Estimation of error sources for optical head tracking in cranial radiation therapy

P. Grüning[1], P. Stüber[1,2], L. Richter[1,2], O. Blanck[3], R. Bruder[1], A. Schweikard[1]

[1] Institute for Robotics and Cognitive Systems, University of Lübeck, Lübeck, Germany
[2] Graduate School for Computing in Life Science, University of Lübeck, Lübeck, Germany
3 Department for Radiation Oncology, University Hospital of Lübeck, Lübeck, Germany

Contact: stueber@rob.uni-luebeck.de

Abstract:

There is a growing demand for high-accuracy and frameless solutions for cranial radiation therapy. Among different approaches for intra-fractional head tracking using X-Ray or MV imaging or Cone Beam CT, optical head tracking in particular promises high spatial and temporal resolution with a minimum of system latency and no additional dose exposition. It may therefore be ideal for motion-compensated or high-accuracy cranial radiation therapy. Nevertheless, up to now optical systems lack accuracy and are therefore only found in prototypes or test setups.
Using a consumer-grade optical rangefinder, we have built a test setup to systematically quantify critical error sources for tracking systems based on triangulation. Subsequently, we present and discuss potential solutions to minimize the error.

Keywords: Head tracking, cranial radiation therapy, image-guidance

1 Purpose

In modern cranial radiation therapy, there is a trend towards image-guided treatment with a maximum of accuracy. State-of-the-art X-ray based systems provide high accuracy [1,2] but are, in many cases, time-consuming and cost-intensive. In addition, the exact positioning of the patient is often too difficult for everyday clinical practice. Furthermore, these systems use ionizing radiation. Therefore, optical systems become more attractive [3].
The main objective of our research is the development of a time- and cost-saving, non-invasive system for accurate positioning of the patient before and during radiation therapy. Moreover, movements of the patient must be detected in real-time to allow for real-time motion compensation.
To ensure the accuracy of such a device, it is necessary to systematically evaluate its tracking capabilities. As a start, we record measurements with a movement tracker that is mainly known in the field of consumer electronics, the Microsoft Kinect, which uses the Primesense PS1080 range sensor technology. In its commonly known applications (e.g. as a game controller), the Microsoft Kinect showed the capability to detect objects and trace movements of objects in real-time [4]. It employs a structured light camera, which gains spatial information by successively scanning an area with a laser pattern. As this is an interesting approach, we investigate whether its accuracy is sufficient for medical applications. As a result, our systematic evaluation gives an insight to the possibilities that can be expected, if using a triangulation setup for direct head tracking. Additionally, we also emphasize the particular problems of this technology.

2 Methods

The basic setup for this measurement was a human head phantom made of Styrofoam, which was mounted to an industrial robot (Adept Viper s850, Adept Technology, Inc., Livermore, CA, USA), serving as a ground truth. The robot moved to over 1500 positions with known locations. At each position, data was acquired and the distances of the movements were later compared to the ground truth. In this way, we are able to systematically evaluate the tracking capabilities of the Microsoft Kinect. Furthermore, the amount and location of noise was calculated. However, in contrast to off-the-shelf tracking systems (e.g. the Polaris system), the Primesense sensor does not provide the spatial position and orientation of a tracker marker. Instead, it results in a 3D point cloud representing the surface of the scanned object. Generally, a point cloud R can be described as a set of points $p_i \in R$, where each point represents a set of different features $p_i = \{f_1, f_2, \dots, f_n\}$. In our case, those features are the

3D position with respect to the origin of a 3D-camera's head and the RGB-coded color. Further data processing must be therefore applied to gain the head pose over time. For this direct head tracking, a template is matched upon the captured data. The Microsoft Kinect data is range filtered to an area the object is presumed, ignoring unnecessary points to decrease the computation time. The distance between the template and the acquired data can be quite large. Therefore, surface normals are estimated for both point clouds, approximating a plane using points in a predefined radius for each normal. With this information, the template is roughly matched upon the sensor data. Subsequently, an Iterative-closest-points (ICP) algorithm is employed to precisely match both data sets. The ICP-algorithm iteratively computes a rigid transformation to fit one set of points to another with a given accuracy or maximum number of iterations. The template is now successively transformed and a rigid transformation matrix is obtained for every scan. Based on the current position of the template and the estimated transformation, assertions can be made about the head's pose.
The depth sensor consists of an infrared-laser projector (λ=830nm, P < 60mW), which is emitting a known speckle pattern. An infrared-Sensor (IR-sensor) is able to collect the reflected light of the sent pattern. The resolution is 640x480pixel and the sensor's angle is 57° horizontal and 43° vertical, providing 11Bit of information with 30 frames per second. In addition, a processor (PrimeSense PS1080-A2-Chip) computes the distances in the captured image.
The processor's memory includes a reference image I_{ref}, capturing a known, usually planar, object at a certain distance Z_o. The distance estimation for a new Image is done by comparing parts of it to I_{ref} with the obligation to find the best match. Deviations in the scaling of the speckle pattern can lead to assumptions about the spatial change in the z-direction. Dark areas, respectively pixels which fall below a certain threshold, are regarded as shadow areas and are thus of no interest for further processing. Considering triangulation, a change in z-direction δZ produces a proportional speckle shift in the x-direction δX which is described as:

$$\delta X \cong \delta Z \, S/Z_o$$

whereas S is the distance of projector to sensor.
As we were able to obtain a 3D point cloud for any object we scan, we used our robot-based setup to evaluate the sensor. To systematically analyze the accuracy and performance of the PrimeSense sensor system, we mount a human head phantom to the robot's end effector. We positioned the PrimeSense sensor opposite to the robot, facing the mounted head phantom. Subsequently, we moved the effector with the attached head phantom to a set of roughly 1500 well known poses within a 3D grid spaced at 20 mm. At each pose, we tracked the head phantom with the sensor. To evaluate the stability of the measurements, we recorded 8 point clouds at each position.
Further, those point clouds were transformed into 2D range images, containing each point's z-location as the pixel value with a resolution of 2mm for each pixel. This format allowed the calculation of mean and standard derivation images. The latter gives information about the distribution of noise within each position.
In a next step, the accuracy of the movement detection was evaluated. The main problem was to simplify the set of points provided by each data set to a single point representing the head's position precisely. To minimize the error of the evaluation itself, this single point should correspond to every measured point cloud in the same manner. Two possible solutions were used:

1. For every mean image, a center spot was calculated by computing the mean vector of the points that could be found in all of the 8 images of the position. In several locations, especially on the sides of the head, whole parts of the point cloud can appear and disappear from picture to picture. A sudden gain of points on a particular side of the image can dramatically change its center. Leaving out those areas provided a consistent centre spot for every position.

2. Since the measurement focused on translations and the head itself did not rotate while varying its position, the assumption could be made that the phantom's nose tip was always the area nearest to the camera. Therefore, the 20 nearest points for each range image were selected and from those, outliers were removed. The medial pixel represented the nose tip for a single range image and for the 8 scans of each position, a grand mean was calculated.

After estimating those particular points, the movement from one point to another was calculated, using the Euclidian norm, and the difference from the ground truth was estimated.
Further, an ICP-evaluation was done. For every captured point cloud, a transformation was calculated from its preceding position. With the estimation of the rotation angles and the length of the translation, it was possible to examine whether the movement was recognized correctly.

3 Results

The estimated standard derivation (SD) images showed an accumulation of relatively big error pixels in areas of a high pixel value gradient (Fig.1). Those regions were, for example, the side of the nose as well as the corners of the face. Pixels which could not be consistently found in all 8 images were not regarded, because presence and absence of certain pixels are hard to quantify. However, those pixels are another error source and all object edges are showing those problems. In a certain perspective, whole areas, respectively body parts, can alternately appear and disappear. The noise for planar areas was scattered equally with an average of less than 0.4 mm. Regarding the spatial distribution of a position's mean SD, a small enhancement to 0.2 mm could be detected in the areas that were farer than 790 mm away from the camera. On average the SD is 0.4 mm (Fig.2).
The Microsoft Kinect camera was able to detect movements with an accuracy of 1.7 mm +/- 2.7 mm, using the center point of the mean image (Fig.3). The detection of the tip of the nose brought an accuracy of 1.7 mm +/- 3.6 mm. The spatial distribution showed no noticeable areas. In addition, the estimated points representing the tip of the nose were compared to each other for every position. The average of the maximum error was 1 mm +/- 1.6 mm. Due to noise, the highest estimated value was 650 mm. The median error was 0.4mm. For the above mentioned measurements, the mean percentage of outliers was 5.24% +/- 1.6%. Those average 80 outliers were mostly caused by noise speckles that did not belong to the object, but were not removed adequately by the implemented filter methods. Those located in an area nearer to the camera than the object tremendously changed the outcome of the computations.
With an average error of 7.5 mm +/- 10 mm, the ICP algorithm could not keep up with the preceding results. It is very likely that 20 mm distance is too far to get a sufficient matching and that the algorithm is mislead by local minima. The Euclidian norm of the three averaged rotation angles was 0.015°. This means, the method succeeded in identifying, that the movement was only a translation. Regarding the 3D-error-plot (Fig.5), the large errors were equally distributed and showed no particular accumulation.

4 Conclusion

In general, a motion tracking system based on triangulation, like the Microsoft Kinect, might be applied for head-tracking during radiation therapy. However, the presented setup is not optimal. The used sensor is only a standard consumer electronics device and thus, there is a lot of potential for an accuracy increase, by using specialized equipment, for example.

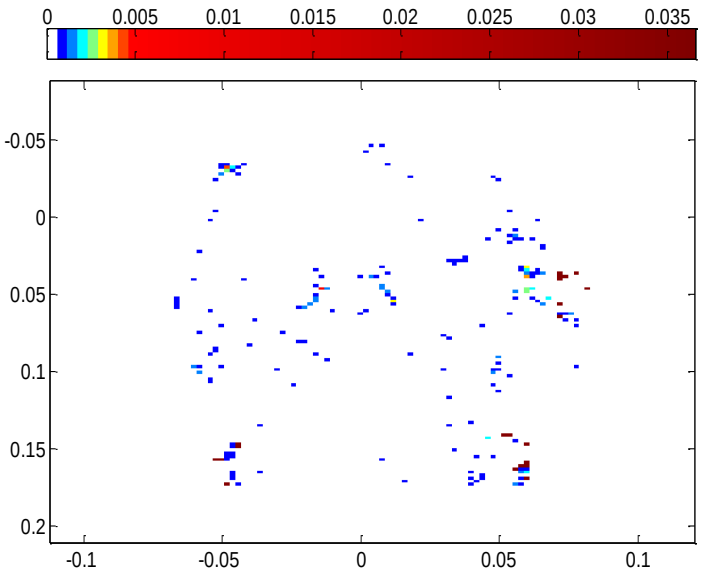

Figure 1: SD Image from one position. A mean Image is computed from 8 different range images and subsequently, the SD is estimated. It shows that concerning the noise of a position, larger errors are in the areas of high z-range gradients (e.g. the side of the nose). All coordinates in [m], color bar values in [mm]

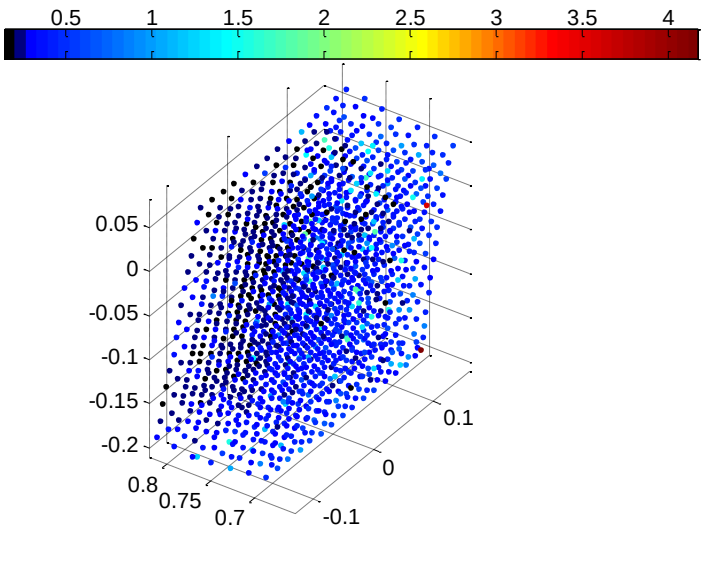

Figure 2: Spatial distribution of the average SD per Image. The outcome enhances in the range of 790 mm to 810 mm. All coordinates in [m], color bar values in [mm]

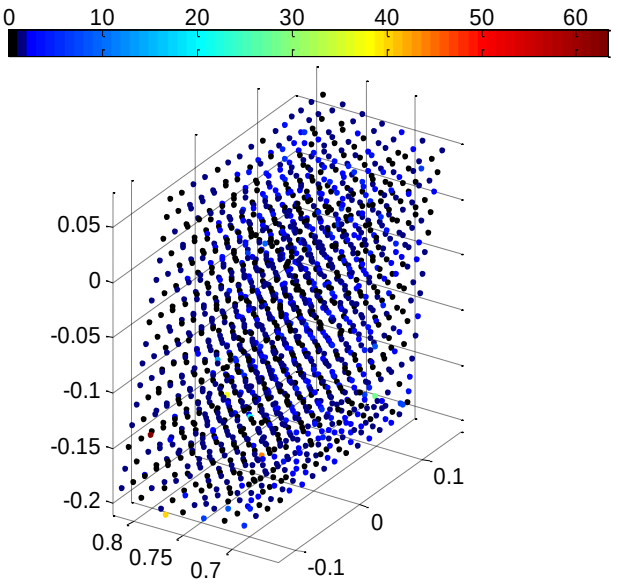

Figure 3: Averaged deviation from the 20 mm movement estimated with the center spot of the mean images. All coordinates in [m], color bar values in [mm].

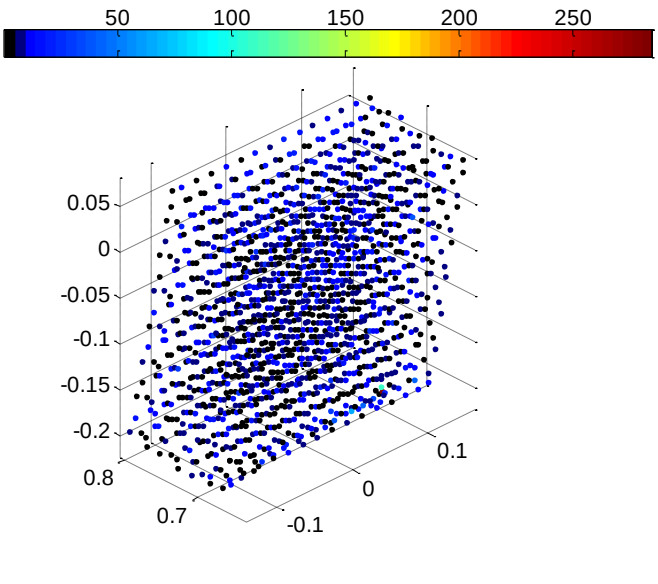

Figure 4: Spatial distribution of the averaged Iterative closest points (ICP) deviation. The plot shows no signs of a particular area where the errors accumulate. All coordinates in [m], color bar values in [mm].

Nevertheless, our measurements give an insight into the capabilities of all range finders based on the triangulation principle. We have found a set of common problems: First of all, no information is gained from shadowed areas, leaving a constrained data field to the user. Second, when facing a high depth gradient high noise can be regarded including an inconsistent number of points. Third, there is an occurrence of noise speckles which do not belong to the object. Positive aspects are, that the camera showed no signs of an accumulation of errors in a certain location or a spatial distortion, but it is substantial to consider the camera's ground noise. It was not possible to achieve an accuracy better than 1.7 mm, which clearly not fulfills the localization requirements for radiation therapy. As for the software, the ICP algorithm could not compete with our other methods, as it is too susceptible to noise and too unstable for a distance of 20 mm. Still, the ICP remains an essential part of data processing, since it can track the head with full six degrees-of-freedom. To reduce the ICP error, smart templates are needed which on the one hand, avoid noisy areas to increase the robustness of the calculation, but on the other hand, need to contain distinguishable landmarks. Moreover, data filtering and preprocessing should be used, for example by calculating averaged data. Further investigations are needed, in both hard- and software, to overcome the evaluated problems.

In summary, we presented a first setup that allows for a systematic analysis of error sources with tracking devices based on structured light. Even though current systems do not fulfill the desired accuracy, our proposed idea might be applicable with the next generation of depth sensors.

References

[1] M.J. Murphy, S.D. Chang, I.C. Gibbs, Q.T. Le, J. Hai, D. Kim, D.P. Martin, J.R. Adler Jr, *Patterns of patient movement during frameless image-guided radiosurgery*, Int J Radiat Oncol Biol Phys, 2003

[2] D.A. Jaffray, J.H. Siewerdsen, J.W. Wong, A.A. Martinez, *Flat-panel cone-beam computed tomography for image-guided radiation therapy* , Int J Radiat Oncol Biol Phys, 2002

[3] T. Moser, G. Habl, M. Uhl, K. Schubert, G. Sroka-Perez, J. Debus, K. Herfarth, C.P. Karger, Clinical *Evaluation of a Laser Surface Scanning System in 120 Patients for Improving Daily Setup Accuracy in Fractionated Radiation Therapy*, Int J Radiat Oncol Biol Phys, 2012

[4] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, A. Fitzgibbon, *Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera*, ACM Symposium on User Interface Software and Technology, 2011