Group 12: dVRK Stereo Camera Calibration and Model Registration

Seminar Paper Critical Review

Selected Paper: Azad P., Asfour T., Dillmann R. "Stereo-Based vs. Monocular 6-DoF Pose Estimation Using Point Features: A Quantitative Comparison." In: *Dillmann R., Beyerer J., Stiller C., Zöllner J.M., Gindele T. (eds) Autonome Mobile Systeme 2009.* Informatik aktuell. Springer, Berlin, Heidelberg

Project Background

The goal of our project is to accurately register the known and unknown surfaces of the phantom to the da Vinci Patient Side Manipulator (PSM) utilizing stereo camera system. To do so, first we want to commence a hand-eye calibration between the PSM and the stereo camera. Then we need to register the surfaces of the phantom to the PSM through the stereo camera system.

Once the hand-eye calibration is complete, we need to detect the phantom and derive its pose to register the surface to the PSM. For registering a known surface, we know the dimensions and the shape of the phantom model and we can use this information to determine the location and the orientation of the phantom. This process is essentially identical to pose estimation of an object of interest and pose estimation can be done using both single camera system or stereo camera system.

Paper Selection and why

I did a research on the degree of accuracy that can be achieved using stereo camera system and what kind of benefits it has over single camera system when we are using RGB images. During the research, I encountered a paper that did a quantitative comparison between the stereo-based and monocular 6 degree of freedom pose estimation and decided to review the paper titled "Stereo-Based vs. Monocular 6-DoF Pose Estimation Using Point Features: A Quantitative Comparison".

Paper Summary

This paper discusses the importance of accurate pose estimation of objects in 3D space, especially for robotic manipulation applications. It compares the two different approaches to computing a 6-DoF pose: monocular and stereo-based pose estimations. The paper presents the theoretical and practical drawbacks and the limits of monocular approaches based on 2D-3D correspondences. It also shows the experimental evaluation of both approaches. At the end, it concludes that the stereo-based approach performs superior in terms of robustness and accuracy with little additional computations.

Theoretical Accuracy Comparison

The paper compares the theoretically achievable accuracy of pose estimation methods based on 2D-3D correspondences to 3D calculations using stereo triangulation. Because this paper is for humanoid robot applications, the examples use real setup of humanoid robot

(ARMAR-III). The camera focal length is assumed to be 4 mm, resulting in approximately 530 pixels ($f = f_x = f_y$) computed by the calibration procedure. The stereo camera system has a baseline (b) of 90 mm and the principal axes are assumed to be parallel.

This paper utilizes the formula derived in other papers, one of which is to determine the relative error in the estimated z_c-coordinate from a pixel error of Δ pixels (where z_c is z-coordinate in the camera coordinate system and u is the projected size of the object) [2]:

$$\frac{z_c(u)}{z_c(u + \Delta)} - 1 = \frac{\Delta}{u} \quad (1)$$

According to this formula, the error depends on the projected size of the object. As the project size got bigger, the error got smaller. Assuming the object surface and the image plane run parallel and that the largest distance between a feature pair of an object is set as 100 mm, $u = \frac{f * x_c}{z_c} \approx 70$. A pixel error of $\Delta = 1$ would already lead to a total error of the z_c-coordinate of $75\ cm * \frac{1}{70} \approx 1\ cm$. This calculation is for perfect conditions, but in reality, the projection of the object is skewed, making the projected size smaller. A projected size of 50 pixels and an effective pixel error of $\Delta = 1.5$ results in error greater than $2\ cm$. Because the depth accuracy depends on the pixel errors in the current view and the learned view, the error is amplified.

When using a calibrated stereo camera system, the depth only depends on the current view. A relative error in the estimated z_c-coordinate of a stereo system depends on a disparity error of Δ pixels (where d is the disparity between the left and right camera image) [2]:

$$\frac{z_c(d)}{z_c(d + \Delta)} - 1 = \frac{\Delta}{d} \quad (2)$$

The error no longer depends on the projected size of the object and it depends on the disparity. The disparity is calculated to be $d = \frac{f * b}{z_c} \approx 64$. For most stereo camera setups, the correspondences between the left and right camera images can be computed with subpixel accuracy, so for this calculation $\Delta = 0.5$ is assumed, giving total error of only $75\ cm * \frac{0.5}{64} \approx 0.6\ cm$.

According to the theoretical calculations, the position accuracy achieved by stereo camera system was higher by a factor of approximately 2 ~ 3. Furthermore, the accuracy and stability is drastically lower for the 2D-3D point correspondences, which is how monocular pose estimation is done.

6-DoF Pose Estimation

The conventional approaches (monocular approach) to pose estimation are based on 2D-3D point correspondences and this cannot achieve a sufficient accuracy and robustness for grasping and other robotic manipulation applications. They often become instable when the effective resolution of the object decreases and thereby the accuracy of the 2D feature point positions decreases [2].

In the paper, an algorithm for stereo-based pose estimation is presented. The algorithm first determines the set of interest points within the given calculated 2D contour of the object in the left camera image. Then for each calculated point, find the correspondences in the right camera image along the epipolar line. 3D point for each correspondence is calculated. Lastly, one could fit a 3D model of the object into the calculated 3D point cloud, deriving the rotation and the translation matrices.

The paper suggests two variants of the last step of the algorithm: one that fits a formulated 3D representation (a geometric 3D model) of an object into the derived point cloud and one that performs an alignment based on 3D-3D point correspondences.

Experimental Evaluation

In the paper, the accuracies of the monocular and stereo-based pose estimation are compared in several experiments. For the first experiment, the authors used a recognition and 2D localization sequences presented in [3], which included SIFT feature descriptors, Hough transform, RANSAC, and least squares homography estimations.

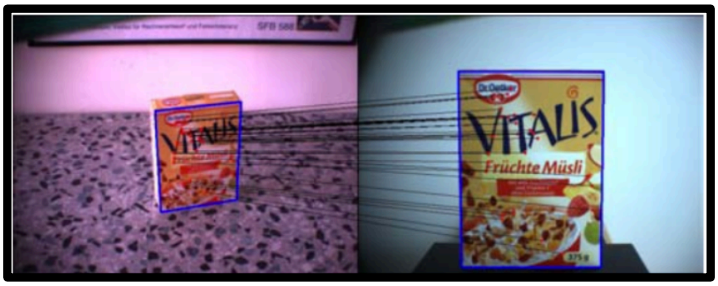

Figure 1 – Correspondences / SIFT feature matching. Blue box illustrates the result of 2D localizations (Hough Transform, RANSAC, Homography Estimation process). Right: training image; Left: input image

First experiments simulated the wide angle stereo camera pair of the humanoid robot (ARMAR-III) to measure the estimation errors under ideal conditions and having accurate ground truth information. The errors of the z-coordinate can be found in Figure 2. In addition, 1000 random poses were evaluated and the results are shown in Figure 3.

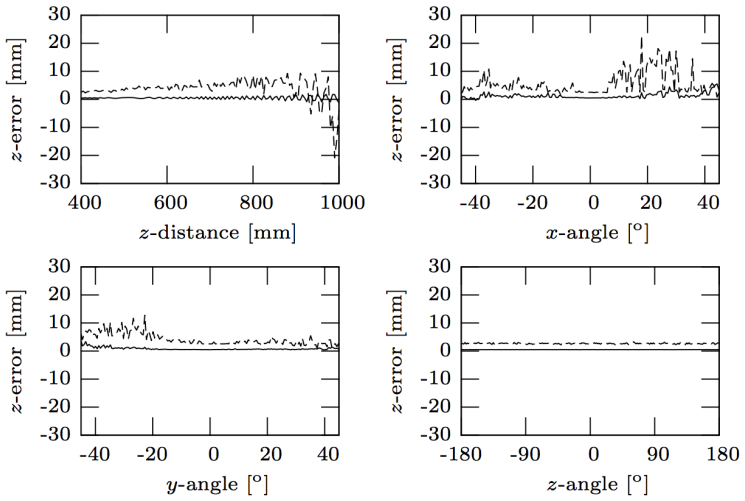

Figure 2 – Errors of the z-coordinate. Results of the simulation experiments. The solid line indicates the result of the proposed stereo-method and the dashed line the result of monocular pose estimation.

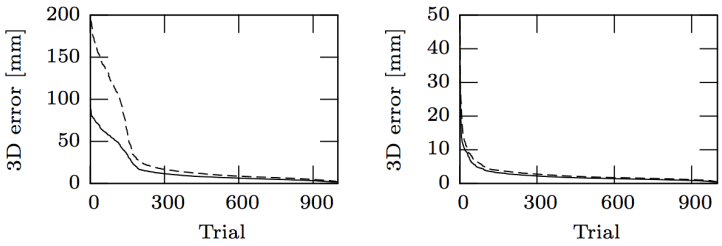

Figure 3 – Accuracy of 6-DoF pose estimation for 1000 random trials. The errors are sorted in decreasing order. The solid line indicates the average error, the dashed line the maximum error. The 3D error was measured on the basis of sampled 3D surface points. Left: monocular method; Right: proposed stereo-based method

Figure 4 – Result of 6-DoF pose estimation. Left: monocular approach (when it is instable); Right: Stereo-based approach.

	x	y	z	θ_x	θ_y	θ_z
Proposed method	0.23	0.42	0.39	0.066	0.17	0.10
Conventional method	0.24	0.038	1.52	0.17	0.29	0.13

Table 1 – Standard deviation for the estimated pose of a static object (calculated for 100 frames). Units in [mm] and [degrees]. Chose situation when monocular approach does not become instable.

The standard deviation of the z-coordinate amounts to 1.52 mm for the monocular approach and only 0.39 mm for the stereo-based approach.

The runtime of the stereo approach was 6ms for a single object.

Assessment

The paper provides a detailed analysis and comparison between the monocular and stereo-based pose estimation. It shows how the stereo-based approach is significantly more robust and more accurate. The results clearly showed the higher stability and significantly lower error values for the stereo-based method, especially in the z-coordinates. We can be assured that using stereo camera system will provide multiple benefits such as higher accuracy and higher stability in the results.

The paper went in depth about the different kinds of toolkits (such as Integrating Vision Toolkits) they used to derive the optimal results, making it easier for the readers to assess the validity of the work and, if needed, recreate the experiments. The references and sources that the authors use also have detailed evaluations of the algorithms and toolkits they used in their experiments, allowing the readers to know which tools to choose for better efficiency (for example, in [3], it compares the computation time of the Harris corner detector given by Integrating Vision Toolkit and OpenCV 1.0 and suggests IVT since its computation time is 10ms whereas OpenCV takes 17ms. Highly optimized implementations by keyetech can achieve Harris corner detection within 5ms).

Only thing I wonder is the breadth of the poses taken when the 1000 random poses were evaluated. There is no information about which degree of freedom and how many DoF were manipulated and to what extent the DoF were moved around.

Reference:

[1] Azad P., Asfour T., Dillmann R. “Stereo-Based vs. Monocular 6-DoF Pose Estimation Using Point Features: A Quantitative Comparison.” In: *Dillmann R., Beyerer J., Stiller C., Zöllner J.M., Gindele T. (eds) Autonome Mobile Systeme 2009.* Informatik aktuell. Springer, Berlin, Heidelberg

[2] P. Azad. *Visual Perception for Manipulation and Imitation in Humanoid Robots.* PhD thesis, Universit ̈at Karlsruhe (TH), Karlsruhe, Germany, 2008.

[3] P. Azad, T. Asfour, and R. Dillmann. "Combining Harris Interest Points and the SIFT Descriptor for Fast Scale-Invariant Object Recognition.” In *IEEE/RSJ In- ternational Conference on Intelligent Robots and Systems (IROS),* St. Louis, USA, 2009.