A Metric for Tracking Robustness in Real-Time Panorama Acquisition

Christopher Coffin* Sehwan Kim† Tobias Höllerer‡
University of California, Santa Barbara

ABSTRACT

To test the performance of tracking systems on a high level it is necessary to perform both quantitative and qualitative analyses. This is particularly true for concepts such as robustness that lack a clear-cut quantitative definition. Addressing the difficulty inherent in collecting user evaluation data, we present a metric which maps the quantitative evaluation of systems to a close approximation of qualitative robustness of the experience. This largely reduces the need to perform qualitative analysis of tracking robustness. We motivate the need for this metric through an analysis of four orientation tracking systems used for the construction of environment maps. This initial analysis demonstrates that ground truth error does not directly reflect the results of a qualitative analysis. We then show that our proposed metric is able to use the quantitative analysis of the systems to correctly approximate the relative robustness of the systems in our initial evaluation. Therefore our proposed metric is able to estimate qualitative evaluation results, without the need for an additional user study. Our metric requires only a set of representative video sequences along with ground truth data for each frame in terms of yaw, pitch, and roll.

Keywords: Robustness metric, vision-based tracking, real-time panorama acquisition, expert evaluation, camera pose relocalization

Index Terms: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality; I.4.8 [Image Processing and Computer Vision]: Scene Analysis

1 INTRODUCTION

Despite a large increase in the number of tracking systems in recent years, there is still no simple method for comparing the performance of multiple tracking systems at a high level. This is particularly true with respect to obtaining a qualitative sense of robustness. Any such comparisons must currently be carried out by asking a large number of users to rate or rank the performance of the systems tested. Additionally, the data from such user studies cannot be reused and a new study must be performed each time a new system is introduced.

We are interested in determining the qualitative robustness of tracking systems. However, robustness itself is a nebulous term, particularly in relation to computer vision. Robustness of an algorithm in computer science is defined as the ability for the algorithm to continue to function despite abnormal input. We similarly define robustness as the ability of vision-based tracking systems to function in the presence of sudden or severe changes in the input video such as changes in lighting, camera movement, occlusions, etc. For a qualitative sense of robustness, we are interested in how the system appears to perform while under the control of users.

The contribution of this paper is in providing a metric which estimates qualitative ratings of the robustness of tracking systems consistent with such user evaluations, but without requiring additional user studies beyond those performed in this paper. The proposed metric is uncomplicated in design, and can be easily reused for evaluating future orientation tracking systems. Our metric requires only a set of representative video sequences along with ground truth data for each frame in terms of yaw, pitch, and roll. By representative we mean that the video sequences should cover the spectrum of conditions over which the tracking is said to be robust. For example, if a system is robust against lighting changes, then the input videos should contain a collection of movements over which the lighting changes considerably.

Our goal is to reduce the need for case-specific user studies in order to analyze the qualitative performance of systems. In order to determine if this is possible, we must first understand the relationship between qualitative and quantitative performance. We therefore begin by presenting both the qualitative and the quantitative performance of four tracking systems in section 3. Our results clearly indicate that direct comparison of raw quantitative data is not sufficient to gain an understanding of the qualitative perception of robustness. To address this issue, we derive a metric which can be used in order to perform such a mapping. This metric consists of two parts, a classification of the error, and an equation to translate the classified error into an estimation of qualitative robustness. We explain both the thresholds for classification and the parameters of the equation in section 4. In section 5, we then demonstrate the effectiveness of our metric by correctly predicting the qualitative performance of a system we did not use in establishing the metric, using only the measured quantitative performance of the system. Finally, we conclude with a discussion of our results and the direction of future work in section 6.

2 RELATED WORK

A wide variety of tracking systems using various kinds of sensors have been investigated for Augmented/Mixed Reality (AR/MR) applications. There has also been extensive evaluation of low-level interest point detectors, feature descriptors [16] [13] [12] and camera pose techniques. Our goal is complementary to and very different from these evaluations as we focus on classifying performance on a much higher level.

Even though there have been several metrics for measuring tracking errors, there is still a lack of research regarding metrics for assessing the robustness of tracking approaches. That is, most of the tracking methods for AR applications developed their own error metrics or employed conventional metrics, such as reprojection errors, comparison with ground truth and overlay of synthetic models. Satoh et al. used a robotic arm to obtain ground truth for measuring registration error [15] [11] [6]. Most metrics make no consideration of visual appearance of tracking results and user experience. Recent work in the field of AR has visited the issue of robustness [14], but even though the authors used a special error metric, it was not used to assess the robustness of the system but for detecting tracking failures only. Klein et al. also performed research to improve the robustness of SLAM systems by using edge features and an inter-frame rotation estimator [10]. The general robustness of a system involves more than just the average system error, and most of the existing evaluations have been on the error of a single system, while we are interested in a metric which is more generally applicable. Our work is complementary to the laudable efforts of the TrakMark working group [17], and the data we are

*e-mail: ccoffin@cs.ucsb.edu
†e-mail: skim@cs.ucsb.edu
‡e-mail: holl@cs.ucsb.edu

20th International Conference on Artificial Reality and Telexistence (ICAT2010), 1-3 December 2010, Adelaide, Australia
ISBN: 978-4-904490-03-7 C3450

making publicly available [1] are meant to further some of the same goals.

Some of our analysis is similar in form to work on empirical evaluation done for qualifying computer vision work. While we are not aware of any high level robustness evaluations of camera pose, there has been some work in classifying other tracking algorithms. In particular, Bowyer et al. [1] presented an overview of several methods for empirical evaluation of computer vision algorithms. Our methodology is similar to both the second and third categories they describe. Our work shares some similarities with what they term independent evaluation in that we do not have a favored technique. Our expert evaluation is also akin in spirit to the evaluation performed by Heath et al. [7] as our evaluation has a qualitative component in the human definition of what is robust tracking. However, in our case the evaluation is not the final result. Instead, we are interested in generating a metric from both the user evaluation information we collect and a large set of data with known ground truth so that similar user evaluations may be circumvented in future robustness comparisons of tracking methods. As long as evaluators are able to provide sets of input streams that represent the robustness challenges that they are considering along with ground truth tracking information for these sequences, our metric will enable comparison of different tracking algorithms with respect to their robustness, without requiring additional user studies.

This work extends and builds upon a previous overview evaluation of a number of tracking systems [3].

3 Motivating Evaluation

In this section we motivate the need for a metric of robustness based on quantitative observation through the results of a study comparing tracker performance of several real-time computer vision systems for panorama acquisition. For this study, we present a quantitative analysis of these systems based on distance to ground truth. We also present a qualitative analysis based on both the output panoramas and a live user evaluation. Finally, we demonstrate that ground truth error alone does not provide sufficient insight into the perceived robustness of the system.

For this analysis we examined the performance of four variations on an existing orientation tracking system, Envisor [4]. We analyzed regular Envisor, Envisor with constant recovery, Envisor with selective recovery, and Envisor with constant recovery and pre-scanning. A detailed discussion of these methods can be found in [2]. Briefly, the base system of Envisor uses a frame to frame feature-based tracking system. With regular Envisor, recovery is only possible within a narrow region around the last tracked position via regular feature-based tracking. Envisor with constant recovery is similar to regular Envisor but uses a keyframe-based recovery method which is always actively trying to recover. As a result, it can quickly recover to any camera pose already seen during acquisition. However, the constant recovery may result in some additional jitter or jumps (due to keyframe alignment). Envisor with selective recovery is similar to the constant recovery method, but only attempts to use the keyframes for recovery if it suspects that tracking has failed. This results in a smaller amount of jitter than found in the constant recovery method. Envisor with constant recovery and pre-scanning is our ideal case. It uses the constant recovery method but with a full set of previously acquired keyframes surrounding the user. This greatly reduces the chance of tracking failure, but the system may suffer from a small amount of jitter due to the keyframe alignment.

We chose these systems for several reasons. First, our familiarity with the systems allows us to verify that the results regarding their performance are realistic. Second, their performance range from 'fragile' to 'quite robust', which provides desirable differences for our evaluation. Finally, we are sufficiently familiar with the methods featuring constant and selective recovery to know that their qualitative results should generally be very similar. This is useful, as we should be able to distinguish not only very poor and very robust systems, but also systems which are similar in terms of robustness. Note that the goal of this paper is not to demonstrate the superiority of a particular tracking system; the exact tracking solutions used are of less importance than the ability of users to distinguish the systems by their robustness.

[1] http://tracking.mat.ucsb.edu/

3.1 Quantitative Evaluation

In order to accurately perform a quantitative evaluation of our tracking systems it was necessary to collect a data set containing ground truth with respect to the orientation of the camera. For our study we collected video data with very accurate ground truth by mounting a camera on a pan tilt unit (PTU). The PTU allows for a previously constructed path file to be replayed precisely and allows us to obtain a highly accurate estimate for the orientation of each frame of video. Specifically, we used a PTU-D46 pan tilt unit [5] from Directed Perceptions. The PTU has an upper speed limit of 300°/second and a resolution of 0.0514°. We are therefore able to replay motion data precisely and at the speeds recorded. This allows us to retain motion blur and other real-life imaging artifacts during playback.

The orientation paths used as input to the PTU were collected from the head movements of a large set of users performing both searching and exploration tasks. More detailed information on these paths will be discussed in Section 6. For the purpose of our evaluation we used 45 different paths each a minute long. The data was replayed in both indoor and outdoor locations to provide a larger variation between samples for the evaluation. We are making this motion path data available for download, alongside with the captured panoramas in various places under various acquisition conditions. [1]

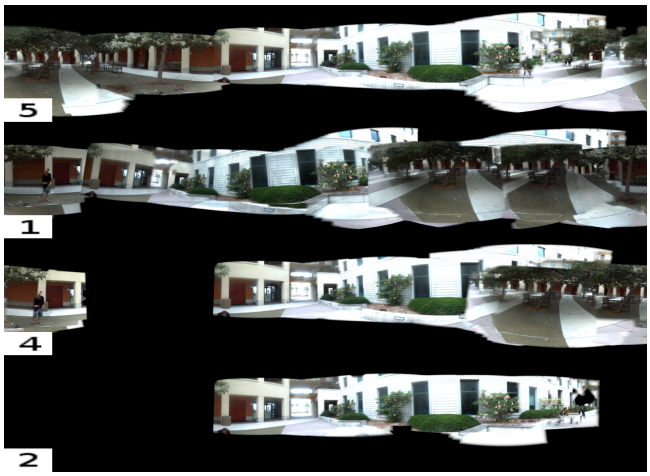
The camera used for the capture of the panoramas was a Point-Grey DragonFly2, which delivers 640×480 pixel RGB frames at 30 Hz. We used Zhang's calibration technique to measure the camera's intrinsic parameters in a one-time offline calibration procedure [18] [8]. In addition to the focal length and principal point, lens distortion parameters were also measured which were used to undistort each frame.

3.2 Qualitative Evaluation

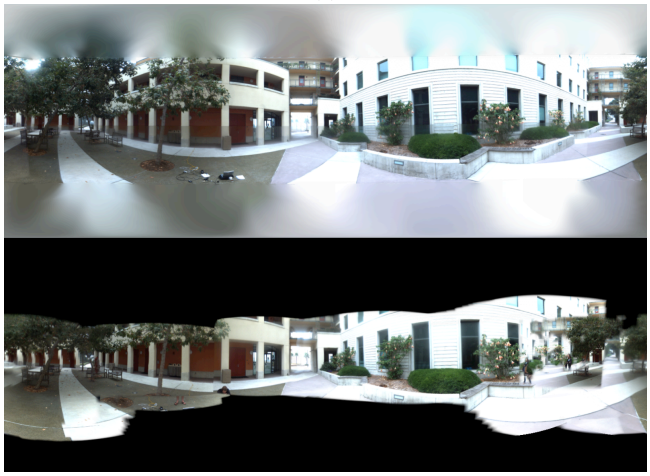
The qualitative analysis of system robustness consisted of two parts, an analysis of the resulting output panoramas and a live expert evaluation of the tracking systems. The analysis of the generated panoramas allows for a large breadth of samples (different places and acquisition conditions) to be examined, while expert analysis provides confirmation of the trends seen in the analysis of the results, using live system exploration in one place.

Panorama Evaluation For this evaluation, all of the video samples collected were used as input for each of the four systems in order to obtain output panoramas. Given 45 input sequences this resulted in 180 panoramas being used for the evaluation. Each of these panoramas was ranked based on the size and quantity of any noticeable visual artifacts. The evaluations were performed by six domain experts (researchers in the fields of computer vision and augmented reality).

To allow the expert evaluators to rate the panoramas, we designed the ranking interface seen in Figure 1. The experts ranked the panoramas in sets. Each set contains four panoramas of the



(a)



(b)

Figure 1: The interface used for the panorama evaluations: a) users were asked to rank panoramas generated by each of the four variations of Envisor. b) users were able to click on an item to compare it to a ground truth panorama of the scene.

same scene as seen in Figure 1(a). All panoramas in a set are generated using the same motion path. The only difference is the tracking method used to generate the panorama with one panorama for each tracking solution. The order of the panoramas inside each set was chosen randomly for every viewing. The presentation order of the sets themselves was also random.

The six experts were asked to evaluate every panorama in each set. To evaluate a panorama the experts selected it by clicking. They then assigned a value between one (worst) and seven (best) by pressing the corresponding number key on the keyboard. Once assigned, the selected score was displayed on the left-hand side of the panorama. Note that these values could be changed and were only fixed once the experts moved to the next set in the series. Additionally, experts were able to compare each panorama to a ground truth panorama of the scene as shown in Figure 1(b).

Live Evaluation As an additional metric beyond evaluation of the resulting environment maps, we had five expert users evaluate each system in a live demonstration (3 people were chosen from the previous set of six, but we allotted several days in between the two studies). These users were all experts in the field of vision-based tracking, and therefore they were asked to simply rate the

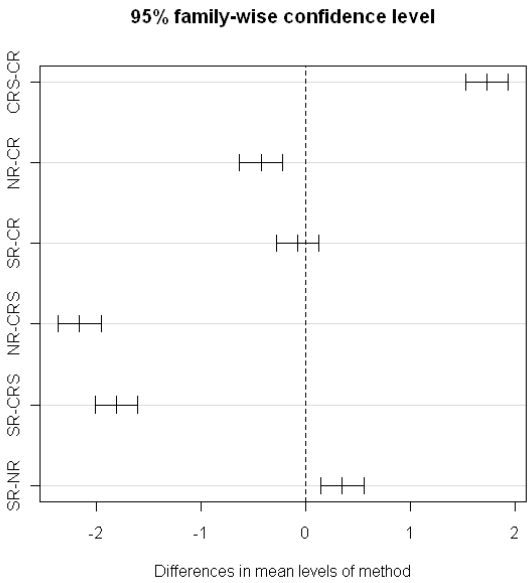
Figure 2: The interface shown to the expert evaluators as they determined the robustness of the systems. The evaluators rotated a camera on a tripod at speeds of their choice. The yellow rectangle in the center represents the current camera image. Part of the currently captured panorama data is visible outside of the yellow rectangle. All evaluations of the systems were recorded on paper after the test.

robustness of the systems as they perceived them based on their experience. In order to ensure a fair comparison, we had each user rank each system four times for a total of 16 randomly ordered runs per user. To eliminate the effect of initial lack of comparison as the experts determined a scale for their ratings, we used only the last two iterations. The ordering of the tracking methods was randomized. For each individual test the evaluators were given at most a minute to examine the tracking performance of the systems.

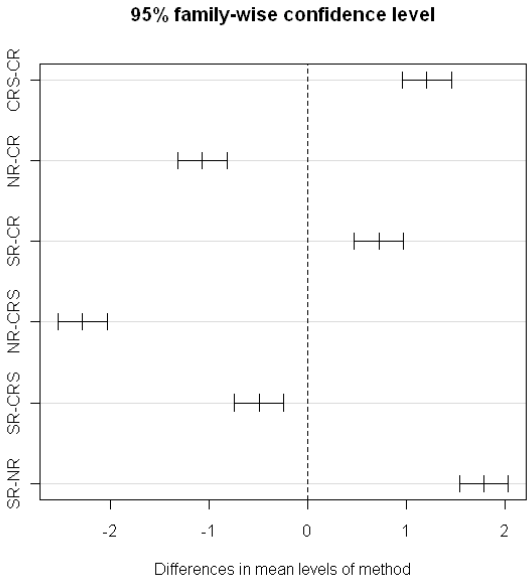
Every evaluated system used the same interface to represent the constructed panorama, seen in Figure 2. This interface shows a 90° vertical field of view (FOV) of the environment map, with new frames posted directly into the environment map without blending. After each evaluation the users were asked to rank the system from one (worst) to seven (best). We normalized and averaged the results for both panorama and live evaluation, accounting for subjects' relative differences in rating.

3.3 Evaluation Results

The average ratings of our evaluators for each set of result environment maps were very consistent for each of the methods. Performing an analysis of variance single factor test with the independent variable being the method and the dependent variable being the ratings resulted in a residual of 1:1940 with F = 572 and $p < 0.0001$. From the evaluation we can see that the different methods performed with significant difference among all sequences. The results from a set of corresponding Tukey Post-hoc evaluations is shown in Figure 3. Note that every tracking system was significantly pairwise different from each other in terms of the quality of the generated panoramas, with the exception of constant recovery compared with selective recovery in the indoor case. Among all methods, these two are most similar to each other in terms of their robustness. This is important as it indicates two things. First, the methods could generally be distinguished, and second, the qualitative performance of the selective recovery method is very close to that of the constant recovery method. This relationship is also clearly reflected in the results from the live evaluation, but is not seen in the raw ground truth error. Table 1 summarizes the evaluation results from the panorama evaluations, the live tests, and lists average distance to ground truth in degrees for each of the tracking systems over all the input sequences.



(a)



(b)

Figure 3: Results of a Tukey multiple comparison of means given an ANOVA comparison for a) indoor and b) outdoor sequences showing that the evaluations of the panoramas are statistically different for each pair of systems with the exception of constant and selective recovery. Note that these results are also reflected in the live evaluation. (CRS: Envisor with constant recovery and pre-scanning, SR: Envisor with selective recovery, CR: Envisor with constant recovery, NR: original version of Envisor (No Recovery))

Table 1: First row, average distance to ground truth in degrees. Second row, ratings assigned to the panorama output data (scale one (worst) to seven (best)). Third row, the robustness ratings from the live evaluation (scale one (worst) to seven (best)). (NR: original version of Envisor (No Recovery), CR: Envisor with constant recovery, SR: Envisor with selective recovery, CRS: Envisor with constant recovery and pre-scanning)

	NR	CR	SR	CRS
Distance to ground truth (°)	26.75	8.08	16.38	3.27
Panorama evaluation	2.03	3.12	3.54	5.41
Live evaluation	1.63	3.95	4.03	6.05

Note that for both of the qualitative analyses seen in Table 1, the relative distance between the selective recovery method and the constant recovery method was very small. This distance is not only much greater in the ground truth error, but also reversed, indicating that there is not a direct linear mapping between tracking error and qualitative robustness. This result clearly indicates that a more indirect mapping is needed.

4 Formulation of the Robustness Metric

In this section we discuss the formulation of our proposed metric. To reiterate, the purpose of this metric is to map a quantitative analysis of tracking systems to a qualitative estimation of the robustness of those systems. The input for this metric is the per frame tracking error (distance from ground truth) provided by the quantitative analysis.

Our metric involves two steps. First we classify the error generated by the systems, and second we determine the robustness of the systems using the classified error percentages as input to our metric. In this section we begin by describing the thresholds used to classify the error. We then present our proposed metric and discuss its formulation.

4.1 Classification of Error

We classify tracking errors into three main regions as shown in Figure 4: ‘acceptable’, ‘recoverable’ and ‘irreparable’ tracking regions. The regions we propose are based on the premise that some levels of error affect the perception of robustness in very different ways. For example, we suggest that time spent with tracking in the ‘irreparable’ tracking region is disproportionately more harmful to the perception of robustness than time spent in the ‘recoverable’ tracking region. Similarly, our assumption is that the few very noticeable errors are much worse for the sense of robustness than many more nearly imperceptible errors.

Note also that the definition of robustness discussed in the introduction is tied closely to the idea of a breaking point. That is, robustness is focused on the ability of systems to continue to operate. The thresholds and regions we define are based around breaking points, with the regions separated into distinct tiers of failure. This is also the reasoning behind the naming convention applied to the regions.

We have defined our thresholds to capture what we consider to be three variations of error. The lowest level of error lies in the ‘acceptable’ region, which we define to be errors which are largely unnoticed by users. The values which lie in the ‘recoverable’ region are then clearly noticeable by users. We have defined the errors in the ‘irreparable’ region to be errors in which the tracking is either lost and frame to frame tracking has ceased, or the error is exceptionally noticeable. The exact boundary of the regions are defined

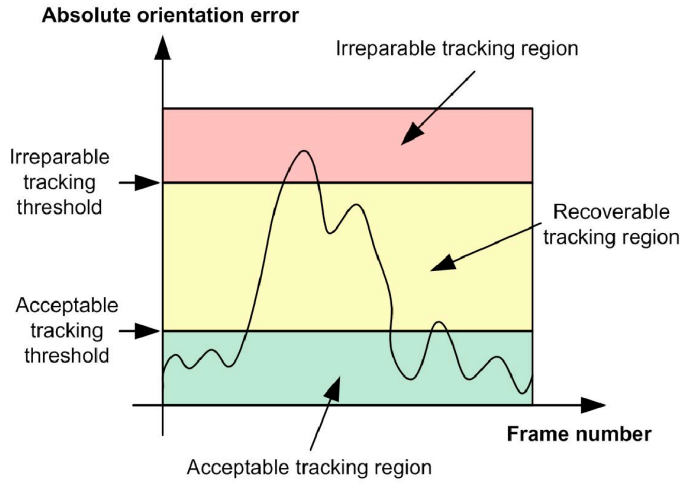


Figure 4: An illustration of acceptable, recoverable and irreparable tracking regions, and acceptable and irreparable tracking thresholds for an absolute orientation error graph

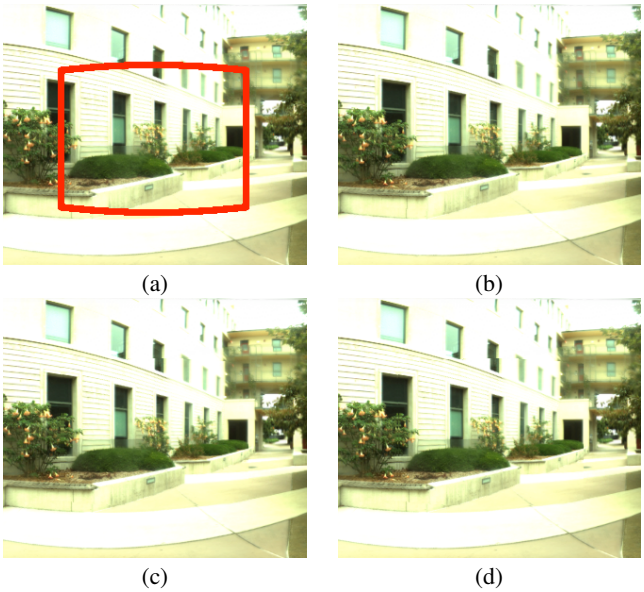


Figure 5: Tolerable range of the acceptable threshold (a) view frustum area (b) 0.3° shift (c) 0.5° shift (d) 0.7° shift

by two thresholds, the 'acceptable' tracking threshold and the 'irreparable' tracking threshold. We will define how we derived the values for these thresholds in the remainder of this section.

Acceptable Tracking Threshold The acceptable tracking threshold is effectively an upper bound on the errors which can occur in a system without causing any easily noticeable errors in the output. This threshold is based on the premise that for every application there is a certain amount of tracking error which is small enough that it does not greatly affect the qualitative evaluation of the system.

We are interested in orientation tracking and the resulting constructed panorama. Therefore, we base our acceptable tracking threshold on the distance in frame to frame error allowable before the errors become noticeable.

In order to obtain an estimate for the maximum angular distance which remains inconspicuous in a panorama, we selected several sections of a full environment map as seen in Figure 5 and manually introduced errors by shifting the images off by a range of degrees. The actual value depends on the size and quality of the generated environment map, and even the detail in the observed scenes would make some amount of difference to the evaluator's ability to notice errors in the results. An empirical evaluation for our example application showed that for a sphere map of 1536×512 pixels a shift of around 0.5° at the equator is negligible.

Note that this threshold is domain specific. As an example of an alternative, an adaptation of this metric focusing on AR displays might set this lower bound to a level at which augmentations have a drift or offset which becomes a distraction to users.

Irreparable Tracking Threshold The irreparable tracking threshold is an upper bound on the frame to frame error at which normal tracking breaks and some recovery method is needed. This threshold is based on the simple premise that there is a significant difference between the time spent with lost tracking and time spent with tracking which is generally operable, but contains some errors (potentially a constant offset or otherwise systematic tracking deviation).

For our analysis the tested systems all share the same code base, and therefore, the irreparable threshold for all systems was set to the point at which Envisor [4] using only frame to frame feature tracking is just about able to run successfully.

To determine this value we collected a number of samples using the previously mentioned PTU and camera configuration. This data consisted of video taken of an indoor environment, with a generous number of features to enable tracking. In each video the camera is rotated around the vertical axis at an exact rotational speed, using the camera mounted PTU as described above. At this time only yaw rotational data has been collected. We consider this generally sufficient as performance for combinations of yaw and pitch are generally similar, and high speed roll is seldom a large issue in AR. This testing determined the irreparable tracking threshold for our system to be 56°/second. As our tracking system runs at 48.08ms/frame this gives a maximum distance per frame of about 2.69°.

While the acceptable tracking threshold is based on the general application area (panorama acquisition), this bound is dependent on the specific systems tested. Therefore the value used in this study is useful for the tracking systems tested, but new values need to be derived on a per system basis. We are making all data sets used in our testing publicly available to ease the process of determining this value for future systems.

4.2 The Robustness Metric

Using the thresholds discussed previously, we are now able to determine the percentage of time each tracking system spends in each classification region on average. This percentage is represented in Eq. 1 by N_T, N_A, N_R and N_I, where N_T is the number of total frames, and N_A, N_R and N_I denote the numbers of frames belonging to the acceptable tracking region, recoverable tracking region, and irreparable tracking region, respectively.

Our work is based on the assumption that the time spent in each region contributes differently to the overall sense of the robustness of the system. To reflect this in our metric, we have included weighting factors (α, β, and γ). The higher the value of the weighting factor the more detrimental the time spent in that region is to the sense of robustness R.

$$R = 1 - (\alpha \cdot \frac{N_A}{N_T} + \beta \cdot \frac{N_R}{N_T} + \gamma \cdot \frac{N_I}{N_T}) \quad (1)$$

In order to determine optimal values of the weighting factors to approximate our observations for R from the qualitative user evaluations, we performed a brute force optimization testing a large range of values for the weights. The optimization worked by cycling through all sensible value combinations for the weights in small discrete steps. For each weight combination, for each panorama, we then determined the estimated robustness using those weights. This provides a level of error based on the difference between the average of the normalized ratings of the expert users for that panorama and the estimated robustness using our metric, R. To determine the optimal set of weights we then minimized the sum of squared differences over all the panoramas. Note that, the panorama scores were placed in a range of zero (worst) to one (best) for this evaluation, by subtracting one (the lowest score) and then scaling by a factor of one over six (the difference between the highest and lowest scores). This scaling was done to avoid having an arbitrary value of seven for the highest score produced by our metric.

In order to obtain a full range for computing the weighting factors, we added an additional artificial data set which is the performance produced by an ideal system. This contributes two things: it serves to ensure that there are sufficient samplings from the low-error end of the spectrum in order to ensure the α values have sufficient information to converge on a meaningful value. Intuitively, it also serves to provide a grounding to the metric to ensure that as the performance increases, we do in fact approach a perfectly robust system.

Note that for determining the weighting factors we used data from only three of the four methods. These were Envisor with no recovery, Envisor with constant recovery, and Envisor with constant recovery and pre-scanning. Using only three of the four methods to derive the weights allows us to evaluate the accuracy of our metric using the fourth system.

We chose not to use the data from Envisor with selective recovery, as it can be considered the most non-trivial case of mapping the quantitative to qualitative data as seen in Table 1. Therefore, if our metric is able to map the quantitative analysis of this method to the qualitative evaluation, then it is a very strong indication of the usefulness of our metric.

For our evaluation of the first three methods, over all 135 data sets (again, each data set is obtained as a combination of many inputs from 6 experts), we obtained weightings of 0.030, 0.56, and 0.83 for α, β, and γ, respectively. We limit the obtained results to two significant digits as these values are somewhat flexible; additional data sets would help to further refine our estimates.

While the proposed values were determined using a single application area, namely panorama construction, the weights themselves are designed to be general, and we predict that they will not need to be recomputed when we will branch out to different domains. As they are based on the time spent in each tracking region, they can be reused provided the values for the thresholds have been determined as previously discussed.

5 Evaluation of the Metric

To demonstrate how the performance of the four different versions of Envisor fits into our three-region model, we show histogram distributions of time spent in the acceptable, recoverable and irreparable regions for each tracking method. From Figure 6, we can observe that Envisor with constant recovery and pre-scanning carries out more stable and accurate tracking compared to the other methods. On the other hand, the original version of Envisor shows many input frames classified in the irreparable tracking region.

Note that as mentioned previously, we did not use the data from Envisor with selective recovery when determining the weights of our metric. We can therefore give an estimate of how well our metric works, by determining the difference between the evaluations of

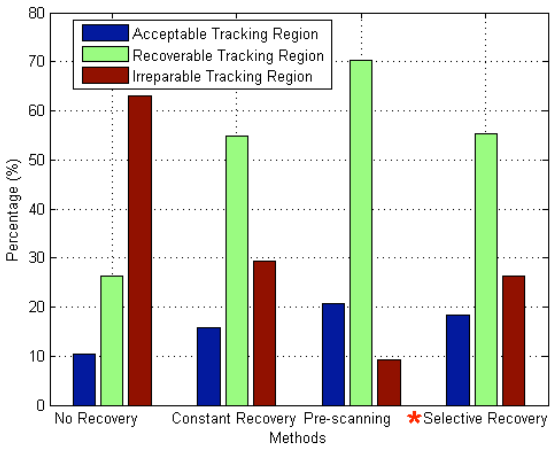


Figure 6: Histogram distributions of time spent in acceptable, recoverable and irreparable regions for each tracking method. 45 video sequences of indoor and outdoor scenes were tested for each method. Each sequence was about one minute (1800 frames) long. Selective recovery is starred as the data was not used for the original formulation of our metric, but was later used to test its validity.

Table 2: First row, the mean measurement of robustness for each system over all users and over all datasets. Second row, the robustness ratings from the live evaluation, Third row, the projected robustness generated by our metric, Fourth row the scaled projection from our metric for illustrative purposes (multiplied by seven). (NR: original version of Envisor (No Recovery), CR: Envisor with constant recovery, SR: Envisor with selective recovery, CRS: Envisor with constant recovery and pre-scanning). Note that the scale for the evaluations is from one (worst) to seven (best). The projected value from our metric does not have units, so it is important to consider how closely the projected values reflect the relative distance between rankings.

	NR	CR	SR	CRS
Panorama evaluation	2.03	3.12	3.54	5.41
Live evaluation	1.63	3.95	4.03	6.05
Projection from metric	0.33	0.44	0.47	0.52
Scaled Projection from metric	2.28	3.11	3.27	3.67

the experts for the robustness of this system and the result of the projected robustness of our metric based on the ground truth data sets.

The results of our projections are listed in Table 2. Note that apart from differences in absolute values and slight proportional differences for constant recovery with pre-scanning, the projections were very accurate, and in all cases the relative differences were well captured. In the case of pre-scanning, our results can most likely be explained by the fact that it was relatively the best of the tested methods despite the fact that as seen in Figure 6 it spends most of its time in the recoverable region. Note that the scale of our evaluation is more compact than that of the original panorama and live evaluation results. This is due to the inclusion of an artificial dataset of a perfect system in the construction of the weights. Again this was done to ensure the metric converges meaningfully toward a value of one as systems become more robust.

The final averages are listed alongside the average scores for the panorama evaluations and the average scores from our metric given absolute orientation error from ground truth. The fact that the live user evaluation which was obtained separately from the panorama evaluation, matches so closely to the metric and panorama evalu-

ation data sets, is a strong indication that our metric is useful (cf. Table 2).

6 Discussion

We have presented a method for mapping the quantitative performance of systems to the qualitative analysis of their performance, with the goal of eliminating the need for performing subsequent user evaluations in the comparison of orientation tracking systems.

While our metric can estimate the qualitative robustness of orientation tracking systems, there are still requirements in terms of the quantitative data needed. Specifically, we require that the user have some estimate for the upper level of the performance of their system. This requirement is necessary to determine the irreparable tracking threshold for our analysis. In order to ease this requirement we have made the speed tests used in our analysis publicly available [1] . Note that this data is also generally useful for determining the qualitative robustness of orientation tracking systems.

We also require that the user have representative input sequences with associated ground truth in order to determine the quantitative performance of their systems relative to ground truth. In order to ease this requirement and provide an example for a range of different environments over which robustness could be evaluated, we have also made the data used in this evaluation publicly available. This data consists of two parts. First, we provide the video data and associated ground truth used in the evaluation of our metric, which can be used directly to test the performance of additional systems. Secondly, we provide the head orientation information which we collected for this study. The orientation data we provide can be replayed with a PTU at a later time under the conditions being evaluated. For example if users are interested in testing the performance of their tracking systems under changes in illumination, they will be able to replay the motion paths under various lighting conditions.

We believe this orientation data is generally helpful as the movement of the camera should reflect realistic movements of human users, as opposed to simple or random movement paths. This is true for both our metric and any quantitative analysis. The 45 sets of orientation data used in the analysis presented in this paper were obtained from a larger data set consisting of samples from 23 participants over a set of nine tasks. The participants were campus students with little to no experience with AR. Each subject was given a small monetary compensation for their participation. The participants were asked to perform their tasks while wearing a hat with an attached orientation tracker (InterSense InertiaCube2 [9]). Between each run the tracker was calibrated to ensure that the motion data collected accurately matched the view of the participant. This was accomplished by having the students boresight an object at a known height and distance from their starting position.

Of the nine tasks, five varied in duration and four were a minute long. For the purposes of evaluating our metric we used only the minute long samples. This reduced our pool to 92 usable sets. Of the four tasks, three are essentially the same observational task repeated multiple times. This essentially means there were two types of tasks, one a casual exploration task, and the other a searching task. From these we randomly selected 45 sets to be used in further evaluations.

As mentioned before, we used these 45 sets of realistic head motions for play-back in different locations to acquire our test panoramas. We should mention a minor disadvantage of the PTU system we employed for this purpose. It stems from a very small lag in the number of quickly executed commands. Therefore there was a small amount of filtering applied to the data from the users. The effects of such filtering are minimal, however, and this smoothing only affects how accurately the input path given to the PTU matches the original user's head motion. It does not affect the accuracy of the ground truth values obtained, as the exact position of the PTU was sampled for each frame.

7 Conclusions and Future Work

In this paper we introduced a new metric for determining the qualitative robustness of systems from a quantitative analysis of their performance. Our metric requires only a set of representative video sequences along with ground truth data for each frame in terms of yaw, pitch, and roll. We provided evidence of the usefulness of our metric through the correct prediction of the relative robustness of an uninvolved tracking system.

Future work will evaluate the metric using additional systems. While we believe that all of the tracking systems presented are sufficiently different for evaluating our metric, it will be helpful to analyze the performance of externally developed systems.

For future work we would also like to evaluate the system over additional application domains, such as overlaying and displaying annotations in the scene. In this case, assuming Envisor was used for tracking, the only value necessary to change would be the acceptable tracking threshold. This is assuming that there is a different level of quality expected when displaying annotations than when constructing an environment map.

An additional goal of future work is to adapt our metric to be useful for a 6 degree of freedom (DoF) tracking solution. For this, we will need to collect additional 6 DoF ground truth data which could be acquired e.g., by using a robotic system. The central idea of the metric should be extensible to 6 DoF. At the base of the metric we are interested in classifying the quality of tracking and understanding the relationship of those classifications to qualitative robustness. The general idea of classifying the error is certainly flexible enough to extend to 6 DoF. However, the exact implementation needs to be clearly defined. Currently, we imagine two possible options. First we could attempt to formulate the error for both position and orientation in terms of a single value. Alternatively we could use the values from both errors separately and attempt to perform the optimization of the weighting factors over this tuple. The second option is more descriptive, although it would require a much larger set of data in order to determine the values of the weights.

The idea of the thresholds is also generally applicable, and could be applied to 6 DoF work. It is reasonable to assume that there are thresholds for both position and orientation errors under which the misregistration is largely unnoticed. Similarly, there are also appropriate levels of error at which we can place the upper threshold, although this error may be slightly more difficult to quantify. Particularly, positional error may have to be defined relative to some scale, as nearby objects are more influential to the tracking quality than objects far away from the user.

Acknowledgements

This work was supported in part by a research contract with KIST through the Tangible Space Initiative Project, and NSF CAREER grant #IIS-0747520. The authors also wish to thank Stephen DiVerdi for his development of the original Envisor system and his continued help.

References

[1] K. Bowyer and P. Phillips. Overview of work in empirical evaluation of computer vision algorithms. In *Empirical Evaluation Techniques in Computer Vision*, pages 1–11, Los Alamitos, CA, USA, 1998. IEEE Computer Society Press.

[2] C. Coffin, S. Kim, and T. Höllerer. Evaluation of four methods for real time panorama acquisition. Technical Report 01, Santa Barbara, CA, 2010. http://www.cs.ucsb.edu/research/tech_reports/reports/2010-01.pdf.

[3] C. Coffin, S. Kim, and T. Höllerer. Evaluation of tracking robustness in real time panorama acquisition. In *IEEE Virtual Reality Conference (VR)*, pages 259 –260, Jan 20-24 2010. poster.

[4] S. DiVerdi, J. Wither, and T. Höllerer. Envisor: Online environment map construction for mixed reality. In *IEEE VR*, pages 19–26, 2008.
[5] DPerception. http://www.dperception.com, June 2009.
[6] S. Gauglitz, T. Hollerer, P. Krahwinkler, and J. Rossmann. A setup for evaluating detectors and descriptors for visual tracking. In *ISMAR*, pages 185–186, 2009.
[7] M. D. Heath, S. Sarkar, T. Sanocki, and K. W. Bowyer. Robust visual method for assessing the relative performance of edge-detection algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1338–1359, 1997.
[8] Intel. Open source computer vision library. http://www.intel.com/technology/computing/opencv/, May 2007.
[9] InterSense. http://www.intersense.com/, June 2009.
[10] G. Klein and D. Murray. Improving the agility of keyframe-based SLAM. In *Proc. 10th European Conference on Computer Vision (ECCV'08)*, pages 802–815, Marseille, October 2008.
[11] S. Lieberknecht, S. Benhimane, P. Meier, and N. Navab. A dataset and evaluation methodology for template-based tracking algorithms. In *IEEE International Symposium on Mixed and Augmented Reality 2009*, pages 145 –151, Oct. 19-22 2009.
[12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, Oct. 2005.
[13] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *Intl. Journal of Computer Vision*, 73(3):263–284, 2007.
[14] G. Reitmayr and T. W. Drummond. Going out: Robust tracking for outdoor augmented reality. In *Proc. Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'06)*, pages 109–118, October 22–25 2006.
[15] K. Satoh, K. Takemoto, S. Uchiyama, and H. Yamamoto. A registration evaluation system using an industrial robot. In *ISMAR'06: Proc. 5th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 79–87, Washington, DC, USA, 2006. IEEE Computer Society.
[16] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Intl. Journal of Computer Vision*, 37(2):151–172, 2000.
[17] TrakMark. Benchmark test schemes for ar/mr geometric registration and tracking methods. http://trakmark.net/, Aug. 2010.
[18] Z. Zhang. A flexible new technique for camera calibration. *Transactions on PAMI*, 22(11):1330–1334, 2000.