A HAND POSE TRACKING BENCHMARK FROM STEREO MATCHING

Jiawei Zhang[†], *Jianbo Jiao*[†], *Mingliang Chen*[†], *Liangqiong Qu*[†‡⋆], *Xiaobin Xu*[†], *and Qingxiong Yang*[§]

Department of Computer Science, City University of Hong Kong, Hong Kong, China[†]
State Key Laboratory of Robotics, Shenyang Institute of Automation, CAS, Shenyang, China[‡]
University of Chinese Academy of Sciences, Beijing, China[⋆]
University of Science and Technology of China, Hefei, China[§]

ABSTRACT

In this paper we establish a long-term 3D hand pose tracking benchmark[1]. It contains 18,000 stereo image pairs as well as the ground-truth 3D positions of palm and finger joints from different scenarios. Meanwhile, to accurately segment hand from stereo images, we propose a novel stereo-based hand segmentation and depth estimation algorithm specially tailored for hand tracking here. The experiments indicate the effectiveness of the proposed algorithm by demonstrating that its tracking performance is comparable to the use of an active depth sensor under various of challenging scenarios.

Index Terms— stereo matching and hand pose tracking

1. INTRODUCTION

Vision-based hand pose tracking can be applied a series of scenarios including human computer interaction systems. A literature survey has been proposed in [1]. The challenging difficulties include high-dimensional articulated structure, severe self-occlusion and chromatically uniform appearance which can be solved through involving depth information. The traditional way to estimate depth is from either active depth sensor or passive stereo. However, active sensors can be interfered by other active sources such as the sun or another active sensor. Moreover, active sensor has relatively high power consumption and is not suitable for mobile devices. Alternatively, the depth can be obtained from passive sensors. But it is slow and the depth estimates are noisy and unstable especially when the scene lacks of texture.

At present, existing hand tracking datasets [2–7] are captured with active depth sensors. In this paper, we resolve hand pose tracking problem using depth information obtained from passive stereo. To evaluate the performance of passive stereo for hand pose tracking, a new benchmark is proposed. The data set is simultaneously captured by a Point Grey Bumblebee2 stereo camera and an Intel Real Sense F200 active depth camera. We manually label positions of finger joints and center of palm in depth images. Our benchmark contains six environments with different difficulty levels for hand segmentation and disparity estimation. It is hard to track hand poses with self-occlusions or global rotations, and thus we capture two sequences for every environment without and with these two tracking difficulties. Our benchmark has thus 12 different sequences and every sequence contains 1,500 stereo pairs and depth images.

Before performing tracking, the hand region should be segmented in advance. An active depth sensor can provide accurate depth information which simplifies the hand segmentation. However, it is difficult to segment hands using inaccurate depth from passive stereo. We use efficient color-based skin detection method [8] for hand segmentation and find the limitation occurs under unconstrained environments (e.g., different lighting conditions and backgrounds). To adapt to different environments, we capture an online training sequence with waving hand before tracking. The adaptive Gaussian Mixture Model (GMM) [9, 10] is then used to perform foreground/background segmentation and the foreground is treated as skin color. The skin and non-skin histogram models can then be computed and we use the skin color probability for hand segmentation.

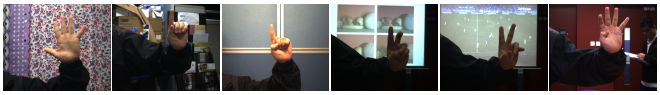
There are a lot of stereo matching methods [11–14]. However, their performance is more unstable and noisy than the active sensors. We propose a new stereo algorithm for hand tracking here. To achieve real-time performance for hand tracking, the proposed stereo is based on efficient traditional local stereo matching [15]. The skin color probability is used as the guidance of the guided image filter [16] for matching cost aggregation to increase the robustness around textureless regions. Since some background regions may have a similar color to skin and have high skin probability, a robust hand segmentation method is proposed by using confidence-guided combination of color based hand segmentation and depth from stereo matching. The experiments show that the proposed stereo method improves the tracking performance.

To evaluate hand tracking from stereo sequences, we implement two hand pose tracking methods [3, 17]. The experiments show that tracking using the proposed stereo matching can achieve comparable performance relative to active depth cameras.

The contributions of this paper are:

- a hand pose benchmark with 18,000 stereo image pairs;
- a robust stereo matching specially designed for 3D hand pose tracking and achieved comparable performance to active depth sensors.

[1]Stereo hand pose data set can be downloaded from https://sites.google.com/site/zhjw1988/



(a) *B1* (b) *B2* (c) *B3* (d) *B4* (e) *B5* (f) *B6*

Fig. 1. Six different environments in the benchmark.

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Two different types of hand poses in the benchmark.(a)-(d) are simple counting pose. (e)-(h) are difficult random poses.

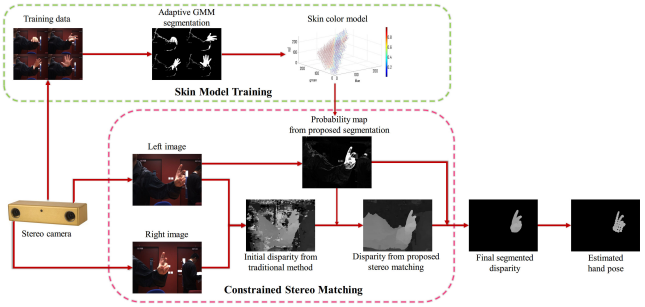


Fig. 3. Framework of the proposed stereo based 3D hand pose tracking method.

2. STEREO-BASED HAND TRACKING BENCHMARK

As shown in Fig.1, the proposed benchmark contains sequences with 6 environments to evaluate stereo-based hand pose tracking. Indoor environments are usually textureless which significantly increase the difficulty for passive stereo. Highlight (*B3*) and shadow (*B4*, *B5*) are also quite challenging for both stereo matching and skin color modeling. Besides static backgrounds, we also capture 3 sequences with dynamic backgrounds including PowerPoint presentation (*B4*), video playing (*B5*) and people walking (*B6*).

Since self-occlusion and global rotation are two major challenges in hand tracking, we capture two sequences for every environment with two different poses as shown in Fig.2. One captures simple counting poses with slowly moving fingers as can be seen from Fig.2 (a)-(d). The other is supposed to be much more difficult for hand pose tracking. The hand/fingers move randomly with severe self-occlusions and global rotations as shown in Fig.2 (e)-(h). Counting and random poses are designed to be similar for all the 6 environments to ensure a fair comparison.

For quantitative comparison, we capture the stereo and depth images from a Point Grey Bumblebee2 stereo camera and an Intel Real Sense F200 active depth camera simultaneously. Camera calibration [18] is performed in advance to obtain the parameters of the cameras. We manually label the ground-truth positions of finger joints and palm centers in depth images. Our benchmark has a total of 12 sequence and each sequence contains 1,500 frames.

3. STEREO-BASED HAND POSE TRACKING

This section presents the details on the proposed stereo based hand pose tracking method. The framework is summarized in Fig.3.

3.1. Training based hand modeling

As discussed above, segmentation should be performed before tracking. Unlike the segmentation methods adopted with active depth cameras, it is hard to obtain accurate depth from passive stereo. Hand segmentation from color is difficult. Some of the background colors could be similar to skin. In addition, skin color may also vary under different lighting conditions. It is difficult to construct a generic color model which is suitable for all scenarios.

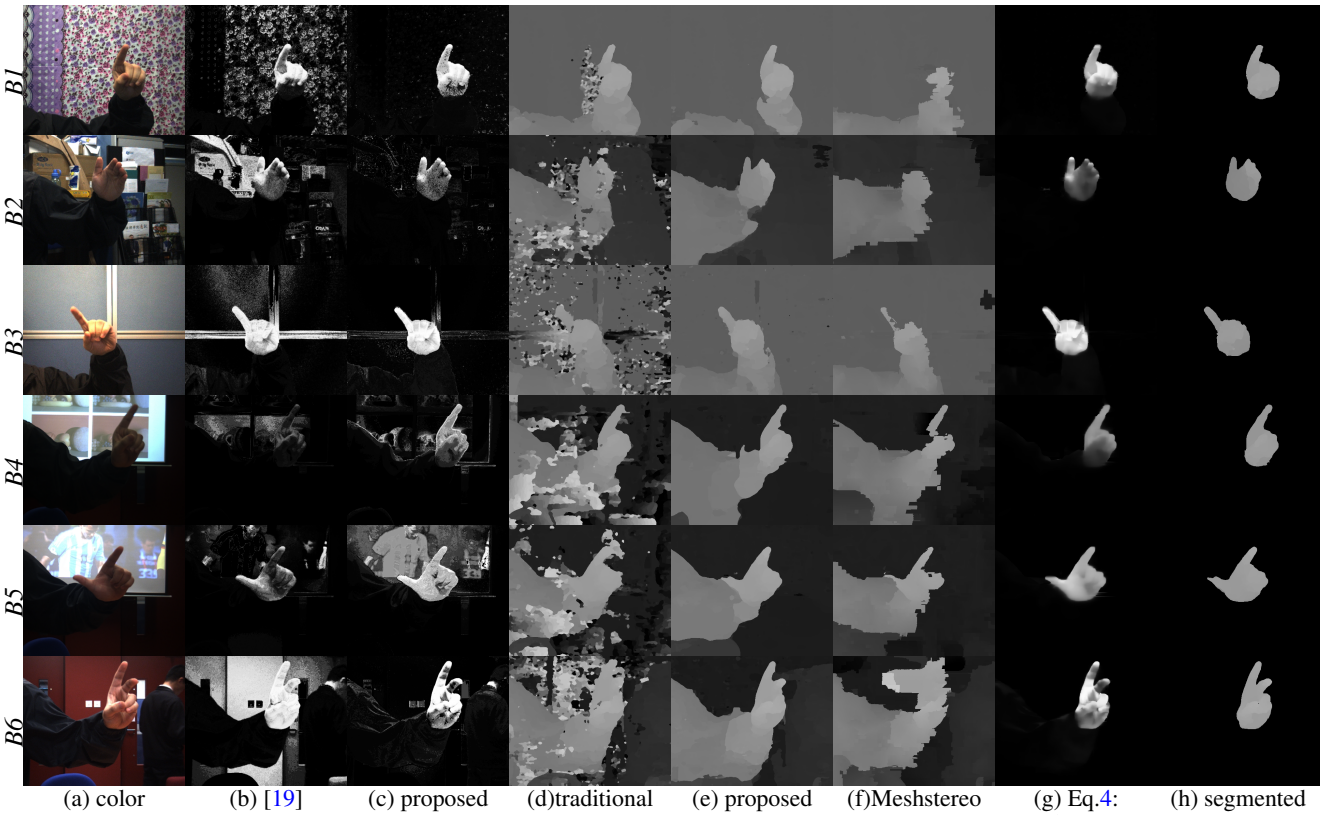
To address these problems, an online training based skin color detector is proposed. A training sequence is captured just before tracking. Adaptive GMM [9, 10], which is a real-time background modeling method, is adopted to segment foreground hand from the background. The hand should keep waving (for a few seconds) in the training sequence to make sure that it is detected as foreground.

After foreground segmentation, the foreground objects are assumed to be the hand which have a specific skin color. The color histogram for the hands $\mathbf{H}^h$ and the histogram for all the images $\mathbf{H}^i$ in the training video sequence are computed. Then the skin color probability is

$$P^s(c) = \mathbf{H}^h(c)/\mathbf{H}^i(c) \quad (1)$$

where c represents a color candidate.

Fig.4 (b)-(c) compares the skin color probability $P^s(c)$ from [19] and the proposed training method. Different from the proposed method, the generic skin color model in [19] is trained by plenty of images from the Internet. Fig.4 (b)-(c) show that the proposed model can better separate skin region from the other objects. It is simply because the generic skin color model in [19] is trained from plenty of images, and thus it treats more colors as skin (e.g., *B2*, *B3* and *B6*). However, skin colors normally only dominate a small region in the color space for a specific scenario. In addition, the generic skin color model [19] provides unsatisfied skin detection result in dark environments like *B4*. It is likely because this type of lighting condition seldom appears in its training data set. As a result, [19] assigns very low skin probability to shadows (on the hand). The proposed hand detector is more robust than [19] mainly because a specific skin color model is trained for every individual scenario. *B5* in Fig.4 (b) shows that the hand probability from the proposed method is relatively high in the background since its background also has fast motion objects and adaptive GMM treats them as foreground. However, this problem can be solved by considering the depth information from the proposed stereo as shown in Fig.4 (e)(g)(h). The details will be presented in the following subsections.



(a) color | (b) [19] $P^s(c)$ | (c) proposed $P^s(c)$ | (d)traditional stereo | (e) proposed stereo | (f)Meshstereo [14] | (g) Eq.4: $P^h(c,d)$ | (h) segmented disparity

Fig. 4. Intermediate results of the proposed hand tracking framework.

3.2. Constrained stereo matching

A traditional stereo matching algorithm does not make any specific assumption on the scene to be captured. It performs well on textured/synthetic scenes as shown in the first row in Fig.4 (d). However, its performance may drop dramatically in a real-world indoor environment where most of the objects contain large textureless regions as can be seen from the last two rows in Fig.4 (d).

The performance of existing stereo algorithms resides on the sufficient texture. However, the region around hand is relatively smooth for stereo matching. Meanwhile, the boundary between hand and background is not clear. Both of these factors prevent existing stereo matching algorithms to seek accurate correspondence. On the other hand, accurate hand segmentation is important for hand tracking. So the proposed stereo algorithm only needs to maintain (i) the depth accuracy of the hand and (ii) a clear depth difference between the hand and the background objects.

Here we propose an stereo matching which is specially designed for hand tracking. Due to the view angles, low texture, and illumination changes, it is difficult to estimate accurate depth in some regions based on the stereo image pair. We classify these regions as occluded pixels and unstable pixels. The occluded pixels only appear in one view (left or right) of the stereo pair, like the background part near the hand edge. These pixels are detected using the left-right consistency check. Meanwhile, the unstable pixels (due to lack of texture, specularity, etc.) are detected based on matching cost confidence [20].

In our algorithm we denote d as a depth/disparity candidate, a new matching cost $\mathbf{N}_p$ at pixel p that excludes the contribution of occlusions is computed as follows:

$$\mathbf{N}_p(d) = \begin{cases} 0 & \text{if } p \text{ is occluded,} \\ \mathbf{M}_p(d) & \text{otherwise.} \end{cases} \quad (2)$$

in which $\mathbf{M}_p$ denotes the original matching cost at pixel p from Census transform [21]. Instead of the reference color image, the skin probability $P^s(c)$ estimated from the model proposed in Sec.3.1 is used as the guidance image for cost aggregation on the new matching cost and let $\mathbf{N}_p^F$ denote the aggregated cost at p by guided filter [16]. As shown in Fig.4(c), most of the non-skin regions are very dark and thus the guidance image filter kernel is very large around those regions. As a result, the corresponding aggregated cost values are quite stable inside these regions. Let $D_p^{\mathbf{N}}$ denote the depth obtained from $\mathbf{N}_p^F$ by winner-take-all. $D_p^{\mathbf{N}}$ is normally over-smoothed around non-skin regions due to the adoption of the large filter kernel. However, the depth accuracy requirement is low on non-skin regions.

Additionally, $D_p^{\mathbf{N}}$ is the intermediate depth estimation but not the final result. It is only used to adjust the original matching cost:

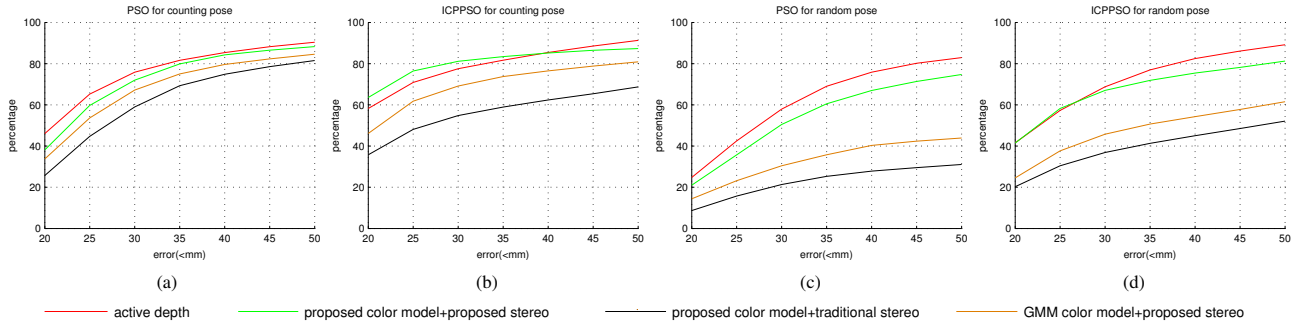


Fig. 5. Average percentage of all joints that have an maximum error less than x mm under six different environments.

$$\mathbf{M}'_p(d) \leftarrow \begin{cases} \alpha|d - D_p^{\mathbf{N}}| & p \text{ is occluded,} \\ \mathbf{M}_p(d) + \beta|d - D_p^{\mathbf{N}}| & p \text{ is unstable,} \\ \mathbf{M}_p(d) & \text{otherwise.} \end{cases} \quad (3)$$

$\mathbf{M}_p$ is the original matching cost, and $\mathbf{M}'_p$ is the cost after the adjustment. α and β are two constants determining the contribution of the intermediate depth $D_p^{\mathbf{N}}$ for occluded and other unstable pixels respectively.

The new matching cost is also filtered using guided image filter with the reference color as the guidance to compute the final depth/disparity map in Fig.4(e). Note that most of the noisy depth estimated in Fig.4(d) are removed from Fig.4(e) although the background depth is slightly over-smooth. However, the depth accuracy on the hand region is well-preserved because the filter kernel is relatively small when cost aggregation is performed on $\mathbf{N}_p$ and the pixels inside the hand region are mostly stable pixels.

Although the state-of-the-art stereo methods like Meshstereo [14] may also produce "clean" backgrounds in Fig.4 (f), it is very slow and the performance around the hand region is obviously lower than the proposed method.

3.3. Hand segmentation

In this paper, a pixel is inside the hand region if its skin color probability $P^s(c)$ is high and its depth d is close to the hand depth in the previous frame. Under this assumption, a hand probability for each pixel can be defined as

$$P^h(c, d) = P^s(c)N(d; \mu_d, \sigma_d), \quad (4)$$

where $N(d; \mu_d, \sigma_d)$ is a Gaussian distribution. The mean μ_d is the average hand depth in the previous frame and the standard deviation σ_d is fixed to 150mm in all the conducted experiments. Finally, we assume a pixel to be inside the hand region if $P^h(c, d) > 0.1$.

Some hand segmentation results with disparities are presented in Fig.4(h). When the background is highly-textured (e.g., *B1*), the disparity from traditional stereo is sufficiently accurate. However, applying traditional stereo on *B6* results in many disparity noises in the background due to the lack of texture. The constrained stereo matching algorithm proposed in Sec.3.2 can obtain globally-smooth (although may not be very accurate) disparity estimates in the background region and thus is very useful in hand segmentation. For some specific backgrounds like *B5*, although trained skin color model is not good enough, accurate hand segmentation could still be obtained with the help of the disparity estimates from the proposed stereo matching algorithm.

4. EXPERIMENTS

This section presents quantitative comparison on hand tracking (PSO [17] and ICPPSO [3]) using the proposed passive stereo system and other configurations. We set the hyper parameters α and β of the proposed constrained stereo as 2 and 0.5 in all the experiments. The experimental results demonstrate that the proposed stereo can achieve tracking performance comparable to the active depth camera. The average percentages of all joints that have a maximum error less than a threshold over different environments are plotted in Fig.5 including simple counting and difficult random poses. The green and red curves are from the proposed stereo and Intel F200 active depth camera. In all the figures, they are close to each other which means passive stereo is suitable for hand pose tracking and its performance is comparable to the active depth cameras. To demonstrate the effectiveness of training based hand modeling and constrained stereo matching in the proposed stereo, they are replaced by generic GMM skin color model [19] (brown curves) and a traditional stereo (Census transform for matching cost computation and the guided image filter for cost aggregation) (black curves) respectively. According to Fig.5, the tracking performance drops without using the proposed method. Please see supplemental material for visual tracking results.

5. CONCLUSIONS

In this paper, we develop a benchmark for evaluating hand pose tracking on passive stereo. Unlike existing benchmarks, it contains both stereo images from a binocular stereo camera and depth images from an active depth camera. It has a total of 12 video sequences and each sequence has 1,500 frames. A novel stereo-based hand segmentation algorithm specially designed for hand tracking is proposed to estimate an accurate hand depth and its performance is demonstrated to be comparable to the active depth cameras under different scenarios.

6. REFERENCES

[1] Emad Barsoum, "Articulated hand pose estimation review," *arXiv:1604.06195*, 2016.

[2] James Steven Supancic III, Gregory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan, "Depth-based hand pose estimation: methods, data, and challenges," in *ICCV*, 2015.

[3] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun, "Realtime and robust hand tracking from depth," in *CVPR*, 2014.

[4] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt, "Interactive markerless articulated hand motion tracking using rgb and depth data," in *ICCV*, 2013.

[5] Danhang Tang, Tsz-Ho Yu, and Tae-Kyun Kim, "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," in *ICCV*, 2013.

[6] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun, "Cascaded hand pose regression," in *CVPR*, 2015.

[7] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *TOG*, 2014.

[8] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis, "A survey of skin-color modeling and detection methods," *PR*, 2007.

[9] Chris Stauffer and W Eric L Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999.

[10] Zoran Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *ICPR*, 2004.

[11] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *TPAMI*, 2013.

[12] Michael Bleyer, Christoph Rhemann, and Carsten Rother, "Patchmatch stereo-stereo matching with slanted support windows.," in *BMVC*.

[13] Qingxiong Yang, Liang Wang, Ruigang Yang, Henrik Stewénius, and David Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *TPAMI*, 2009.

[14] Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui, "Meshstereo: A global stereo model with mesh alignment regularization for view interpolation," in *ICCV*, 2015.

[15] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, 2002.

[16] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *TPAMI*, 2013.

[17] N. Kyriazis I. Oikonomidis and A. A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect," in *BMVC*, 2011.

[18] Jean-Yves Bouguet, "Camera calibration toolbox for matlab," https://www.vision.caltech.edu/bouguetj/calib_doc/, 2004.

[19] Michael J Jones and James M Rehg, "Statistical color models with application to skin detection," *IJCV*, 2002.

[20] Geoffrey Egnal, Max Mintz, and Richard P Wildes, "A stereo confidence metric using single view imagery with comparison to five alternative approaches," *Image and vision computing*, 2004.

[21] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *ECCV*. 1994.