FINGERTIP DETECTION BASED ON PROTUBERANT SALIENCY FROM DEPTH IMAGE

Yuseok Ban[1,2,*], *Minglei Li*[2], *Lei Sun*[2], *Qiang Huo*[2]

[1] School of Electrical & Electronic Engineering, Yonsei University, Seoul, Republic of Korea
van@yonsei.ac.kr

[2] Microsoft Research, Beijing, China
{v-mingll, lsun, qianghuo}@microsoft.com

ABSTRACT

We propose a new approach for detecting a protuberant region from a depth image by leveraging a notion of protuberant saliency which describes how much the protuberant region stands out from its surroundings in the depth image. An intuitive and simple method is designed to calculate protuberant saliency from depth image, which can be used effectively together with the nearness information to detect the tip of a protuberant object such as a fingertip. We evaluate and compare our method with several state-of-the-art saliency methods for fingertip detection. Experimental results demonstrate that our method outperforms the comparing methods in terms of detection accuracy, being more robust against the rotation and isometric deformation of a fingertip region, the scale and depth noise issues, and the low resolution of a depth image.

Index Terms— Fingertip detection, Protuberant saliency, Depth, Isometric deformation, Rotation, Scale, Noise

1. INTRODUCTION

Finger gesturing is one of the most intuitive ways to convey a user's intention in natural user interface. As the number of products using finger gestures increases rapidly, the performance of detecting fingers becomes crucial. In response to this need, many techniques have been suggested to address the challenge of the many Degrees of Freedom (DoFs) of a finger, resulting in visually different shapes and sizes. Recent works mostly comprise (i) the model-based algorithm using an explicit model to recover a hand pose and (ii) the classification/regression-based algorithm to find a mapping to a predefined set of hand pose configurations [1]. Most of the existing methods are based on segmenting a hand region out from a background [1, 2] or classifying different hand posture modes [2]. On the other hand, few researches have been carried out on relatively simple but straightforward techniques which can directly indicate fingertips.

Motivated by human's ability to promptly identify noticeable regions based on visual attention, a variety of visual saliency approaches have been suggested to various preprocessing tasks such as detecting the location of a salient object as the rectangular region in a scene, segmenting out a conspicuous object along its contour from a background, detecting keypoints from an object, etc. New progress of efficient 3D sensing technology has inspired many studies on how to incorporate 3D information (i.e., depth, point cloud, and mesh) into visual saliency to benefit various computer vision tasks. Particularly, depth information has many advantages such as easiness for acquisition, low computational cost, robustness to illumination, and unambiguous encoding of surface metric dimensions [3]. Depth information can be used to calculate the degree of depth saliency of an object. However, previous approaches [4, 5, 6, 7, 8] have considered depth to be either an auxiliary cue for a simple fusion strategy or prior information to preliminarily assist general saliency models. Few studies have delved deeply into extracting the salient feature of depth perception based on depth data itself.

Recently, Borji et al. [8] gives a review of general state-of-the-art saliency methods. Actually, the influence of depth on visual saliency has been studied in the literature [4, 9], and the works of Lang et al. [4] and Peng et al. [7] elaborated the latest progress on saliency methods using depth. Starting from the very beginning of depth saliency model, Ouerhani et al. (OU) [10] adopted weighted sum of depth, depth mean curvature, and depth gradient to obtain conspicuous locations in a depth image. Zhong et al. (JND3D) [11] proposed 3D Just Noticeable Difference (JND) model based on the difference in conspicuity between pop-out and concave regions and the conspicuity of an inconsecutive depth region. Niu et al. (SS) [12] combined the two approaches of using the disparity change abruptness among different graph-based segmented regions and both minimizing the vergence-accomodation conflict [13] and maximizing the influence of negative disparity which denotes an object popping out from the screen.

In this paper, we propose first a novel method to compute the degree of saliency of a depth region using the concept of *protuberant saliency*. Then, we introduce an intuitive and simple, yet effective method to detect a fingertip in a depth image. Extensive experiments are conducted to demonstrate the effectiveness of the proposed approach.

[*] Yuseok Ban contributed to this work during his internship in the Speech Group, Microsoft Research Asia.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

2. OUR APPROACH

2.1. Calculating Protuberant Saliency from Depth Image

The *protuberance* can be explained by the definition of protuberant objects with protuberant shapes (e.g., fingertip, gunpoint, and tip of a stick). A protuberant region varies with the isometric deformation, out-plane rotation, scale, and low resolution of noisy depth, resulting in shape and depth level changes and inconsistently extracted features [14]. The purpose of our method is to consistently indicate the protuberant saliency regardless of the aforementioned issues. To do so, we propose a rotation and scale invariant filter to calculate protuberant saliency on each pixel in a depth map. The filter has the same pattern of the Local Binary Pattern (LBP) [15] using a different calculation of filter responses based on circularly symmetric neighbors set as inference points (see Fig. 1).

To get scale invariance, the radius $r(d_c)$ of the filter is dynamically determined according to the normalized depth value d_c of the center point as in Eq. (1). The parameters are obtained by fitting the sum of exponentials function to physical statistics with $\eta = 0.05, A = 242, B = 0.6499, C = -0.00079$, and $D = 0.0011$.

$$r(d_c) = \eta(Ae^{Cd_c} + Be^{Dd_c}) \tag{1}$$

In this case, we use eight inference points along the circle of the filter with clockwise indices denoted. We compute the filter response on each inference pixel by subtracting the center value d_c from d_i, where Eqs. (2) and (3) use the sign function $\delta(\cdot)$ to facilitate leveraging the filter responses.

Given filter responses on inference pixels around a center pixel, we find that using the number of the positive filter responses p_c to represent the protuberance is very distinctive. However, this does not fully describe the protuberance and also suffers from depth noises. Therefore, we import the number of zero-crossings z_c, which occur when two consecutive filter responses do not share the same sign, to support p_c and inhibit depth noises, where $rem(i, j)$ is remainder after dividing i by j (see Eqs. (4) and (5)).

$$h_i = \delta(d_i - d_c) \tag{2}$$

$$\delta(\Delta d) = \begin{cases} 1 & \text{if } \Delta d > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$p_c = \sum_{i=0}^{7} h_i \tag{4}$$

$$z_c = \sum_{i=0}^{7} |h_{rem(i,8)} - h_{rem(i+1,8)}| \tag{5}$$

By jointly considering the two terms p_c and z_c, a novel protuberance measure function is proposed as in Eq. (6), which intuitively implies that ”the larger the p_c or the smaller the z_c is, the larger the protuberance is”. From the equation, we can see that this function is rotation and scale invariant with inhibition on depth noises.

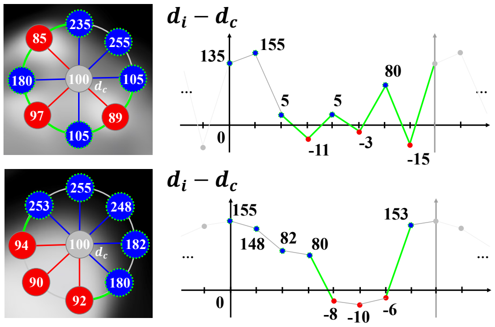


Fig. 1. Illustration of two different protuberances which share the same number of *positive filter responses* (blue-colored).

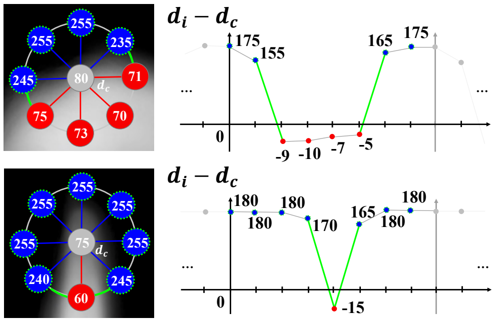


Fig. 2. Illustration of two different protuberances which share the same number of *zero-crossings* (green-colored).

$$f_c = \begin{cases} \frac{p_c}{2^{z_c/2-1}} & \text{if } 0 < z_c \leq 8, \\ p_c & \text{if } z_c = 0. \end{cases} \quad \text{where } 0 \leq z_c \leq 8 \tag{6}$$

2.2. Fingertip Detection using Protuberant Saliency

In common with Lang et al.’s observation that a few interesting objects occupy the majority of the fixations [3], a few points will take up the majority of the fixations among the points within an object. They also observed in their 3D eye fixation statistics that human preferentially fixates at closer depth ranges [3]. Thereby, we introduce a geometry energy that is designed to meet the balanced need of protuberance and nearness, namely the more protuberant or closer a depth region is, the more conspicuous a point is. The geometry energy E_c at the center point consists of a protuberance term F_c and a nearness term G_c as in Eq. (7). A saliency level can be directly applied for F_c, while G_c is obtained by reversing and normalizing d_c in depth image as in Eq. 8, where $h(\cdot)$ is a normalization function to ensure the depth ranges between 0 and 255. Consequently, a geometry energy map is obtained by collecting the geometry energy level at each pixel.

$$E_c = |F_c \cdot G_c| \tag{7}$$

$$G_c = h(255 - d_c) \;\; \text{where } 0 \leq d_c \leq 255 \tag{8}$$

The protuberant saliency is effective enough to detect the tip of an object simply by finding the maximum among the geometry energy levels in a geometry energy map.

3. EXPERIMENT

3.1. Experimental Setup

Dataset. We introduce a new depth image dataset for fingertip detection by collecting around 356,000 images from 42 subjects including 21 females and 21 males. The subjects have been given no prior instruction except the two tasks, finger pointing (DB1-FP) and finger writing in-the-air (DB1-FW). Kinect v2 has been leveraged to capture the depth image of a hand region of which size is adaptively rescaled as in Eq. (1). Depth intensity is median filtered (3 by 3 sized) for denoising. Finally, every ground truth of a pixel-wise fingertip location is manually labeled. Also, we used the portion of MSRA hand pose (DB2-HP) depth image dataset [16] that resembles DB1-FP. Moreover, the qualitative tests of feasibility on gunpoint and tip of a stick data are provided as shown in Fig. 3.

Evaluation Metrics. We compute the distance error of every estimated location from the ground truth location in millimeters and adopt Positive Predictive Value (PPV) curve [17]. The PPV curve is obtained via incrementally altered False Positive thresholds. (In the rescaled data according to the physical distance from hand to sensor based on Eq. (1), 1 pixel corresponds to 2.588 millimeters.)

Hardware Configuration. We used a desktop computer with Intel(R) Core(TM) i7-4770 CPU @ 3.4GHz, RAM 16.0 GB.

3.2. Experimental Results and Discussion

Effectiveness. Fig. 3 shows that the saliency of a protuberant region can be indicated by protuberant saliency despite the isometric deformation and rotation of the protuberant regions around fingertip, gunpoint, and tip of a stick. High value is assigned for protuberant region, while low value is computed for relatively even region. Also, our method can consistently extract the protuberance with low resolution of depth. More detailed results regarding the geometry energy for fingertip detection are provided in Fig. 4. By simply finding the maximum among the levels in the geometry energy map, the fingertip in a depth image can be accurately detected. Fig. 5 shows the visual comparisons of different saliency methods. It is worth observing that our method effectively indicates and highlights the protuberance of a fingertip region. Especially in more challenging conditions (e.g., fingertip pointing forward or extremely outward), our method outperforms the others. OU is biased on depth edges by leveraging the mean curvature and gradient information from a depth image, while SS highly depends on depth abruptness among graph-based depth segments. Meanwhile, JND3D is rather weighted towards the depth contrast and edges. However, our method effectively indicates the highest saliency on protuberant region (e.g., fingertip) by exploiting the inherent protuberance information encoded in the structure of depth data. Our method implies two intuitions: (i) the smaller the number of zero-crossings is the larger the protuberance is (see Fig. 1), and (ii) the larger

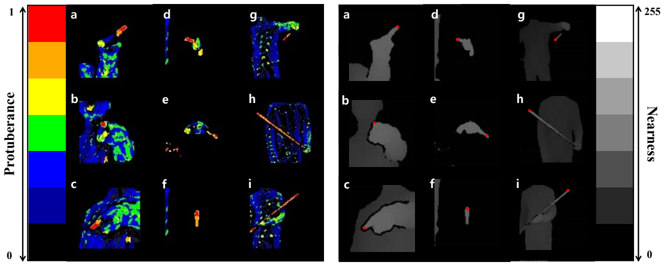
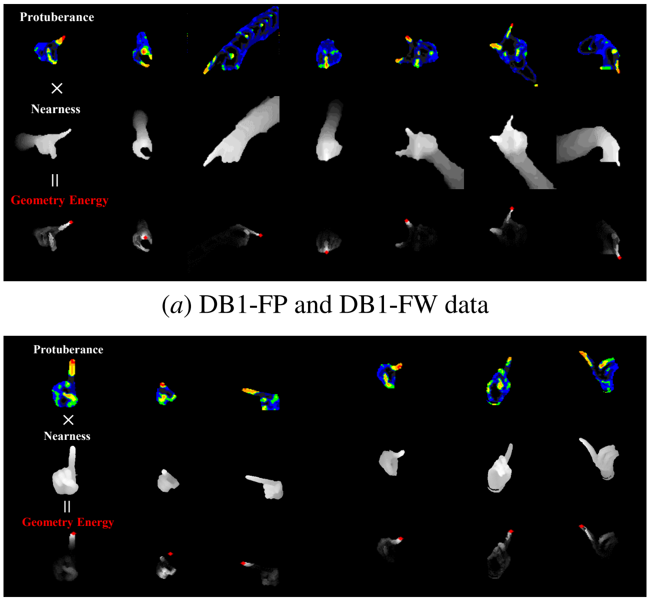


Fig. 3. Examples of (*a-c*) fingertip, (*d-f*) gunpoint, and (*g-i*) tip of a stick for (*left*) the protuberant saliency and (*right*) the detection result of a protuberant region.



(*a*) DB1-FP and DB1-FW data

(*b*) DB2-HP data

Fig. 4. Experimental results of fingertip detection, (*top*) the protuberance term, (*middle*) the nearness term, and (*bottom*) the detection result marked with a red point.

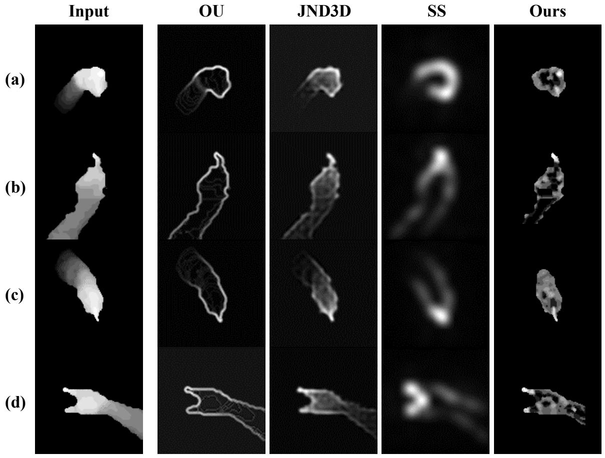


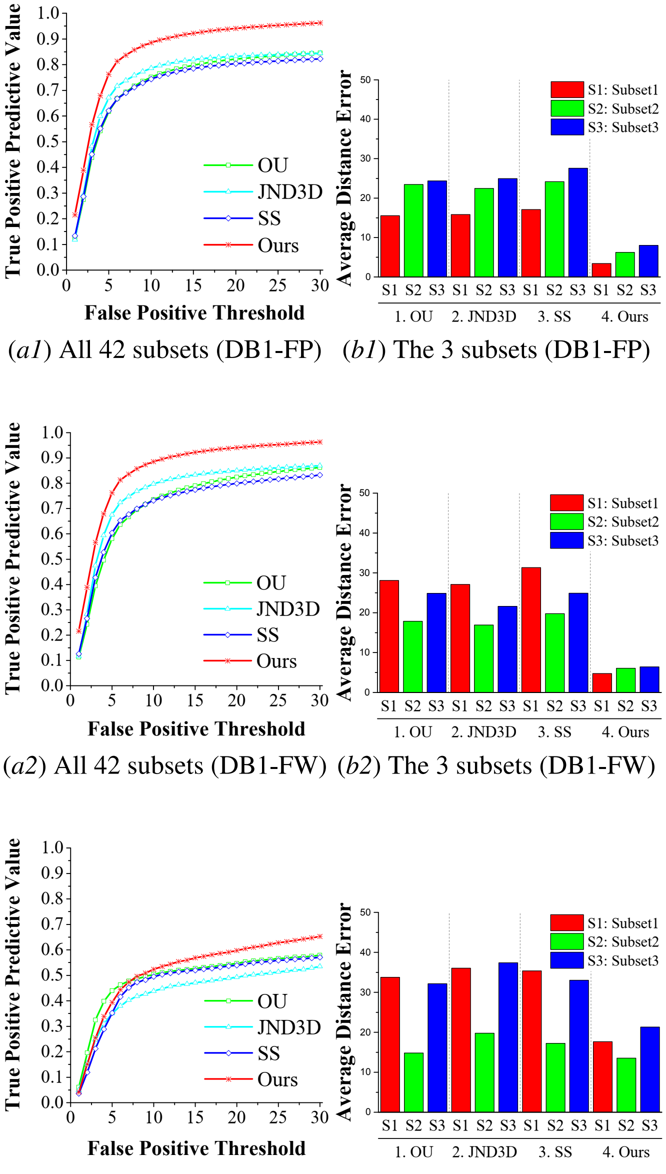
Fig. 5. Comparative examples of saliency results in gray scale representation on DB1-FP and DB1-FW data.

the number of positive filter responses is, the larger the protuberance is (see Fig. 2). The two intuitions mutually provide confidence levels to each other so as to indicate the degree of protuberance in a gradual manner. Fig. 6 reports the PPV curves via varying False Positive thresholds of all subsets in (*a*) and the average distance errors of three particular subsets with challenging conditions in (*b*). As can be seen, our method outperforms the other methods in respect of both PPV and distance error. This observation intuitively conveys that our method effectively works for detecting fingertip region as protuberant by successfully emphasizing the protuberant region in a depth image. The other comparing methods suffer from performance degradation under the condition of large rotation and isometric deformation of a fingertip included in Subset3 as depicted in Fig. 6(*b*). Our method also provides better results under gently moving condition of a fingertip. An important note of the results is that the other methods fail to clearly distinguish the fingertip region when pointing straight forward. Similarly, many of the previous works which are based on hand segmentation [1, 2, 18, 19, 20, 21] also have difficulty dealing with the fingertip pointing forward. On the contrary, our method directly and effectively calculates the protuberance of a fingertip not only pointing outward but also pointing forward.

Efficiency. We observe that our method runs fast enough for real-time computation by achieving around 47 FPS (see Table 1 comparing the average computational time by each method). The efficiency results from the simplicity of leveraging the filter responses, i.e., using the numbers of positive filter responses and zero-crossings.

4. CONCLUSION

In this paper, we propose the protuberant saliency of depth data for detecting the protuberant depth region such as the fingertip in a depth image. The protuberant saliency is based on an intuitive and novel way of leveraging filter responses for calculating the protuberance of a depth region in a gradual manner. Based on the comparative experiments, our proposed method shows the robustness against the rotation and isometric deformation of a depth region as well as the noisy low resolution depth information and scale variation. Also, the simplicity of our method enables fast computation for the practical uses based on depth data in various scenarios such as smart phone [19, 22], smart car [23, 24], smart home [25], etc. Yet, we observe that protuberant saliency still has limitation in generalization capability as it only focuses on the feature of protuberance. For future work, we plan to generalize our method by exploring other useful features extracted from depth data itself and to leverage advanced detection methods for improving fingertip detection accuracy and performing more complicated hand-gesturing tasks.



(*a1*) All 42 subsets (DB1-FP) (*b1*) The 3 subsets (DB1-FP)

(*a2*) All 42 subsets (DB1-FW) (*b2*) The 3 subsets (DB1-FW)

(*a3*) All 6 subsets (DB2-HP) (*b3*) The 3 subsets (DB2-HP)

Fig. 6. Comparison of experimental results evaluated on (*1*) *DB1-FP*, (*2*) *DB1-FW*, and (*3*) *DB2-HP* data by (*a*) the PPV curve and (*b*) the average distance error of a subset.

Table 1. Comparison of computational time for fingertip detection (unit: milliseconds).

Method	Language	CPU Time	
		Mean	STD
OU	Matlab	14.33	4.1
JND3D	Matlab	196.05	45.7
SS	Matlab	21.93	6.9
Ours	Matlab	21.27	6.7

5. REFERENCES

[1] Chaoyu Liang, Yonghong Song, and Yuanlin Zhang, “Real-time fingertip detection based on depth data,” in *ACPR*, 2015, pp. 443–447.

[2] K. Li and X. Zhang, “A new fingertip detection and tracking algorithm and its application on writing-in-the-air system,” in *CISP*, 2014, pp. 457–462.

[3] Yulan Guo, Mohammed Bennamoun, Ferdous Ahmed Sohel, Min Lu, and Jianwei Wan, “3D object recognition in cluttered scenes with local surface features: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2270–2287, 2014.

[4] Congyan Lang, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan S. Kankanhalli, and Shuicheng Yan, “Depth matters: Influence of depth cues on visual saliency,” in *ECCV*, 2012, pp. 101–115.

[5] Ali Borji, Hamed Rezazadegan Tavakoli, Dicky N. Sihite, and Laurent Itti, “Analysis of scores, datasets, and models in visual saliency prediction,” in *ICCV*, 2013, pp. 921–928.

[6] Arridhana Ciptadi, Tucker Hermans, and James M. Rehg, “An in depth view of saliency,” in *BMVC*, 2013, pp. 1–11.

[7] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji, “RGBD salient object detection: A benchmark and algorithms,” in *ECCV*, 2014, pp. 92–109.

[8] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li, “Salient object detection: A benchmark,” *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[9] J. M. Wolfe and T. S. Horowitz, “What attributes guide the deployment of visual attention and how do they do it?,” *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[10] Nabil Ouerhani and Heinz Hügli, “Computing visual attention from scene depth,” in *ICPR*, 2000, pp. 1375–1378.

[11] Rui Zhong, Ruimin Hu, Yi Shi, Zhongyuan Wang, Zhen Han, Lu Liu, and Jinhui Hu, “Just noticeable difference for 3d images with depth saliency,” in *PCM*, 2012, pp. 414–423.

[12] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu, “Leveraging stereopsis for saliency analysis,” in *CVPR*, 2012, pp. 454–461.

[13] B. Mendiburu, “Innovation and intellectual property rights,” in *3D movie making: stereoscopic digital cinema from script to screen*, chapter 2, pp. 11–33. CRC Press, 2012.

[14] Chaoyu Liang, Yonghong Song, and Yuanlin Zhang, “Hand gesture recognition using view projection from point cloud,” in *ICIP*, 2016, pp. 4413–4417.

[15] Marko Heikkilä, Matti Pietikäinen, and Janne Heikkilä, “A texture-based method for detecting moving objects,” in *BMVC*, 2004, pp. 1–10.

[16] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun, “Realtime and robust hand tracking from depth,” in *CVPR*, 2014, pp. 1106–1113.

[17] Mohammad H. Jafari, Shadrokh Samavi, S. Mohamad R. Soroushmehr, Hoda Mohaghegh, Nader Karimi, and Kayvan Najarian, “Set of descriptors for skin cancer diagnosis using non-dermoscopic color images,” in *ICIP*, 2016, pp. 2638–2642.

[18] Yanguo Zhao and Zhan Song, “Hand posture recognition using approximate vanishing ideal generators,” in *ICIP*, 2014, pp. 1525–1529.

[19] Lukas Prasuhn, Yuji Oyamada, Yoshihiko Mochizuki, and Hiroshi Ishikawa, “A hog-based hand gesture recognition system on a mobile device,” in *ICIP*, 2014, pp. 3973–3977.

[20] Michel Abboud, Abdessalam Benzinou, Kamal Nasreddine, and Mustapha Jazar, “Robust statistical shape analysis based on the tangent shape space,” in *ICIP*, 2015, pp. 3520–3524.

[21] Henrique Weber, Cláudio Rosito Jung, and Dan Gelb, “Hand and object segmentation from RGB-D images for interaction with planar surfaces,” in *ICIP*, 2015, pp. 2984–2988.

[22] Phawis Thammasorn, Sukitta Boonchu, and Aram Kawewong, “Real-time method for counting unseen stacked objects in mobile,” in *ICIP*, 2013, pp. 4103–4107.

[23] Paul K. J. Park, Baek Hwan Cho, Jin Man Park, Kyoobin Lee, Ha Young Kim, Hyo Ah Kang, Hyun Goo Lee, Jooyeon Woo, Yohan Roh, Won Jo Lee, Chang-Woo Shin, Qiang Wang, and Hyunsurk Ryu, “Performance improvement of deep learning based gesture recognition using spatiotemporal demosaicing technique,” in *ICIP*, 2016, pp. 1624–1628.

[24] Robert Nesselrath, Mohammad Mehdi Moniri, and Michael Feld, “Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions,” in *IE*, 2016, pp. 190–193.

[25] Guillaume Plouffe and Ana-Maria Cretu, “Static and dynamic hand gesture recognition in depth data using dynamic time warping,” *IEEE Trans. Instrumentation and Measurement*, vol. 65, no. 2, pp. 305–316, 2016.