Measurement science for 6DOF object pose ground truth

Roger Eastman, Jeremy Marvel, Joseph Falco, and Tsai Hong
National Institute of Standards and Technology
Reastman@Loyola.edu
Jeremy.Marvel@nist.gov
Joseph.Falco@nist.gov
Tsai.Hong@nist.gov

Abstract

Users of perception systems in industrial manufacturing applications need standardized, third-party ground truth procedures to validate system performance before deployment. Many manufacturing robotic applications require parts and assemblies to be perceived, inspected, or grasped. These applications need accurate perception of object pose to six degrees of freedom (6DOF) in X, Y, Z position with roll, pitch, and yaw. A standardized 6DOF ground truth system should include test procedures, algorithms, artifacts, fixtures, and measurement equipment. Each of these must be openly documented so manufacturers, vendors, and researchers can recreate and apply the procedures. This article reports on efforts to develop an industrial standard for 6DOF pose measurement. It includes the design of test methods using a laser-tracker, an aluminum pose fixture, and a modular, medium density fiberboard (MDF) pose fixture.

1. Introduction

Robust perception is critical for industrial robotics [1]. Currently, most manufacturing robots operate blindly in highly predictable environments, performing their tasks by grasping parts presented in rigid fixtures. To operate adaptably and flexibly in unstructured environments, robots need reliable, accurate, non-contact sensing to perceive the identity, location, and pose of objects. Furthermore, users in manufacturing need convincing verification of a vision system's reliability and accuracy before deployment.

For manufacturers, this verification is best established through third-party, well-documented standards that can be used to validate vendor claims, carry out acceptance testing against user requirements, and support research and development. Perception system vendors often report performance figures, but this is usually with vaguely documented, proprietary testing methods that are not easy to compare across systems. Public, well-documented test procedures developed through consensus in standards committees offer manufacturers a familiar and solid basis for evaluation of perception systems and algorithms.

ASTM Working Group 31638, under Subcommittee E57.02 on Test Methods, is working on an ASTM Work Item (WK) WK31368, "Test Method for Evaluating the Performance of Systems that Measure Static, Six Degrees of Freedom (6DOF), Pose." This effort is intended to establish required measurement science including test procedures, artifacts, metrics, and recommended equipment for evaluating static six degrees of freedom (6DOF) pose perception systems. 6DOF pose perception means locating an object to position (X, Y, Z) and orientation (pitch, roll, and yaw). The test procedure for 6DOF will compare 6DOF estimates computed by a system under test to 6DOF ground truth values from a reference standard system with established accuracy. The test will be conducted under controlled conditions with user specified artifacts and characteristics, so as to compute performance metrics. The proposed 6DOF standard also requires a mathematical framework to compute these metrics.

Establishing a ground-truth measurement system as a reference standard requires extensive work and the use of measurement tools and techniques in validating its uncertainty. As a general rule, a ground-truth system (GTS) must be at least an order of magnitude more accurate than the system under test (SUT). As a practical matter, the reference standard system may be very expensive to purchase and/or apply. It is used primarily to establish the performance of a transfer standard system that can be used in the field. This leads to traceability standards that establish a solid chain of comparisons from reference standard, to transfer standard, and finally, to the system under test.

For industrial applications, standard test procedures have advantages over ground-truth databases. A fixed database of imaged objects or scenes, annotated with ground truth, can be exceedingly useful for researchers to fine-tune and compare algorithms. However, the data is tied to a fixed set of sensors and objects. In contrast, a test procedure allows researchers, developers, and users to generate new data with new sensors, new artifacts, and new conditions. The procedures are sensor agnostic and remain relevant as sensor technology advances. Industrial users can evaluate vision systems and algorithms on their own unique objects under environmental conditions found in their facilities. Moreover, because they are using a

standardized process, they can have more confidence in the resulting data. Manufacturers can ask vision system vendors to report performance using consensus procedures and metrics. If needed, the manufacturers can use them to generate proprietary data, with proprietary objects and conditions, for internal use.

This paper reports on efforts to establish ground-truth reference and transfer standards for 6DOF pose. Different measurement tools and procedures are presented and compared based on accuracy, measurement uncertainty, and expense. The tools include (1) a laser tracker, which is highly accurate but expensive and complex to operate, (2) an aluminum fixture, which is less accurate and less flexible, but also less expensive and less complex to operate, and (3) a medium density fiberboard (MDF) system, which shares many characteristics with the aluminum system but is even less expensive and can be duplicated easily. The goal of this paper is to develop practical measurement techniques that can assist manufacturers, vendors, and researchers in evaluating perception systems for industrial robotics.

2. Related work

The established literature of 3D pose estimation is extensive. Providing an overview of its breadth, however, is beyond the scope of this report. Instead, the focus of this study is on the test methods used to evaluate such systems.

2.1. Ground Truth

Establishing ground truth is a challenging task. Yet, it is the most crucial component of evaluating the performance of a pose measurement system. The evaluation of many novel systems is based on human-annotated ground truth with respect to some percentage confidence value. Such practices are common with data sets for which accurate ground truth is either nonexistant or is otherwise impractical or impossible to define. Examples of such scenarios include video feeds of human data (e.g., [2]) and large scale scene modeling (e.g., [3]). Other approaches utilize the power and control of modern simulation technologies to effectively generate ground truth and sensor data virtually (e.g., [4]). Such methods can provide perfect ground truth and can be used to easily evaluate robustness to random noise. Unfortunately, they are currently limited in their fidelity and the sensor types they can model.

With rigid, well-defined parts on smaller scales, ground truth may be established by direct measurement using tools with known, rigid transformations to an established coordinate frame origin (e.g., using coordinate-measurement machines as described in [5], high-precision laser measurement systems [6], or motion capture devices [7]). Ground truth may also be provided by means of controlling how test artifacts are presented to systems under test. A fixture or robotic mechanism can present an object in a pose known to the party conducting the test, but unknown to the SUT and its operator. The ground truth pose is given in advance by the fixture configuration. A fixture which allows more poses, more precisely, will be superior. In [8], a high-precision, pan-tilt mechanism was used to present objects in known poses. In [9], the authors detailed the design and evaluation of fixture-based ground truth with known uncertainties. The systems from [9] are discussed briefly in Section 3.

2.2. Evaluation metrics

In 3D, the pose accuracy of rigid objects is traditionally evaluated by assessing the relative Euclidean distance of the object's established coordinate origin as a measurement offset (or error) from the ground truth. Values are reported as a distance error measurements (e.g., [10]), or as a percentage of some maximum permissible error for a sensor system (e.g., [11]). Provided a known transformation exists between the ground truth coordinate frame and that of the system under test, absolute X, Y, and Z coordinate errors may also be evaluated (e.g., [12]). Similarly, given a known, rigid transformation from the ground truth to the system under test, rotational errors may be evaluated in a number of representations - though axis-angle representations and Euler angles (e.g., [13]) are most common for static systems. When a rigid tranformation between the ground truth and system under test is not known, however, one can assess errors only in terms of the relative measured transformations from one pose to another (e.g., [10]). For relative pose assessment, one makes a measurement at one lcoation, then moves the object a known displacement, and makes a measurement at the new location. The differences in the measured poses can be compared to the known poses. Both absolute and relative pose measurement evaluations are provided in WK31638.

2.3. Artifacts

The selection of artifacts for system evaluation has an impact on the performance of model-based and shape-based pose measurement systems. Objects may be selected for their utility in an application, or for particular shape or surface properties. Because WK31638 is written for general application with the needs and applications of users in mind, it does not define any artifacts for use with the test method. Ongoing standards efforts, however, are looking into the possibility of generating or adopting a set of artifacts that are representative of the common uses of pose measurement systems [15].

For standard artifacts, we may want a range of shape and surface features. It is helpful, but not essential, that reference objects provide unambiguous views and minimize symmetry. Symmetrical geometrical shapes such

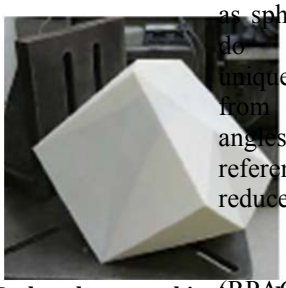

Figure 1. Reduced pose-ambiguity cuboctahedron (RPAC) [Permission granted by the author [14] to use the image

as spheres and cubes do not provide unique solutions from all viewing angles. One such reference object, the reduced pose ambiguity cuboctohedron (RPAC) [14] (Figure 1), minimizes pose measurement errors over all poses.

3. Ground truth systems

Three ground-truth systems were described in [9]. One is a highly accurate, sensor-based system suitable for use as a reference standard. The other two are fixture-based systems suitable for use as transfer standards. A brief overview of each is provided here. The sensor-based ground truth provides the highest accuracy; but, it is labor intensive and expensive. In contrast, the two fixture-based systems are cheaper and provide fast and convenient means for generating ground truth; but, they suffer from increased measurement uncertainty.

3.1. Sensor-based system: laser tracker

The laser tracker system is shown with its active target in a testing configuration in Figure 2. It determines distance by measuring time of flight, or interference, of a laser beam returned from a small, compliant mirrored target. The 3DOF position (X, Y, and Z) is given by distance combined with angular pointing direction. The orientation component is given by adding a separate, active, gimbaled target that rotates the target to point back to the base unit. (The base unit is on the left in Figure 2 while the active target is the cross-shaped unit on the right). The active target is limited to measuring (1) static 6DOF poses with a precision of ±3 arc-seconds in angle (±0.0008º) and (2) an average positional error of 15 µm with uncertainty of 10 µm at 2.0 m. The active target requires direct line-of-sight with the laser beam, limiting 6DOF tracking to only one object at a time per sensor. Another limitation is the 1.4 kg weight and size of the active target. This limits where it can be attached.

3.2. Fixture-based system: GT2011

To compensate for the single-target limitation, setup complexity, and cost of the laser tracker system, a portable aluminum mechanical fixture GTS (GT2011) was developed to support several manufacturing part artifacts simultaneously. GT2011, shown in Figure 3, was designed to (1) generate repeatable ground truth artifact poses, and (2) then provide this pose data in the form of known homogeneous transformation matrices for algorithm evaluation. The aluminum construction provides stiff transformations and limits wear from repeated use.

Table 1. Measurement Accuracy

	Laser Tracker	GT2011	GT2012
Mean Translation Error (mm)	0.015	0.4905	0.5794
Translation Error Variance (mm)	0.0053	0.1257	0.3854
Mean Rotation Error (degrees)	0.03 (active target)	0.1522	0.1686
Rotation Error Variance (degrees)	0.0007	0.0312	0.3124

Table 2. Utility of the Three Ground Truth Systems

	Laser Tracker	GT2011	GT2012
Max number of objects/scene	1	4	Unlimited*
Range (m)	0 – 80	0.6 – 2.0	Unlimited*
Range (XY)	± 320º azimuth -60º – 77º elevation	0 – 0.25 m	Unlimited*
Cost (US$)	$150 000	$4000	$400

* Theoretical; due to the modular design of the fixture, the larger the area spanned by the objects over the fixture, the greater the pose uncertainties.

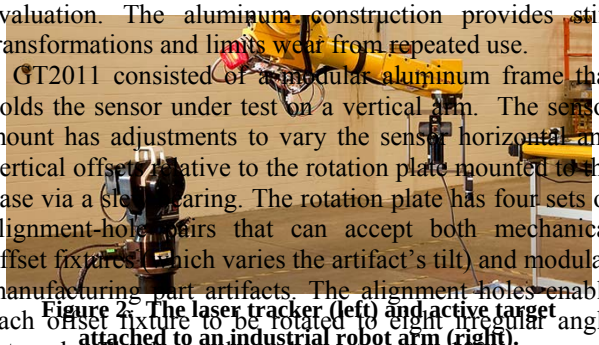

Figure 2. The laser tracker (left) and active target attached to an industrial robot arm (right).

GT2011 consisted of a modular aluminum frame that holds the sensor under test on a vertical arm. The sensor mount has adjustments to vary the sensor horizontal and vertical offset relative to the rotation plate mounted to the base via a slewing bearing. The rotation plate has four sets of alignment-hole pairs that can accept both mechanical offset fixtures (which varies the artifact's tilt) and modular manufacturing part artifacts. The alignment holes enable each offset fixture to be rotated to eight irregular angle intervals. The base plate also rotates at 10º increments using a ball plunger quick lock mechanism. Up to four artifacts can be placed on the rotation plate simultaneously. The fixture's provides comparatively high accuracy, but has limited range.

3.3. Fixture-based system: GT2012

The range and number of position offsets for GT2011 is limited by the rotational base, which constituted most of the construction cost. In contrast, the MDF mechanical fixture system (GT2012, Figure 4) was designed to increase modularity while decreasing production cost. GT2012 was designed using a lightweight, low-cost material and was produced using third-party manufacturing services. The open-source design can be kept on file at third-party services so users can order it on demand.

Figure 3. The GT2011 fixture base plate with a attached sensor and rotation plate with object.

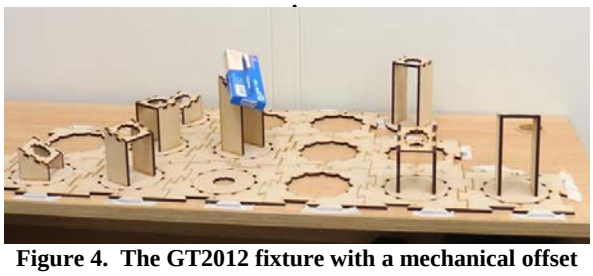

Figure 4. The GT2012 fixture with a mechanical offset configured for metrology testing.

The GT2012 system is constructed from 6.4 mm MDF using a laser cutter. Its modular design allows it to be assembled similar to a puzzle, enabling scalability from simple to complex artifact groupings. Each base piece accepts (1) a fixture assembly containing two rotational keys, each containing twelve angular increments and (2) an angular tower for adjusting Z offset, roll, pitch, and yaw of a mounted artifact. Each mechanical offset fixture integrates two rotation keys: one to integrate with the modular expansion components and the other to accommodate individual artifact mounting and rotation. Each key contains twelve rotation increments of 30º, and a preset angular tilt angle.

In contrast with the GT2011 design, the tolerances of GT2012 are far less rigid, and the material properties of MDF allows for faster wear (and increased measurement uncertainty) from repeated use. Relative positioning errors of the ground truth are attributed to the laser cutting process, which produced a kerf of approximately 0.2 mm.

3.4. Comparing the three ground truth systems

The performances and utilities of the three GTSs are summarized in Tables 1 and 2. The laser tracker is the most accurate, but also most expensive and complex. GT2012 is the least expensive and easiest to distribute, but is significantly less accurate. The usefulness of each GTS depends on the required accuracy for the intended application, such as part acquisition or inspection.

4. Applications of 6DOF ground truth

We have used the ground truth systems to support research challenges and the development of standard test procedures. Both applications use the same framework.

We represent 6DOF pose by a 4x4 homogenous transformation matrix (Equation 1):

$$\mathbf{P} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & t_x \\ r_{2,1} & r_{2,2} & r_{2,3} & t_y \\ r_{3,1} & r_{3,2} & r_{3,3} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \qquad \textbf{Eq. 1}$$

Using the test procedures [16] under development for ASTM, a selected test object is to be placed in one of 32 randomly allocated test poses within a predetermined volume, viewed by both the system under test (SUT) and the GTS as in Figure 5. A sensor-based GTS measures the pose of the SUT and Reference Object (RO) as matrices P_S^G and P_O^G, respectively, while the SUT measures the pose of the RO as P_O^S.

A sensor-based solution requires a homogeneous transformation registering the GTS to the SUT in space for absolute pose evaluations (e.g., [17]). In comparison, for a fixture-based GTS, the transformation from the fixture to the object is given inherently by the fixture set-up. There is an extra step to estimate the position of the fixture base relative to the SUT. If unknown, a relative pose procedure can be used. Fiducial marks can be put on the fixture for a separate GTS, or the SUT, to compute the SUT to fixture transformation. This only has to be done once for the SUT, as individual objects can then be swapped out.

Almost any object can be used as a reference object, but standard artifacts allow for consistent reporting of results. Figure 6 shows a modular set of standard artifacts designed to represent a variety of shapes and surface features, along with a base plate on which the artifacts can be arranged in different ways. These objects were machined from aluminum with high tolerance; but the models can be constructed from less expensive material or 3D printed. The CAD models of the objects can be openly distributed.

A test report should give the conditions of the test and the test results. The conditions include the date, time, test duration, operator(s), location, system settings, ambient conditions, test object descriptions, and test sequence. The results will include the raw measurements and derived error metrics.

The error metrics use the GTS and SUT measurements compared for each simultaneous measurement event. The current draft is for static pose, so time synchronization is not required. From the GTS measurements we can compute the SUT measure $\hat{P}_O^S$ as (Equation 2):

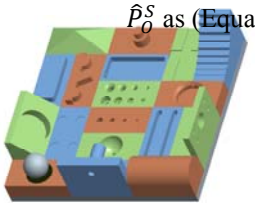

Figure 6. Standard part artifacts on modular base

$$\hat{P}_O^S = (P_S^G)^{-1} P_O^G \qquad \textbf{Eq. 2}$$

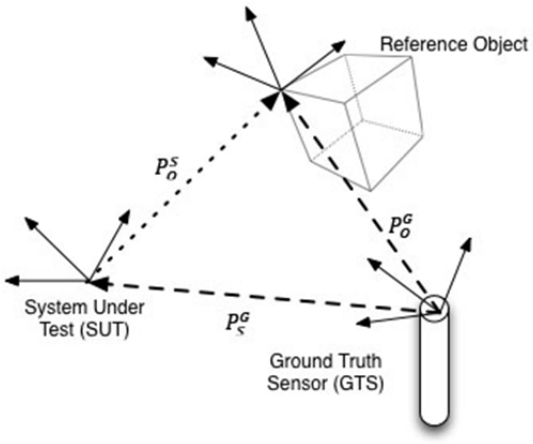

Figure 5. Relationship of sensors and coordinate systems

The pose error can be computed as the difference between P_O^S and $\hat{P}_O^S$ both for rotational error and translational error. For translation, the error e_t can be computed by standard Euclidian distance between the translation components of P_O^S and $\hat{P}_S^G$, while for rotation the error e_r can be computed by Equation 3 using the rotational components (Equation 3):

$$e_r = \cos^{-1}\left(\frac{trace\left(\mathbf{R}\hat{\mathbf{R}}^T\right) - 1}{2}\right) \qquad \textbf{Eq. 3}$$

For a single object, the draft 6DOF standard requires that 4 to 30 measurement pairs be taken from the GTS and SUT. For efficiency, an F-statistic is used to determine if the variance is converging to a stable value. If the variance, s_n^2, is stable after $n \leq 30$ pairs, the data collection can be halted and the errors reported.

The translation and rotation errors can then be used in four reporting statistics:

- The expected average pose errors, $E(e_t)$ and $E(e_r)$, as compared to the vendor specified performance limits on the expected average errors, $\delta_{avg,t}$ and $\delta_{avg,r}$ (Average Error Test),
- The maximum average pose errors, $max(e_t)$ and $max(e_r)$, as compared with the vendor's specified maximum average errors, $\delta_{max,t}$ and $\delta_{max,r}$ (Maximum Permissible Error),
- The variance of the average pose errors, s_t^2 and s_r^2, as compared with the vendor's specified limit on the variance of the average errors, σ_t and σ_r (Precision Error Test),
- The p^{th} quantile of the average pose error as compared with the upper bound on the vendor-specified quantile of the average error (Quantile Error test).

Each of the reporting statistics can be used to determine if the SUT is operating under the performance limits given by the manufacturer. For the Average Error Test, if the SUT manufacturer reports an average translation error of $\delta_{avg,t}$, it can be tested against the reported average error. If Equation 4 is satisfied, then the device is operating outside the reported performance limit.

$$\frac{E(e_t) - \delta_{avg,t}}{\sqrt{s_n^2/n}} > t_{\alpha,\nu} \qquad \textbf{Eq. 4}$$

The test statistic $t_{\alpha,\nu}$ is compared against the cumulative distribution of the probability density function (PDF) of the t-distribution with probability $\alpha = 0.05$ and degrees-of-freedom $\nu = n - 1$, n the number of poses..

This general test procedure was applied during the Solutions in Perception Challenge (SPC) held at the 2011 IEEE International Conference on Robotics and Automation. The procedure used the GT2011 GTS with machined artifacts given applied surface textures.

This general test procedure supports the ASTM Committee E57.02 "Test Method for Evaluating the Performance of Systems that Measure Static, Six Degrees of Freedom (6DOF), Pose" (Work Item ASTM WK31638). This draft standard provides detailed guides on the conditions and conduct of the test procedure.

5. Conclusions

In this paper, we presented an openly distributed, standardized, sensor agnostic, ground-truth-based test procedure for the evaluation of 6DOF perception systems. We discussed three supporting ground-truth measurement systems: a laser-tracker based system; GT2011, a low-cost machined aluminum fixture system; and, most recently, GT2012, a laser-cut, MDF fixture. The laser-tracker ground truth system is used to evaluate the 6DOF pose in Cartesian space. The two, fixture-based systems are intended to provide *a priori* pose data based on known transformations from a reference position via mechanical offsets relative to a given sensor under test.

The ground truth systems support sensor-agnostic test procedures currently being integrated by the ASTM E57.02 standards committee for 6DOF static pose system evaluation. Combined with open sourced artifacts, open sourced fixtures, and openly documented test procedures, researchers, manufacturing users, and vendors can use the standard to collect comparable data.

6. Acknowledgements

The authors would like to thank Elena Messina for her leadership to make this research possible and thank Geraldine Cheok and Mili Shah for their technical support in this study.

7. Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] Georgia Institute of Technology, *et al*. 2009. A Roadmap for US Robotics: From Internet to Robotics. http://www.us-robotics.us/reports/CCC Report.pdf 21 May, 2009.

[2] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose

estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1-8, 2008.
[3] S. Hsu, S. Samarasekera, R. Kumar, and H. S. Sawhney. Pose estimation, model refinement, and enhanced visualization ssing video. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 488-495, 2000.
[4] R. J. Murphy, *et al*. Pose estimation of known objects during transmission tomographic image reconstruction. IEEE Transationcs on Medical Imaging. 25(10):1392-1404, 2006.
[5] M. Góra and M. Maniowski. Verification of 6DOF platform with wire-based sensors for spatial tracking. The Archive of Mechanical Engineering, LVIII(2):157-174, 2011.
[6] T. Chang, *et al*. Methodology for evaluating static six-degree-of-freedom (6DoF) perception systems. Proceedings of the Performance Metrics for Intelligent Systems Workshop, 2010.
[7] K. Satoh, K. Takemoto, S. Uchiyama, and H. Yamamoto. A registration evaluation system using an industrial robot. Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality, 79-87, 2006.
[8] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3D recognition and pose using the viewpoint feature histogram. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2155-2162, 2010.
[9] J. A. Marvel, J. Falco, and T. Hong. Ground truth for evaluating 6 degrees of freedom pose estimation systems. Proceedings of the Performance Metrics for Intelligent Systems Workshop, 69-74, 2012.
[10] J. A. Marvel, T. H. Hong, and E. Messina. 2011 Solutions in Perception Challenge performance metrics and results. Proceedings of the Performance Metrics for Intelligent Systems Workshop, 59-63, 2012.
[11] International Organization of Standardization. ISO 1:1998. Geometrical Product Specifications (GPS) – Inspection by measurement of workpieces and measuring equipment – Part 1: Decision rules for proving conformance or non-conformance with specifications. Geneva, Switzerland.
[12] M-Y Liu, *et al*. Fast object localization and pose estimation in heavy clutter for robotic bin picking. International Journal of Robotics Research, 31(8):951-973, 2012.
[13] I. K. Park, M. Germann, M. D. Breitenstein, and H. Pfister. Fast and automatic object pose estimation for range images on the GPU. Machine Vision and Applications, 21(5):749-766, 2010.
[14] C. English, G. Okouneva, P. Saint-Cyr, A. Choudhuri, and T. Luu. Real-time dynamic pose estimation systems in space: Lessons learned for system design and performance evaluation. International Journal of Intelligent Control and Systems, 16(2):79-96, 2011.
[15] C. M. Light, P. H. Chappell, and P. J. Kyberd. Establishing a standardized clinical assessment tool of pathological and prosthetic hand function: Normative data, reliability, and validity. Archives of Physical Medicine and Rehabilitation, 83(6):776-783, 2002.
[16] D. MacKinnon, A. Beraldin, T. Hong, and J. Marvel, "Coordinate Metrology and the Proposed E57.02 Static Pose Determination Standard", The Journal of the Coordinate Metrology Systems Conference (CMSC), 8(1):4-8, 2013.
[17] M. Shah. Comparing two sets of corresponding six degree of freedom data. Computer Vision and Image Understanding, 115(10):1355-1362, 2011.