Using a Panoramic Camera for 3D Head Tracking in an AR Environment

Björn Giesler, Tobias Salb, Rüdiger Dillmann,
IAIM, University of Karlsruhe (TH), Germany
{giesler,salb,dillmann}@ira.uka.de

Tim Weyrich,
ETH Zürich

Abstract
For Augmented Reality using a pair of transparent 3D glasses, a precise and fast method for head tracking is required, to determine the user's position and direction of gaze in all six degrees of freedom. The methods currently available require expensive external sensors and have small working areas and/or other limitations. We propose a method that uses a panoramic camera that is mounted directly on the user's head, combined with cheap, easily mountable passive artificial landmarks. The panoramic camera uses a paraboloid mirror, which allows for interesting algorithmic simplifications. The system has been tested both in simulation and in reality and shows promising results.

Keywords:
Augmented reality, head tracking, panoramic camera.

1 Introduction

Augmented Reality (AR, for short) is the layering of Virtual Reality elements (such as 3D models or markers) over a viewer's image of the real world. This can be achieved in a number of ways, the most popular of which is to have the user wear a pair of semi-transparent 3D goggles. In Karlsruhe, we are constructing an AR system for human-robot interaction, to make it easy for the user to immediately see the interpretation of his or her actions by the computer. To be able to overlay elements of simulation with the view through the 3D glasses, it is necessary for the system to very precisely keep track of the viewer's head position and direction of gaze.
The tracking must meet real-time requirements to prevent lagging of the virtual elements. Since it is very difficult to model human head motions, the benefit of using motion-prediction techniques is very limited. For real-world usability, the

tracking should also be able to cover as wide an area as possible, should be very easy to set up, require little or no modification to the environment and, last but not least, should not be prohibitively expensive.
Some existing approaches to this problem make use of external camera systems, such as the commercially available *POLARIS* [NDI02]. Such systems, while often very precise, have a limited working space and are therefore more suitable for usage in stationary applications, such as surgical aid [Sal00]. Others require extensive modification to the environment, such as the University of North Carolina's *HiBall* tracking system [WB+99].
Our novel approach consists of a panoramic camera that is affixed to the AR glasses and tracks artificial environmental features. This approach, using only a single camera, requires relatively little computation and therefore comes close to real-time requirements; it is also inexpensive. We are currently tracking artificial targets that can be mounted very quickly and easily, so very little environmental modification is necessary.

2 Properties of the System Components

We have decided to use a system that uses artificial landmarks instead of natural environmental features, because we consider 6-DOF position reconstruction using natural features to be still too difficult and most of all too costly in terms of processing time to meet with the required real-time constraints. Therefore, we can design both the features that we are recognizing and the sensors that we recognize them with to be perfectly adapted to each other.

2.1 Properties of the Panoramic Camera

A panoramic camera is a system using a CCD camera taking pictures of a convex mirror that reflects a distorted view of the environment. This mirror can be a half-sphere, cone, paraboloid or any other regular convex body. Using panoramic cameras for position reconstruction is not in itself a novel approach; such cameras have been successfully used, for example, for position estimation of mobile robots (with a conical mirror: [FS+98], [FS+97]; with a spherical mirror: [G+V99]). However, so far work has been limited to two-dimensional reconstruction. For three dimensions, a novel approach is needed, and the camera should fulfill certain conditions.

Position reconstruction from a single image and known landmarks is essentially triangulation, since information about distance to the landmarks is not known. For triangulation, it is necessary to take multiple bearings from a single point in space. That means that the *rays of sight* that hit the landmark centers

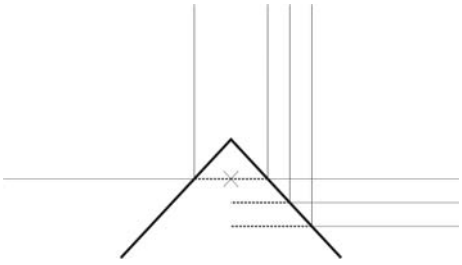

Figure 1: Deflection of rays of sight by a conical mirror.

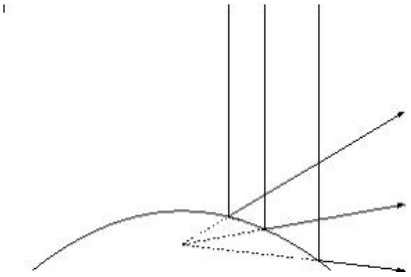

Figure 2: Deflection of rays of sight by a parabolical mirror

have to emanate from, or meet in, a single point.

For two-dimensional position reconstruction, it is sufficient to locate environmental features that lie in the plane of reconstruction, or in a plane parallel to that. A conical mirror is perfectly suited for this application; as can be seen from **figure 1**.
To make three-dimensional position reconstruction possible from a single image, it is necessary to find environmental features distributed in all three dimensions around the camera; it must be possible to take bearings toward multiple non-coplanar environmental features.
A parabolical mirror has this important property (see **figure 2**): All rays of sight that hit the mirror parallel to the mirror's symmetry axis are reflected in such a way that they all seem to originate in the focal point of the paraboloid. Therefore, we can relate all bearings to the focal point.

A parabolical mirror does have the disadvantage that it distorts images of objects in a non-trivial way; while a circle reflected in a conical mirror becomes an easily-matchable ellipsis, a paraboloidal mirror distorts it into an egg shape. Furthermore, the center of gravity of the pixel set that forms an object's image does not coincide with the object's center of gravity in cartesian space. It is therefore not trivial to match an object in the camera image or even its center of gravity with numerical methods.
However, figure 2 shows also that in a picture taken of a parabolical mirror, there is a simple correlation between azimuth and elevation of a point in space and the image coordinates of its picture. Since we are reconstructing the position by triangulation, we can perform all our calculations in *ray space*, that is, the vector space spanned by all rays of sight emanating from the mirror's focal point. Therefore, the reconstruction of object shapes should take place in ray space as well. We achieve this by conic matching; the method is described in section 3.1.

The camera currently in use has a mirror covering an angular area of 360° azimuth and 60° elevation. The mirror's image is taken by a 640x480 pixel CCD and encoded as a PAL signal. The resolution is very low considering the large section

Figure 3: A 360° image and its cartesian reconstruction. The image resolution is 640x480 pixel; this can be seen in the coarse reconstruction results, especially towards the top of the picture.

of the environment that is depicted; **figure 3** shows an image taken with the camera and its cartesian reconstruction. However, simulation and experiments both show that the low resolution mainly poses problems for actual landmark recognition; the camera resolution has only marginal effect on the accuracy of the position reconstruction if sufficiently many landmarks are successfully recognized.

2.2 Properties of the Artificial Landmarks

To meet with the real-time requirements outlined in section 1, landmarks should be designed in such a way that we can find them very easily and quickly. We are currently using circular landmarks in the primary colors red and blue, as shown in **figure 4**. The landmarks are not uniquely coded in any way, because owing to the limited resolution of the panoramic camera they should be extremely simple in structure to be easily recognizable even if their image is only a few pixels large.

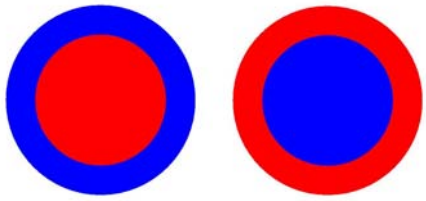

Figure 4: Circular landmarks in red with a blue border and blue with a red border. The landmarks are easily recognizable and have a defined color border circle that is important in conic matching.

The landmark colors make it easy to find the exterior and interior of the landmarks by examining only the red and blue channels of the RGB image stream (thus working on two separate binary images) and finding blobs. This results in two sets of blobs, one for each channel. We then calculate the size of each blob's bounding box and compare the bounding boxes in the red channel with those in the blue channel. If a 'red' bounding box lies within a 'blue' one or vice versa, and the sizes of the bounding boxes fulfil, within certain limits, the known size ratio of the inner and outer circles, a landmark has been found. **Figure 5** shows the basic process of landmark identification.

3 Landmark Recognition and Pose Reconstruction

The outlined process makes it possible to quickly find landmarks in the camera image. However, their *precise* location and shape in ray space is still not known. The camera's limited resolution necessitates the development of some scheme to achieve sub-pixel accuracy. As stated in section 2.1, the camera's parabolic mirror makes simple center-of-gravity calculation difficult; but in the course of this work, a method has been developed that is both more precise and makes excellent use of the camera's properties.

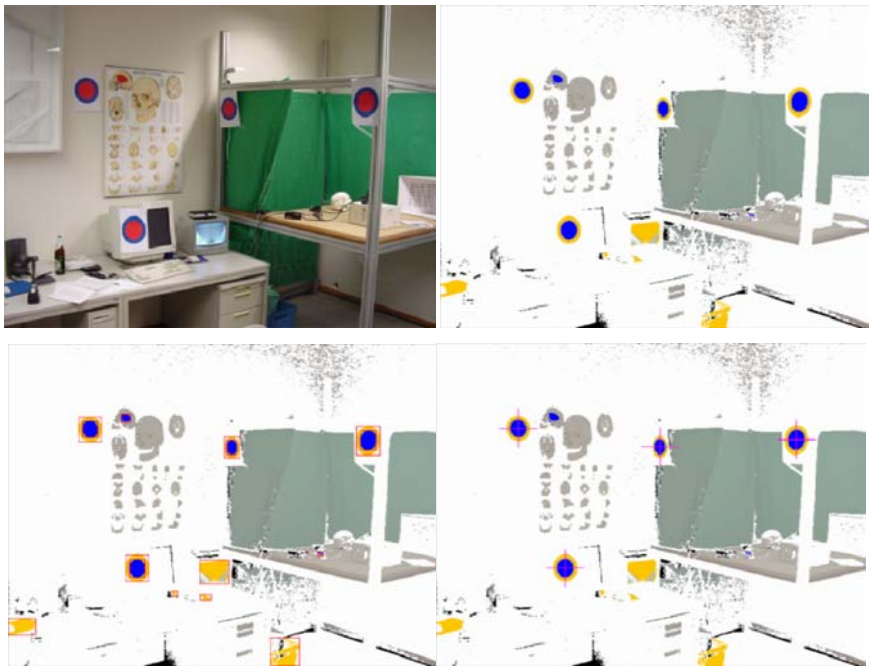

Figure 5: The process of finding landmarks in the camera picture. Top left: Original image. Top right: Finding red and blue areas. Bottom left: Calculating bounding boxes. Bottom right: Detecting landmarks where correct bounding box ratios are found.

3.1 Conic matching

Since working with the 2D representation of the landmarks is not practical due to the distorting properties of the parabolic mirror, we have decided to use the landmarks' 3D shape. Therefore, the border pixels are detected where a landmark's outer ring adjoins to the inner circle. It is known that these pixels must (within some margin of error) lie on a cone with an elliptical cross-section whose origin is the optical center of the camera. Therefore, rays from the optical center are constructed that go through the border pixels, and a least-squares error

minimization algorithm is used to construct a conic through the optical center that approximates the rays most closely. The conic's center ray is then used as a bearing for triangulation.

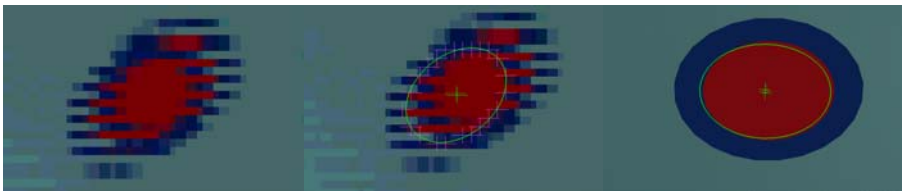

Figure 6: Border pixel detection and conic matching. The first and second images show the distorted camera picture; the jagged edges are due to the interleaved video signal. The third image shows the landmark in "real-world", i.e. cartesian coordinates. The two crosses mark the two possible normal vector roots.

3.2 Different Landmark Centers

As can be seen in **figure 6**, one final ambiguity must be resolved: The conic matching process delivers a very good estimate of the landmark's perimeter, but not of the landmark's center. If the landmark is not seen straight-on, there are two possible surface normals and therefore two possible centers: Let H be the set of all planes whose section with a given conic C is circular. Then H consists of two

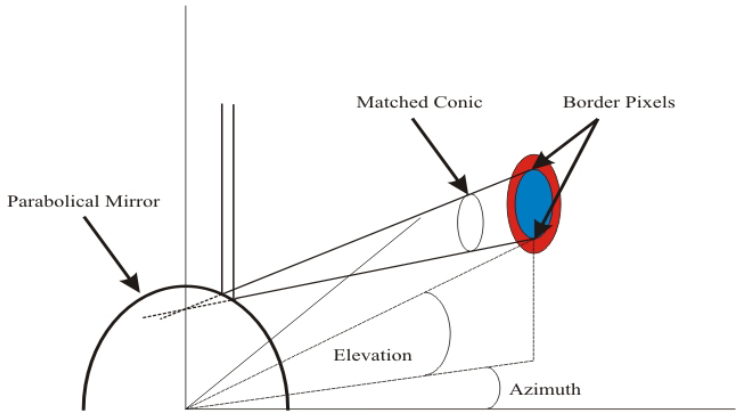

Figure 7: Schematic of conic matching. Elevation and azimuth of the rays toward the border pixels are extracted from the mirror image and are used to create a conic using a least-squares method. The resulting conic describes the landmark in ray space except for its distance, which is not needed for triangulation.

plane lots $\mathrm{H}_i = \{H : H \parallel H_i, H \cap H_i = \varnothing\}, i \in \{1,2\}$ created by two planes H_1 and H_2 going through the coordinate origin. There are especially two "candidate planes" $H_{c1}, H_{c2} \in \mathrm{H}$ that contain the landmark's center. The parallel projections of H_{c1} and H_{c2} onto a plane perpendicular to C are identical and elliptical; since all we see is this projection, we cannot tell which of the two planes contains the landmark. Since we are only interested in the center for bearing, this normally would not matter; however, we do not really see a *parallel* but a *perspectivic* projection of $H_{c1} \cap C$ or $H_{c2} \cap C$, and these are not identical but slightly distorted; therefore, the center is slightly shifted as well (see **figure 7**).
To solve this ambiguity, we use a heuristic that employs the fact that most walls in human-habited areas are perpendicular to each other; therefore, if multiple landmarks can be seen, their candidate planes can be determined by selecting a combination of candidate planes over all landmarks that are either coplanar or perpendicular (within a margin of error).

Up to this point, all calculation takes place in ray space, that is, the three-dimensional rotational coordinate system rooted at the optical center of the camera. Due to lack of stereo information, it is not possible to determine the distance to the landmark at this point; therefore, all rays are represented as infinite lines. This method delivers a high sub-pixel accuracy while operating directly on the camera image, which is given in ray space. No conversion to the cartesian coordinate system is necessary, which makes this approach both fast and accurate.
Unfortunately, the method only works for landmarks that are close enough for a sufficiently large number of border pixels to be detected. Therefore, we use a simple center-of-gravity approach as a fallback if the landmark's image occupies less than 16x16 pixels. If the image is so small, the distorting effect of the parabolic mirror can be neglected.

3.3 Pose reconstruction

After the landmarks have been detected and identified, the viewer's pose can be reconstructed: As stated above, the angles of the landmark normals are now known, and therefore the angle that the camera takes relative to each landmark is known as well. When the positions of the landmarks in cartesian space are known, the reconstruction becomes trivial. Theoretically three recognized landmarks would suffice to determine the camera's pose, but if more can be found in the image, the resulting over-determined system can be used to enhance accuracy.
The pose is reconstructed via translatoric and rotatoric adaption. The adaption steps are also performed in ray space and are repeated until the resulting error lies below

a certain threshold[1]. The translatoric adaption calculates the approximate section point of the viewing rays of the landmark centers, and then matches the viewer's position to that point. The rotatoric adaption uses the method described by B.K.Horn in [Hor87].

4 Results and further work

The system has been implemented initially in a simulation environment, to be able to experiment with several types of virtual cameras and use the results in choosing the optimal real camera. Therefore, we present two sets of experimental results; one for the simulation system and one for the actual camera.

4.1 Results in simulation

In [Wey01], a simulation system has been developed that transforms a 3D scene into a simulated camera image. This system has made it possible to test the effects of changes in camera resolution and interlaced/non-interlaced sensors on the accuracy of position reconstruction. **Figure 8** shows an image obtained from the camera simulation.
The simulation environment allows the recording and playback of trajectories in the simulated 3D scene. In playback, the position reconstruction works on the simulated camera image. **Figure 9** shows a recorded trajectory from two different angles. The graphs on the right display the rotational and translational errors corresponding to the reconstructed trajectory.

[1] It can be shown that the employed method does not converge in some rare cases. Therefore, the reconstruction process is aborted after a maximum of 48 repetitions.

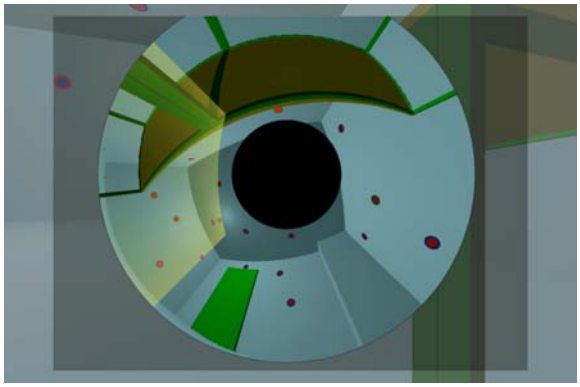

Figure 8: A simulated camera image. The image is superposed with a conventional, cartesian view of the simulation environment, i.e. the view that a user carrying the camera upon his/her head would see. The image segment representing this view in the simulated panoramic camera is marked as a bright segment in the camera image.

The graphs show that the rotational error on this (fairly typical) run is always well below one degree, and the translational error below 5mm. These results were achieved by simulating a NTSC camera of 640x480 pixels interleaved resolution. Moving toward a simulated camera of 1024x768 pixels and "progressive scan" technology improves the medium errors to 4.25mm (10.9% improvement) and 0.14° (4.4% improvement). This shows that an increase in resolution alone does not significantly increase the accuracy of the method.

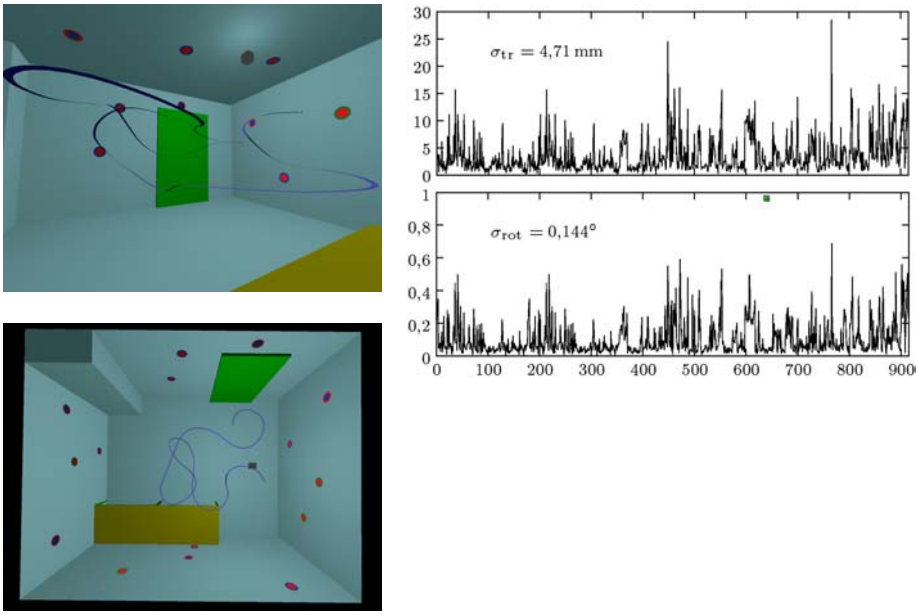

Figure 9: A recorded trajectory in the simulation environment (left) and the translational and rotational errors of the position reconstruction working on the simulated camera image.

4.2 Results with a real camera

The actual panoramic camera is an NTSC camera yielding pictures at 640x480 pixel interleaved resolution, equalling the camera simulation. The results with this real camera mirror those achieved with the simulated one; however, image noise (which was disregarded in simulation) poses a large problem; the color-separated red and blue images are too noisy to effectively recognize the more distant landmarks. **Figure 10** shows the results of a test run with this real camera; the dashed lines mark a time segment when the camera was moved out of the range of several landmarks. Naturally, in this time the error increased well below the tolerable measure, which reflects in the medium accuracy over this test run.

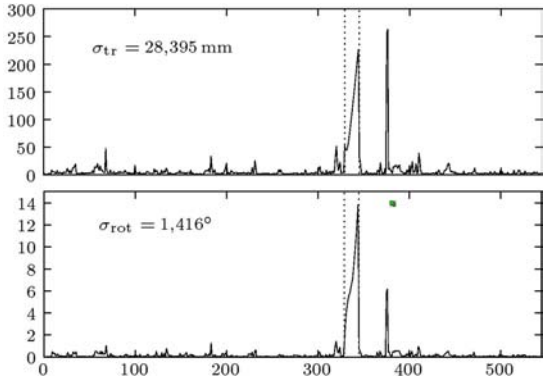

Figure 10: Rotational and translational errors derived from a trajectory followed with the actual camera. The dashed lines mark a time segment where only three landmarks were seen.

4.3 Further work

The camera module used at the moment is far too noisy to be really useful for position reconstruction in larger areas. Therefore, our main focus right now is twofold: First, we are examining image enhancement techniques to investigate whether images from the existing camera can be improved sufficiently; secondly, other more adequate camera modules are being evaluated.
Once a sufficiently good camera has been found, the system will be enhanced to support automatical calibration, so that the landmark positions do not have to be initially known. Afterwards, we will use the system as the main head tracking component in the AR system used at the University of Karlsruhe for robot programming by demonstration.

5 References

[FS+98] M.O.Franz, B.Schölkopf, H.A.Mallot, H.H.Bülthoff, A.Zell. *Navigation mit Schnappschüssen.* In: Proceedings of the 20th DAGM Symposium, Springer Verlag, Berlin 1998.

[FS+97] M.O.Franz, B.Schölkopf, H.H.Bülthoff. *Homing by Parameterized Scene Matching.* In: Proc. of the 4th European Conference on Artificial Life, MIT Press, Cambridge 1997.

[G+V99] J.Gaspar, J.S.Victor. *Visual Path Following with a Catadioptric Panoramic Camera.* In: Proceedings of the 7th International Symposium on Intelligent Robotic Systems (SIRS'99), University of Coimbra, 1999.

[Hor87] B.K.Horn. *Closed-form solution of absolute orientation using unit quaternions.* In: Journal of the OSA, Issue 4, April 1987.

[NDI02] Northern Digital Inc. *Product website for the Polaris Optical Tracking System.* At http://www.ndigital.com/polaris.html

[Sal00] T. Salb et al. *Intraoperative presentation of surgical planning and simulation results using a stereoscopic see-through head-mounted display.* In: Proceedings of Stereoscopic Displays and Virtual Reality Systems VII, part of SPIE / Photonics West 2000, San Jose, CA, January 2000

[Sal01] T. Salb, O. Burgert, T. Gockel, B. Giesler, R. Dillmann. *Comparison of tracking techniques for Intraoperative Presentation of medical data using a see-through head-mounted display.* In: Proceedings of Medicine Meets Virtual Reality 9, Newport Beach, CA, January 2001

[WB+99] G.Welch, G.Bishop et al. *The HiBall Tracker: High-Performance Wide-Area Tracking for Virtual and Augmented Environments.*In: Proceedings of the ACM VRST 99, University College London, December 20-22, 1999.

[Wey01] T.Weyrich. *Entwicklung eines Kopfverfolgungssystems auf der Basis einer Panoramakamera und künstlicher Landmarken.* Master's Thesis, University of Karlsruhe (TH), 2001.