A Dynamic Programming Approach for Fast and Robust Object Pose Recognition from Range Images

Christopher Zach[1], Adrian Penate-Sanchez[2], Minh-Tri Pham[1]
[1]Toshiba Research Europe Ltd, Cambridge, UK. [2]Universitat Politècnica de Catalunya, Barcelona, Spain.

Recognizing objects and estimating their poses from a depth sensor using depth data alone is more difficult than using both depth and color data [1, 3] since depth data is far less discriminative than color data in their appearance. Traditionally, this problem is approached using either global or local object representations. Global methods [6, 9] accumulate votes via a Hough transform and then select the pose with the largest number of votes. They suffer when strong occlusions are present. Local methods [2, 5, 7] detect features and obtain invariant descriptors of the regions around them. However, since depth data is usually uniformative, these features are hardly repeatable.

This paper fits into local approaches. However, we opt for a dense computation of features and descriptors in order not to rely on unstable points. As depth data are only reliable and accurate in smooth regions, we use surface points and normals as features, and sampled occupancy grids as descriptors. Rather than predicting poses directly based on feature correspondences, we follow [1, 8] in predicting "object coordinates" (i.e. 3D vertices on the object of interest) and computing more certain and accurate poses from multiple correspondences. Unlike [1, 8] that learn a random forest for prediction, we treat the object coordinate hypotheses as unknown (or latent) states and employ the methodology of inference in graphical models in order to rank the set of putative object coordinates.

In our graphical model, each pixel s in the query range image is associated with a few putative object coordinates X_s. We use the Hamming distance between the descriptor extracted at s and the ones returned by the (approximate) nearest neighbor search for X_s as unary potential $\phi_s(X_s)$. If p and q are two pixels in the query range image, and $\hat{X}_p$ and $\hat{X}_q$ are the respective back-projected 3D points induced by the observed depth, and X_p and X_q are putative correspondences reported at p and q, then a necessary condition for $\hat{X}_p \leftrightarrow X_p$, $\hat{X}_q \leftrightarrow X_q$ being inlier correspondences is that the Euclidean distance between $\hat{X}_p$ and $\hat{X}_q$ does not deviate substantially from the one between X_p and X_q. We use the deviations to play the role of pairwise potentials:

$$\psi(X_p, X_q; \hat{X}_p, \hat{X}_q) \stackrel{\text{def}}{=} \begin{cases} \Delta^2(X_p, X_q; \hat{X}_p, \hat{X}_q) & \text{if } |\Delta(X_p, X_q; \hat{X}_p, \hat{X}_q)| \le \sigma \\ \infty & \text{otherwise.} \end{cases}$$

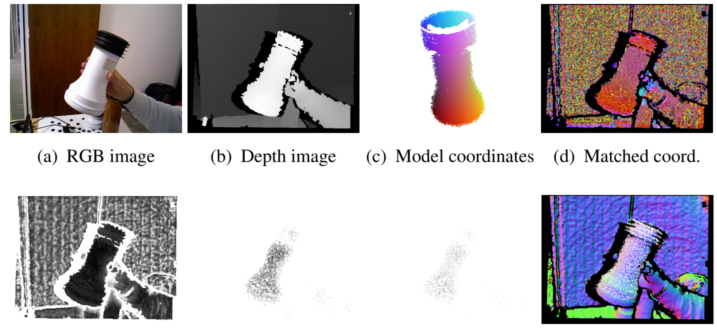
with $\Delta(X_p, X_q; \hat{X}_p, \hat{X}_q) \stackrel{\text{def}}{=} \|\hat{X}_p - \hat{X}_q\| - \|X_p - X_q\|$. σ is the maximum noise or uncertainty level expected from the depth sensor and matching procedure.

Rigid pose estimation requires at least three (non-degenerate) point-to-point correspondences via the Kabsch algorithm or Horn's method [4]. However, random sampling three putative correspondences is very inefficient, since the inlier ratio is very small. Instead, we use the graphical model to generate promising sets of three correspondences (up to 2000) by ranking minimal sample sets. We propose to compute min-marginals to quickly discard outlier contaminated minimal sample sets. Let $\{p, q, r\}$ be a set of (non-collinear) pixels in the query image, let X_s, $s \in \{p, q, r\}$ range over the putative object coordinates, then the negative log-likelihood (energy) of states (X_p, X_q, X_r) according to our graphical model is

$$E_{pqr}(X_p, X_q, X_r) \stackrel{\text{def}}{=} \phi_p(X_p) + \phi_q(X_q) + \phi_r(X_r) + \psi(X_p, X_q; \hat{X}_p, \hat{X}_q) + \psi(X_p, X_r; \hat{X}_p, \hat{X}_r).$$

reprenting a tree rooted at p. We use belief propagation on many small trees to compute min-marginals efficiently.

Like most approaches in the literature, output hypothesized poses are then refined using an ICP-like approach on a subset of model points, and evaluated using a robust fitting cost.



(a) RGB image (b) Depth image (c) Model coordinates (d) Matched coord.
(e) Feature distance (f) Self-consistency (g) Pose score (h) Overlaid model

Figure 1: Method overview: (a) input RGB image (for illustration purpose only); (b) input depth image; (c) view on the trained CAD model with color coded object coordinates; (d) best matching object coordinates for the input to illustrate the level of false positives; (e) the corresponding minimal feature distances, which also serve as unary potentials; (f) the smallest min-marginals per pixel; (g) the geometric pose scores after pose refinement; and (h) points of the model superimposed according to the best pose estimate.

Implementation of this method is described in the paper. Most steps can be trivially parallelized. Our conclusion is that the method obtained state-of-the-art detection rates with a much lower computational cost. We also do not rely on a computationally expensive training phase.

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Proc. ECCV*, volume 8690, pages 536–551, 2014.

[2] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *Proc. CVPR*, pages 998–1005, 2010.

[3] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *Proc. ICCV*, 2011.

[4] Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *JOSA A*, 4(4):629–642, 1987.

[5] Andrew E. Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):433–449, 1999.

[6] Jan Knopp, Mukta Prasad, Geert Willems, Radu Timofte, and Luc Van Gool. Hough transform and 3D SURF for robust three dimensional classification. In *Proc. ECCV*, pages 589–602, 2010.

[7] Ajmal S. Mian, Mohammed Bennamoun, and Robyn Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1584–1601, 2006.

[8] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proc. CVPR*, pages 2930–2937, 2013.

[9] Oliver Woodford, Minh-Tri Pham, Atsuto Maki, Frank Perbet, and Bjorn Stenger. Demisting the hough transform for 3d shape recognition and registration. In *Proc. BMVC*, pages 32.1–32.11, 2011.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.