Robust Statistics for 3D Object Tracking

Peter Preisig
Institute of Robotics and Intelligent Systems (IRIS)
Swiss Federal Institute of Technology, ETH
Zurich, Switzerland

Danica Kragic
Computational Vision and Active Perception
Centre for Autonomous Systems
CSC, KTH, Stockholm, Sweden

***Abstract*— This paper focuses on methods that enhance performance of a model based 3D object tracking system. Three statistical methods and an improved edge detector are discussed and compared. The evaluation is performed on a number of characteristic sequences incorporating shift, rotation, texture, weak illumination and occlusion. Considering the deviations of the pose parameters from ground truth, it is shown that improving the measurements' accuracy in the detection step yields better results than improving contaminated measurements with statistical means.**

I. Introduction

Methods for 3D object tracking are frequently used in robot-control, augmented reality and medical applications [2], [4], [5], [6], [7], [8]. Different approaches dealing with sensitivity to occlusion and noise have been presented [1], [3], [9], [10], [11] but no extensive evaluation of their efficiency is given.

Although there have been examples of appearance based pose tracking systems [12], most of the current tracking systems are based on wireframe models. One of the first such systems, RAPID [13], represents a 3D object as a set of control points which lie on the high contrast edges. An extension to this work has been presented in [1] where robust improvements have been considered. In particular, RANSAC [14] was used to detect outliers among the edges detected in each control points. It was concluded that the use of a robust estimator improved the overall results but the evaluation considered objects with with very simple texture. In [16], an iterative minimization is used to find the pose transformation that aligns a set of lines and ellipses, after which the pose is integrated using a Kalman filter. However, no extensive performance evaluation is given.

Similarly to the above, we consider only object boundaries in the tracking process. Hence, a wireframe model of the object is constructed and used in each frame to estimate the relative motion of the object. The main objective of this work is to enhance the robustness of our tracking system by evaluating both statistical and image filtering approaches in a number of ambient conditions that typically degrade performance, e.g. partial and heavy occlusion, photometric changes, incorrect matching, textured background. Our tracking system achieves robustness using a large number of measurements for a single pose hypothesis. It relies on a motion model to account for pose change between consecutive frames using normal displacements. We show that there are a number of improvements that can be applied to the original tracking system and show system's performance regarding these. The particular contribution of the paper is the evaluation of both statistical and image filtering techniques in a number of challenging settings.

The tracking approach follows the ideas proposed in [3]. It relies on the estimation of normal flow and preserves the rigid structure of the object. The positions of features are represented both in the image plane as well as in the 3D space. When the new estimate of the object's pose is available, the visibility of each edge feature is determined and internal camera parameters are used to project the model of the object onto the image plane. In points along visible edges, the perpendicular distance to the nearby edge is determined using a one-dimensional search.

II. Robust Improvements

The robustness of a tracking system depends on the quality of the detection- and the pose estimation steps. Hence, we have considered methods that improve these two steps. Shortly, the incremental change in pose between two consecutive frames is estimated by minimizing the squared error between the transformed feature positions and the actual feature positions, $\mathbf{d}$ as follows as proposed in [3]:

$$\begin{aligned} \mathbf{O}_i &= \sum_p \mathbf{d}^p (\mathbf{L}_i^p \cdot \mathbf{n}^p) \\ \mathbf{C}_{ij} &= \sum_p (\mathbf{L}_i^p \cdot \mathbf{n}^p)(\mathbf{L}_j^p \cdot \mathbf{n}^p) \\ \alpha_i &= \mathbf{C}_{ij}^{-1} \mathbf{O}_j \end{aligned} \tag{1}$$

$\mathbf{L}_i^p \cdot \mathbf{n}$ represents the magnitude of the edge normal motion that is observed at each node and α_i represents the quantity for each Euclidean motion. This least-squares estimation is vulnerable to instabilities caused by the presence of outliers. Two standard techniques used in such cases are RANSAC [14] and M-Estimators, [18]. The 1D-search algorithm used for normal flow estimation is computationally efficient but has the disadvantages that, due to the discrete nature of images, the stepping is performed in a limited number of directions. This problem can be eliminated by using higher dimensional and orientation dependent filters.

Therefore, the following improvements to the original algorithm have been evaluated:

- **M-Estimators** weight measurements depending on parameters taken from the whole sample. As M-estimates or influence functions several different functions with corresponding weights have been proposed such as Metric trimming, Metric Winsorizing, Tukey's biweight or Iterative Reweighted Least Squares (IRLS). Here, M-estimators are implemented by modifying the LS algo-

rithm and introducing the corresponding weighting factor in (Eq.1).

- **RANSAC** was first introduced in [14]. It repeatedly chooses subsets of random samples, uses these to estimate model parameters and then evaluates the parameters using the complete data set.
- **Improved edge detection** have been considered compared to the normal flow estimates in four major directions as originally proposed. Here, gradient masks are calculated dependent on the orientation of the estimated edges.

III. Experimental Evaluation

The performance of the algorithms was evaluated through a number of valid measurements in a benchmarking sequence incorporating textured background, shift, rotation and occlusion. The number of valid measurements represents the quotient between the number of successful searches for a derivative belonging to an edge in every frame and the number of totally performed searches. In the presented plots, we denote this number as the signal-to-noise ratio. It is an efficient measure that gives an idea of the behavior of the algorithm, i.e. a sudden drop in the ratio can be interpreted as loss of track whereas a high ratio stands for smooth and constant tracking. Tracking performance considering all six pose parameters have been evaluated.

For the chosen benchmarking sequence two situations are characteristic, Fig.1. Around frame 378 the hand acts as strong edge versus the bright background and almost all algorithms miss-interpret it as an object edge. In the signal-to-noise plots this can be recognized by a slight drop in the ratio. The second and more obvious incident occurs around frame 432. Due to weak illumination and rich texture of the tracked object the foremost edge is nearly invisible. The right back edge of the object is at the same time aligned with the front one and due to rotation also shifts into its search region. The right edge is strong, as it appears against the bright background, and most algorithms lose track of the real front edge and mistake it with the right edge. This means that the front, back and upper right edge of the model are coplanar. As the object continues to move, these edges tend to stick to the static and even stronger edges of the chair in the background and cause the tracker to break down. The break-down is clearly visible as a big drop of the ratio, Fig. 2.

A. Original version

Regarding the original implementation, as presented in [3], the performance mainly depends on the length of the search region *s* and the value of the derivative threshold used to detect edges, Fig. 2. It can be seen that a lower threshold as well as a larger search region result generally in a higher signal-to-noise ratio. All algorithms break down around frame 468 but the algorithms with threshold 13 and search length above 6 do not miss-interpret the hand as an edge (frame 378). Therefore, parameters Th=13 and s=8 have been selected in further evaluation.

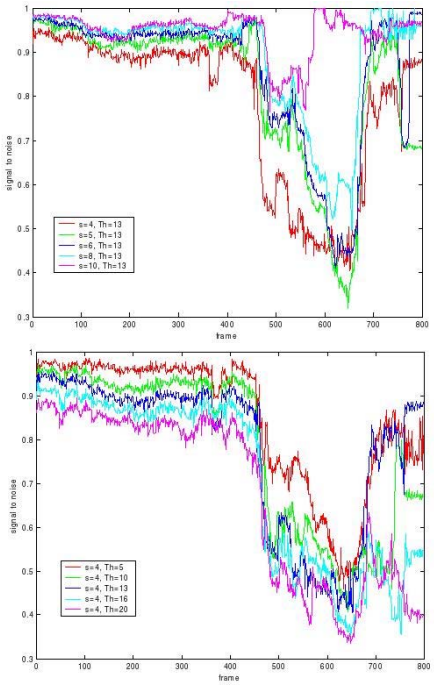

Fig. 2. Signal-to-noise ratio for the original algorithm: left) varying the size of the search region, and right) varying the derivative threshold.

B. M-estimators

We have implemented and evaluated the performance using four different M-estimators: Metric trimming, Metric Winsorizing, Tukey's biweight or Iterative Reweighted Least Squares (IRLS), [18]. Out of all four approaches, Tukey estimator yielded best results and only this one has been considered for further investigation. The two important parameters here are search length, s and the influence function parameter, L. For comparison, the values of the original method are also shown Fig.3. The two plots show that the results obtained with the estimator do not differ much from the original values. This is since the limitation of the search path already acts as a metric-trimming estimator. Implementations with a search length s=8 can cope better with the problem of the hand while the influence of the parameter R cannot be clearly seen from the plots, although it exists. The algorithm performed smoothest with parameter R=2. Therefore, R=2 and s=8 have been chosen for further evaluation.

C. RANSAC

We have evaluated RANSAC using three different approaches. In the first approach, random subsets of measurements were generated from only three independent (orthogonal) edges. The goal here was to minimize the random sample size. However, results from this approach were highly unstable since not all the motion directions were explicitly modeled. The model stretched in these directions and caused the tracker to break down. To stabilize the situation, a fourth edge have been incorporated into the former subset. This improved the performance of the tracker and it indicated that the subset

Fig. 1. Example frames from the benchmarking sequence: frames 90, 126, 198, 324, 378, 432, 468, 594

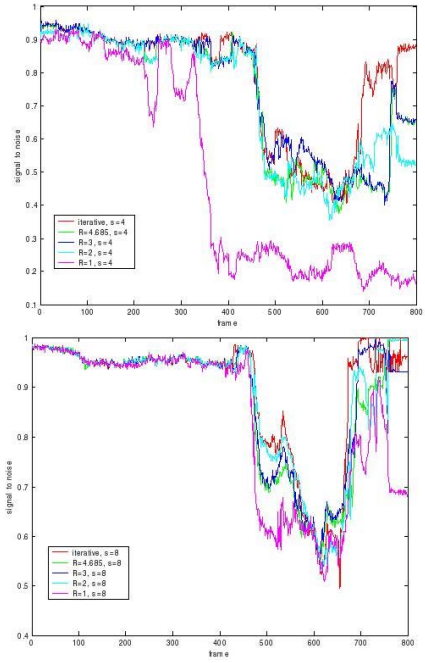

Fig. 3. Signal-to-noise ratio for the M-estimator based algorithm: up) varying the size of the search region, and down) varying the value of the influence function. "Iterative" stands for the original algorithm.

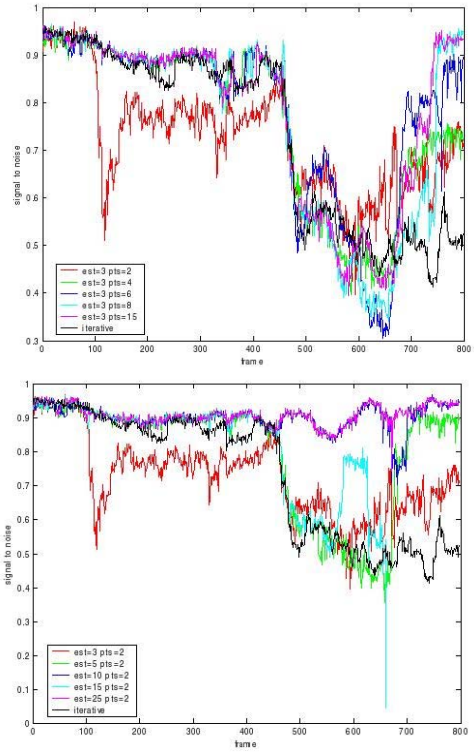

Fig. 4. Signal-to-noise ratio for the RANSAC based algorithm: up) varying the number of points per edge, and down) varying the number of estimation steps.

should be taken from all available edges to maximize the stability.

Given random subsets, a new projective matrix was computed and its accuracy was evaluated on the whole data set via a cost function. As the cost function, the total deviation of the new projected model to the available object features has been taken. For testing purposes the algorithm has been executed a predefined number of times and the best pose has been selected. In the parameter optimization two parameters have been considered. The first one was the number of points taken per visible edge (pts) and the second one the number of estimation steps (est), Fig.4. The plots display that the variation of the number of points per edge has small influence on the performance. For all values of pts the results are comparable to the ones yielded by the original algorithm. The hand is interpreted as an edge (frame 378) and the tracker fails around frame 468. In contrast, the number of estimation steps is of significant importance. From the right plot it can be seen that higher estimation numbers (est=10 and est=25) allow the algorithm to cope with both problems and lead to increased stability of the tracker. For further evaluation, the parameters pts=2 and est=10 have been chosen.

D. Edge detection

From the original code, Roberts edge detector was exchanged with Kirsch detector. Using the current pose of the object, the new masks were computed and used in the detection of the displacements between consecutive frames. The computation of the displacements was again performed with a search for the closest maximum. An earlier attempt to use total maxima in the whole search regions turned out to be less efficient. For the Kirsch detector, the crucial parameter is the mask size. One test has also been performed with a higher search length of s=8 and mask size=9, but this did not give significantly better results, Fig. 5. The plot reveals that an increased number of filter elements leads to higher stability of the tracker. This leads again to a miss-interpretation of the hand but prevents the algorithm from loosing the foremost edge and failing with the chair. As high mask sizes are computationally expensive, the size of 9, which equals 81 filter elements, has been chosen as the optimal parameter.

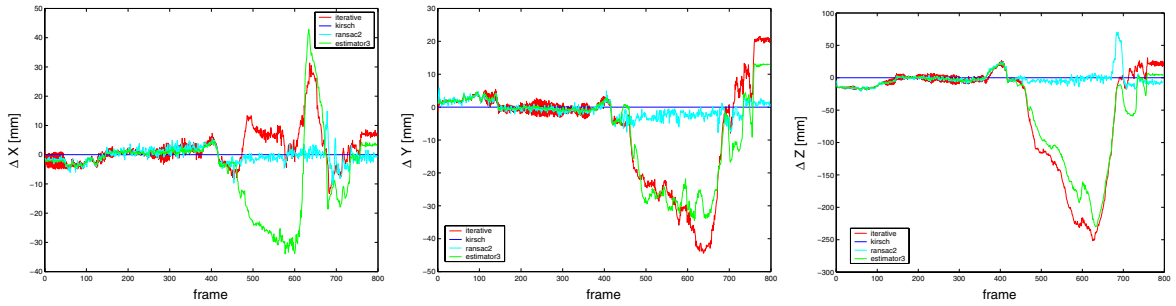

Fig. 6. Relative deviation of pose using different tracking algorithms: only translations are shown. The pose yielded by the Kirsch algorithm is used as ground truth.

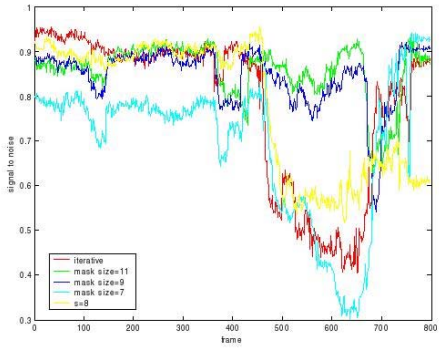

Fig. 5. Signal-to-noise ratio for the original algorithm using Kirsch detector.

E. Comparison to the original sequence

We have also evaluated the performance of the original algorithm compared to the performance with the above presented improvements: M-estimator, RANSAC and edge detection. The original sequence, shown in Fig. 1, contains object movements in all directions and background edges. The translational movement at the end of the sequence have been carried out relatively fast and causing most of the approaches to fail. The original algorithm miss-interprets the hand around frame 378 as object edge, looses the front edge around frame 432, aligns it with the right object edge, gets stuck to the leg of the chair in the background (frame 468). Fig.6 shows the object pose parameters plotted versus the parameters yielded by the Kirsch algorithm, because this was the only algorithm that has successfully completed tracking.

F. Textured Background

We have tested the performance of the algorithms with a complex background. The object undergoes both translational and rotational motion, Fig. 7. It can be seen that all algorithms had difficulties with this sequence, Fig. 8. Only the Kirsch-based algorithm remained tracking and it is therefore taken as ground truth.

G. Occlusion and Textured Background

The performance was also tested for a case when the object is static but there is another object moving in front of it and covering the four strong upper right edges, Fig. 9. It can be seen that all approaches successfully maintain tracking, Fig. 10. The original algorithm has troubles at the point of heaviest occlusion (around frame 180) and then switches the upper two model edges onto the upper object edge and stays stable in this pose. The fluctuation of the pose estimation of other approaches increases with the degree of occlusion but remains stable. The smoothest and most precise tracking is reached with the Kirsch operator.

H. Heavy Occlusion

The last testing sequence aimed to analyze the performance in a case of complete occlusion. Therefore, four edges of the static box have been completely covered, Fig. 11. Again, the ground truth values are constant. The original algorithm and M-estimator algorithm show almost the same behavior, Fig.12. The occluding box causes them first to shrink in vertical direction, then the upper model edges align to the occlusion edges, stick to them and finally cause the tracker to fail. RANSAC based algorithm, first shrinks during the occlusion and then sticks to it. At around frame 230 it flips back to a pose almost coincident with the object pose and stays stable. Finally, Kirsch based algorithm aligns with the occluding edge, shrinks and stretches with it, but completely switches back in frame 290 to the original pose and again, performs best.

IV. Conclusion

We have evaluated the performance of a model-based tracker using M-Estimators, RANSAC and an improved edge detector. These improvements decrease the deviation of estimated pose and allow the tracker to work under challenging ambient conditions. The best recovery after a partial loss of track was achieved with the Kirsch-algorithm. Given the evaluation, we can conclude that for a 3D model based tracker, it is better to concentrate on obtaining good measurements with improved detection (Kirsch-filter) than to enhance the matching step with statistical means (Estimators, RANSAC and Histograms).

References

[1] M. Armstrong and A. Zisserman, "Robust object tracking," in *Proceedings of the Asian Conference on Computer Vision*, vol. I, pp. 58–61, 1995.

Fig. 7. Example images from a sequence with a textured background: frames 18, 90, 126, 162, 180,216, 252, 306.

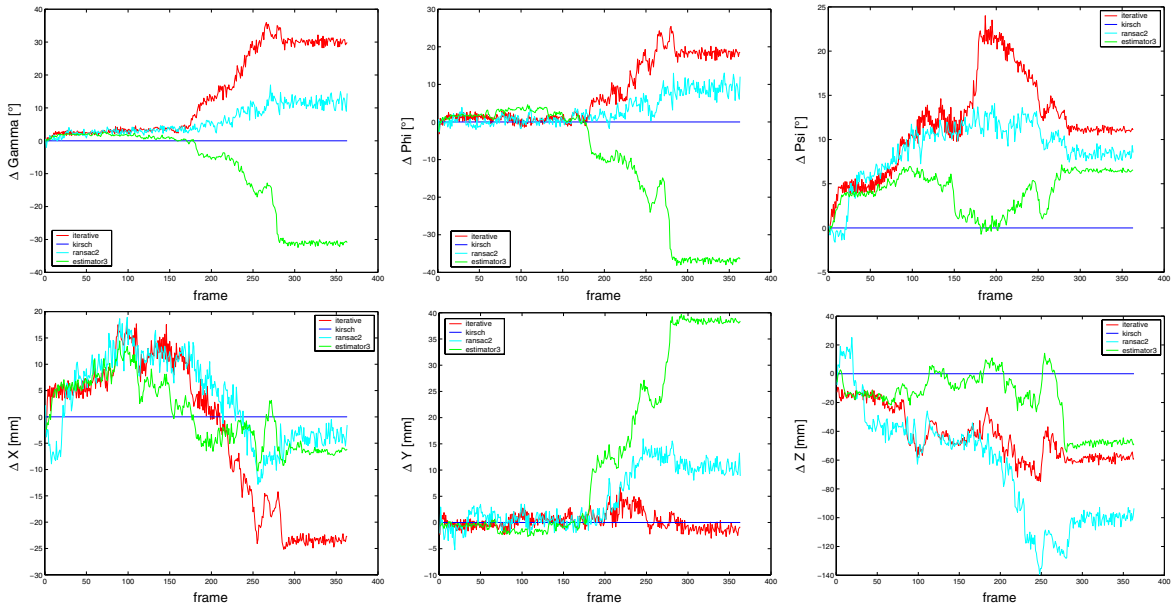

Fig. 8. Deviation of pose for a sequence with textured background: upper) rotations around X, Y and Z, lower) translations along X, Y and Z.

Fig. 9. Example images from a sequence with occlusion and textured background: frames 108, 126, 162, 180, 198, 216, 243,252.

[2] D. Koller, K. Daniilidis, and H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes," *International Journal of Computer Vision*, vol. 10, no. 3, pp. 257–281, 1993.

[3] T. Drummond and R. Cipolla, "Real-time tracking of multiple articulated structures in multiple views," in *Proceedings of the 6th European Conference on Computer Vision, ECCV'00*, vol. 2, pp. 20–36, 2000.

[4] D. Lowe, *Perceptual Organisation and Visual Recognition*. Robotics: Vision, Manipulation and Sensors, Dordrecht, NL: Kluwer Academic Publishers, 1985. ISBN 0-89838-172-X.

[5] P. Wunsch and G. Hirzinger, "Real-time visual tracking of 3D objects with dynamic handling of occlusion," in *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'97*, vol. 2, pp. 2868–2873, 1997.

[6] M. Vincze, M. Ayromlou, and W. Kubinger, "Improving the robustness of image-based tracking to control 3D robot motions," in *Proceedings of the International Conference on Image Analysis and Processing*, pp. 274–279, 1999.

[7] E. Marchand and F. Chaumette, "Feature tracking for visual servoing purposes," *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 53–70, 2005.

[8] V. Kyrki and D. Kragic, "Integration of model-based and model-free cues for visual object tracking in 3d," in *IEEE International Conference on Robotics and Automation, ICRA'05*, pp. 1566–1572, 2005.

[9] O. Chum, J. Matas, and J. Kittler, "Locally optimized ransac," in *Pattern Recognition, Lecture Notes in Computer Science 2781*, pp. 236–243, 2003.

[10] D. Kragic and H. Christensen, "Confluence of parameters in model based tracking," *Proceedings. IEEE International Conference on Robotics and Automation, ICRA'03*, vol. 3, pp. 3485 – 3490, September 2003.

[11] H. Wang and D. Suter, "Robust adaptive-scale parametric model estimation for computer vision," *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 26, no. 11, 2004.

[12] F. Jurie and M. Dhome, "Real time tracking of 3D objects: an efficient and robust approach," *Pattern Recognition*, vol. 35, pp. 317–328, 2002.

[13] C. Harris, "Tracking with rigid models," in *Active Vision* (A. Blake and A. Yuille, eds.), ch. 4, pp. 59–73, MIT Press, 1992.

[14] M. Fischler and R. Bolles, "Random sample concensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. ACM*, vol. 24, pp. 381–395, 1981.

[15] D. G. Lowe, "Robust model-based motion tracking through the integration of search and estimation," *Int. J. of Comp. Vis.*, vol. 8, no. 2, pp. 113–122, 1992.

[16] P. Wunsch and G. Hirzinger, "Real-time visual tracking of 3-d objects with dynamic handling of occlusion," in *IEEE Int. Conf. on Robotics and Automation, ICRA'97*, (Albuquerque, New Mexico, USA), pp. 2868–2873, Apr. 1997.

[17] J. Selig, *Geometrical methods in robotics*. Springer-Verlag, New York, 1996.

[18] P. Huber, *Robust statistics*. No. QA276.H785, ISBN 0-471-41805-6 in

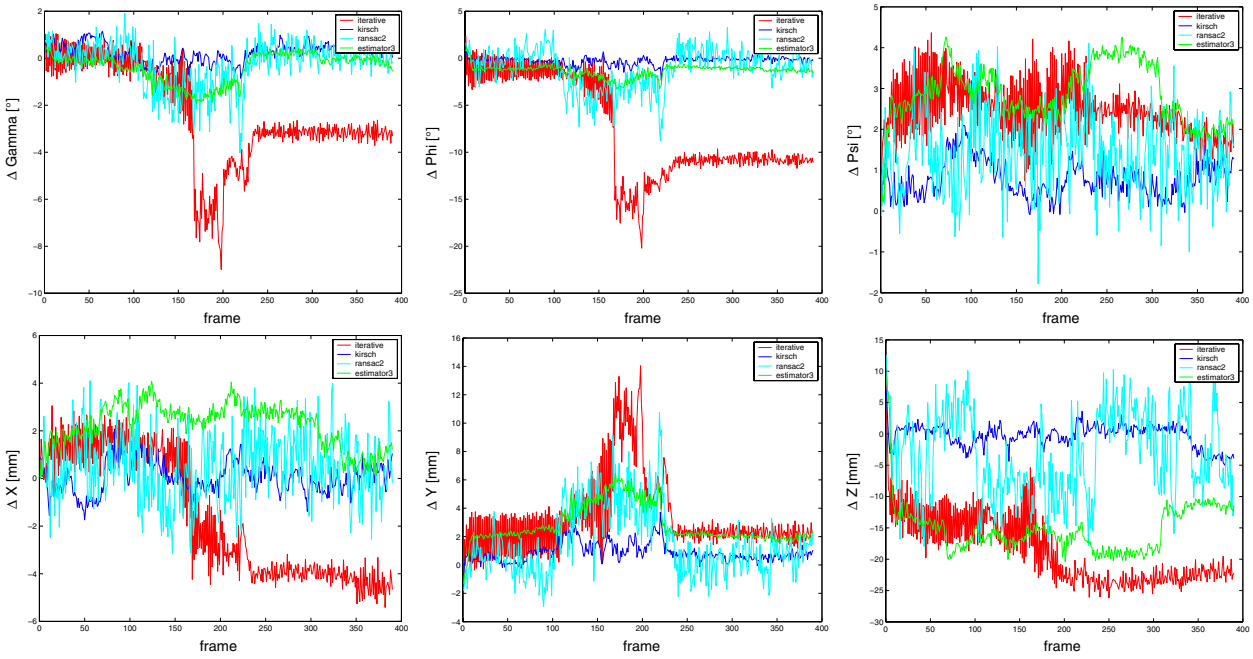

Fig. 10. Deviation of pose for textured background sequence with occlusion: upper) rotations around X, Y and Z, lower) translations along X, Y and Z.

Fig. 11. Example images from a sequence with occlusion and textured background: frames 18, 72, 108, 144, 180, 216, 252, 270.

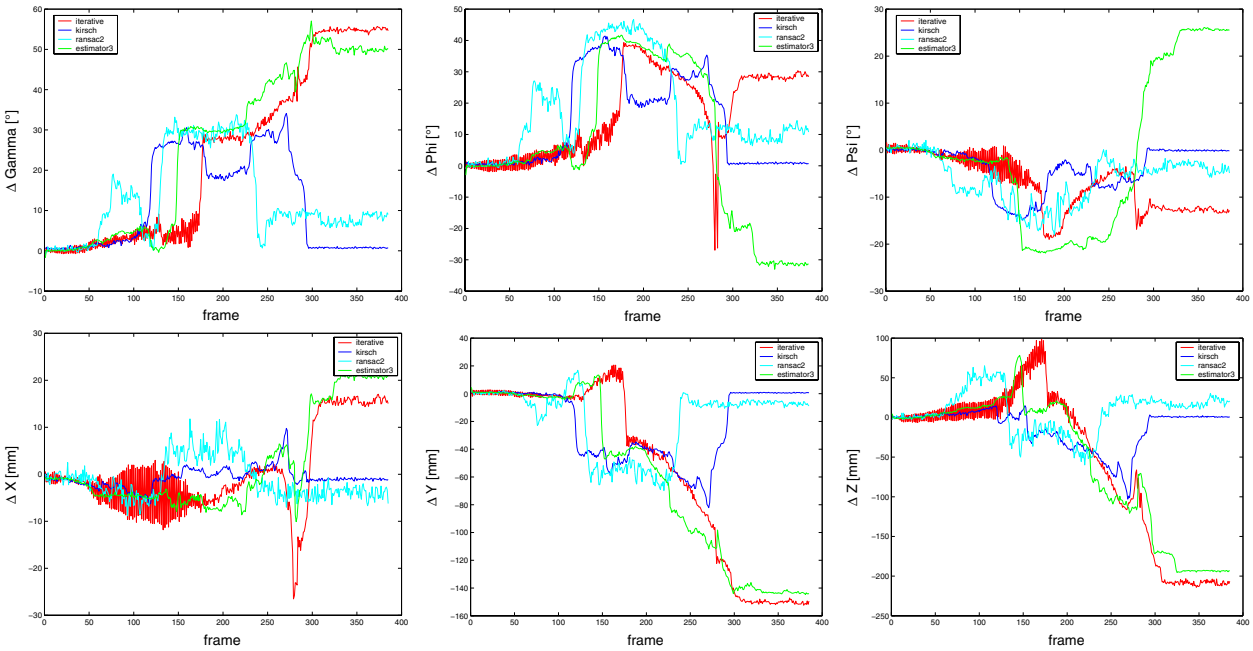

Fig. 12. Deviation of pose for a heavy occlusion sequence: upper) rotations around X, Y and Z, lower) translations along X, Y and Z.

Wiley series in probability and mathematical statistics, USA: John Wiley and Sons, Inc., 1981.