Unsupervised Learning of Complex Articulated Kinematic Structures combining Motion and Skeleton Information

Hyung Jin Chang, Yiannis Demiris
Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom.

Learning the underlying kinematic structure of articulated objects is an active research topic in computer vision and robotics. RGB-D sensors-based human/hand skeleton estimation methods have been successfully presented [6, 7], but the methods are designed for specific target skeletons and computationally demanding pre-training step required. Also the results are typically skeletons and not kinematic structures. Many algorithms which recover an articulated structure from 2D tracking data have shown automatic detection of articulated motion types [2, 9] and building kinematic chains [1, 5, 9], but they have been applied to relatively simple articulations only. Our target is to find a kinematic structure of arbitrary objects with highly articulated motion capabilities. Furthermore, most of the existing kinematic structure generation methods [1, 9] use motion information only. Such techniques miss global refinement steps that enforce topological or kinematic constraints, and as such can produce highly implausible structures. On the other hand, articulated structure estimation from shape [10] has been presented, but such estimation method cannot represent kinematic structures.

In this paper, we present a novel framework for complex articulated kinematic structure estimation from 2D feature points trajectories. We combine motion and skeleton information for generating elaborate and plausible kinematic structure (see Figure 1). We assume that an articulated object is composed of a set of rigid segments and the structure represents the connections between segments. The 2D feature point set is defined as $X = \{x_1, x_2, \ldots, x_N\}$ where $N$ is the total number of points, and the trajectories are represented as $x_i^f$, with $f = 1, \ldots, F$ as sequence index and $F$ as the number of frames. To express motion segments, we use $S_k$ for the disjoint set of points belonging to the $k^{th}$ segment where $k = 1, \ldots, c$, and $c$ as the total number of segments, and $y_k$ denotes a centre position of segment $S_k$ obtained by averaging its points.

It is difficult to estimate the precise number of motion segments ($c$) especially when the motions are highly articulated and the input data is noisy. In order to cope with these complicated cases, we present an iterative fine-to-coarse inference strategy with randomized voting (RV) method [3], which adaptively estimates an upper-bound number of initial motion segmentation. We also propose an adaptive object boundary ($\delta\Omega$) generation method from sparse feature points $X^f$ based on support vector data description [8] with a novel optimal kernel parameter selection method using sample margins [4].

A skeleton of an object, $Y(\Omega)$, is defined as a set of all centre points of maximal circles contained in an object $\Omega$, which is a medial axis:

$$Y(\Omega) = \{p \in \Omega | \exists q, r \in \delta\Omega, q \neq r : dist(p, q) = \lambda \cdot dist(p, r)\}. \quad (1)$$

The skeleton contains both shape features and topological structures of the original objects. Using the obtained object boundary, the distance function $(\Psi(p))$ of $\Omega$ is defined as $\Psi(p) = \min_{q \in \delta\Omega} (dist(p, q))$ for all points $p \in \Omega$.

To generate the kinematic structure, we utilise a graphical model $G = (V, E)$ to determine the topological connections between motion segments. All the motion segment centres $y_1, \ldots, y_F$ are treated as nodes $V$ in a complete graph. The proximity $E(y_k, y_l)$ between segment $y_k$ and $y_l$ is defined as

$$E(y_k, y_l) = \text{median}_{f=F} \{\|y_k^f - y_l^f\|_2 \times \|\Psi_f\| - \|y_k^f - y_l^f\|_2\} \quad (2)$$

which is a combination of geodesic distance in skeleton distance transform and moving velocity difference. Given the distance function $\Psi$, a geodesic distance between two points $p$ and $q$ is defined as follows:

$$\xi(p, q; \Psi_f) = \min_{\Gamma \in \mathcal{P}_{p,q}} \sum_{i \in \Gamma} \frac{1}{\Psi_f(p_i)} \quad (3)$$

where $\Gamma$ is a path connecting the two points and $\mathcal{P}_{p,q}$ is the set of all possible paths. Thus the Equation (3) defines the minimum distance between two points in the object region via the skeletal topology path. The proposed proximity measure separates segments that are topologically apart and move with different velocity. Two segments with small edge weight have a large possibility to be connected. We generate the graph’s minimum spanning tree as the kinematic structure of the object. However, the initially generated structure is highly contorted, because many small motion segments deviate from the median axes. So we further perform structure smoothing by an iterative merging procedure guided by the skeleton distance function.

We introduce new challenging sequences which are composed of highly articulated and concurrent motions. Our experiments show that the proposed method outperforms state-of-the-art methods quantitatively and qualitatively. While previous work needed manual interventions, we could find plausible motion parts and skeletons adaptively without tuning parameters.

Acknowledgement: This work was supported in part by the EU FP7 project WYSIWYD under Grant 612139.


