Evaluation of four methods for real time panorama acquisition

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

Abstract

In this paper, we present an analysis of four orientation tracking systems used for construction of environment maps. This paper focuses on the methodology used to collect data for a detailed analysis of these tracking methods. This analysis consists of three parts. Two methods of qualitative analysis are used, one based on the panoramas generated by each system, and another based on an expert evaluation of a live demo. A ground truth evaluation of the systems is also performed. Finally, we present an analysis of the performance of these methods, and a comparison of the results of each test.

1 Introduction

In this paper we present an analysis of the robustness of four existing orientation tracking systems. These methods are variations on an existing orientation tracking system Envisor [2]. A detailed discussion of these methods can be found in Section 2. In order to obtain an accurate understanding of the robustness of each system, we performed a quantitative analysis, an analysis of the performance output and a live evaluation. For the quantitative analysis, we collected distance to a known ground truth over a large set of input videos. However, ground truth error alone does not provide insight into the perceived robustness of the system. We obtain this through two qualitative analyses of the systems.

The first qualitative analysis was based on the final output of the systems, in our case, a set of environment maps. The second qualitative analysis focused on the results of a live expert evaluation of each system. While the analysis of the results allows for a larger breadth with respect to samples, expert analysis provides confirmation of the trends seen in the analysis of the results.

Robustness in general terms is the quality of being able to withstand stresses, pressures, or changes in procedure or circumstance. A system is generally considered robust if it is able to cope well with input variations with a minimal amount of damage, alteration, or loss of functionality. Robustness of an algorithm in computer science is described loosely as the ability for the algorithm to continue to function despite abnormal input.

We define robustness for the purpose of computer vision tracking as the ability to cope with input irregularities, such as fast motion, sudden turns, blurred imagery, over or under exposed areas, as well as other artifacts leading to the loss of tracking. Similarly to fuzz testing, difficult input values need to be given to the tracking system in order to form an evaluation of robustness. Ideally, the data should contain a range of samples of varying difficulties for the situations being evaluated, e.g., lighting changes, occlusions, fast movements, etc.

2 Related Work

Up to now, a wide variety of tracking systems using various kinds of sensors have been investigated for augmented/mixed reality applications. There has also been extensive evaluation of lower level interest point detectors, feature descriptors [10] [9] [8] 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.

3 Systems Evaluated

In this section we describe each of the methods we used for our evaluation in greater detail.

3.1 Envisor: Online Environment Map Construction

DiVerdi et al. presented Envisor, a system for online construction of environment maps in new locations [2] [3]. In order to construct an environment map, they use a vision-based hybrid orientation tracking mechanism which provides relatively drift-free orientation registration. The tracking mechanism is composed of two phases, one of which is frame-to-frame relative rotation tracking using Shi and Tomasi's feature detector [11] and a pyramidal version of Lucas and Kanade's optical flow algorithm [7]. The second phase is landmark-based absolute orientation tracking which adds landmarks to the frame to frame feature tracking system to combat drift during long tracking runs. Using the orientation tracking techniques described here, they construct an environment map of the surrounding scene online and automatically. That is, the tracked video is projected into a cubemap frame by frame. However, small gaps are likely to occur unless the user is very careful about complete coverage. Thus, a texture diffusion technique to blend surrounding pixels into those gaps is applied, reducing their visual impact.

3.2 Envisor with Constant Recovery

The original version of Envisor is vulnerable to several factors, such as varying lighting conditions, drastic speed changes, or insufficient texture information on surrounding scenes. That is, there are many cases where tracking is lost. Thus, we extend Envisor by presenting methodology for camera orientation relocalization, using virtual keyframes for online environment map construction [6]. Instead of relying on real keyframes from incoming video, we enable camera orientation relocalization by employing virtual keyframes which are distributed strategically within an environment map. After shading correction, we relocalize camera orientation in real-time by comparing the current camera frame to virtual keyframes. While expanding the captured environment map, we continue to simultaneously generate virtual keyframes within the completed portion of the map as descriptors to estimate camera orientation. We implement our camera orientation relocalizer using a fragment shader for real-time applications. With the help of the pose relocalization, we can enable a user to generate an environment map independently of the camera path the user chooses to trace over the environment. That is, we can recover from tracking failures simply by returning to a general area that was previously captured.

3.3 Envisor with Pre-Scanning

While the addition of tracking recovery is an improvement, the performance of the live environment mapping process is still subject to imperfections and robustness problems. In order to span the whole bandwidth of robustness from medium-poor to excellent, we were interested in adding a highly robust system to our set of methods. We are aware of very recent efforts in the research community to

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

derive a new method for the panorama acquisition problem that are far less computationally expensive and therefore faster and more reactive than the systems used in this evaluation [12]. However, since we cannot currently have access to the method, we decided to add a third version of Envisor, creating higher robustness by use of additional a-priori information. This version of Envisor is the same as the Envisor with constant recovery except that we use pre-scanning of the entire environment to generate virtual keyframes in advance. This is akin to generating an a-priori model of the environment for model-based tracking. One problem of Envisor with constant recovery is that if it cannot generate valid virtual keyframes, recovery might not produce a good panorama image. Thus, for more stable and accurate tracking, we generate virtual keyframes and the corresponding absolute orientations in advance, and make use of them during the actual tracking procedure. This strategy is applied for the creation of panoramas from sample input camera motions as well as for the live expert evaluation. The system behaves exactly like Envisor with constant recovery, only more robustly. In our experiments, to generate accurate a-priori virtual keyframes and the corresponding absolute orientations, we use a camera on a pan tilt unit by moving it from -15° to 15° in pitch and from -180° to 180° in yaw.

3.4 Envisor with Selective Recovery

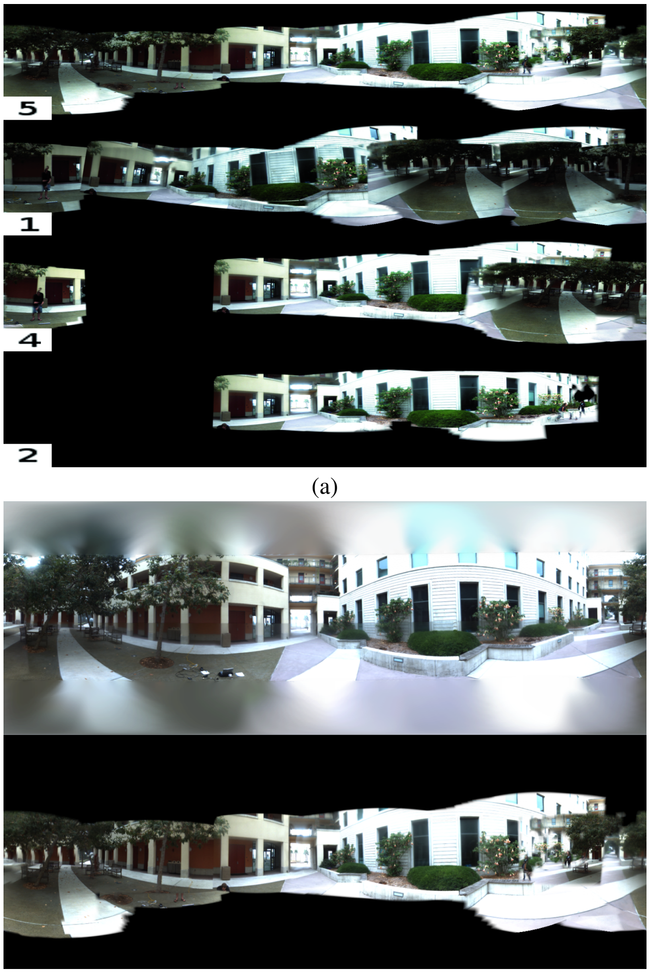
While Envisor with constant recovery is intended to recover its camera pose with virtual keyframes at every frame, this new variant, Envisor with selective recovery, attempts recovery only when it detects that tracking is lost. In this variant, we also replaced the SURF [1] feature descriptor for landmark features stored in the sphere map with faster but less invariant image patches. Intuitively, we expect this method to have slightly better performance than constant recovery as it is a bit more economical with regard to computing resources.

4 Data Collection

In order to get a meaningful data set for our evaluation, we collected head orientation information from 23 casual users over a set of 9 tasks. Of the 9 tasks, 5 had consisted of largely varying durations and amount of area covered. This reduced our pool to 92 usable sets. From these we randomly selected 45 sets for which we would record ground truth information. The samples in our collection of 45 were taken from users performing two distinct tasks. The subjects 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 [5]). Between each run the tracker was calibrated in order to ensure that the motion data collected accurately matched the view of the participant. This was done by having the students boresight an object at a known height and distance from their starting position.

For the first task, the participants were asked to look around them for a full 60 seconds in order to observe as much as possible about their surroundings. After this they were asked a question about their surroundings as motivation for continued searching. This task was repeated 3 times. They were informed that the questions asked would not be related, in order to avoid, as much as possible, a question changing the student's focus. For example, our first question was "How many trees are around you?" We made it clear that the questions would not be similar to each other.

The second set of data was collected from a simple search task. In this task the object the user was looking for was not present in the scene. While some users became suspicious, many were not as the previous search tasks had all contained the sought items. All of participants continued to search their surroundings for a full minute, at which time they were asked to stop. The data collected is therefore representative of a casual user going about both observational and search tasks.



(a)

(b)

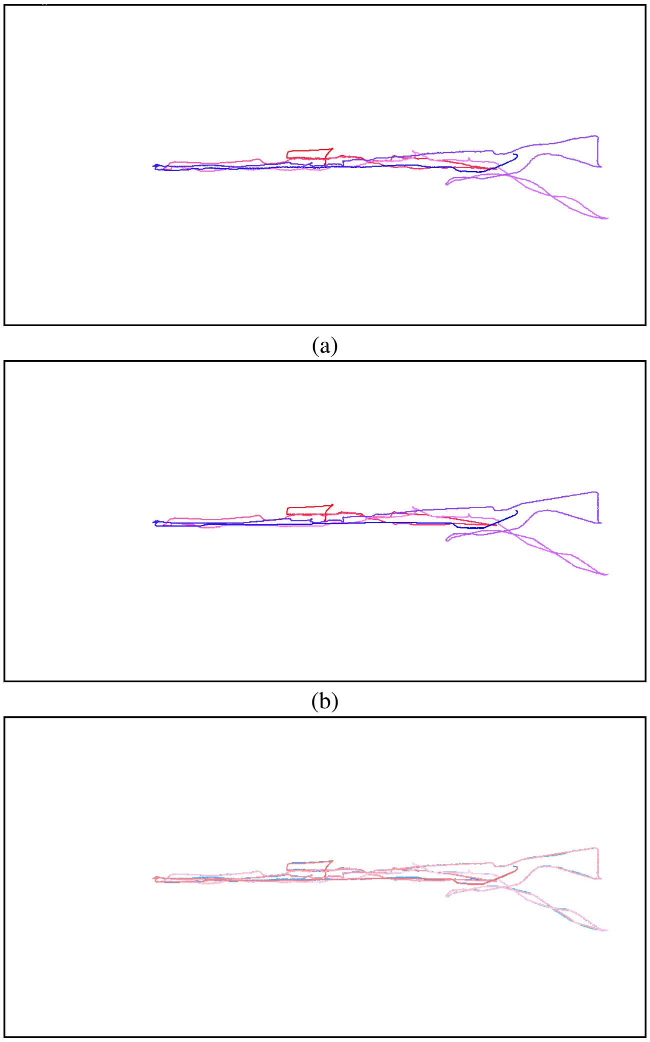
Figure 1: Users were asked to rank panoramas generated by each of the four methods used, a) the interface they used b) users were able to click on an item to compare it to a ground truth panorama.

5 Evaluation

In this section we discuss the details of each of the three evaluations performed on the systems tested.

Tracking Error The ground truth values for each video were obtained by mounting a camera on a PTU-D46 pan tilt unit [4] from Directed Perceptions. The pan tilt unit has an upper speed limit of 300° per second and a resolution of 0.0514°. We are therefore able to not only replay head movement at matching positions but also at the speeds recorded. This allows us to retain motion blur and other real-life imaging artifacts. We are also able to replay the same orientation information in multiple locations, providing a sizable representative set of tracking environments. A measure of absolute tracking error was then obtained by using each video as input to the tracking systems and comparing the resulting positional updates with the ground truth input to the PTU.

The only disadvantage of the PTU stems from some lag in the number of quickly executed commands. Therefore there was a small amount of filtering applied to the data from the users. Results of such filtering are minimal, and were applied before the video samples were captured by the PTU, and therefore do not affect the



(a)

(b)

(c)

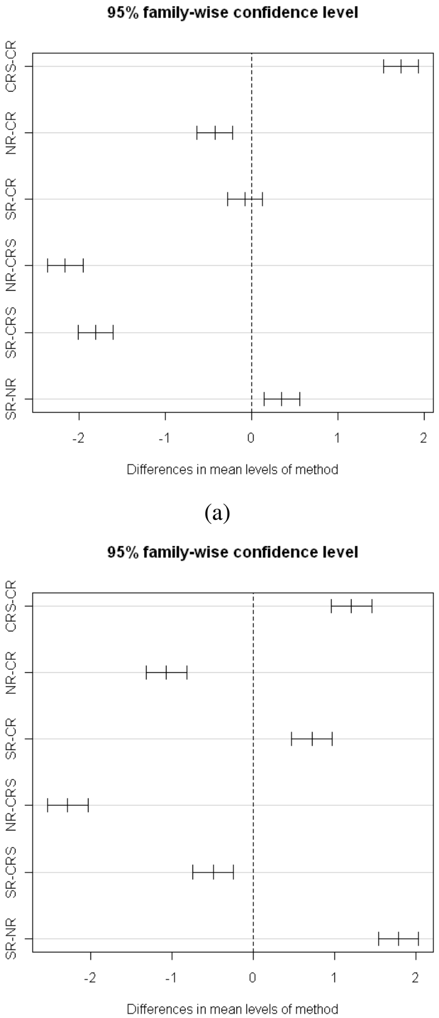
Figure 2: a) Yaw and pitch values for a single task represented in a sphere map, with points ranging from red to blue with respect to time along the path. b) the results of filtering on the path seen in a). c) an overlay of the difference between the two plots with the filtered path in red and the original in blue.

accuracy of the ground truth values obtained. A sample of the applied filtering can be seen in Figure 2.

Result Evaluation For the analysis of the panoramas produced by each method, we designed a simple ranking program, with the user interface seen in Figure 1. For each data set, every expert was shown the panoramas generated by every method simultaneously. They then selected each panorama and rated them on a scale of 1 to 7. The assigned scores were then displayed on the left-hand side of the associated panorama. To assist in ranking, users were able to compare each panorama to a ground truth panorama as seen in Figure 1(b).

We found that the evaluations were very consistent among users. We then normalized the results of each users data by subtracting each vote by the minimum vote for that user and dividing the difference by their overall range of votes. From this data we were then able to determine an average ranking over all users, for each method applied to each data set.

The average ratings of the users for each set of environment maps was very consistent over 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 $p < 0.0001$. The results from a set of corresponding Tukey Post-hoc evaluations are shown in Figure 3. The graph shows pairwise comparisons between the scoring of each method with distance from the center line indicating increased difference between the methods. Methods not interesecting the vertical line are significantly different. Note that all methods were significantly pairwise different with the exception of constant recovery compared with selective recovery in the indoor case. This is important as it implies users were able to differentiate methods as more or less robust.



(a)

(b)

Figure 3: Results of a Tukey multiple comparison of means given an ANOVA comparison for a) indoor and b) outdoor showing that the evaluations of the panoramas are statistically different for each method with the exception of constant and selective recovery, which fits with our expectations. (NR: original version of Envisor (No Recovery), CR: Envisor with constant recovery, SR: Envisor with selective recovery, CRS: Envisor with pre-scanning)

Live Evaluation For the live evaluation, we had 5 expert users evaluate each system in a live demonstration. In order to ensure a fair comparison, we had each user rank each method four times for a total of 16 randomly ordered runs. The evaluators were asked to rank each system from 1 to 7, and then we normalized and averaged the results for this evaluation similar to the panorama evaluation results.

Table 1: First row, relative distance to ground truth in degrees. Second row, ratings assigned to the panorama output data (scale 1 poor to 7 perfect). Third row, the robustness ratings from the live evaluation (scale 1 poor to 7 perfect). (NR: original version of Envisor (No Recovery), CR: Envisor with constant recovery, SR: Envisor with selective recovery, CRS: Envisor with 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

6 Results

We present a comparison of each of the methods in Table 1. The values of orientation error are measured in degrees while the values for each of the evaluations are on a scale of one to seven with seven being the best.

Clearly the original version of Envisor is the worst of the tested methods, and Envisor with pre-scanning performs the best. However the selective and constant recovery methods are not so clearly differentiated. From the qualitative evaluations alone, selective recovery is the better of the two methods, however this is not the case based on the measure of absolute orienation error.

While the values of the results are not directly comparable due to differences in units, the relative distances are useful. These distances are very similar for both of the evalution methods, with selective and constant recovery being very closely related with respect to the other two methods. The difference between the qualitative and quantitative measurements implies that there is not a direct linear mapping between error and qualitative robustness. Therefore determining one value from ther other requires a more complex mapping.

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] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[2] S. DiVerdi, J. Wither, and T. Höllerer. Envisor: Online environment map construction for mixed reality. In *IEEE VR*, pages 19–26, 2008.

[3] S. DiVerdi, J. Wither, and T. Höllerer. All around the map: Online spherical panorama construction. *Computers and Graphics*, 33(1):835–846, 2009.

[4] DPerception. http://www.dperception.com, June 2009.

[5] InterSense. http://www.intersense.com/, June 2009.

[6] Reference suppressed due to double blind reviewing.

[7] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[8] 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.

[9] 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.

[10] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Intl. Journal of Computer Vision*, 37(2):151–172, 2000.

[11] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, pages 593–600, 1994.

[12] Daniel Wagner, personal communication, Sept. 2009.