Articulation Estimation Using Depth Sensing

Suren Kumar
Mechanical and Aerospace Engineering
State University of New York at Buffalo
Buffalo, NY, USA
Email: surenkum@buffalo.edu

Vikas Dhiman
Electrical Engineering
University of Michigan
Ann Arbor, MI, USA
Email: dhiman@umich.edu

***Abstract*—**

- **Detect distinctly moving clustered points/voxels/objects in a scene.**
  - **Use Kinect fusion to create a static map.**
  - **Use some kind of noise threshold to detect object movement independent of camera movement.**
  - **Trigger algorithm (may be use RANSAC ?? etc.) that will segment out the object that just moved. The object should be spatially clustered and should be explained by the same rigid 3D motion.**
  - **Maintain a pairwise relative localization graph of the scene.**
- **Semantic reasoning in map update of these objects and their localization. Reason about Physical support and articulated linkage.**
- **Build a 3D reconstruction of these objects.**
- **Find similar unmapped static objects in the scene. May be use Jeff's detection and segmentation code.**
- **Try algorithm for long term mapping (a week) by using auto charging turtlebots in a living room and compare with existing algorithms.**

I. Introduction

Imagine a robot moving in a typical living room environment which encounters indoor objects such as doors, drawers and chairs etc. We posit that in order for the robot to understand, map or interact with such objects, the robot needs to be able to understand the articulation. Pyschophysical experiments on human motion understanding have demonstrated that human first distinguish between competing motion models (translation, rotation and expansion) and then estimate the motion conditioned on motion model [1].

II. Related Work

A. Historical Perspective

Using image motion to understand motion and structure in the scene is a historically well studied problem in computer vision. Ullman [2] proposed that in non-degenerate cases under orthographic projection, three pictures of four points can determine structure and motion. Tomasi and Kanade [3] formalized Ullmans's idea and proposed one of the influential method to compute camera motion and image structure by tracking features in the images. They proposed factorizing a matrix of feature tracks into motion and shape matrix by enforcing the rank constraint of the rigid body motion and metric constraints of a rotation matrix. Costeira and Kanade [4] extended the factorization idea to segment and recover shape along with motion of multiple moving bodies in the scene.

Yan and Pollefeys [5] further extended this rank and subspace idea to estimate kinematic chains from tracked features. There are certain fundamental limitations which is common to these approaches. First, the reliance on feature tracking methods such as KLT is not suitable for indoor environments which may not have much texture. Furthermore, this feature tracking requirement limits the application of such methods tremendously by i) Not being able to track new parts/objects that enter/exit the scene, ii) Not modelling the entire scene and as a result not exploring dependencies between neighbouring objects, iii) Restricting the ability to assimilate the motion of objects in SLAM like approaches that map the entire scene. Secondly, motion orthogonal to image plane is not modelled [5] as image projection is modelled as affine projection in the most general case.

With the discovery of cheap and commonplace hardware such as Kinect, there is a need to re-examine this structure from motion idea. First, since such hardware already provides depth for a feature point, one already has shape as estimated by traditional structure from motion. Also since depth is available, one can model the motion orthogonal to image plane. Textureless objects can be tracked by adding depth edges to the tracking mix.

[6] Build on existing work on articulation estimation. Add interactive perception where manipulation adds to perception and vice versa. [5] Problem: Analysis and reconstruction of dynamical scenes Method: Estimate the rigid motion subspaces. Extend the method to non-rigid parts by modeling it with linear combination of key shapes. Use motion segmentation (by SVD) to segment feature trajectories for each object. Use n neighbors to estimate local subspace of each point and cluster the subspaces to estimate the cluster of trajectories that form the same metric subspace.

[7] Uses RGBD

[8] [9] [10]

III. Articulation Classification

We consider the problem of motion model identification from point correspondences of motion. In the current work, we consider revolute, prismatic and general motion. Consider motion of two points x_0, x_1 (represented in an inertial frame) on a rigid body at time t_0 and at some subsequent times t_1, t_2. The most general form of rigid body motion of a point can

be represented using a rotation matrix R and an associated translation vector T

$$x_0^{t_1} = R_{t_0}^{t_1} x_0^{t_0} + T_{t_0}^{t_1} \tag{1}$$

where the superscript on the point denotes the time.

A. Prismatic

For points lying on a prismatic joint such as a drawer, rotation w.r.t inertial frame remains the same (R=I), resulting in $x_1^{t_1} - x_0^{t_1} = x_1^{t_0} - x_0^{t_0}$. This is essentially saying that the vector joining two points on a prismatic joint remains the same before and after the motion.

B. Revolute

For further distinction between revolute and general motion, we need information from more than one time step. For points lying on a body undergoing revolute motion such as a door, the points have same translation vector over time. Hence estimating the translation vector from two time steps $T_{t_0}^{t_1} = T_{t_1}^{t_2}$ is a sufficient condition to classify a joint as revolute joint.

C. Plane Constrained Motion

Plane constrained motion is useful for characterizing motion of objects like chair that can be translated on a plane and rotated about the normal to the plane. Let the plane be denoted by an point x_p lying on the plane and $\hat{n}$ being normal to that plane. Consider the case of a rigid body that has point $x_c^{t_0}$ in contact with the ground plane which after undergoing the motion moves to $x_c^{t_1}$. Since $x_c^{t_0}$ and $x_c^{t_1}$ both lie on the ground plane, we have

$$(x_c^{t_0} - x_0)^T \hat{n} = 0 \tag{2}$$

$$(x_c^{t_1} - x_0)^T \hat{n} = 0 \tag{3}$$

$$x_c^{t_1} = R_{t_0}^{t_1} x_c^{t_0} + T_{t_0}^{t_1} \tag{4}$$

By doing algebraic manipulation we get, $(R_{t_0}^{t_1} x_0 + T - x_0)^T \hat{n} = 0$

D. General Rigid Body Motion

For general motion such as a book that can be rotated and translated anywhere in the space both the rotation and translation matrix will be different.

IV. Scene Understanding

Analysis by method such as ours is essential in order to decompose the scene into types of motion that a robot can influence on the scene. For example: Understanding the way a drawer can be opened, fridge door can be opened, what can be moved around in the scene

Other important use cases of motion estimation

- Estimation a joint can induce prior over objects such as revolute joint can induce prior over refrigerator and door
- Visual object identification such as drawer can induce a prior over motion estimation
- Object tracking
- For grasping? – such as how to open a door?

V. Factorization Approach

In this section, we extend the factorization approach as described in [4] to 3-D track data available from a depth camera. Assuming for now that a single object moves relative to a static camera and we track features from frame to frame. Following the notation in the paper, lets represent a point on the object as $p_i^T = [X_i, Y_i, Z_i]^T$ in the camera frame, in the current frame f, the position of the point in homogeneous coordinates can be represented as

$$s_{fi}^C = \begin{bmatrix} p_{fi}^C \\ 1 \end{bmatrix} = \begin{bmatrix} R_f & t_f \\ 0_{1\times 3} & 1 \end{bmatrix} \begin{bmatrix} p_i \\ 1 \end{bmatrix} = \begin{bmatrix} R_f & t_f \\ 0_{1\times 3} & 1 \end{bmatrix} s_i \tag{5}$$

where R_f and t_f are the rotation and translation of the object from current frame w.r.t to the frame in which object points are initially represented. Assuming that we track N features over F frames, one can write

$$\begin{bmatrix} u_{11} & \dots & u_{1N} \\ . & & . \\ u_{F1} & \dots & u_{FN} \\ v_{11} & \dots & v_{1N} \\ . & & . \\ v_{F1} & \dots & v_{FN} \\ w_{11} & \dots & w_{1N} \\ . & & . \\ w_{F1} & \dots & w_{FN} \end{bmatrix} = \left[\begin{array}{c|c} i_1^T & t_{x_1} \\ . & . \\ i_F^T & t_{x_F} \\ j_1^T & t_{y_1} \\ . & . \\ j_F^T & t_{y_F} \\ k_1^T & t_{z_1} \\ . & . \\ k_F^T & t_{z_F} \end{array}\right] \begin{bmatrix} s_1 & . & . & . & s_N \end{bmatrix} \tag{6}$$

where (u_{fi}, v_{fi}, w_{fi}) is the location of feature point in current frame, vectors i_f^T, j_f^T, k_f^T are the rows of the rotation matrix R_f and $(t_{x_f}, t_{y_f}, t_{z_f})$ represent the components of the translation vector at time instant with frame f. Equation 6 can be represented in a accumulated form as

$$\mathbf{W} = \mathbf{MS} \tag{7}$$

where $\mathbf{W}$ represents the accumulation from trajectories of N points tracked over F frames, $\mathbf{M}$ contains all the information about the motion of the object present in the scene and $\mathbf{S}$ contains all the information about the shape of the object. Since rank of product of two matrices can not exceed the minimum of rank of individual matrics, It is clear that the maximum rank of W is 4. Computing singular value decomposition of $\mathbf{W}$, we get

$$\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^T \tag{8}$$

where $U \in R^{3F\times 4}$, and $V \in R^{N\times 4}$ are left and right real singular matrices and Σ is a 4×4 diagonal matrix of singular values. Although if $\mathbf{W}$ was full rank, we would have to consider N singular values but as the rank of W is 4, we only write out the components corresponding to first 4 singular values. Writing the factorization as product of two matrices,

$$\hat{\mathbf{M}} = \mathbf{U}\Sigma^{\frac{1}{2}}, \hat{\mathbf{S}} = \Sigma^{\frac{1}{2}}\mathbf{V}^T \tag{9}$$

A. Motion and Shape Estimation

The factorization as defined in Equation 9 is not unique as any invertible 4×4 matrix A will lead to an alternate solution $\mathbf{M} = \hat{\mathbf{M}}A$, $\mathbf{S} = A^{-1}\hat{\mathbf{S}}$

$$a = 1$$

$$\begin{bmatrix} m_{00}^2 & 2m_{00}m_{01} & 2m_{00}m_{02} & m_{01}^2 & 2m_{01}m_{02} & m_{02}^2 \\ m_{10}m_{00} & m_{10}m_{01}+m_{11}m_{00} & m_{10}m_{02}+m_{12}m_{00} & m_{11}m_{01} & m_{11}m_{02}+m_{12}m_{01} & m_{12}m_{02} \\ m_{20}m_{00} & m_{20}m_{01}+m_{21}m_{00} & m_{20}m_{02}+m_{22}m_{00} & m_{21}m_{01} & m_{21}m_{02}+m_{22}m_{01} & m_{22}m_{02} \\ m_{00}m_{10} & m_{00}m_{11}+m_{01}m_{10} & m_{00}m_{12}+m_{02}m_{10} & m_{01}m_{11} & m_{01}m_{12}+m_{02}m_{11} & m_{02}m_{12} \\ m_{10}^2 & 2m_{10}m_{11} & 2m_{10}m_{12} & m_{11}^2 & 2m_{11}m_{12} & m_{12}^2 \\ m_{20}m_{10} & m_{20}m_{11}+m_{21}m_{10} & m_{20}m_{12}+m_{22}m_{10} & m_{21}m_{11} & m_{21}m_{12}+m_{22}m_{11} & m_{22}m_{12} \\ m_{00}m_{20} & m_{00}m_{21}+m_{01}m_{20} & m_{00}m_{22}+m_{02}m_{20} & m_{01}m_{21} & m_{01}m_{22}+m_{02}m_{21} & m_{02}m_{22} \\ m_{10}m_{20} & m_{10}m_{21}+m_{11}m_{20} & m_{10}m_{22}+m_{12}m_{20} & m_{11}m_{21} & m_{11}m_{22}+m_{12}m_{21} & m_{12}m_{22} \\ m_{20}^2 & 2m_{20}m_{21} & 2m_{20}m_{22} & m_{21}^2 & 2m_{21}m_{22} & m_{22}^2 \end{bmatrix} \begin{bmatrix} a_{00} \\ a_{01} \\ a_{02} \\ a_{11} \\ a_{12} \\ a_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{10}$$

REFERENCES

[1] Shuang Wu, Hongjing Lu, and Alan L Yuille. Model selection and velocity estimation using novel priors for motion patterns. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1793–1800. Curran Associates, Inc., 2009.

[2] Shimon Ullman. *The interpretation of visual motion.* MIT Press, 1979.

[3] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[4] João Paulo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

[5] Jingyu Yan and M. Pollefeys. Automatic kinematic chain building from feature trajectories of articulated objects. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 712–719, June 2006.

[6] Roberto Martin Martin and Oliver Brock. Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 2494–2501. IEEE, 2014.

[7] Dov Katz, Moslem Kazemi, J. Andrew (Drew) Bagnell, and Anthony (Tony) Stentz. Interactive segmentation, tracking, and kinematic modeling of unknown 3d articulated objects. In *Proceedings of IEEE International Conference on Robotics and Automation*, May 2013.

[8] Jürgen Sturm, Kurt Konolige, Cyrill Stachniss, and Wolfram Burgard. 3d pose estimation, tracking and model learning of articulated objects from dense depth video using projected texture stereo. In *Robotics: science and systems*, volume 2010, 2010.

[9] Xiaoxia Huang, Ian Walker, and Stan Birchfield. Occlusion-aware reconstruction and manipulation of 3d articulated objects. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1365–1371. IEEE, 2012.

[10] Sudeep Pillai, Matthew Walter, and Seth Teller. Learning articulated motions from visual demonstration. In *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014.