On Bayesian Inference for Embodied Perception of Object Poses

Zoltán-Csaba Márton and Serkan Türker

I. Introduction and Motivation

While there exists no generally accepted theory explaining human brain functions, there are large scale projects aiming to elucidate it in the near future [1]. There are several theories on how the brain perceives and reacts to the world, and these provide promising research directions also for robotics. The one most relevant to the integration approach presented here is the Bayesian brain principle [2], which assumes a continuous update of hypotheses about the world, and correcting them through sensory information. The authors of [3] detail how ideas related to artificial neural networks might explain the real network of the brain's neurons and its capability for inference, adaptation and plasticity, using Bayesian filtering in a neurobiologically plausible way.

In this work, we explore how discrete Bayes filters (specifically histogram filters [4]) can improve perception capabilities, while holding specific benefits for robotic applications. Sensing is an important component of embodied agents as feedback about the body's and the environment's state allows for control and learning strategies. Manipulation tasks are heavily aided by object recognition, for which visual perception is probably the most important modality [5]. Conversely, perception is itself aided by locomotion and manipulation skills. Piaget discovered for example that humans "calibrate" their near and far vision by reaching and locomotion, respectively [6]. In the robotics domain as well, interaction and multiple viewpoints (spatio-temporal integration) aid model learning and recognition [7].

However, perception systems are struggling to reach the level of a two year old child [8], with the pose estimation systems for example just coming into the reach of allowing robust tool use. In [9] the careful inching forward of current robots is contrasted with a person running through a crowd (without injuring anyone), highlighting not only the agility of humans, but also the understanding of the dynamics of the surrounding that is unmatched by robotic perception. Perception, however, is identified as one of the key challenges that need to be solved generally, for a multitude of research endeavors. In contrast to perception in industrial settings, one of the impediments is, as described by [8], that vision (in an unknown and "un-cooperating" environment) is an inverse problem, where the perceived image needs to be understood without sufficient information (see Figure 1).

There have been many advances in theories of perception [11], [12] and cognition, with applications to the large goal of AI [13]. The cognitivist and emergent philosophies disagree on how high level concepts should be represented/learned, however, getting feedback from the outside world and reacting to it seems to be an important aspect of cognition [14]. At least that is how we can intuitively judge intelligence (noting however the issues brought up by Searle's Chinese room argument). The idea that a cognitive agent needs to be embodied to learn from gathered experiences has been around since Turing [15]. He proposed equipping computers "with the best sense organs that money can buy" and teaching them in order to pass his famous test [16]. Similarly, the authors in [17] and [14] argue for embodiment, and present different paradigms on how to approach the learning and grounding of new information. The common coding theory [18] argues that perception and action are treated together in the brain, one generating the other in order to achieve a desired configuration. This calls for a closer integration for perception and action than the classic perception-cognition-action loops. This research direction is starting to gain momentum both in the robotic vision and action planning domains (e.g. inspired by mirror neurons).

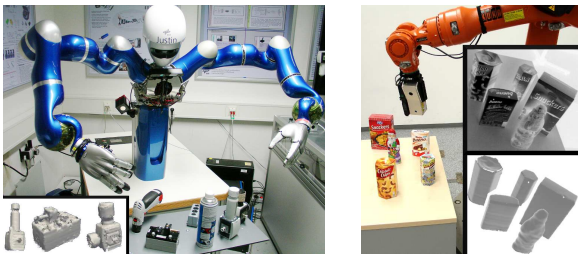


Fig. 1. Application scenarios for model-based object pose estimation in industrial and household settings. For more details on the setup see [10].

We address object pose estimation, an important prerequisite for model based robotic grasping, that uses pre-computed grasp points [19]. For an overview of the previous works in this field, please see the related work sections of [10], [20]. While in our previous work we discussed how to model objects [10], [21] and estimated the pose of previously learned objects [20], [22], here the focus is on integrating multiple views and priors about possible errors. We build on our work on merging different information sources for classification [23], [24], and multi-view/interactive recognition [10], [25], [26], extending them to pose estimation. We present the principles and the design of experiments for integrating different views, methods, error/symmetry models, priors, and enable the selection of disambiguating actions.

II. Bayesian Filtering for Pose-Estimation

Histogram filters are related to particle filters, but have a fixed set of particles that cover the whole search-space, representing discretized cells of it. Thus, they avoid the problem of particle depletion, at the cost of the rigidity of the particles, which can result in the need of a huge number of them. In the case of pose estimation, the search-space is the 6 degrees of freedom (DOF) pose space SO(3). As this is quite large, several pose estimation methods separate the translational and rotational parts, as we discussed in [20].

Here we abstract away from the pose estimation methods, and treat only their output as estimates of the pose, and integrate the rotation estimates using a histogram filter. (As the translational part is in the 3D Euclidean space, it is intuitive to deal with.) This way we can consider the

We thank Manuel Brucker, Simon Kriegel and Christian Rink. Contact: Institute of Robotics and Mechatronics, German Aerospace Center (DLR). {zoltan.marton, serkan.tuerker}@dlr.de

measurement (and movement) model in the updating step, and can consider detections from multiple viewpoints and the results of multiple detectors at once. As errors from different views have a large chance of being uncorrelated, their combination increases the overall accuracy, as we shown in [10]. Additionally, if multiple detectors are used, and the mistakes they make amongst themselves are somewhat uncorrelated, their combination should also improve performance (for example using accuracy and confidence weighted voting), as we showed in the case of classification tasks [23].

To represent the rotations, we evenly divided the space of quaternions, and selected those cells that contain unit quaternions. This does not result in a perfectly even sampling, but the area of SO(3) that falls into each cell was estimated, and used as a uniform prior. We found that a 32x32x32x16 division of the 4D quaternions with dimensions between ± 1 (considering only half of w values is enough) allows for an accuracy of 4-6 degrees, and resulted in merely 68414 cells.

To test the idea, several Asus Xtion frames were captured using an industrial robot, where ground truth object poses were known and the changes in camera positions could be accurately measured. More information on the data/method can be found in [10] – here we compared against the simple pose clustering employed there. Fig. 2 shows the results for the three industrial objects being more accurate using the histogram filter. Additionally, it offers the advantage of implicit handling of symmetries and estimator biases through an estimated error model, as discussed in the next section. To compare against a particle filter implementation, we performed 20 runs of it per object as well.

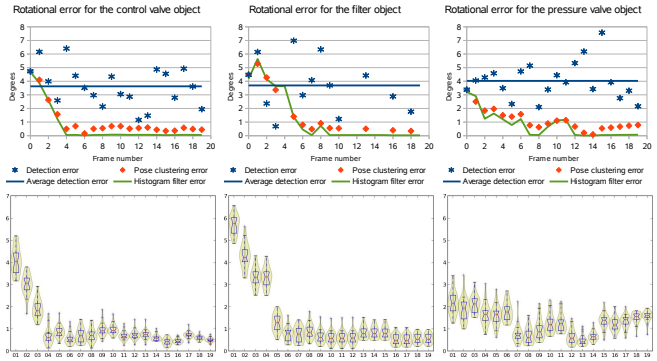


Fig. 2. *Top:* improvements in pose estimation for the three industrial objects from Fig. 1. *Bottom:* particle filter errors (frame 0 had extreme variations).

III. Error Model Estimation

As discussed earlier, results can be improved further if not the standard Gaussian measurement error model is considered, but ground truth data is used to estimate it. We currently do this by simulating scans of the objects from each pose in the histogram, and record the number of confusions by the method. However, to speed up the computations for our initial test, we perform the evaluation on a downsampled set of 8480 rotations (therefore we use a 16x16x16x8 division)[1].

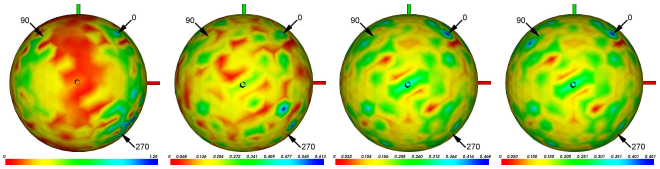
The most important question is the required number of tests needed for each of the 8480 poses, in order to guarantee a narrow confidence interval for the cells holding no detections. As we can treat the tests falling into such a cell as coming from a binomial distribution, we can compute the Jeffreys interval for them given a significance level. In this case this is equivalent to the Bayesian credible interval, and depends on the Beta distribution. The non-informative Jeffreys prior is $\texttt{Beta}(0.5, 0.5)$, which gets updated using Bayes' theorem after each trial. In the case of the Beta distribution this fortunately yields another Beta distribution: $\texttt{Beta}(0.5, n + 0.5)$, where n is the number of steps (trials). Thus, we can find the minimum number of trials, such that the confidence/credible interval for the 0 value is [0,5%] as:

$$\min\{n | n \in \mathbb{N}, \int_0^{0.05} \texttt{Beta}(0.5, n + 0.5) \geq 1 - \alpha\}$$

Since we test multiple cells for confusions (8480 or 68414), we need to perform Bonferroni correction as well. For $\alpha = 5\%$ this results in 250 trials for each of the 8480 evaluated rotations, which is manageable, and allows for mistakes of up to 6 (for 68414 possible detections) in 95% of the cases for the cells that were not hit during evaluation.

For our initial experiments we required a fast pose estimation method, s.t. we can generate the confusion matrices as quickly as possible. Therefore, we used PCL's registration module to create a RANSAC-based feature correspondence estimator (similar to [27], using SHOT feature correspondences as they worked best in [22]). Thus we can directly balance runtime and accuracy through the number of iterations (i.e. the desired probability of success). To obtain results quickly, we built the error model based on a low quality (but fast) configuration. After simulating the scans from each of the 8480 poses we used 5 as pseudocount.

First results are shown Fig. 3, visualized by the location of the $(1, 1, 1)$ vector after the detected or predicted rotation is applied to it (color coded according to the density). The detected orientations are rather bad (mostly rotated around Z), but still, the prediction is reasonable. Due to object symmetries and pose estimator bias, the distribution of the predicted orientations shows multiple peaks with 90° offset.



(a) *23 detections* (b) *after frame 0* (c) *after frame 12* (d) *after frame 23*

Fig. 3. Histogram filtering (H.F.) results using an estimated error model. An industrial filter object is used, having four similar sides (perpendicular to each other and parallel to the Z axis). Probabilities of orientations are visualized as a color coded projection to the 3D sphere – the maximum value (blue) is 0.613%, 0.468% and 0.451%, respectively in the H.F. results. Due to object symmetries and pose estimator bias, the distributions show multiple peaks (roughly 90° rotated around the Z axis) at the first and last H.F. step. The correct solution is at 0° (maxima starting with step 12).

As a conclusion, the learned error model did capture the properties of the method, but it remains to be seen how it can deal with variations that are not easy to simulate (e.g. occlusions, different surface types, etc.). The representation, however, can be used to consider object symmetries when selecting the final orientation, and different priors on the orientation can be employed (e.g. if we expected the object's pose to be physically stable and not simply uniformly distributed – according to cell area). Additionally, the estimated probability distribution and the error models can be used for selecting the best method [28], and in theory also for next-best-view selection (similarly to [29]). These capabilites are of key importance for embodied agents to reach their full potential in visual perception. and we presented a statistical analysis for their design and an experimental evaluation.

[1]We plan to capture the confusions of the 8480 rotations with the original 68414 rotations, and obtain the final error model by interpolating the results.

REFERENCES

[1] A. P. Alivisatos, M. Chun, G. M. Church, R. J. Greenspan, M. L. Roukes, and R. Yuste, "The Brain Activity Map Project and the Challenge of Functional Connectomics," *Neuron*, vol. 74, no. 6, pp. 970–974, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0896627312005181

[2] K. Doya, S. Ishii, A. Pouget, and R. P. N. Rao, *Bayesian brain: Probabilistic approaches to neural coding*. The MIT Press, 2007.

[3] K. Friston and K. Stephan, "Free energy and the brain," *Synthese*, vol. 159, no. 3, pp. 417–458, December 1 2007.

[4] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MA, USA: MIT Press, 2005.

[5] D. Lynott and L. Connell, "Modality exclusivity norms for 423 object properties," *Behavior Research Methods*, vol. 41, no. 2, pp. 558–564, 2009.

[6] B. J. Grzyb, V. Castelló, and A. P. del Pobil, "Reachable by walking: inappropriate integration of near and far space may lead to distance errors," in *Proceedings of the Post-Graduate Conference on Robotics and Development of Cognition*, J. Szufnarowska, Ed., September 2012, pp. 12–15.

[7] A. Pronobis, O. M. Mozos, B. Caputo, , and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research (IJRR)*, vol. 29, no. 2-3, pp. 298–320, Feb. 2010.

[8] R. Szeliski, *Computer Vision: Algorithms and Applications*, ser. Texts in Computer Science. New York, NY, USA: Springer-Verlag New York, Inc., 2010.

[9] M. Künzel, Ed., *Multimodale Sensorik – Konzepte zur Umwelterkennung und -modellierung*. AUTONOMIK Begleitforschung, VDI/VDE Innovation + Technik GmbH, 2012, http://www.autonomik.de/documents/Autonomik_Weissbuch.pdf.

[10] S. Kriegel, M. Brucker, Z.-C. Marton, T. Bodenmüller, and M. Suppa, "Combining Object Modeling and Recognition for Active Scene Exploration," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, Nov. 3-8 2013.

[11] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc., 1982.

[12] S. E. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, Massachusetts: MIT Press, 1999.

[13] R. Brooks, *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: The MIT Press, 1999.

[14] D. Vernon, "Cognitive vision: The case for embodied perception," in *Image and Vision Computing*. Elsevier, 2005.

[15] A. M. Turing, "Intelligent machinery," *Machine Intelligence*, vol. 5, 1970, different sources cite 1947 and 1948 as the time of writing.

[16] ——, "Computing Machinery and Intelligence," *Mind*, vol. LIX, pp. 433–460, 1950.

[17] L. Steels and R. A. Brooks, Eds., *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., 1995.

[18] R. W. Sperry, "Neurology and the mind-body problem," *American Scientist*, vol. 40, pp. 291–312, 1952.

[19] C. Ferrari and J. Canny, "Planning optimal grasps," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, May 1992.

[20] C. Rink, Z.-C. Marton, D. Seth, T. Bodenmüller, and M. Suppa, "Feature Based Particle Filter Registration of 3D Surface Models and its Application in Robotics," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, Nov. 3-8 2013.

[21] Z. C. Marton, D. Pangercic, N. Blodow, and M. Beetz, "Combined 2D-3D Categorization and Classification for Multimodal Perception Systems," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1378–1402, September 2011.

[22] A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli, and M. Vincze, "Tutorial: Point Cloud Library – Three-Dimensional Object Recognition and 6 DoF Pose Estimation," *IEEE Robotics & Automation Magazine*, vol. 19, no. 3, pp. 80–91, September 2012.

[23] Z.-C. Marton, F. Seidel, F. Balint-Benczedi, and M. Beetz, "Ensembles of Strong Learners for Multi-cue Classification," *Pattern Recognition Letters (PRL), Special Isssue on Scene Understandings and Behaviours Analysis*, 2012.

[24] O. M. Mozos, Z. C. Marton, and M. Beetz, "Furniture Models Learned from the WWW – Using Web Catalogs to Locate and Categorize Unknown Furniture Pieces in 3D Laser Scans," *IEEE Robotics & Automation Magazine*, vol. 18, no. 2, pp. 22–32, June 2011.

[25] K. Hausman, F. Balint-Benczedi, D. Pangercic, Z.-C. Marton, R. Ueda, K. Okada, and M. Beetz, "Tracking-based interactive segmentation of textureless objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, May 6–10 2013, *Best Service Robotics Paper Award finalist*.

[26] Z.-C. Marton, F. Balint-Benczedi, O. M. Mozos, D. Pangercic, and M. Beetz, "Cumulative Object Categorization in Clutter," in *2nd Workshop on Robotics in Clutter, at Robotics: Science and Systems*, 2013.

[27] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning Point Cloud Views using Persistent Feature Histograms," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, Nice, France, Sept. 22-26 2008.

[28] T. Gao and D. Koller, "Active classification based on value of classifier," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[29] F. Callari and F. P. Ferrie, "Active Object Recognition: Looking for Differences," *International Journal of Computer Vision*, vol. 43, no. 3, pp. 189–204, 2001.