Image and Vision Computing 79 (2018) 63–75

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Top-down model fitting for hand pose recovery in sequences of depth images☆

Meysam Madadi[a,b,*], Sergio Escalera[a,c], Alex Carruesco[d], Carlos Andujar[d], Xavier Baró[a,e], Jordi Gonzàlez[a,b]

[a]Computer Vision Center, Edifici O, Campus UAB, Bellaterra (Barcelona) 08193, Catalonia, Spain
[b]Dept. of Computer Science, Univ. Autònoma de Barcelona (UAB), Bellaterra 08193, Catalonia, Spain
[c]Depl. Mathematics and Informatics, Universitat de Barcelona, Catalonia, Spain
[d]ViRVIG-Moving Research Group, UPC-BarcelonaTech, Spain
[e]Universitat Oberta de Catalunya, Catalonia, Spain

ARTICLE INFO

Article history:
Received 9 October 2017
Received in revised form 18 May 2018
Accepted 12 September 2018
Available online 21 September 2018

Keywords:
Hand pose recovery
Shape description
Depth image
Hand segmentation
Temporal modeling

ABSTRACT

State-of-the-art approaches on hand pose estimation from depth images have reported promising results under quite controlled considerations. In this paper we propose a two-step pipeline for recovering the hand pose from a sequence of depth images. The pipeline has been designed to deal with images taken from any viewpoint and exhibiting a high degree of finger occlusion. In a first step we initialize the hand pose using a part-based model, fitting a set of hand components in the depth images. In a second step we consider temporal data and estimate the parameters of a trained bilinear model consisting of shape and trajectory bases. We evaluate our approach on a new created synthetic hand dataset along with NYU and MSRA real datasets. Results demonstrate that the proposed method outperforms the most recent pose recovering approaches, including those based on CNNs.

1. Introduction

Hand pose recovery has attracted great interest in recent years due to the availability of affordable depth cameras. Depth sensors have allowed researchers to use non-invasive, accurate approaches to hand pose estimation, which are more robust to illumination and color changes than standard RGB cameras. These features have lead to significant advances in multiple applications including human-computer interaction, virtual reality, robot learning and gesture recognition, just to name a few [4,5,7,8].

Although recent hand tracking approaches based on depth cameras achieve high performance for some applications, there are still several open challenges to tackle, such as finger self-occlusion, hand-body occlusions, low resolution/noisy depth images, and above all, the inherent complexity of modeling hand motion due to its highly articulated nature. Available datasets mainly provide front-face hand deformations, which are not suitable to compare state-of-the-art approaches against hard cases with large occlusions. To the best of our knowledge, little attention has been paid to incorporate temporal motion information in hand pose recovery problems. As an example, Oikonomidis et al. [17] only initialized the model using previous frame.

In this paper, a solution to the problem of hand pose recovery in depth image sequences is proposed. The solution combines both spatial and temporal information in a top-down strategy. We present a system for efficient hand pose recovery in non-controlled settings involving self-occlusions. Based on current trends towards minimizing pose parameters in the space of nearest candidates [23,39], we exploit an effective shape descriptor to extract such nearest candidates. As in [28] we estimate each object part separately while reducing the search space. We first extract palm joints, which provide a basis for fingers, using nearest candidates. Following [20] we define an efficient objective function and then minimize parameters of each finger model to fit with its appearance. Our function is different from [20] since they extract fingertips while we accurately segment fingers. Thanks to this objective function we get a fast convergence to the finger model parameters while handling occluded parts.

Motivated by [43], our estimated joints are applied in a sequence of frames to minimize parameters of a trained bilinear model [1] consisting of shape and trajectory bases. This process further refines

☆ This paper has been recommended for acceptance by Catherine Pelachaud.
* Corresponding author.
E-mail address: mmadadi@cvc.uab.es (M. Madadi).

https://doi.org/10.1016/j.imavis.2018.09.006
0262-8856/

the estimation of occluded parts. Fig. 1 shows our method pipeline: nearest neighbors extraction, hand segmentation, single-frame pose recovery, and temporal pose recovery. Our approach has proven to be more robust under large viewpoint sets and complex hand poses than state-of-the-art approaches when data is balanced for different viewpoints and poses. To evaluate our method under such situations, we created a synthetic dataset with +600 K hand pose samples for single-frame pose recovery and +1 M frame sequences for temporal pose recovery, with high deformations and occlusions in both learning and test sets. We call this dataset *SyntheticHand*. Although, egocentric datasets have been recently introduced [22], hand-object interaction is not within the scope of this paper. Though, we evaluate on real datasets like NYU [36] and MSRA [28], and obtain comparable results on both model-based and data-driven approaches.

The rest of the paper is organized as follows. Section 2 reviews state-of-the-art works in the field. Section 3 presents the proposed system. Results are shown in Section 4, and finally, Section 5 concludes the paper.

2. Related work

The field of hand pose estimation has become very active due to the use of depth sensors. A comprehensive survey on existing methods can be found in [8] and [41]. In this section we focus on those approaches most related to our contribution. Hand pose estimation methods can be roughly divided into model-based methods and data-driven methods [6,19].

Model-based techniques consider an a priori 3D hand model whose pose is determined over time by some tracking procedure [12,20,23], like the Particle Swarm Optimization presented in [17]. Hand model can take a simple shape by using cylinders and spheres [17] or be defined in a parametric space learned by some priors [9]. Unfortunately, these approaches require some kind of accurate initialization, and due to the fast motion and non-rigid nature of hands, together with finger self-occlusions, it is still a challenge for single-hand trackers to correctly maintain the state of an animated 3D hand model over time. In recent works, while some works propose more advanced hand models [35], others try to sample hypotheses by physical constraints [21]. In model based approaches, designing an efficient energy function is important to guarantee a global solution with minimum energy. In this sense, Taylor et al. [34] use a complex function including surface discrepancy and normal vector consistensy, constraints on pose parameters, temporal sequence smoothness, self-intersection and fingertip checking. Minimization of this function in such a high-dimensional parametric model is tractable using gradient-based optimization which needs a differentiable function.

On the other side, data-driven methods directly predict at each frame the pose of the hand by learning depth and image features [28]. Contrary to using hand trackers, which lead to model drift over time, single-frame detection methods are initialized at each frame, thus recovering more easily from estimation errors [23]. Multiple procedures based on Random Forests (RF) have emerged including Hough Forests [39], Random Decision Forests [11] and Latent Regression Forests [30], as detailed in [8]. Unfortunately, the number of occluded joints is commonly bigger in hands than in human bodies. As a result, techniques based on RF usually require huge training sets, and some kind of viewpoint estimation is needed in order to improve performance [32]. Some data-driven works analyze the hand in the space of nearest shapes in order to reduce the search space [23] or approximate unknown pose parameters through matrix factorization [3].

Following current trends in Computer Vision, although both the architecture and weight initialization of a neural network strongly determine its performance, CNN-based techniques continuously improve the state-of-the-art accuracy on different benchmarks. Tompson et al. [36] optimized an inverse kinematic approach based on joint heatmaps generated by CNN for 2D joints estimation. Oberweger et al. [15] optimized a shallow CNN based on embedded space of hand

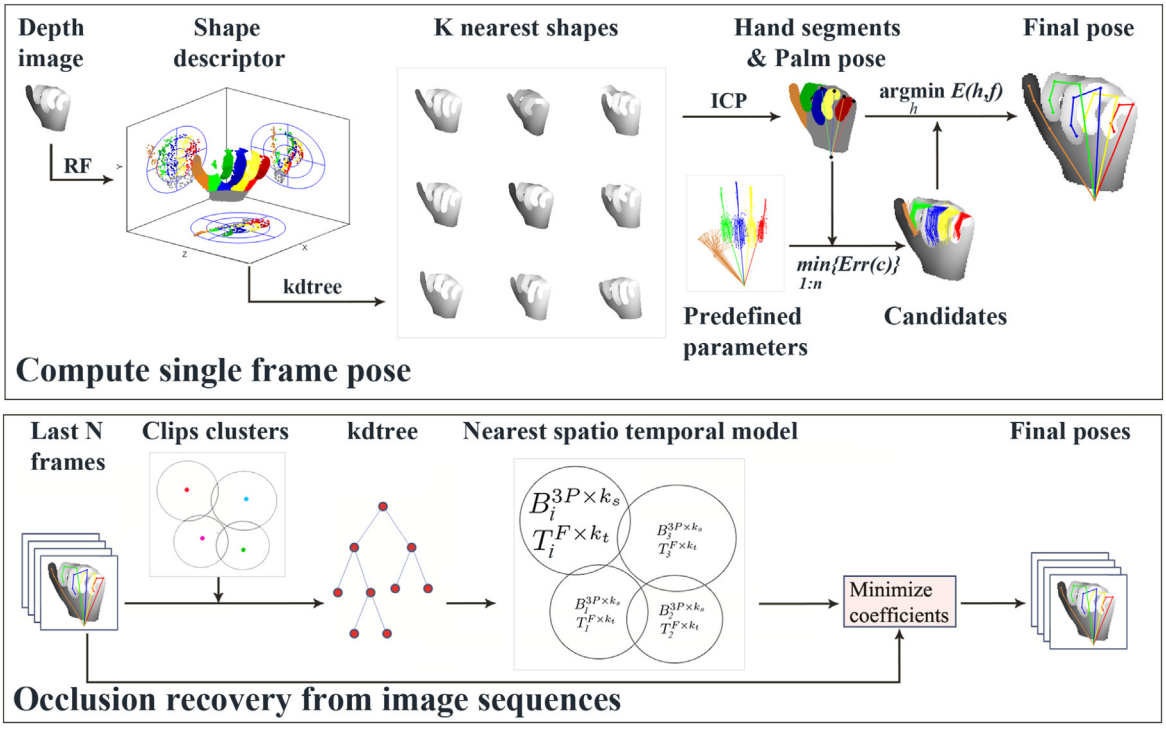

Fig. 1. Diagram of the proposed method. In the first step, a single-frame hand pose is estimated. First palm joints and finger segments are recovered through nearest shapes. Then finger models are fitted using extracted candidates. In the second step, temporal data is incorporated to refine first step estimation.