Stereo-based vs. Monocular 6-DoF Pose Estimation using Point Features: A Quantitative Comparison

Pedram Azad, Tamim Asfour, Rüdiger Dillmann

Institute for Anthropomatics, University of Karlsruhe, Germany

Abstract. In the recent past, object recognition and localization based on correspondences of local point features in 2D views has become very popular in the robotics community. For grasping and manipulation with robotic systems, in addition accurate 6-DoF pose estimation of the object of interest is necessary. Now there are two substantially different approaches to computing a 6-DoF pose: monocular and stereo-based. In this paper we show the theoretical and practical drawbacks and limits of monocular approaches based on 2D-3D correspondences. We will then present our stereo-based approach and compare the results to the conventional monocular approach in an experimental evaluation. As will be shown, our stereo-based approach performs superior in terms of robustness and accuracy, with only few additional computational effort.

1 Introduction

Accurate pose estimation of objects in 3D space is an important computer vision task, in particular for robotic manipulation applications. For a successful grasp, in particular accurate estimation of the depth is crucial. In the recent past, the recognition and pose estimation of objects based on local point features has become a widely accepted and utilized method. The most popular features are currently the SIFT features [10]; followed by the more recent SURF features [5], and region-based features such as the MSER [13]. Object recognition frameworks using such features usually operate on a set of computed feature correspondences, either by simply counting feature correspondences or by also exploiting the spatial relationships of the feature points, as proposed in [10].

Operating on the 2D localization result of such a framework, the common approach for 6-DoF pose estimation of objects computes the rotation and translation of the object in 3D space on the basis of 2D-3D point correspondences. The traditional method for this is the POSIT algorithm [7]. A more recent method for estimating a 6-DoF pose on the basis of 2D-3D point correspondences, which also succeeds for coplanar point sets, is presented in [11], and was used throughout our comparative experiments. Monocular approaches are popular for augmented reality applications as presented e.g. in [9].

Such methods all have in common that the full pose of the object is computed on the basis of a monocular image. This means that in particular the

distance of the object to the camera, namely the z-coordinate in the camera coordinate system, is derived from the scaling i.e. the size of the object in the image. Furthermore, the computation of out-of-plane rotations on the basis of 2D-3D correspondences is sensitive to small errors in the 2D feature positions.

One possibility for improving the accuracy of a pose estimate is the application of an edge-based optimization step exploiting the projected contour of the object. Such an optimization utilizes essentially the same methods as applied for edge-based rigid object tracking (e.g. [12]). A hybrid approach fusing texture, edge, and color information in an Iterated Extended Kalman Filter (IEKF) is proposed in [14]. However, an edge-based improvement of an object pose estimate always requires the projected contour of the object to be prominent in the image, which is often not the case.

In order to overcome the abovementioned problems, we have developed an approach that exploits the benefits offered by a calibrated stereo system, operating on top of the 2D recognition and localization result based on feature correspondences, as introduced in [3]. In this paper, we will show in theory and in practice that our stereo-based approach is more robust and more accurate compared to conventional monocular approaches based on 2D-3D point correspondences. Note that it is neither an accepted fact nor obvious that monocular pose estimation based on 2D-3D point correspondences performs inferior compared to stereo-based pose estimation[1].

In [6], related work on 3D object tracking is presented, which uses the KLT tracker [15] for tracking features in order to save computation time. A monocular approach using 2D-3D point correspondences and a stereo-based pose estimation method using 3D-3D point correspondences are presented. Although it is experimentally shown that the stereo-based approach is more accurate, a thorough analysis is not performed. Our measurements will show, when monocular and stereo-based pose estimation achieve comparable results and when the monocular approach deteriorates. In particular, we will show that the monocular approach suffers from instabilities for planar objects in the presence of skew and that the stereo-based approach achieves a significantly greater depth accuracy at far distances of the object. Furthermore, our stereo-based approach achieves maximum accuracy by exploiting model fitting rather than relying on point correspondences only.

In Section 2, the maximally achievable accuracy of monocular and stereo-based pose estimation is compared in theory. Our stereo-based 6-DoF pose estimation method is explained in detail in Section 3. The two approaches are compared in simulation and in real-world experiments in Section 4, ending with a conclusion in Section 5.

[1] Also note that not any stereo-based approach performs superior. For instance, performing 2D localization in the left and right camera image independently and then fusing the results is a – theoretically and practically – suboptimal approach in terms of accuracy.

2 Accuracy Considerations

In this section, the theoretically achievable accuracy of pose estimation methods based on 2D-3D correspondences will be compared to 3D calculations using stereo triangulation. As an example, values from a real setup on the humanoid robot ARMAR-III [1] are used. The task of localizing an object at a manipulation distance of approx. 75 cm for subsequent grasping is considered. Lenses with a focal length of 4 mm are assumed, resulting in approx. $f = f_x = f_y = 530$ (pixels) computed by the calibration procedure. The stereo system has a baseline of $b = 90$ mm; the principal axes of the cameras are assumed to run parallel.

As shown in [2], a pixel error of Δ pixels leads to a relative error in the estimated z_c-coordinate of:

$$\frac{z_c(u)}{z_c(u+\Delta)} - 1 = \frac{\Delta}{u}\,. \tag{1}$$

This shows that the error depends – in addition to the pixel error – on the projected size of the object: The greater the projected size u, the smaller the error. For the calculation of the pose on the basis of feature points, u is related to the farthest distance of two feature points in the optimal case. For an object whose feature pair with the farthest distance has a distance of 100 mm, it is $u = \frac{f \cdot x_c}{z_c} \approx 70$, assuming the object surface and the image plane run parallel. A pixel error of $\Delta = 1$ would already lead to a total error of the z_c-coordinate of $75\,\text{cm} \cdot \frac{1}{70} \approx 1\,\text{cm}$ under in other respects perfect conditions.

In a realistic scenario, however, objects often exhibit out-of-plane rotations, leading to a skewed image. This skew not only causes a smaller projected size of the object but also a greater error of the feature point positions. A projected size of 50 pixels and an effective pixel error of $\Delta = 1.5$ would already lead to an error greater than 2 cm. Note that the depth accuracy not only depends on the pixel errors in the current view, but also in the learned view, since the depth is estimated relative to the learned view.

In contrast, when exploiting a calibrated stereo system, the depth is computed on the basis of the current view only. As shown in [2], a disparity error of Δ pixels leads to a relative error in the estimated z_c-coordinate of:

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

where d denotes the disparity between the left and right camera image. Eq. (2) shows that the error does not depend on the projected size of the object, as it is the case in Eq. (1), but instead depends on the disparity d: The greater the disparity, the smaller the error. For the specified setup, the disparity amounts to $d = \frac{f \cdot b}{z_c} \approx 64$. For typical stereo camera setups, the correspondences between the left and the right camera image for distinctive feature points can be computed with subpixel accuracy. For this, usually a second order parabola is fitted to the measured disparity/correlation pairs and the two neighbors. In practice, a subpixel accuracy of at least 0.5 pixels is achieved easily by this approach. Together with Eq. (2) this leads to a total error of only $75\,\text{cm} \cdot \frac{0.5}{64} \approx 0.6\,\text{cm}$.

Judging from the presented theoretical calculations, the position accuracy that can be achieved by stereo vision is higher by a factor of approx. 2–3. Although for fine manipulation of objects, e.g. grasping the handle of a cup, the lower estimated accuracy of methods relying on 2D-3D correspondences is problematic, for many other applications it might be sufficient.

However, the real errors arising from pose estimation on the basis of 2D-3D point correspondences can hardly be expressed by theoretic formulas. The accuracy and stability of such approaches dramatically depends on the spatial distribution of the feature points and their accuracy.

3 6-DoF Pose Estimation

Conventional approaches to 6-DoF pose estimation, which are based on 2D-3D point correspondences, cannot achieve a sufficient accuracy and robustness. In particular, they tend to become instable when the effective resolution of the object decreases and thereby also the accuracy of the 2D feature point positions (see Section 4 and [2]). In this section, we present our approach, which exploits the benefits of a calibrated stereo system. As will be shown, this leads to a significantly higher robustness and accuracy, and succeeds also at lower scales of the object.

The idea is to compute a sparse 3D point cloud for the 2D area that is defined by the transformation of the contour in the training view by means of the homography estimated by the 2D recognition and localization procedure (see [4,2]). Given a 3D model of the object, this model can be fitted into the calculated point cloud, resulting in a 6-DoF pose. The general approach is summarized in Algorithm 1.

Algorithm 1 CalculatePoseTextured(I_l, I_r, C) $\rightarrow R, \boldsymbol{t}$

1. Determine the set of interest points within the calculated 2D contour C of the object in the left camera image I_l.
2. For each calculated point, determine a correspondence in the right camera image I_r by computing the *Zero Normalized Cross Correlation* (ZNCC) along the epipolar line.
3. Calculate a 3D point for each correspondence.
4. Fit a 3D model of the object into the calculated 3D point cloud and return the resulting rotation R and the translation $\boldsymbol{t}$.

Essentially, two variants of Step 4 in Algorithm 1 are possible: Fit an analytically formulated 3D representation (or a high-quality mesh) of an object into the point cloud, or perform an alignment based on 3D-3D point correspondences. For applying the first variant, the object or a substantial part of the object, respectively, must be represented as a geometric 3D model.

For applying the second variant, 3D points must be calculated for the feature points from the training view in the same manner as throughout recognition, i.e. by computing stereo correspondences and applying stereo triangulation. A set of 3D-3D point correspondences is then automatically given by the filtered set of 2D-2D point correspondences resulting from the homography estimation. If applicable, the first variant should be preferred, since it does not rely on the accuracy of the feature point positions. However, even the second variant is expected to be more robust and more accurate than the conventional approach, since it does not suffer from the instabilities that are typical for pose estimation based on 2D-3D point correspondences.

In the case of cuboids – as used throughout the comparative experiments from Section 4 – a 3D plane is fitted into the sparse 3D point cloud, which is obtained by computing stereo correspondences for the interest points within the 2D contour of the object in the left camera image (details are given in [2]). Then, the intersection of the 3D plane with the estimated 2D contour is calculated to obtain 3D contour points. In the case of a cuboid, this can be easily achieved by intersecting the 3D straight lines through the corner points and the projection center of the left camera with the computed plane. To provide the result as a rigid body transformation, finally the rotation R and the translation $\boldsymbol{t}$ must be computed that transform the points of the 3D object model from the object coordinate system to the world coordinate system. When using 3D-3D point correspondences without fitting a 3D primitive, this transformation is calculated automatically. Otherwise, the searched rigid body transformation can be computed on the basis of 3D-3D point correspondences between the calculated 3D contour points and the corresponding 3D model points. For this, the minimization method from [8] is used. In the case of a rectangular planar surface, it is sufficient to use the four corner points.

4 Experimental Evaluation

In this section, the accuracies of monocular and stereo-based pose estimation are compared in several experiments. For recognition and 2D localization, the features and the recognition pipeline presented in [4] were used. The system was implemented using the Integrating Vision Toolkit (IVT)[2]. The company keyetech[3] offers highly optimized implementations (e.g. Harris corner detection within less than 5 ms).

In the first experiments, the wide angle stereo camera pair of the humanoid robot ARMAR-III (as specified in Section 2) was simulated, allowing to measure the estimation errors under perfect conditions and having accurate ground truth information. The errors of the z-coordinate are shown in Fig. 1, as they show the weak points of the monocular approach. For this purpose, the object of interest was moved along (resp. rotated around) a single degree of freedom for each plot. In addition, 1,000 random poses were evaluated; the results are shown in Fig. 2.

[2] `http://ivt.sourceforge.net`

[3] `http://www.keyetech.de`

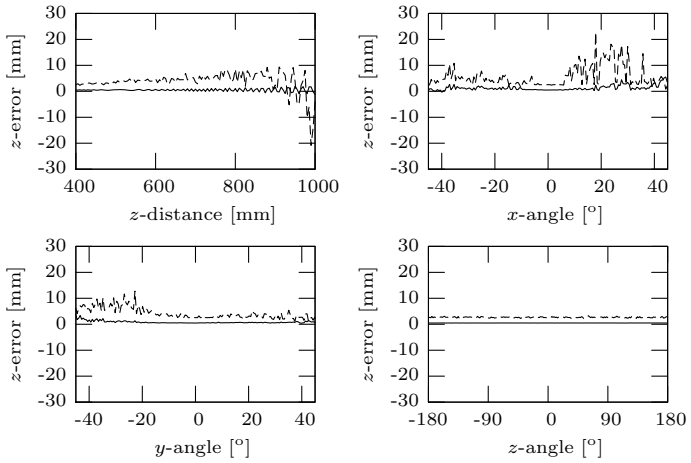

Fig. 1. Results of the simulation experiments. The z-errors are plotted depending on changes in a single degree of freedom. The solid line indicates the result of the proposed method, the dashed line the result of monocular pose estimation.

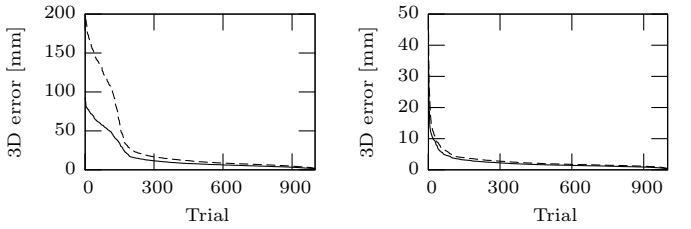

Fig. 2. Accuracy of 6-DoF pose estimation for 1,000 random trials; the errors are sorted in decreasing order. The solid line indicates the average error, the dashed line the maximum error. The 3D error was measured on the basis of sampled 3D surface points. Left: using the monocular method. Right: using the proposed stereo-based method. Note the different scaling of the vertical axis.

In Fig. 3, a situation is shown in which the monocular approach becomes instable. In Table 1, the standard deviations over 100 frames for experiments with a static object are given. As can be seen, the standard deviation of the z-coordinate amounts to 1.52 mm using the monocular approach, in contrast to 0.39 mm when using the stereo-based approach.

The runtime of the 6-DoF pose estimation procedure amounts to approx. 6 ms for a single object using the specified stereo setup (3 GHz single core CPU). The only computational expensive task here is the correlation procedure for the interest points belonging to the object. The runtime can be reduced easily by taking into account the correlation results of neighbored interest points.

Fig. 3. Result of 6-DoF pose estimation. Left: using the conventional monocular approach. Right: using the proposed 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 deviations for the estimated poses of a static object. The standard deviations have been calculated for 100 frames. The units are [mm] and [°], respectively. Note that a situation was chosen in which the monocular approach does *not* become instable.

5 Discussion and Outlook

We have compared monocular 6-DoF pose estimation based on 2D-3D point correspondences to our stereo-based approach. After discussing both approaches, it was shown that our stereo-based approach is significantly more robust and more accurate. The greatest deviations between the two approaches could be observed in the z-coordinate.

Various grasping experiments with the humanoid robot ARMAR-III have proved the applicability of our stereo-based pose estimation method [16].

In the near future, we plan to investigate the performance of our stereo-based approach for objects of arbitrary shape, in particular evaluating the improvement that can be achieved by the fitting of 3D primitives in addition to pose estimation using 3D-3D point correspondences.

Acknowledgment

The work described in this paper was partially conducted within the EU Cognitive Systems projects PACO-PLUS (IST-FP6-IP-027657) and GRASP (IST-FP7-IP-215821) funded by the European Commission, and the German Humanoid Research project SFB588 funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft).

References

1. T. Asfour, K. Regenstein, P. Azad, J. Schröder, N. Vahrenkamp, and R. Dillmann. ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 169–175, Genova, Italy, 2006.
2. P. Azad. *Visual Perception for Manipulation and Imitation in Humanoid Robots*. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2008.
3. P. Azad, T. Asfour, and R. Dillmann. Stereo-based 6D Object Localization for Grasping with Humanoid Robot Systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 919–924, San Diego, USA, 2007.
4. P. Azad, T. Asfour, and R. Dillmann. Combining Harris Interest Points and the SIFT Descriptor for Fast Scale-Invariant Object Recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, St. Louis, USA, 2009.
5. H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision (ECCV)*, pages 404–417, Graz, Austria, 2006.
6. C. Choi, S.-M. Baek, and S. Lee. Real-time 3D Object Pose Estimation and Tracking for Natural Landmark Based Visual Servo. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3983–3989, Nice, France, 2008.
7. D. F. DeMenthon and L. S. Davis. Model-Based Object Pose in 25 Lines of Code. In *European Conference on Computer Vision (ECCV)*, pages 123–141, Santa Margherita Ligure, Italy, 1992.
8. B. K. P. Horn. Closed-form Solution of Absolute Orientation using Unit Quaternions. *Journal of the Optical Society of America*, 4(4):629–642, 1987.
9. V. Lepetit, L. Vacchetti, D. Thalmann, and P. Fua. Fully Automated and Stable Registration for Augmented Reality Applications. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 93–102, Tokyo, Japan, 2003.
10. D. G. Lowe. Object Recognition from Local Scale-Invariant Features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1517, Kerkyra, Greece, 1999.
11. C.-P. Lu, G. D. Hager, and E. Mjolsness. Fast and Globally Convergent Pose Estimation from Video Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(6):610–622, 2000.
12. E. Marchand, P. Bouthemy, F. Chaumette, and V. Moreau. Robust Real-Time Visual Tracking using a 2D-3D Model-based Approach. In *IEEE International Conference on Computer Vision (ICCV)*, pages 262–268, Kerkyra, Greece, 1999.
13. J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *British Machine Vision Conference (BMVC)*, volume 1, pages 384–393, London, UK, 2002.
14. G. Taylor and L. Kleeman. Fusion of Multimodal Visual Cues for Model-Based Object Tracking. In *Australasian Conference on Robotics and Automation (ACRA)*, Brisbane, Australia, 2003.
15. C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, USA, 1991.
16. N. Vahrenkamp, S. Wieland, P. Azad, D. Gonzalez, T. Asfour, and R. Dillmann. Visual Servoing for Humanoid Grasping and Manipulation Tasks. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Daejeon, Korea, 2008.