Pose Estimation from a Single Depth Image for Arbitrary Kinematic Skeletons

Daniel L. Ly[1], Ashutosh Saxena[2] and Hod Lipson[1]
[1]School of Mechanical and Aerospace Engineering, [2]Department of Computer Science
Cornell University, Ithaca, NY
dll73@cornell.edu, asaxena@cs.cornell.edu, hod.lipson@cornell.edu

***Abstract*—We present a method for estimating pose information from a single depth image given an arbitrary kinematic structure without prior training. For an arbitrary skeleton and depth image, an evolutionary algorithm is used to find the optimal kinematic configuration to explain the observed image. Results show that our approach can correctly estimate poses of 39 and 78 degree-of-freedom models from a single depth image, even in cases of significant self-occlusion.**

I. Introduction

Being able to estimate three-dimensional pose of an articulated articated object, such as a robot or human, is important for a variety of applications (eg. [8]). While recent technological advances have made capturing depth images both convenient and affordable, extracting pose information from these images remains a challenge—even when the kinematic structure of the target is provided. Popular approaches often rely domain specific knowledge and extensive training, thus providing little generality to arbitrary skeletons where little or no training data exists.

This paper presents results on estimating poses of an arbitrary kinematic skeleton from a single depth image without prior training. The pose estimation is defined as a model-based estimation problem and an evolutionary algorithm is applied to find the optimal pose. Rather than using a priori beliefs or pre-trained models, this algorithm extracts the most likely configuration based solely on the kinematic structure to explain the observed depth image (Fig. 1).

II. Related Work

The vast majority of pose estimation research focused specifically on the human kinematic skeleton. Recent surveys [5, 6] describe two primary directions: pose assembly via probabilistic detection of body parts and example-based methods. For example, Shotton et al. [7] described a particularly successful approach to human pose recognition that builds a probabilistic decision tree to first find an approximate pose of body parts, followed by a local optimization step. While this technique is fast and reliable, it relies on significant training: 24000 core hours of training on 1 million randomized poses. A primary limitation of these techniques is their reliance on domain specific information regarding human kinematics which does not generalize to arbitrary skeletons without explicit and additional training.

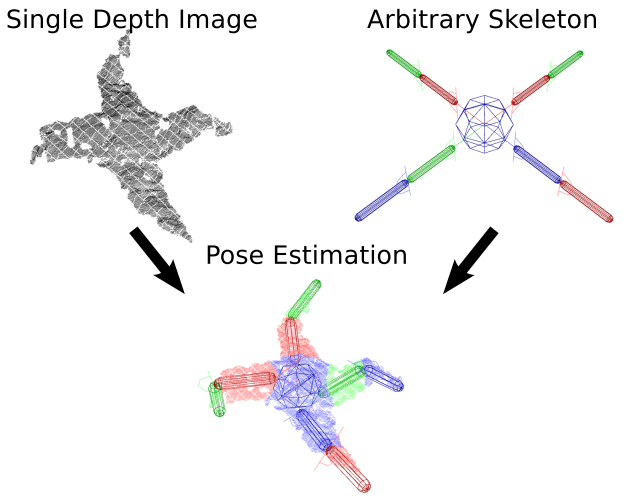

Fig. 1. A diagram illustrating the process of inferring poses from a single depth image using an arbitrary skeleton. Given raw data from a depth image (top left) and a parameterizable skeleton (top right), our algorithm the set of parameters that pose the skeleton to best explain the data.

In comparison, Gall et al. [3] used motion capture with markerless camera systems to find poses of complex models, such as those generated from animals and non-rigid garments. However, this approach required laser scanned visual-hulls which were mapped, by human experts, to an underlying kinematic structure.

In an alternative approach, Katz et al. [4] inferred relational representations of articulated objects by tracking visual features, but is limited to planar objects and requires interactions to infer the underlying structure.

III. Pose Estimation via Evolutionary Computation

The pose estimation is defined as an optimization problem:

$$s^* = \underset{s(\boldsymbol{\theta})}{\operatorname{argmin}} \frac{1}{N} \sum_{n=0}^{N} \ln\left(1 + \frac{||\vec{p}_n - \vec{p}^*(\boldsymbol{\theta}, \vec{p}_n)||}{\sigma}\right) \quad (1)$$

where $s(\boldsymbol{\theta})$ is skeleton model with parameters $\boldsymbol{\theta}$, $\vec{p}_n$ is a point from the observed depth image, $\vec{p}^*$ is the closest point on the model to the point $\vec{p}_n$ given the parameters $\boldsymbol{\theta}$, and σ is the standard deviation of point distances in the depth image. This

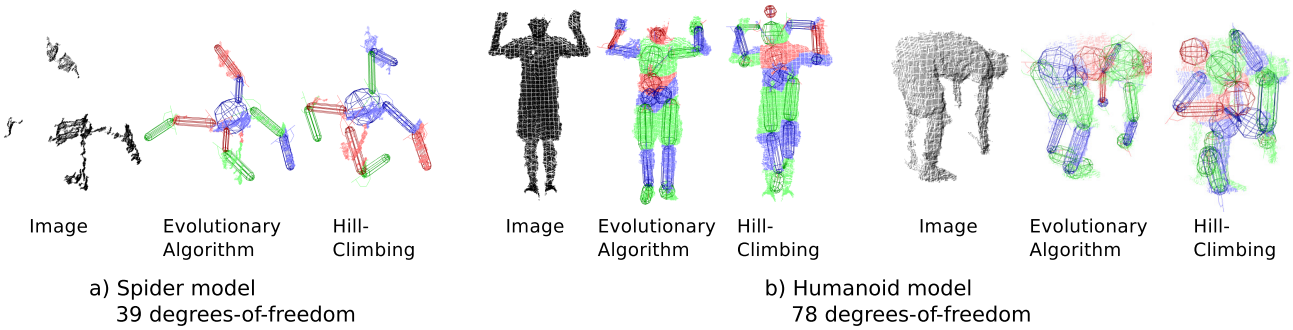

Fig. 2. Selected examples of estimated poses for the spider and humanoid model. The raw depth image, and poses inferred by our algorithm and a hill-climbing baseline are shown for each example.

optimization problem is challenging as it is non-convex with numerous local optima.

Therefore, we use an evolutionary algorithm to find the pose parameters. The evolutionary algorithm is a population-based, heuristic algorithm that iteratively selects and combines solutions to produce increasingly better models [2]. The skeleton is represented as an acyclic graph of links, with parameterizable joint angles and length. Traditional evolutionary operators are used: random mutations are applied to the parameters and recombination swaps branches of links between parents to produce offspring.

IV. Experiments and Results

We captured a data set of an articulated robot using a Kinect camera's depth sensor [1]. Images of two drastically distinct subjects were captured: the first is a spider model is a based on a quadruped robot with 8 links resulting in 39 degrees-of-freedom, while the second is a humanoid model consisting of 17 links amounting to 78 degrees-of-freedom. For the spider model, we arranged a robot in four distinct poses, and collected five images ranging in inclination angles was taken per poses; resulting in a total of twenty depth images. There were multiple examples of self-occlusion in the data set. For the humanoid model, eight images were taken of four subjects, totaling to 32 images. The images in both data sets were pre-processed with background subtraction.

We ran the learning algorithm for 10^9 objective function evaluations, which is approximately 10000 iterations. On a single core 2.8GHz Intel processor, this required approximately 30 and 70 minutes of computational effort per image for the spider and humanoid models, respectively.

Results of the evolutionary algorithm indicate successful pose estimation for both the spider and humanoid models, even with significant self-occlusion (Fig. 2). Quantitatively, our model placed links in the correct position with an accuracy of 99% and 84% for the spider and humanoid model, respectively. Qualitatively, on a scoring survey with a scale of 5, spiders scored 4.9, while the humanoid model achieved a 4.1 score.

Compared results to a hill-climbing baseline, our learning algorithm produced mariginally superior results for the low-dimensional spider model. However, for the high-dimensional humanoid model, there is a sharp contrast in performance. The evolutionary approach is able to consistently infer a reasonable approximation to the model, while the hill-climbing approach is often caught in local optima that are drastically different than the ground truth. The results indicate that for high-dimensional problems with overlapping workspaces, the proposed learning method is vastly superior to determining pose information.

Acknowledgements

This work was supported in part by NIH NIDA grant RC2 DA028981, NSF CDI Grant ECCS 0941561, and DTRA grant HDTRA 1-09-1-0013. D. L. Ly thanks NSERC for their support through the PGS program. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the sponsoring organizations.

References

[1] Microsoft Corp. Kinect for Xbox 360.

[2] K. A. De Jong. *Evolutionary Computation.* MIT Press, 2009.

[3] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenbahn, and H.P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009.

[4] D. Katz, Y. Pyuro, and O. Brock. Learning to manupulate articulated objects in unstructured environments using a grounded relational representation. In *RSS*, 2008.

[5] T. Moeslund, A. Hilton, and V. Kruger. Survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2-3):90–126, 2006.

[6] R. Poppe. Vision-based human motion analysis: an overview. *CVIU*, 108(1-2):4–18, 2007.

[7] J. Shotton, A. Fitzgabbon, M. Cook, T. Sharp, M. Finnochio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011.

[8] J. Y. Sung, C. Ponce, B. Selman, and A. Saxena. Human Activity Detection from RGBD Images. In *AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011.