Pose estimation from a single depth image for arbitrary kinematic skeletons

Daniel L. Ly

***Abstract*— This paper presents a general approach 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 skeletal configuration to explain the observed image. Results show that this approach can correctly pose a 23 degree-of-freedom model from a single depth image, even in cases of significant self-occlusion.**

I. INTRODUCTION

A fundamental issue in a multitude of robotic application is the automated, three-dimensional pose estimation of an articulated object. While recent technological advances have made capturing depth images convenient and affordable, extracting pose information from these images remains a challenge – even when the kinematic structure of the target is provided.

Previous approaches often rely domain-specific knowledge and extensive training, thus providing little generality to arbitrary skeletons where little or no training data exists. These approaches do not use the kinematic skeleton directly, but instead transform the kinematic pose information to an intermediate representation for comparison, such as human body part recognition [1] or referencing a known an external hull [2]. Despite the accuracy and success of these algorithms, they are unable to generalize beyond the narrow scope imposed by the fundamental assumptions that arise from these intermediate representations.

A technique that abstracts skeletal information and extracts poses from arbitrary depth images would have a profound affect the design and training of complex, articular robotic systems. For example, the general definition of a head, torso and four limbs readily describes a wide range of kinematic skeletons, from humanoid structures to quadrupedal forms. The ability to extract a pose given an arbitrary skeleton and depth image pair without any prior assumptions will provide a key analysis tool for robots.

This paper presents a novel approach to 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).

D. L. Ly is with the Department of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14853, USA dll73@cornell.edu

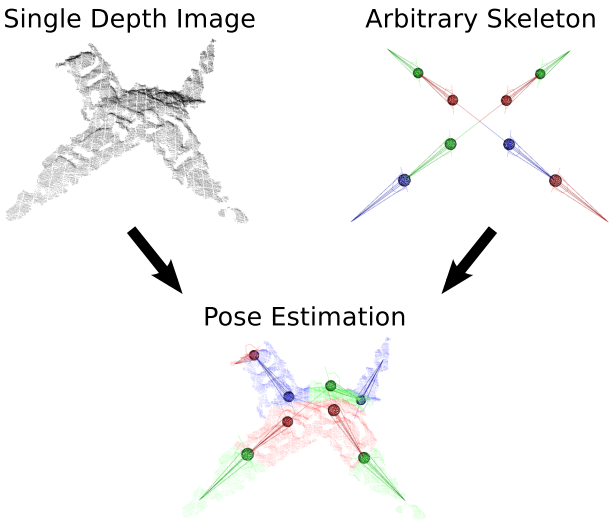

Fig. 1. Inferring pose information from a single depth image and an arbitrary skeleton.

II. RELATED WORK

The vast majority of pose estimation research focused on specifically the human kinematic skeleton. Recent surveys [3], [4] describe two primary directions: pose assembly via probabilistic detection of body parts and example-based methods. Pose assembly attempts to reconstruct the pose by first identifying body parts using pairwise constraints including aspect ratio, scale, appearance, orientation and connectivity. In contrast, example-based methods compare the observed image with a database of samples. A primary limitation of these techniques is their reliance on domain-specific information regarding human kinematics. For pose assembly, direct assumptions are made regarding the feasibility of the constraints while example-based methods extrapolate information from the existing database.

Shotton et al. 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 [1]. This approach forms the basis for a real-time implementation on commercial hardware. While this technique is fast and reliable, it relies on significant training exclusive to the humanoid skeletal structure: 24 hours on 1000 cores of training on 1 million randomized poses.

In comparison, Gall et al. used motion capture with markerless camera systems to find poses of complex models generated from animals and non-rigid garments [2]. How-

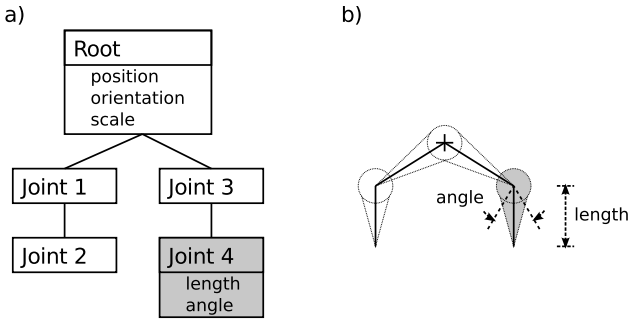

Fig. 2. a) The acyclic graph representation of the skeleton and b) the corresponding visual depiction. In addition to the line segments, directed cones are added to the visual depiction for clarity. Joint 4 is highlighted to show the joint angle and length parameters.

ever, this approach required laser scans of the initial subject to provide a definitive mapping between the skeleton and the point cloud distribution. Expert knowledge was used to define how the external hull moved with respect to the underlying skeleton.

In an alternative approach, Katz et al. inferred relational representations of articulated objects by tracking visual features [5]. While this work does not focus on pose estimation directly, it presents a framework to extract kinematic information from an unknown object using computer vision. However, it is limited to planar objects and requires interactions to infer the underlying structure.

Finally, teaching by demonstration has been shown to be an efficient and natural method to transfer knowledge to robots. Riley et al. used imitation to achieve human-like behaviour in highly-complex, humanoid robots [6] while Kober at al. explored how to use demonstrations to learn motor primitives and tackle complex dynamics problem via reinforcement learning [7]. Although, illustrate the potential uses for automated pose estimation in a robotics setting, current teaching by demonstration implementations relies on predefined transformations between the teacher and student and there have been no attempts to generalize to arbitrary teachers.

III. POSE ESTIMATION VIA EVOLUTIONARY COMPUTATION

This section begins with a formal description of the kinematic models and a definition of the pose estimation problem. A evolutionary computation framework is then presented, followed by a discussion of two techniques to aid in scalable performance: coevolution of rank predictors and age-fitness Pareto optimization.

A. Skeleton representation

Skeletons are represented as a collection of parametrized joints in an acyclic graph structure (Fig. 2). The root node represents a frame of reference that describes the position of the origin, orientation and scale – this corresponds to seven degrees-of-freedom (DOF) for a three-dimensional space. The position and orientation are unbounded while the scale is loosely constrained, allowing to sweep several orders of magnitude.

Every subsequent child represents a joint, which is abstracted mathematically as a line segment of zero thickness. Each joint is described by two free parameters: joint length and joint angle. Both parameters are constrained between two predefined bounds, and linear interpolation or SLERP [8] is used to interpolate between the bounds, accordingly. By setting identical bounds for the upper and lower limits, joint parameters can be effectively removed. For example, although each joint has only one degree of rotation, complex joints such as ball and sockets can be obtained by cascading multiple zero length joints.

B. Pose estimation

The pose estimation problem is defined in general optimization framework:

$$s^* = \underset{s}{\operatorname{argmin}}\, E(s(\theta), \mathbf{p}) \tag{1}$$

where $s(\theta)$ is skeleton model with parameters θ, $\mathbf{p}$ is the collection of points from the observed depth image and $E(\cdot)$ is an objective function to measure the error between the model and the data. Thus, pose estimation is fundamentally an attempt to find a skeleton configuration that best matches the observed data.

However, determining a suitable metric of fit for an arbitrary skeleton model and point cloud data pair is a non-trivial task. Previous work in pose estimation relied on known geometric information of the skeleton model. The relationship between the skeleton and expected external hull is defined a priori, and the point cloud is compared to the existing hull. While this approach allows for accurate results and a simple error metric, its reliance on known external hulls makes it suitable for arbitrary skeletons.

Consequently, we present an objective function to measure the error between the an arbitrary skeleton and a point cloud. This metric is designed to be as general as possible and it only requires that the kinematic structure can be represented as a series of line segments. The objective function is defined as follows:

$$E(s(\theta), \mathbf{p}) = \sum_{n=0}^{N} \log\left(1 + \frac{||p^* - p_n||^2}{\sigma^2}\right) \tag{2}$$

where σ^2 is the variance in the positions of the point cloud data and p^* is the the closest point on the skeleton, which is defined as a series of nested arguments. Since each joint is represented as a line segment, any point on that joint j is represented by interpolating between the two end points $p_{j,i}$ and $p_{j,f}$:

$$p_j = \lambda p_{j,i} + (1 - \lambda) p_{j,f} \tag{3}$$

where $\lambda \in [0, 1]$ is a interpolation parameter. For a given joint, the closest point to the data is then readily defined as:

$$p_j^* = \underset{p_j}{\operatorname{argmin}}\, ||p_n - p_j||^2 \tag{4}$$

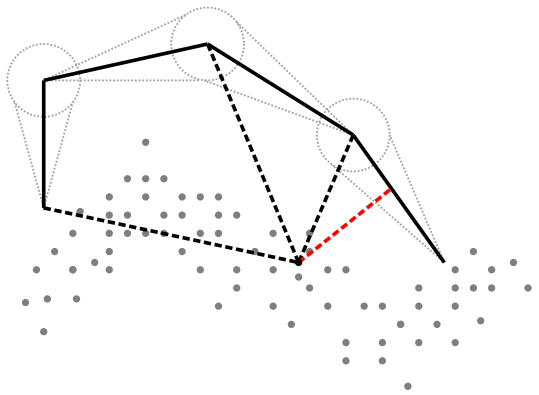

Fig. 3. A visualization of the error metric evaluated on a single point cloud datum. The distance between the point cloud datum and the nearest point on the joint segment is computed for each joint in the skeleton, indicated by the dashed lines. Of these distances, the shortest length (highlighted) is used for the error calculation (Eq. 2).

Finally, the closest point on the skeleton is defined by iterating across all of the joint segments in the structure:

$$p^* = \underset{p_j^*}{\operatorname{argmin}} \left|\left|p_n - p_j^*\right|\right|^2 \tag{5}$$

This piecewise definition is illustrated in Fig. 3. The objective error is computed using a logarithmic error since it selects for the conditional mean and, thus, is more robust to noise than selecting for the conditional mean or median for squared or absolute error, respectively.

An essential feature of this objective function is its data-centric, as opposed to model-centric, definition. The objective function only increases the error for points that are not well explained by the model. However, it is important to note that the error is defined only by the closest joints, which are in turn determined relative to the data. Thus, if a joint is not associated with any data points, it is able to move freely without any affect on the objective error.

The major benefit of this data-centric definition is its ability to deal with partial occlusion in an elegant manner. For single depth images of articulate robots, self-occlusion is often a crippling issue for pose estimation. By avoiding a model-centric error, there is no inherent penalty for positioning occluded joints where no data exists. For partial occlusions, this approach can often lead to the good models by positioning and obstructing a single joint so that the data points are explained by the remainder of joints.

While the objective function had advantageous geometric properties, it has many undesirable properties from an machine learning or optimization perspective. The function is not convex and is densely populated with local optima. Furthermore, the large number of nested arguments to determine the closest point can result in sharp discontinuities in the parameter space of the skeleton. As a result, an evolutionary algorithms is the preferred approach for this difficult optimization problem.

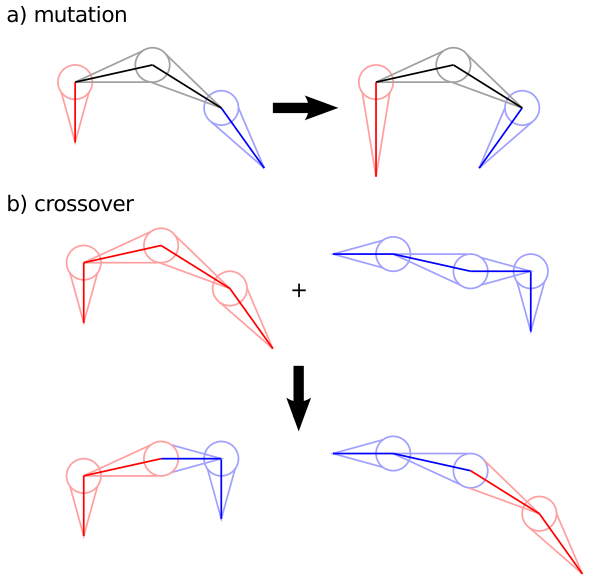

Fig. 4. A visualization of the a) mutation and b) crossover operators. In this example, mutation increased the length of the red joint and changed the joint angle of the blue joint. For crossover, the root was selected as the crossover point and the joint chains were swapped accordingly to produce offspring.

C. Evolutionary algorithm

An evolutionary algorithm (EA) is proposed to determine the optimal pose estimation parameters. EAs are a stochastic, population-based, heuristic algorithms that iteratively selects and combines solutions to produce increasingly better models. Traditionally, EAs are framed as to maximize the fitness of an individual rather than minimize an error metric, which is often achieved by simply negating the sign of the metric.

In the pose estimation, the genotype, or solution encoding, of an individual is the underlying parameters of a given kinematic acyclic graph described in Section III-A.The population is initialized with a collection of randomly generated individuals. For each individual, the root node position is initialized on an existing point in the depth image, selected from a uniform distribution, while the orientation is obtained via four independent samples from a Gaussian distribution followed by normalization to a unit quaternion. For the bounded parameters of scale, joint length and joint angle, the interpolation parameter is sampled from 0 to 1 with a uniform distribution.

Th acyclic graph representation is very amenable to the evolutionary processes of mutation and recombination. Stochastic point mutations are applied to each of the parameters in the graph in a similar method to the initialization protocol, but localized to individual nodes (Fig. 4.a). For recombination, a random crossover point is selected for the existing parent pair, and the offspring are produced by swapping subgraphs at the crossover point (Fig. 4.b).

The phenotype, or output behaviour, of an individual is the pose of the kinematic model. The phenotype of each individual is then used to determine its fitness, or how well it explains the data, according to Eq. 2.

An evolutionary approach was the preferred method for this pose estimation problem as it provides numerous benefits. First, EAs have been successfully applied to non-linear, non-convex optimization problems with deceiving fitness landscapes. Next, the population-based dynamics in EAs allow it to search large and high-dimensional search spaces in an efficient manner. Finally, EAs are best suited when the genotype representation allows for local optimization. For pose estimation, two skeletons might optimize different subgraphs and joint chains, and by recombination, their offspring can contain each superior subgraph to resulting in a significantly better model than either parent.

However, due to the complexity of the pose estimation problem, additional modifications to the EA were required to provide scalable performance. A competitive coevolution algorithm using rank predictors is applied to enhance the effective computation while, age-fitness Pareto optimization was used as the selection strategy to maximizing performance while ensuring diversity.

D. Competitive coevolution using rank predictors

A common criticism of evolutionary algorithms and a prohibitive limitation in practice stems from the computationally heavy demands of these algorithms. Often, the primary culprit in the computational requirements arises from calculating fitnesses. In pose estimation, determining the fitness of a single individual requires repeatedly evaluating a local metric. A single depth image can consist of thousands of points and, since most points are nearly identical to its neighbour, computing the fitness of nearby points adds limited information in terms of evolutionary progress but nonetheless requires significant computational resources.

Rather than using the entire large data set, a coarser and lightweight approximation is substituted to alleviate the computational requirements by competitively coevolving predictors. Instead of using the complete depth image, the fitness is measured only on a dynamic subset of the data. The members of this subset, called predictors, are coevolved simultaneously based on the solution population, allowing for evolutionary progress through direct competition. The key to this coevolution technique relies on the systematic method of evolving predictors – predictors are rewarded based on their ability to rank solutions, rather than using fitness measurements directly [9].

In this work, we show that depth images consisting of tens of thousands of points can be can be effectively replaced by a dynamic selection of a hundred points. Since a single fitness computation consists of nested iterations, decreasing the number of evaluated points by a few orders of magnitude results in drastic performance benefits. Effectively, rank predictors allow more generations to be evaluated for the same computational effort, providing greater exploration of the search space. In addition to the reduced computational load, rank predictors also provide indirect performance benefits by focusing the search to the areas of greatest interest.

E. Age-fitness Pareto optimization

A common issue for machine learning algorithms for problems with numerous local optima is that the algorithm often stagnates on a local optima and solutions stop improving. Despite being a population-based algorithm, the entire EA population is capable of prematurely converging on a local optima, failing to make any substantial progress despite expending additional computational effort. This issue is particularly daunting for pose estimation, where a large number of local optima exist. By the nature of the acyclic graph, parameters near the root have a greater affect on the final pose, and thus root initialization with bad conditions almost surely leads to suboptimal solutions.

A popular remedy for dealing with premature convergence is to perform multiple evolutionary searches via multiple times. However, due to the stochastic nature of the algorithm, it is difficult to know when a restart is required. Furthermore, even if the best individuals no longer improve, the remainder of the population may still contain relevant genotypic snippets for future generations and removing the entire population may be computationally inefficient.

One of the best performing heuristics to deal with premature convergence is the application of genotypic age – a measure of how long genotypic material has existed in the population. For every generation, a new random individual is inserted into the population and for every generation an individual exists, its age is incremented. During crossover, an offspring's age is inherited by the maximum age of its parents. The primary role of age is its affect on selection; individuals are selected for the next generation according to a multi-object Pareto front optimization that ensures an individual cannot be removed if it has the best fitness for a given age [10]/

The age-fitness Pareto optimization maintenance of an effective balance of diversity and performance. Individuals that no longer show improvement are susceptible to being replaced, while young individuals are shielded from being unfairly dominated by individuals who had a more time to explore the search space.

IV. EXPERIMENTAL SETUP

A data set of an articulated robot was captured using a Kinect camera's depth sensor [11]. The robot consists of four legs, each with two rotational degrees-of-freedom. The data set consists of four distinct poses, each with ten images across the complete range of inclination angles, totalling to forty single depth images. The variation in inclination angles resulted in numerous of images with self-occlusion. Each image was pre-processed with background subtraction and the images contained between 13,000 and 24,000 points. Note that the evolutionary algorithm inferred the images independently and no information is transferred between runs.

The target kinematic skeleton has eight limbs of unknown length and joint angles, resulting in a 23 DOF model. Note that the size, orientation and position of the model is part of the search problem and no calibration or prior distribution

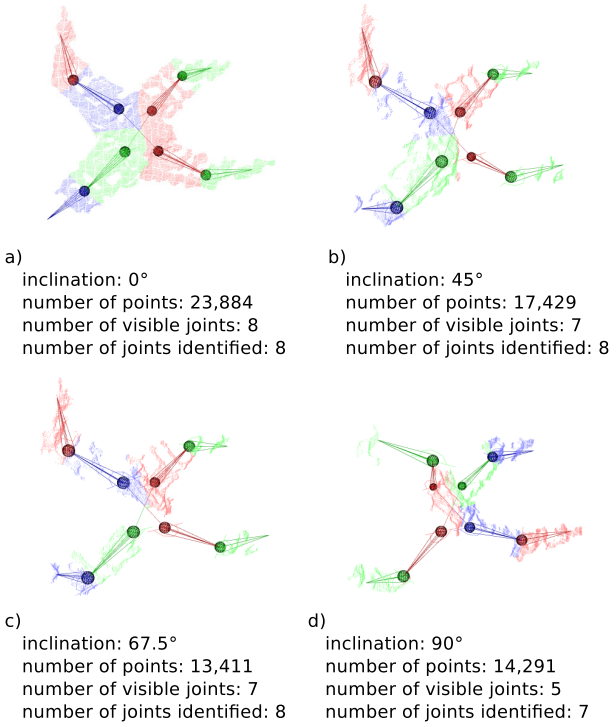

Fig. 5. Selected examples of estimated poses for various inclination angles.

was required. Furthermore, there were no geometric constraints on the limbs, such as enforcing symmetry. Although the kinematic skeleton and robot has the same fundamental structure for this experiment, the evolutionary approach does not require such restrictions and can be readily applied to non-isomorphic skeleton/depth-image pairs.

The evolutionary search had 256 individuals with a mutation probability of 1% and a crossover probability of 50%. There were 16 predictors, each as a subset of 128 points from the depth image. The trainer population consisted of 8 individuals and was updated every 100 generations. In addition to the standard crossover and mutation operators, a greedy hill-climbing subalgorithm was used to tweak to produce minute changes in the model parameters as mutation produced large and unreliable changes.

The evolutionary algorithm was terminated after 2000 generations, which required approximately 5 minutes of computational effort per image on a single core of a 2.2GHz Intel processor. The termination condition was chosen arbitrarily as a conservative estimate of the computational effort required to reach convergence.

V. RESULTS

The EA approach to pose estimation was applied to the data set captured from the eight-jointed robot. Of the 40 single depth images, the EA is able to identify $7.1 \pm .9$ of the eight limbs or, equivalently, achieve an 89% accuracy on limb identification. However, when accounting for the number of visible limbs, the EA is able to achieve a 1.08% accuracy, indicating that can reliably able to find the original pose of the robot, even with significant self-occlusion. Fig. 5 shows a collection of inferred poses across a range of inclination angles. As the camera was brought closer to the horizon, several joints became entirely occluded (Fig. 5.b,c).

The primary failure mode occurred when an end of the skeleton chain was occluded, resulting in a degenerate model (Fig. 5.d). In this case, there is not enough information to reconstruct the upper right joint and the algorithm opted to effectively collapse that joint chain into a single limb.

Informally, while the EA algorithm is effective at determining qualitative properties of the pose information, it is less precise for quantitative measurements, such as finding the exact location of joint positions. The generality of estimating poses of arbitrary skeletons comes at the cost of precision – the EA algorithm poses the skeleton as close to the surface to minimize the objective function. However, for most objects, the kinematic structure often lies behind the surface captured by depth cameras. Nonetheless, the ability to get qualitative pose information from an arbitrary skeleton and depth image pair is a vital development for a variety of robotic applications.

VI. DISCUSSION AND FUTURE WORK

While the preliminary results are promising, the 23 DOF model has kinematic chains of only two joints deep. Investigating how the accuracy of evolutionary algorithm scales with more complex skeletons is essential, as deeper chains produce more local optima. Other avenues of research include inferring poses from non-isomorphic depth images and tracking kinematic information over time.

REFERENCES

[1] 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." *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[2] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenbahn, and H. Seidel, "Motion capture using joint skeleton tracking and surface estimation." *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] T. Moeslund, A. Hilton, and V. Kruger, "Survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[4] R. Poppe, "Vision-based human motion analysis: an overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[5] D. Katz, Y. Pyuro, and O. Brock, "Learning to manupulate articulated objects in unstructured environments using a grounded relational representation," *Robotics: Science and Systems Conference (RSS)*, 2008.

[6] M. Riley, A. Ude, K. Wade, and A. C. G., "Enabling real-time full-body imitation: a natural way of transferring human movement to humanoids," *International Conference on Robotics and Automation (ICRA)*, pp. 2368–2374, 2003.

[7] J. Kober and J. Peters, "Learning motor primitives for robotics." *International Conference on Robotics and Automation (ICRA)*, pp. 2112–2118, 2009.

[8] K. Shoemake, "Animating rotation with quaternion curves," *ACM SIGGRAPH Computer Graphics*, vol. 19, no. 3, 1985.

[9] M. D. Schmidt and H. Lipson, "Predicting Solution Rank to Improve Performance," *Genetic Evolutionary Computation Conference (GECCO)*, pp. 949–956, 2010.

[10] ——, "Age-Fitness Pareto Optimization," *Genetic Programming Theory and Practice*, vol. 8, pp. 129–146, 2010.

[11] M. Corp, Kinect for Xbox 360.