Specifying 3D Tracking System Accuracy
One Manufacturer's Views

Don D. Frantz[1], Stefan R. Kirsch[2] and Andrew D. Wiles[1]

[1]Northern Digital, Inc., 103 Randall Drive, Waterloo, Ontario N2V 1C5, Canada
[2]NDI Europe GmbH, Fritz-Reichle-Ring 2, 78315 Radolfzell, Germany
Email: skirsch@ndieurope.com

Abstract. Manufacturers of 3D tracking systems use a wide variety of statistical measures, assessment protocols and measurement volumes when stating their systems' accuracies. These factors typically differ according to the underlying technologies and the manufacturers' personal preferences and experience, but because of competitive pressures, manufacturers tend to use protocols and statistical measures that emphasize their systems' strengths and provide the best numerical values for comparisons. In addition, since 3D tracking systems generally have errors whose spatial distributions are nonuniform and which seldom follow known analytic distributions, the common practice of using a small number of statistical measures to represent "typical" accuracies for these systems is usually inadequate, and occasionally misleading. This can lead to a form of specmanship that can confuse potential users attempting to select the tracking systems best suited for their specific needs. We discuss some of the key accuracy factors often used to compare tracking systems, and we demonstrate some of the subtleties involved in accuracy specifications that potential customers should be aware of. The example systems cited are all manufactured by NDI.

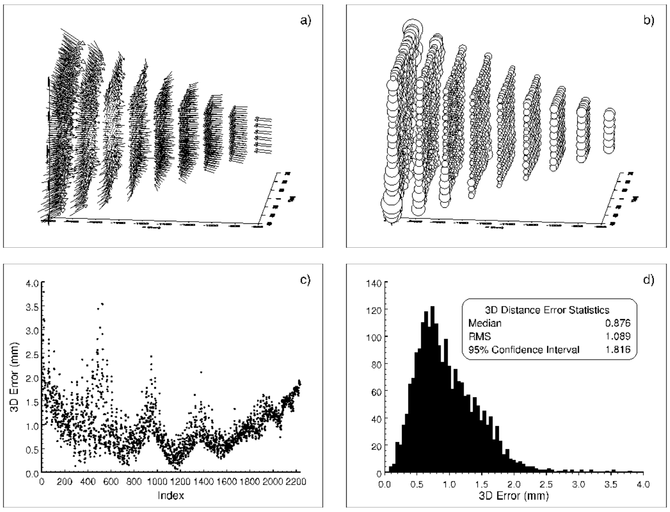
1 Introduction

Spatial tracking systems provide the core technology for many sophisticated image-guided systems that are used by physicians for various procedures such as tool navigation, patient positioning, and treatment planning. The critical performance criterion of these systems is their spatial accuracy, but accuracy assessments of such systems are inherently statistical and typically complicated by their non-uniform error distributions over their operational volumes. Manufacturers of these systems typically provide specifications to potential customers, which they claim fairly represent the performance of their systems. To properly assess the accuracy of a given measurement system, though, at least two items are required: (1) a proper set of characteristic statistics that define its trueness and precision and (2) the specific protocol on which the assessment is based. Unfortunately, manufacturers often present the statistical results only. Occasionally, manufacturers will indicate which standard they have followed, but the results are not useful unless the specific protocol parameters are also provided. In addition, since marketing literature generally strives to simplify performance related information and to reduce the assessment data to a few "representative" measures, users invoking commonly held statistical assumptions can easily over-interpret or even misunderstand the stated accuracy measures. For example, two common statistical measures often used are the root mean square (RMS) error and the mean (average) error. In general, the RMS value is the preferred statistic, since it incorporates both the mean and the standard deviation ($\mathrm{RMS}^2 \approx \mu^2 + \sigma^2$) [1], but some manufacturers prefer to quote mean values, as they are substantially lower for 3D distance errors. Since the measure actually specified by different manufacturers is typically identified in the fine print of a footnote or endnote, users casually comparing specifications for different systems can easily compare numbers directly without realizing that they are fundamentally different statistical measures. We outline in this paper how the condensed "marketing numbers" are typically derived from a particular validation protocol, and how much of the important information required for users to properly assess a system's performance is lacking.

2 Volumetric Calibration Protocols

Tracking systems are generally characterized by comparing the 3D positions generated from their underlying sensor measurements to the corresponding positions obtained from some appropriate reference. Since the characterization data typically cover much of the operational volume, they can also be used to calibrate the systems and assess their accuracy. The resulting spatial error distributions from such a volumetric protocol provide a detailed assessment over the characterization volume, from which statistical quantities such as the mean error, the RMS accuracy, the percentile confidence intervals (CI), the repeatabilities, and various other measures can be derived. For ideal cases where the errors are spatially distributed in a uniform manner and are subject to a known analytic distribution such as a normal distribution, one or two key statistical measures can be used to represent the entire distribution, and so provide typical accuracies. For most 3D tracking systems, though, errors are not uniformly distributed spatially and seldom follow known analytic distributions, which implies that such a small number of statistical measures cannot adequately represent typical system accuracies. Fig. 1 illustrates this by examining the measured errors obtained from a volumetric calibration of a damaged Polaris optical position sensor. (We have chosen this example because the systematic errors dominate the random errors, making the error patterns especially apparent, but the discussion is fully general and applies to systems within specification as well.) The data were obtained by tracking a single marker throughout the operational volume using a coordinate measuring machine (CMM) as a reference. The overall volume RMS 3D distance error is a statistical measure that is commonly used to specify such a system's typical accuracy, but as can be seen in the plots, the range of errors is too large to represent by a single value in any meaningful way. Including the median and 95% CI values would provide a better indication of the distribution, but even

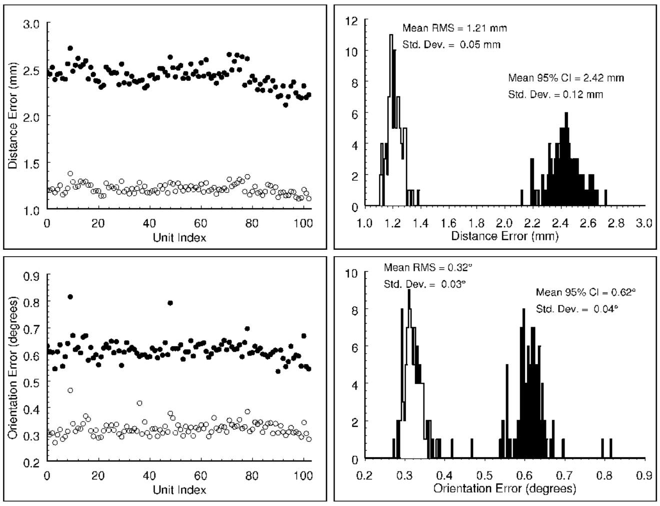
Fig. 1. Inherent data reduction in common statistical measures for a damaged optical position sensor that has a large systematic scale error. For the sequence of plots from the full data set a) to the histogram and final statistical summary d), increasing simplicity and clarity come at the expense of continued loss of information. Plot a) shows the spatial dependence of the 3D error vectors, with the error magnitudes represented by the arrow lengths (the position sensor was located to the right of the grid). In plot b), the 3D errors have been reduced to 1D distance errors, with the error magnitudes now represented by the circle areas. In plot c), the spatial information has been reduced to a measurement index, but the sequential information has been maintained. In plot d), the sequential information is lost, but the nature of the underlying distribution is made evident. The distribution is then reduced to the three statistical measures listed in the box, which is typically all the user gets.



the measured distribution has reduced most of the underlying information, some of which directly affects the resultant statistics. For example, since volumetric calibrations generally require the measured grid positions to be aligned with the reference grid, there is usually a small grid alignment error that is incorporated into the overall calibration error, which is often overlooked. The alignment error in this case is evident in the vector plot, since the error distribution for this scale error when properly aligned actually has the vectors pointing predominately to the right, and their magnitude increasing roughly linearly from the front of the volume (on the right) to the back.

While users are mostly concerned with the performance of their own specific system, manufacturers generally provide specifications for their systems collectively, and this adds another layer of ambiguity. We illustrate this in Fig. 2, where we have presented the overall volume RMS and 95% CI results for a number of Aurora electromagnetic tracking systems [2,3]. We have described the volumetric calibration protocol for the Aurora in detail in a previous article [4], and so here we focus on the variation between systems. As can be seen in Fig. 2, both the overall volume RMS and 95% CI values have substantial variation, which is

Fig. 2. Variation in Aurora accuracy for recently manufactured beta systems. The upper plots show the series variation and corresponding histogram for the overall volume RMS distance errors (open circles and bars) and the overall volume distance error 95% confidence intervals (filled circles and bars), while the lower plots show the analogous results for the orientation errors.

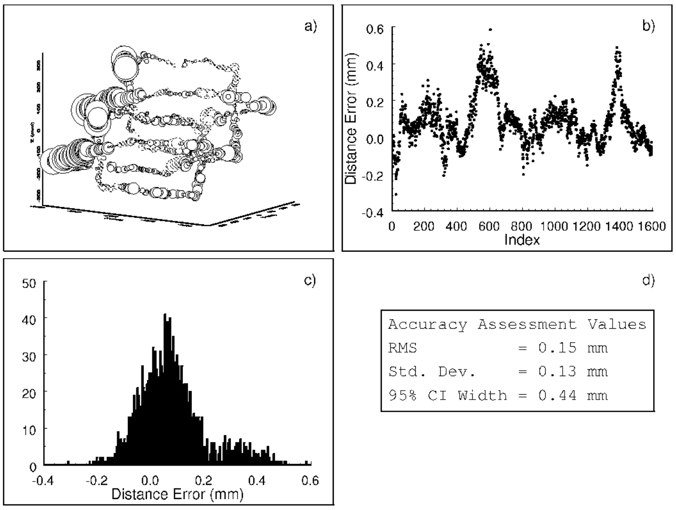


typical of most tracking systems. How should the system accuracy be specified in such cases? Some manufacturers would just quote the mean values as their representative accuracies, which would be acceptable if the relative variation is small, but is disingenuous if the variation is substantial. Other manufacturers would include the standard deviations as uncertainty estimates (e.g., the overall volume RMS accuracy is 1.21 ± 0.05 mm), which would be acceptable if the distribution is approximately normal, but is seldom the case since the distributions generally tend to be skewed to higher errors. The most conservative approach would be to select a threshold value on the high end of the distribution and pass only those systems having lower errors. Another complication arises when several statistical measures are presented, since the distributions for each measure are treated independently, masking the correlations between them.

2.1 Tsunami

Even when manufacturers determine their systems' accuracy performance properly within stated protocols, the results can still differ much from the system's "real world" or application performance. Manufacturers typically select assessment protocols for their specific needs, and these assessments are usually undertaken under laboratory conditions to ensure proper repeatability. Such assessments are not likely to be as relevant for most users' intended uses of the system. It is therefore important for manufacturers to also develop accuracy assessment protocols that are more tailored towards the particular applications most users envisage.

Fig. 3. Results for an in-field calibration of a damaged Polaris optical tracking system. The errors represent the differences between the measured length of a bar and the bar's known length. The data reduction depicted is analogous to the one shown in Fig. 1. The fundamental differences in the error distributions are clear in the plots, but not obvious in the final statistics.



NDI has developed an Accuracy Assessment Kit, which is a tool designed to test the accuracy of Polaris position sensors in the field. This protocol measures the distance between two rigidly attached reference tools and compares their measured length to a pre-determined reference length that was characterized at the factory. The tool software guides the movement of the bar throughout pre-determined regions of the volume to ensure repeatable and reproducible data collections. At the end of each collection, the bar errors are analyzed. Fig. 3 shows a typical collection for a damaged position sensor, and analogous to Fig. 1, crucial information is lost when the data are reduced to the final few statistical values. In particular, the non-uniform spatial dependence is clear in plot a), with the largest errors at the back of the volume, and the fundamentally different types of error in the two protocols are evident in their histograms, where the bar-length errors can take on negative or positive values, while the volumetric distance errors are restricted to positive values.

References

1. Hsu DY: "Spatial Error Analysis, A Unified Application Oriented Treatment", IEEE Press, New York, 1999.
2. Kirsch S et al.: in "Advances in Hadron therapy," Int. Congr. Series 1144, Excerpta Medica, p 269, Elsevier, Amsterdam, 1997.
3. Seiler PG et al.: Phys. Med. Biol. **45** N103–10, 2000.
4. Frantz DD et al.: Phys. Med. Biol. **48** 1–11, 2003.