\{(\mathbf{t_1},\ w_{\mathbf{t_1}}),\ ...,\ (\mathbf{t}_n,\ w_{\mathbf{t}_n})\}$ is the second hand signature with n clusters.

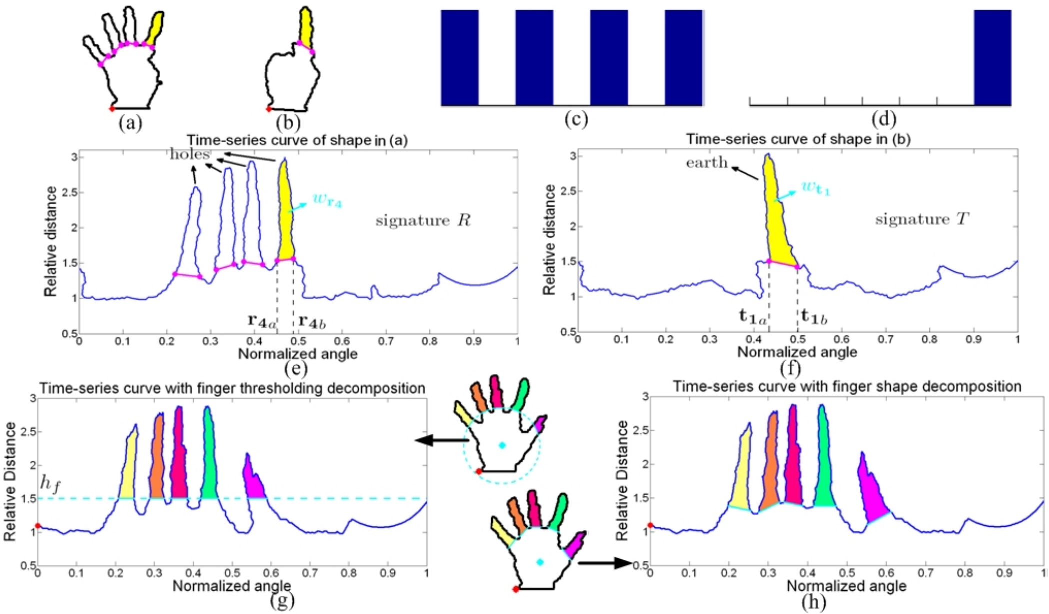


Figure 3: (a) (b): two hand shapes whose time-series curves are shown in (e) (f). (c) (d): two signatures that partially match, whose EMD cost is 0. (e) (f): the signature representations of the time-series curves. (g) (h): two finger detection methods, thresholding decomposition (g) and near-convex decomposition (h).

Now we show how to represent a time-series curve as a signature. Fig.3(e)(f) show the time-series curves of the hands in Fig.3(a)(b) respectively, where each finger corresponds to a segment of the curve. We define each cluster of a signature as the finger segment of the time-series curve: the representative of each cluster $\mathbf{r}_i$ is defined as the angle interval between the endpoints of each segment, $\mathbf{r}_i$=[$\mathbf{r}_{ia}$, $\mathbf{r}_{ib}$], where $0 \leq \mathbf{r}_{ia} < \mathbf{r}_{ib} \leq 1$; and the weight of a cluster, $w_{\mathbf{r}_i} \in (0,1)$, is defined as the normalized area within the finger segment.

$\mathbf{D}$=$[d_{ij}]$ is the ground distance matrix of signature R and T, where d_{ij} is the ground distance from cluster $\mathbf{r}_i$ to $\mathbf{t}_j$. d_{ij} is defined as the minimum moving distance for interval $[\mathbf{r}_{ia},\ \mathbf{r}_{ib}]$ to totally overlap with $[\mathbf{t}_{ja},\ \mathbf{t}_{jb}]$, i.e.:

$$d_{ij} = \begin{cases} 0, & \mathbf{r}_i \text{ totally overlap with } \mathbf{t}_j, \\ \min(|\mathbf{r}_{ia} - \mathbf{t}_{ja}|,\ |\mathbf{r}_{ib} - \mathbf{t}_{jb}|), & \text{otherwise.} \end{cases}$$

For two signatures, R and T, their FEMD distance is defined as the least work needed to move the earth piles plus the penalty on the empty hole that is not filled with earth:

$$\begin{aligned} \text{FEMD}(R,\ T) &= \beta E_{move} + (1-\beta)E_{empty}, \\ &= \frac{\beta \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} + (1-\beta)|\sum_{i=1}^{m} w_{\mathbf{r}_i} - \sum_{j=1}^{n} w_{\mathbf{t}_j}|}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}, \end{aligned}$$

where $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$ is the normalization factor, f_{ij} is the flow from cluster $\mathbf{r}_i$ to cluster $\mathbf{t}_j$, which constitutes the flow matrix $\mathbf{F}$. Parameter β modulates the importance between the first and the second terms. As we can see, E_{empty}, d_{ij} are constants given two signatures. To compute the FEMD, we need to compute the flow matrix $\mathbf{F}$. We follow the definition of the flow matrix $\mathbf{F}$ in EMD, which is defined by minimizing the work needed to move all the earth piles.

4. FINGER DETECTION

Before we can measure the FEMD distance between two hand shapes, we have to obtain the finger clusters in their time-series curves, namely to detect the fingers from the hand shapes. We propose two ways for finger detection:

4.1 Thresholding decomposition

As mentioned before, the time-series curve reveals a hand's topological information well. As shown in Fig.3(g), each finger corresponds to a peak in the curve. Therefore, we can apply the height information in time-series curve to decompose the fingers. Specifically, we define a finger as a segment in the time-series curve, whose height is greater than a threshold h_f. In this way, we can detect the fingers fast. However, choosing a good height threshold h_f is essential.

4.2 Near-convex hand decomposition

Thresholding decomposition is sensitive to the threshold h_f, and it may introduce segmentation errors, e.g., the thumb is incomplete in Fig.3(g). Now we introduce a more accurate finger detection method based on the near-convex shape decomposition scheme [8], as shown in Fig.3(h):

$$\mathbf{min}\ \alpha \parallel \mathbf{x} \parallel_{\mathbf{0}} + (\mathbf{1}-\alpha)\mathbf{w}^{\top}\mathbf{x}, \quad s.t.\ \ \mathbf{Ax} \geq \mathbf{1},\ \mathbf{x}^{\top}\mathbf{Bx} = 0,\ \mathbf{x} \in \{0,1\}^{\overline{n}}.$$

It formulates shape decomposition as a integer optimization problem (details please refer to [8]). By relaxing $0<x_i<1$, the problem becomes a linear programming problem.

5. EXPERIMENTS

5.1 Dataset

We collect a new hand gesture dataset with a Kinect sensor (http://www.ntu.edu.sg/home/renzhou/HandGesture.htm). Our dataset is collected from 10 subjects, and it contains 10 gestures. Each subject performs 10 different poses for the same gesture. Thus in total our dataset has 10 subject × 10 gestures/subject × 10cases/gesture = 1000 cases, each of which consists of a color image and a depth map.

Our dataset is a very challenging real-life dataset, which is collected in uncontrolled environments. Besides, for each gesture, the subject poses with variations, namely the hand changes in orientation, scale, articulation, etc.

5.2 Performance Evaluation on Robustness

First, our hand gesture recognition system is robust to cluttered backgrounds, because the hand shape is detected

	Thresholding Decomposition+FEMD	Near-convex Decomposition+FEMD
Mean Accuracy	90.6%	93.9%
Mean Running Time	0.5004s	4.0012s

Table 1: The mean accuracy and the mean running time of the two proposed methods.

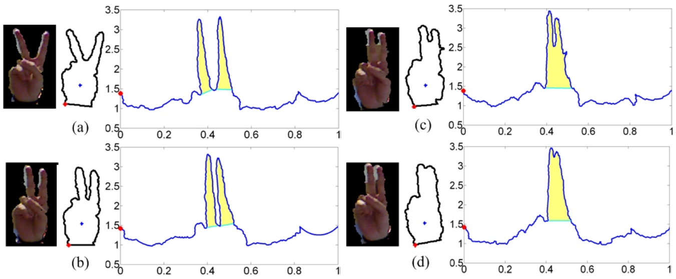


Figure 4: Our system is insensitive to the distortions and articulation.

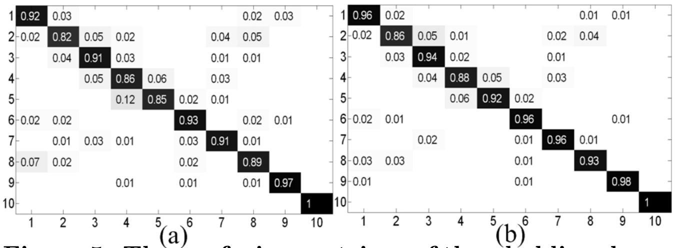
using depth information and the backgrounds can be easily removed. Then, our hand gesture recognition system is robust to orientations changes. The reason is that the initial point and the center point are relatively fixed in each shape. Thus the time-series curves of the hands with different orientations are similar, and their distances are very small. Also, our hand gesture recognition system is robust to scale changes. Because the time-series curve and the FEMD distance are normalized, the hand shapes with scale changes can be correctly recognized as the same gesture.

Furthermore, our hand gesture recognition method is robust to the articulations and distortions due to imperfect hand segmentation. As the proposed FEMD distance metric uses global features (fingers) to measure the dissimilarity, local distortions are tolerable. As for the articulations, Fig.4 shows some examples: the color images show 4 hands of the same gesture; the next columns shows the corresponding hand shapes, and their time-series curves. As we can see, the hand shapes in Fig.4(c)(d) are heavily distorted. However, as illustrated in their time-series curves, by detecting the finger parts (the yellow regions), we represent each shape as a signature whose clusters are the finger parts. Particularly, the signatures of Fig.4(a)(b) have 2 clusters: $\{(\mathbf{r_1}, w_{\mathbf{r_1}}), (\mathbf{r_2}, w_{\mathbf{r_2}})\}$, and the signatures of Fig.4(c)(d) only have 1 cluster: $\{(\mathbf{t_1}, w_{\mathbf{t_1}})\}$. From Section 3.1, we can estimate that $(w_{\mathbf{r_1}}+w_{\mathbf{r_2}}) \approx w_{\mathbf{t_1}}$, and the ground distance d_{11}, d_{21} ≈ 0. According to the definition, we know that the FEMD distances among the 4 shapes ≈ 0. Therefore, our FEMD metric is insensitive to distortions and articulations.

5.3 Accuracy and Efficiency

In Table 1, the mean accuracy and the mean running time of FEMD based on the two finger detection methods are given. The mean accuracy of near-convex decomposition based FEMD (93.9%) is higher than that of thresholding decomposition (90.6%), owing to more accurate finger decomposition. In terms of recognition error, the near-convex decomposition reduces the error rate from 9.4% of the thresholding decomposition to 6.1%, which is a 35% reduction. But on the other hand, the speed of the second method is slower than that of the first one, because of the more complex finger detection algorithm.

Fig.5 shows their confusion matrixes. Compared with the thresholding decomposition based FEMD, the near-convex decomposition based FEMD has less seriously confused categories. And the accuracies in all the classes are improved. Here we fix the near-convex decomposition parameter α=0.5, the FEMD parameter β=0.5, and the thresholding decomposition parameter h_f=0.6.



	1	2	3	4	5	6	7	8	9	10
1	0.92	0.03						0.02	0.03	
2	0.02	0.82	0.05	0.02			0.04	0.05		
3		0.04	0.91	0.03			0.01	0.01		
4			0.05	0.86	0.06		0.03			
5				0.12	0.85	0.02	0.01			
6	0.02	0.02				0.93		0.02	0.01	
7		0.01	0.03	0.01		0.03	0.91	0.01		
8	0.07	0.02				0.02		0.89		
9				0.01		0.01		0.01	0.97	
10										1

(a)

	1	2	3	4	5	6	7	8	9	10
1	0.96	0.02						0.01	0.01	
2	0.02	0.86	0.05	0.01			0.02	0.04		
3		0.03	0.94	0.02			0.01			
4			0.04	0.88	0.05		0.03			
5				0.06	0.92	0.02				
6	0.02	0.01				0.96			0.01	
7			0.02			0.01	0.96	0.01		
8	0.03	0.03				0.01		0.93		
9	0.01					0.01			0.98	
10										1

(b)

Figure 5: The confusion matrixes of thresholding decomposition based FEMD (a) and near-convex decomposition based FEMD (b).

We have built a demo to demonstrate the superiority of our method in real-life applications [7].

6. CONCLUSIONS

Hand gesture recognition for real-life applications is very challenging because of the requirements on its robustness, accuracy and efficiency. In this paper, we presented a robust real-life hand gesture recognition system using the Kinect sensor. A novel distance metric, Finger-Earth Mover's Distance, is used for dissimilarity measure, which treats each finger as a cluster and penalize the empty finger-hole. In order to accurately detect the fingers, we presented two finger decomposition methods: thresholding decomposition and near-convex decomposition. Extensive experiments on a challenging 10-gesture dataset demonstrate that our hand gesture recognition system is accurate, efficient, and robust to articulations, distortions, and orientation, scale changes.

7. ACKNOWLEDGEMENT

This work was supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan.

8. REFERENCES

[1] X. Bai and L. J. Latecki. Path similarity skeleton graph matching. *IEEE Trans. on PAMI*, 30:1–11, 2008.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI*, 24:509–522, 2002.

[3] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang. 3d deformable face tracking with a commodity depth camera. In *Proc. of IEEE ECCV*, 2010.

[4] C. Chua, H. Guan, and Y. Ho. Model-based 3d hand posture estimation from a single 2d image. *Image and Vision Computing*, 20:191 – 202, 2002.

[5] K. Grauman and T. Darrell. Fast contour matching using approximate earth mover's distance. In *CVPR*, 2004.

[6] E. Keogh, L. Wei, X. Xi, S. Lee, and M. Vlachos. Lb_keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. In *Proc. of 32th International Conf. on VLDB*, 2006.

[7] Z. Ren, J. Meng, J. Yuan, and Z. Zhang. Robust hand gesture recognition with kinect sensor. In *Proc. of ACM MM*, 2011.

[8] Z. Ren, J. Yuan, C. Li, and W. Liu. Minimum near-convex decomposition for robust shape representation. In *Proc. of ICCV*, 2011.

[9] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40:99–121, 2000.

[10] J. Shotton, A. Fitzgibbon, and etc. Real-time human pose recognition in parts from single depth images. In *Proc. of IEEE CVPR*, 2011.

[11] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Filtering using a tree-based estimator. In *Proc. of IEEE ICCV*, 2003.

[12] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan. Vision-based hand-gesture applications. *Communications of the ACM*, 54:60–71, 2011.