Markerless Real-Time 3-D Target Region Tracking by Motion Backprojection From Projection Images

Torsten Rohlfing*, *Member, IEEE*, Joachim Denzler, *Member, IEEE*, Christoph Gräßl, Daniel B. Russakoff, and Calvin R. Maurer, Jr., *Member, IEEE*

***Abstract*—Accurate and fast localization of a predefined target region inside the patient is an important component of many image-guided therapy procedures. This problem is commonly solved by registration of intraoperative 2-D projection images to 3-D preoperative images. If the patient is not fixed during the intervention, the 2-D image acquisition is repeated several times during the procedure, and the registration problem can be cast instead as a 3-D tracking problem. To solve the 3-D problem, we propose in this paper to apply 2-D region tracking to first recover the components of the transformation that are in-plane to the projections. The 2-D motion estimates of all projections are backprojected into 3-D space, where they are then combined into a consistent estimate of the 3-D motion. We compare this method to intensity-based 2-D to 3-D registration and a combination of 2-D motion backprojection followed by a 2-D to 3-D registration stage. Using clinical data with a fiducial marker-based gold-standard transformation, we show that our method is capable of accurately tracking vertebral targets in 3-D from 2-D motion measured in X-ray projection images. Using a standard tracking algorithm (hyperplane tracking), tracking is achieved at video frame rates but fails relatively often (32% of all frames tracked with target registration error (TRE) better than 1.2 mm, 82% of all frames tracked with TRE better than 2.4 mm). With intensity-based 2-D to 2-D image registration using normalized mutual information (NMI) and pattern intensity (PI), accuracy and robustness are substantially improved. NMI tracked 82% of all frames in our data with TRE better than 1.2 mm and 96% of all frames with TRE better than 2.4 mm. This comes at the cost of a reduced frame rate, 1.7 s average processing time per frame and projection device. Results using PI were slightly more accurate, but required on average 5.4 s time per frame. These results are still substantially faster than 2-D to 3-D registration. We conclude that motion backprojection from 2-D motion tracking is an accurate and efficient method for tracking 3-D target motion, but tracking 2-D motion accurately and robustly remains a challenge.**

***Index Terms*—Frameless stereotactic radiosurgery, motion backprojection, real-time target tracking, 2-D to 2-D registration, 2-D to 3-D registration.**

I. Introduction

THE CyberKnife (Accuray, Inc., Sunnyvale, CA; see Fig. 1) is a robotic frameless stereotactic radiosurgery system used in cancer therapy [1]. It is an example of an image-guided therapy system that determines the intraoperative patient position by registration of a three-dimensional (3-D) preoperative computed tomography (CT) image to intraoperative two-dimensional (2-D) projection images (see Fig. 2 for examples). In the case of the CyberKnife system, two X-ray projection images are acquired simultaneously using a pair of orthogonal flat-panel amorphous silicon detectors (ASDs). Their locations and projection geometries are constant over time and known with very high accuracy in the intraoperative coordinate system. The projection images are registered to the preoperative CT image, in which the treatment target (typically a tumor) and therapy beams have been defined. The registration yields the current patient pose so that the therapy beams are accurately aligned with their planned position and orientation with respect to the desired target.

In many clinical applications, X-ray image acquisition is repeated at frequent intervals during the intervention to track the patient's motion over time. In the case of the CyberKnife radiosurgery system, for example, cervical spine patients are fitted with a molded Aquaplast (WFR/Aquaplast Corp., Wyckoff, NJ) that stabilizes the head and neck on a radiographically transparent headrest. Thoracic and lumbar spine patients rest in a conformal alpha cradle during CT imaging and treatment. These supports help maintain the general orientation of the anatomy and minimize patient motion. Nonetheless the patient can and does move slightly during treatment. Thus, acquisition of the X-ray images is repeated periodically (typically in 1–min intervals) to follow the patient's motion over time and adapt the beam targeting accordingly. For each new pair of X-ray images, this requires a new registration to the CT image, which is time consuming and, in the presence of large motion, not very robust.

Manuscript received April 15, 2005; revised August 12, 2005. A preliminary version of this paper was presented at the 9th International Fall Workshop Vision, Modeling, and Visualization, November 16–18, 2004, Stanford, CA, USA. The work of T. Rohlfing was supported in part by the National Science Foundation (NSF) under Grant EIA-0104114, "Integrating Soft Segmentation with Intensity-Based Matching for 2-D/3-D Image Data Registration." The work of Ch. Gräßl was supported in part by the European Commission 5th IST Program—Project VAMPIRE. The work of D. B. Russakoff and C. R. Maurer, Jr. was supported in part by the Interdisciplinary Initiatives Program, which is part of the Bio-X Program at Stanford University, under the grant "Image-Guided Radiosurgery for the Spine and Lungs." Only the authors are responsible for the content. This research was performed under a collaboration established with support from the Bavaria California Technology Center (BaCaTec). The Associate Editor responsible for coordinating the review of this paper and recommending its publication was S. Aylward. *Asterisk indicates corresponding author.*

*T. Rohlfing is with the Neuroscience Program at SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025-3493 USA (e-mail: torsten@synapse.sri.com).

J. Denzler is with the Lehrstuhl für Digitale Bildverarbeitung, Fakultät für Mathematik and Informatik, Universität Jena, 07737 Jena, Germany (e-mail: denzler@informatik.uni-jena.de).

C. Gräßl is with the Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, 91058 Erlangen, Germany (e-mail: graessl@informatik.uni-erlangen.de).

D. B. Russakoff is with the Computer Science Department, Stanford University, Stanford, CA 94305-9025 USA (e-mail: dbrussak@stanford.edu).

C. R. Maurer, Jr. is with the Department of Neurosurgery, Stanford University, Stanford, CA 94305-5327 USA (e-mail: calvin.maurer@gmail.com).

Digital Object Identifier 10.1109/TMI.2005.857651

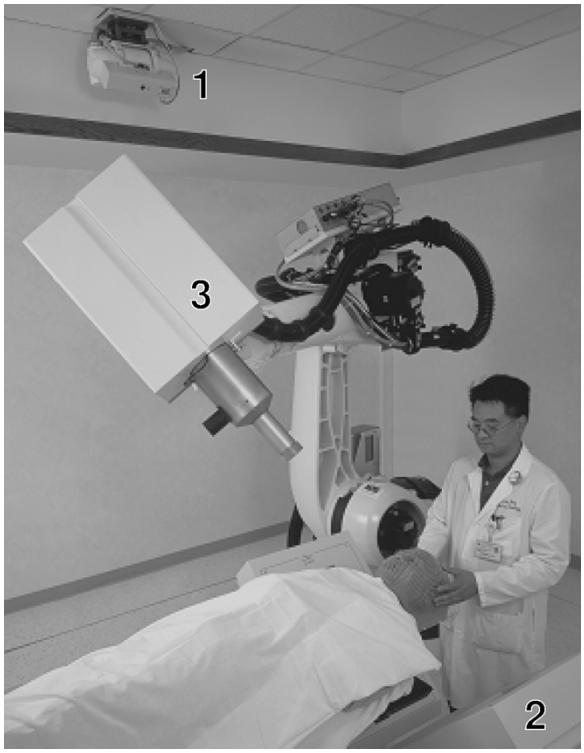


Fig. 1. CyberKnife system with (1) ceiling-mounted X-ray source, (2) ASD, and (3) robot-mounted therapy beam source. A second X-ray imaging system with ceiling-mounted source (not visible) and floor-mounted ASD (partly visible) is installed perpendicular to the first system.

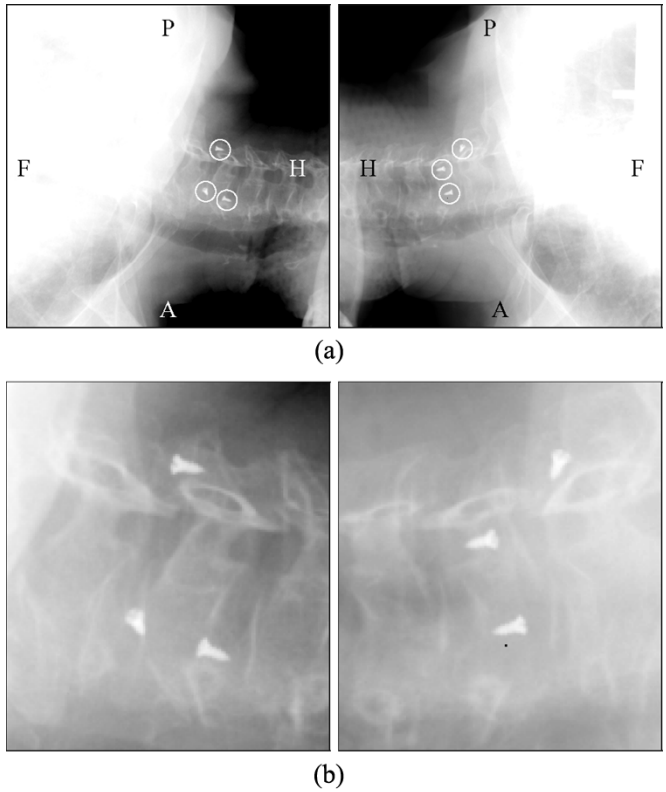


Fig. 2. ASD X-ray images acquired using a CyberKnife system during treatment of a patient with a tumor in the cervical spinal cord. (a) Pairs of images are acquired simultaneously from orthogonal projection directions. Three bone-implanted markers (marked by white circles) used for target tracking in this patient are clearly visible. These markers provide the gold-standard transformations used for validation in this paper. (b) Magnifications of target region and markers from the images in (a).

However, much of the 3-D motion can be deduced from the motion of objects seen in the 2-D projection images, in particular from a pair of orthogonal projections. This principle has been applied in numerous works (see [2] for a recent survey).

For radiosurgery treatment of the spine, the only method used in clinical practice requires bone-implanted markers that can be easily identified and efficiently tracked in the projection images. Such markers can also be implanted in soft tissue, thus allowing, for example, the respiratory tracking of organ motion [3]. Fiducial marker-based methods are in general fast, accurate, and robust. However, artificial fiducial markers require a separate surgical implantation procedure. Marker implantation is not always possible, is often considered too invasive to be clinically acceptable, and entails risk, especially in the cervical spine where the vertebral structures are small and fragile. There is also the issue of whether it is acceptable to leave markers permanently implanted.

Virtually all marker-free target tracking methods are 2-D to 3-D image registration methods, i.e., they solve the alignment problem between the 3-D image and the observed 2-D projection images independent of the concept of motion. To the extent that these methods deal with multiple projection image sets acquired over time, information from previous frames is only implicitly used in the form of the initial 3-D transformation parameters. This is true for methods based on digitally reconstructed radiography (DRR) images [4]–[6], as well as for techniques based on contours [7] and image gradients [8], [9].

We propose in this paper to exploit the observed motion between subsequent projection images to track patient motion. For each projection image, a markerless real-time 2-D region tracking algorithm [10] is used to obtain an estimate of the in-plane motion relative to the previous frame. The in-plane motion estimates from all projections acquired at the same time are then combined consistently into an estimate of the 3-D motion, resulting in an estimate of the new patient-to-image transformation. A schematic illustration of our proposed method is shown in Fig. 3. The backprojected 3-D motion prediction can be used as is, or it can be refined by a full 2-D to 3-D registration. After motion prediction, the registration is started in close proximity of the correct transformation, which improves both its accuracy and its computational efficiency. We evaluate our method with clinical data from 10 patients treated for spinal tumors with the CyberKnife radiosurgery system, but the technique itself is applicable to other treatment systems and clinical applications.

Other groups, such as Sarrut & Clippe [11], have previously suggested using 2-D in-plane transformations to speed up the 2-D to 3-D registration process by precomputing out-of-plane DRR images and applying 2-D in-plane transformations to them during the registration. Our method is different in that it takes the opposite approach. We reverse the direction of inference by directly estimating the 3-D transformation from the observed 2-D motion. Our technique can thereby take advantage of the full X-ray image resolution, as well as the real-time performance of the 2-D region tracking algorithm.

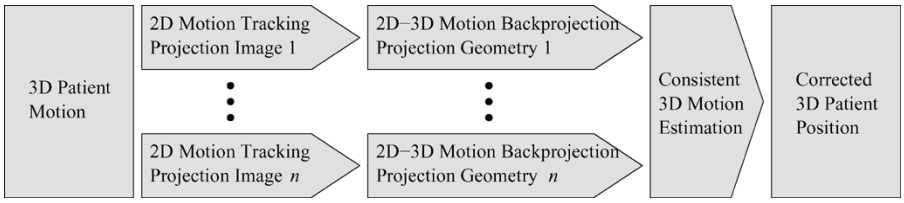


Fig. 3. Schematic illustration of the 3-D motion estimation process. The patient motion leads to 2-D motion in the n projection images. In each projection image, this motion is tracked and backprojected into 3-D space using the known projection geometry. From all n 3-D motion estimates, a consistent 3-D motion estimate is generated, which is then used to update the patient position. In this paper, i.e., for the CyberKnife system, $n = 2$.

Plattard *et al.* [12] evaluated 2-D to 2-D registration of DRR and portal images using mutual information for patient setup in radiation therapy. Their work, however, did not deduce 3-D motion from the observed 2-D transformations. Birkfellner *et al.* [13] presented a method for 2-D to 3-D registration that decouples the in-plane rotation of the projection image from the remaining 5 degrees of freedom of the full 3-D rigid transformation. They did not perform a 2-D to 2-D registration, however, and the registration continued to require computation of DRR images by ray casting. Also, neither Plattard nor Birkfellner considered the tracking of patient motion over time as opposed to setup of a static patient position. In recent work, Brewer *et al.* [14] described a method for tumor tracking using a template correlation-based method for tracking markers in fluoroscopy images. In addition to requiring markers, their method only produced an estimate of the expected maximum range of tumor motion in 3-D. It did not produce an actual estimate of the 3-D target position at any given time.

The present paper is, to the best of our knowledge, the first to use markerless 2-D tracking in X-ray projection images for 3-D target tracking. In particular, our work makes the following novel contributions: 1) backprojection of in-plane motion to 3-D space; 2) closed-form solutions for consistent 3-D translation estimation, and for rotation estimation from images acquired using mutually perpendicular projection geometries; 3) evaluation of a region tracking algorithm for real-time motion tracking in X-ray images and comparison to intensity-based 2-D registration; 4) evaluation of markerless 3-D target tracking performance using clinical data with fiducial marker-based gold standard transformations.

The remainder of this paper is organized as follows. In Section II we develop the general concept of our core contribution, backprojecting 2-D motion into 3-D space, and present expressions for consistent 3-D motion estimates from 2-D motion in multiple projection images. In Section III we review the two techniques for 2-D motion tracking that we evaluated: region tracking using a hyperplane approach, and intensity-based 2-D to 2-D registration. Section IV presents quantitative results obtained using clinical data with fiducial marker-based gold-standard transformations from 10 patients. The paper is concluded by a discussion of the results, their clinical relevance, and possible extensions in Section V.

II. Three-Dimensional Motion From 2-D Motion

A. Problem Statement

Image-guided therapy based on preoperatively acquired 3-D images requires a coordinate transformation that maps the preoperative image coordinates to the coordinates of the physical space of the patient and the treatment room. Often this transformation is assumed to be rigid. This assumption is not correct in general for the spinal application we use for evaluation in this paper, but we avoid this problem by tracking an individual vertebra, which is locally approximately rigid. For the CyberKnife radiosurgery system, and other image-guided therapy systems that use interoperative X-ray images to estimate this transformation, the location and geometry of the X-ray imaging system is known accurately in the physical space (patient) coordinate system from a calibration process. The target and treatment plan on the other hand are defined in the coordinates of the preoperative CT image.

The objective of target region tracking is to maintain an accurate coordinate transformation between patient and image coordinates, taking into account patient motion. We assume that the initial transformation $\mathbf{T}^{(0)}$ is known. From motion observed in the 2-D projection images we estimate the transformation $\mathbf{T}^{(k)}$ at time $k > 0$, i.e., the time when the kth projection image set was acquired.

For two independent projection devices, the result of the tracking for any given frame is a pair of 2-D translation vectors, which quantify the in-plane shift of the tracked region in the projection images, and a pair of rotation angles, which quantify the respective rotations of each tracking region about its center. This motion is always expressed relative to frame $k = 0$.

B. Three-Dimensional Translation Estimation

The projection geometry and mathematical symbols used are illustrated in Fig. 4. For two projections, let $\vec{x}_A$ be the normalized (i.e., $\|\vec{x}_A\|_2 = 1$) 3-D direction vector of detector plane A in the x pixel direction. Analogously let $\vec{y}_A$ be the normalized vector in the y pixel direction of detector A, as well as $\vec{x}_B$ and $\vec{y}_A$ for detector B. In our application, the direction vectors are invariant over time, as the projection imaging devices of the CyberKnife system are installed in fixed locations. However, this is coincidental for our work and not a requirement of the proposed method.

1) Motion Backprojection: The result of the tracking for any given frame is a pair of 2-D translation vectors $\vec{t}_A$ and $\vec{t}_B$, which quantify the in-plane shift of the tracked region in millimeters in projection images P_A and P_B, respectively. From these and the detector orientations we can compute the 3-D motion of the tracked pattern as

$$\vec{d}_A = c_A \left[(\vec{x}_A \quad \vec{y}_A)\vec{t}_A\right] \quad \text{and}$$
$$\vec{d}_B = c_B \left[(\vec{x}_B \quad \vec{y}_B)\vec{t}_B\right]. \tag{1}$$

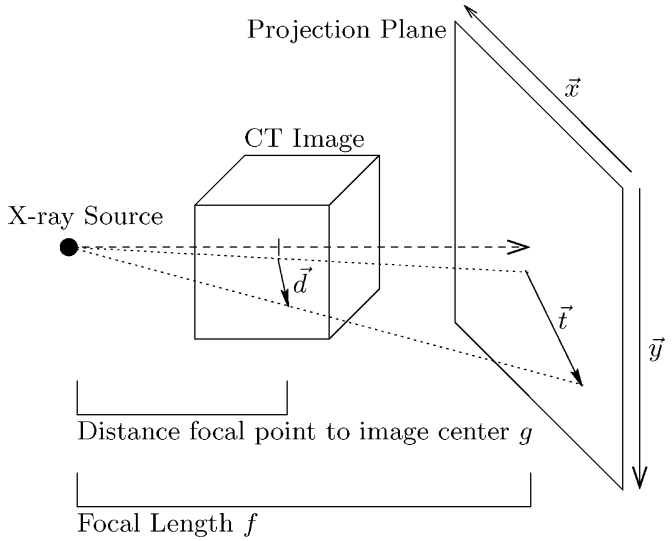


Fig. 4. Projection geometry and notation for translational motion backprojection. The focal length f is the distance between the X-ray source and the projection plane. The object-to-projection plane distance d is the distance between the center of the CT image and the projection plane. The projection plane is spanned in 3-D by the vectors $\vec{x}$ and $\vec{y}$. The 2-D translation vector $\vec{t}$ from tracking is backprojected to yield the 3-D translation vector $\vec{d}$.

The 3×2 matrices $(\vec{x}_A\ \vec{y}_A)$ and $(\vec{x}_B\ \vec{y}_B)$ rotate the 2-D translation vectors $\vec{t}_A$ and $\vec{t}_B$, respectively, from the 2-D X-ray image coordinate system to the 3-D treatment room coordinate system. The coefficients c_A and c_B are linear scaling factors that take into account the perspective effect of the X-ray projection. For projection A this factor is

$$c_A = \frac{g_A}{f_A}. \tag{2}$$

For projection B, the scaling factor c_B is computed accordingly. Note that (2) is only correct on the central (orthogonal) projection ray, at a distance g_A from the focal point. However, for the large focal length in our application ($f_{A/B} \approx 3800$ mm, $g_{A/B} \approx 3000$ mm versus 200 mm X-ray field of view (FOV) and $\leq$500 mm transverse CT FOV) the approximation is sufficiently accurate. The (unlikely) worst case occurs if the region tracked in 2-D corresponds to a region that resides near the boundary of the CT image rather than in its center. For image sizes and geometry as given above, the approximation error would then be 0.07 mm per millimeter of motion observed in the X-ray image.

2) Consistent Translation Estimation: Since for two or more projection geometries, not all of the detector orientations are orthogonal in 3-D, we have to compensate for multiple contributions along the same directions. Let $\vec{e}_x = (1,0,0)$, $\vec{e}_y = (0,1,0)$, and $\vec{e}_z = (0,0,1)$ be the x, y, and z unit column vectors, respectively. When all projection plane direction vectors are added with unit weights, the accumulated contribution in 3-D in direction of the positive x dimension is

$$\begin{aligned} s_x &= \langle \vec{x}_A, \vec{e}_x \rangle^2 + \langle \vec{y}_A, \vec{e}_x \rangle^2 + \langle \vec{x}_B, \vec{e}_x \rangle^2 + \langle \vec{y}_B, \vec{e}_x \rangle^2 \\ &= \vec{e}_x^T (\vec{x}_A \cdots \vec{y}_B) \begin{pmatrix} (\vec{x}_A)^T \\ \vdots \\ (\vec{y}_B)^T \end{pmatrix} \vec{e}_x = \vec{e}_x^T \mathbf{M}\mathbf{M}^T \vec{e}_x \end{aligned} \tag{3}$$

where $\mathbf{M}$ is the matrix that contains all projection plane direction vectors as its columns, i.e., for two projections A and B

$$\mathbf{M} = (\vec{x}_A \quad \vec{y}_A \quad \vec{x}_B \quad \vec{y}_B). \tag{4}$$

Likewise, the contributions s_y and s_z along the y and z directions, respectively, can be expressed. With these, the matrix that normalizes the sum of all directions to unity is

$$\mathbf{N} = \begin{pmatrix} \frac{1}{s_x} & 0 & 0 \\ 0 & \frac{1}{s_y} & 0 \\ 0 & 0 & \frac{1}{s_z} \end{pmatrix}. \tag{5}$$

Using $\mathbf{N}$ and the 3-D in-plane translation vectors $\vec{d}_A$ and $\vec{d}_B$, we can obtain a consistent 3-D translation estimate as

$$\Delta\vec{T} = (\vec{d}_A + \vec{d}_B)\mathbf{N}. \tag{6}$$

As a concrete example, consider the projection geometries of the CyberKnife system, illustrated in Fig. 6, which provided the data for evaluation in Section IV. The two ASD devices of the CyberKnife system have the following direction vectors:

$$\vec{x}_A = (-1,0,0) \quad \text{and} \quad \vec{y}_A = \left(0, -\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right) \tag{7}$$

for projection A and

$$\vec{x}_B = (1,0,0) \quad \text{and} \quad \vec{y}_B = \left(0, \sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right) \tag{8}$$

for projection B. These yield $\mathbf{N} = \mathrm{diag}(1/2, 1, 1)$, so when combining the motion estimates from the two projections, the contributions along the parallel (although oriented in opposite directions) x axes of both projections are averaged. The contributions from the y axes of the projection planes, which are orthogonal with respect to each other and with respect to the x axes, are taken as they are. This is precisely what one would intuitively expect.

C. Three-Dimensional Rotation Estimation From Mutually Perpendicular Projections

For mutually perpendicular projection geometries, a consistent 3-D rotation can be estimated from the 2-D in-plane rotations. This is possible because any in-plane rotation in one projection plane is entirely out-of-plane for any projection perpendicular to the first.

Let ω_A be the in-plane rotation angle of the tracked region around its center for projection A. Then ω_A directly corresponds to a 3-D rotation of the same angle around the central ray of projection A. Note, however, that by rotating around the center ray an additional translational component is introduced. To prevent this, we rotate instead around a rotation axis parallel to the center ray but through the center of the tracked region. Both rotations are equivalent as such, but the latter requires no translational correction. With this in mind, the in-plane rotation for projection A corresponds to a 3-D transformation described by a homogeneous matrix $\mathbf{R}_A = \mathrm{Rot}(\omega_A, \vec{n}_A, \vec{C}_A)$, which is a rotation by ω_A around an axis parallel to the normal vector $\vec{n}_A$

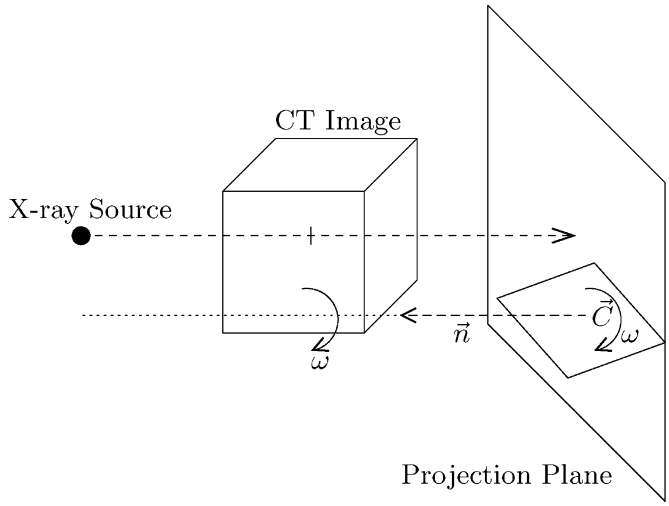


Fig. 5. Projection geometry and notation for rotational motion backprojection. The 2-D in-plane rotation angle ω corresponds to a 3-D rotation around the axis parallel to the plane normal $\vec{n}$ through the center $\vec{C}$ of the tracked region.

of projection image A through the center $\vec{C}_A$ of the tracked region in frame k. See Fig. 5 for a schematic illustration of the situation.

Likewise, let ω_B be the in-plane rotation angle for projection B. The 3-D rotation corresponding to ω_B is computed analogously to the one for projection A. However, since the rotation $\mathbf{R}_A$ has changed the coordinate system, this needs to be taken into account by rotating the next rotation axis and reference point accordingly. The appropriately adapted rotation from projection B is thus $\tilde{\mathbf{R}}_B = \mathrm{Rot}(\omega_B, \mathbf{R}_A\vec{n}_B, \mathbf{R}_A\vec{C}_B)$. The combined rotational component from projections A and B is

$$\mathbf{R} = \tilde{\mathbf{R}}_B \cdot \mathbf{R}_A. \tag{9}$$

When there is a third projection C, perpendicular to both A and B, then a third rotational component $\tilde{\mathbf{R}}_C$ can be defined and incorporated analogously.[1]

D. Consistent 3-D Rigid Transformation

Let the (known) coordinate transformation between the CT image coordinate system and the patient coordinate system at time $k = 0$ be described by the homogeneous matrix $\mathbf{T}^{(0)}$, which defines a mapping $\vec{x} \mapsto \mathbf{T}^{(0)}\vec{x}$. Then the consistent 3-D transformation estimate for a given frame $k > 0$ is computed as

$$\mathbf{T}^{(k)} = \mathbf{R}^{(k)} \cdot \mathbf{X}^{(k)} \cdot \mathbf{T}^{(0)}. \tag{10}$$

Here, $\mathbf{X}^{(k)}$ is the homogeneous matrix that implements the translation $\Delta\vec{T}$ as defined by (6). The matrix $\mathbf{R}^{(k)}$ is the combined rotation matrix as defined by (9). The resulting matrix $\mathbf{T}^{(k)}$ defines the (estimated) transformation from the CT image coordinates to the patient coordinate system at time k.

Note that an alternative method to simultaneously estimate all parameters of the consistent 3-D transformation that best explains the observed 2-D motion was suggested by Sarrut & Clippe [11]. Their method determines the parameters of the 3-D rigid transformation that minimizes the mean squared error between a set of random points in 3-D that are forward projected onto each projection plane and the same points projected and transformed in plane using the observed 2-D motion parameters. While this method is more general than ours (it inherently supports arbitrary numbers of arbitrary projection geometries), it involves an additional iterative optimization step. This can potentially lead to decreased computational performance and convergence issues. Since we limit our evaluation in the paper to data from perpendicular projection geometries, we prefer to use our closed-form solution of the 3-D transformation for this special case.

[1]We are not aware of any such imaging system currently in existence, but it is technically possible and our method would support its projection geometry in a straightforward way.

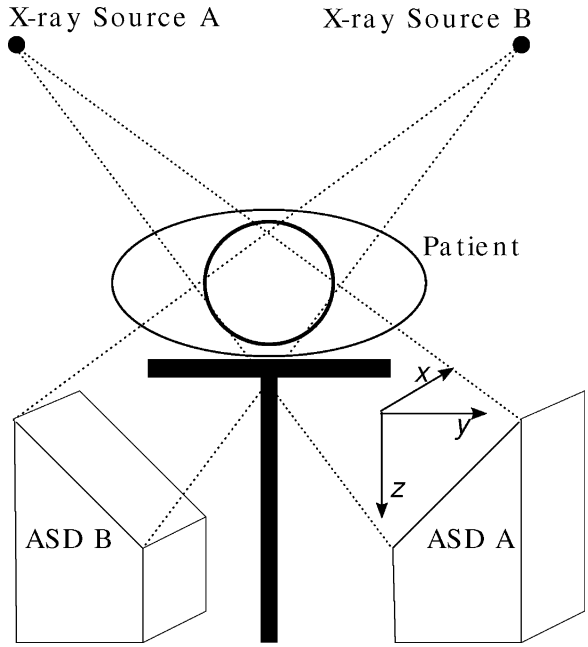


Fig. 6. Schematic illustration of the CyberKnife setup and coordinate system. The two ASDs are mounted perpendicular to each other. See Fig. 1 for a photograph of an actual CyberKnife system.

III. Tracking and Registration Algorithms

A. 2-D Region Tracking

To track objects in the projection images at high frame rates, we use an independent implementation [15] of the hyperplane tracking algorithm introduced by Jurie & Dhomes [10]. The tracking algorithm is trained on a manually drawn region of interest (ROI) in frame 0. Only a relatively small number of so-called template points inside the ROI is actually considered for tracking. The template points can be selected randomly, or based on image features. Out of three different template point selection methods [16] that we evaluated (uniform random distribution, selection based on image gradient, selection based on image intensity variance) we found that the uniform random distribution performed best on X-ray images.

Initial tests to determine the best parameters of the hyperplane tracking algorithm further showed that a larger number of template points did not necessarily result in more accurate motion estimation. Computing the tracking errors at the tips of the bone-implanted markers revealed that this effect originates in the 2-D motion tracking (Fig. 7). One possible explanation for

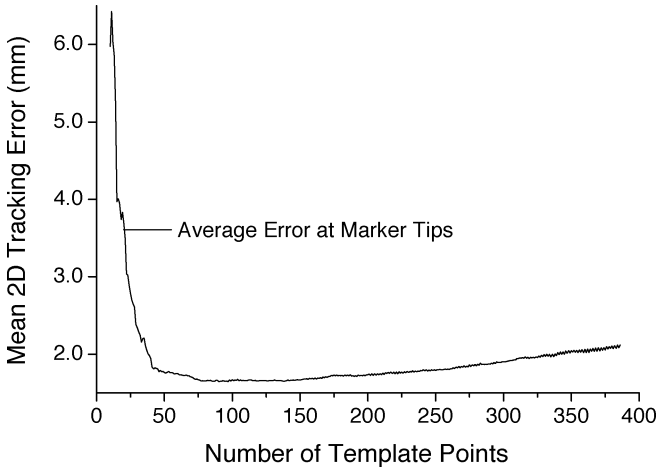


Fig. 7. Tracking error at marker tips in X-ray images from one patient versus number of template points in the hyperplane tracking algorithm. For each template point density, 50 tracking passes were performed using different template point positions (medium tracking region size, uniform random distribution, translation and rotation). The graph shows the average tracking error over all passes at the tips of the bone-implanted markers.

this behavior is that there is no truly rigid relationship between the markers and the target region. It is encouraging, however, that as long as the density of template points does not fall under a certain minimum, the tracking errors is fairly insensitive to the actual density. In order to ensure that we are operating in the flat part of the accuracy curve, we fix the number of template points as 200 for the evaluation in this paper.

After the specification of the ROI and the selection of template points in the first image of the sequence, the reference template is represented by a vector $\vec{r} = (\vec{x}_1, \ldots, \vec{x}_N)^T$, which contains the 2-D coordinates $\vec{x}_i = (x_i, y_i)^T$ of the N template points. The gray-level intensity of a point $\vec{x}_i$ in frame k is given by $f(\vec{x}_i, k)$. Consequently, vector $\vec{f}(\vec{r}, k)$ contains the intensities of template $\vec{r}$ in frame k.

The transformation of the reference template is modeled by $\vec{r}_k = \vec{g}(\vec{r}, \vec{\Delta x}_k)$, where $\vec{\Delta x}_k$ contains the transformation parameters and $\vec{g}(\cdot, \cdot)$ is the function that applies the transformation to the template point coordinates. Template matching can be described as computing the transformation parameters $\vec{\Delta x}_k$ that minimize the least-square intensity difference between the reference template and the current template. To reduce the computational cost of a nonlinear optimization, [10], [17] use a first-order approximation

$$\vec{\Delta x}_{k+1} = \vec{\Delta x}_k + \mathbf{A}\vec{i}_{k+1} \tag{11}$$

with the error vector

$$\vec{i}_{k+1} = \vec{f}(\vec{r}, k_0) - \vec{f}\left(\vec{g}(\vec{r}, \vec{\Delta x}_k), k+1\right). \tag{12}$$

Two approaches have been suggested in the literature for computing the matrix $\mathbf{A}$ in (11). Hager & Belhumeur [17] proposed using a Taylor approximation. Jurie & Dhomes [10] used an initialization stage (i.e., training step) where a number of random motions are simulated and are used to estimate matrix $\mathbf{A}$ by a least-squares estimation. Note that this initialization needs to be performed only for the first frame in the image sequence. For the work described in this paper, we use the hyperplane approach [10], due to its superior basin of convergence.

Tracking is trained on the first frame by applying 1000 random transformations to the reference template. For every transformation the error vector at the template points is calculated according to (12). Also, the parameter vectors of the corresponding transformations are stored. The matrix $\mathbf{A}$ is computed from all error vectors and transformation parameter vectors using a least squares estimation. For more details on background and implementation of the hyperplane tracking algorithm, which are beyond the scope of this paper, the interested reader is referred to Jurie & Dhomes [10].

The quality of tracking depends on the training. If the random transformations are strong, the tracker is able to estimate strong movements but lacks in accuracy. Otherwise if the random transformations are weak, the tracker estimates the movements very accurately, but fails in case of strong movements. A coarse to fine strategy is, therefore, used in a five-level tracker hierarchy. The top level tracker is trained by a maximum transformation of 30 pixels (12 mm) in each direction, the bottom level tracker is trained using a maximum transformation of 3 pixels (1.2 mm).

Typical problems in template tracking are intensity fluctuations caused by the change of illumination. X-ray images are affected similarly, for example by soft tissue motion, or by adjustments of the X-ray dose. Errors in estimation of the transformation parameters may, therefore, occur, as the error vector is computed directly from intensity values. To overcome this problem, we normalize all intensity values that are used in (12) by

$$f_{\mathrm{norm}}(\vec{x}, k) = \frac{f(\vec{x}, k) - f_{\min}(\vec{x}, k)}{f_{\max}(\vec{x}, k) - f_{\min}(\vec{x}, k)} \tag{13}$$

where $f_{\max}(\vec{x}, k)$ and $f_{\min}(\vec{x}, k)$ are the maximum and minimum intensity values of the local neighborhood of a pixel $\vec{x}$ in the image in frame k. For the experiments in this paper, the local neighborhood is a 31×31 square centered at $\vec{x}$. Note that this normalization does not need to be performed on the whole image, but only on the region points which are affected by the tracker.

B. Intensity-Based 2-D to 2-D Registration

For comparison with the region tracking algorithm, we have implemented a rigid 2-D to 2-D registration algorithm based on a successful 3-D to 3-D algorithm [18]. As the registration metrics, we use here two image similarity measures that have previously been used in intensity-based 2-D to 3-D registration [5] and are known to be somewhat effective at registering X-ray images, pattern intensity (PI) [19] and normalized mutual information (NMI) [20].

The PI metric is computed by evaluating the local information in a spherical neighborhood $D_r(i, j)$ of radius r around each voxel in the difference image as

$$E_{\mathrm{PI}} = \sum_{i,j} \sum_{(v,w) \in D_r(i,j)} \frac{\sigma^2}{\sigma^2 + (I_{\mathrm{dif}}(i,j) - I_{\mathrm{dif}}(v,w))}. \tag{14}$$

To allow for global intensity variation, the difference image I_{dif} is computed after global multiplicative intensity normalization using a least-squares scaling factor. For the adjustable parameters of PI we follow Penney *et al.* [5] and use $r = 3$ pixels

as the region radius and $\sigma = 10$ as a numerical stabilizer term. The NMI metric is computed as

$$E_{\mathrm{NMI}} = \frac{H_X + H_Y}{H_{XY}} \tag{15}$$

where we estimate the marginal entropies H_X and H_Y and the joint entropy H_{XY} of both images using discrete histograms [21].

We have also experimented with other similarity measures, namely normalized cross correlation [4], correlation ratio [11], [22], gradient correlation, and difference image entropy, but none of these performed as well as NMI and PI. Likewise, standard mutual information [21], [23] performed similar to, but slightly worse than, NMI. We, therefore, limit our detailed evaluation to the PI and NMI metrics.

The optimal parameters of a 2-D rigid transformation (two translations, one rotation angle) that maximizes the respective similarity measure are determined using a hill climbing algorithm [18]. For each image pair, two optimization passes are performed. To help avoid local optima in the optimization, the first pass operates on images that are blurred with a Gaussian kernel (standard deviation 0.8 mm $= 2$ pixels). The images are only blurred but not downsampled in order to facilitate estimation of the entropy-based similarity measures using discrete histograms. The second pass operates on the original, unblurred images.

C. Two-Dimensional to Three-Dimensional Registration

Three-dimensional motion tracking from 2-D projection images can also be achieved using 2-D to 3-D registration. The most common class of marker-free registration algorithms compare DRR images computed from a 3-D CT image to the actual X-ray images. The pose of the CT image is adjusted until the DRR images best match the X-ray images [4].

The 2-D to 3-D registration algorithm we use here optimizes the PI image similarity measure [24] between actual X-ray images and DRRs. The six parameters of the 3-D rigid transformation that optimizes the similarity measure are computed using a simple but robust hill-climbing algorithm [18]. We also evaluated Powell's direction set method with Brent's line search (modified implementation from Press *et al.* [25]), but found the hill climbing algorithm to fail less often.

Instead of computing DRRs by ray casting, our algorithm employs a progressive attenuation field (PAF) [26]. A PAF is a dynamically growing table of projection values similar to an attenuation field [27], [28], which itself is closely related to a Transgraph [29]. Projection values needed for a DRR but not found in the PAF are computed on demand using a fast ray casting engine [30] that is optimized to take advantage of the SIMD instructions available on the Pentium 4 CPU. The efficiency of the ray casting algorithm is further improved by adaptive ray clipping [31] based on a fast implementation of the Euclidean distance transformation [32]. Every projection value computed by ray casting is added to the PAF for later use, thereby reducing the computational cost of further DRR computation as the tracking proceeds from frame to frame.

Tracking by 2-D to 2-D registration and by 2-D to 3-D registration can be combined in a two-stage algorithm. In particular, we estimate transformations $\mathbf{T}^{(k)}$ using each of the following two combined methods.

1) Two-dimensional to three-dimensional registration of the CT image to the next X-ray projection image frames. Each registration starts with $\mathbf{T}^{(0)}$, which we have found to produce better results than starting with the transformation $\mathbf{T}^{(k-1)}$ computed for the respective previous frame. While it may seem that $\mathbf{T}^{(k-1)}$ is in general closer to $\mathbf{T}^{(k)}$ than $\mathbf{T}^{(0)}$ is, starting registration for frame k at $\mathbf{T}^{(0)}$ is safer since registration might have failed for frame $k-1$.
2) Three-dimensional motion estimation from 2-D tracking, followed by a 2-D to 3-D registration, where the output of the 3-D motion backprojection serves as the starting point for the 2-D to 3-D registration.

D. Three-Dimensional Motion From Fiducial Marker-Based 2-D Motion

As a validation step, we investigate the 3-D target registration error (TRE) [33] that can be achieved from 2-D motion based on the fiducial markers. For each frame and each projection, we perform a point-based rigid registration in 2-D of the markers in that frame to the markers in frame $k = 0$. The resulting 2-D transformation parameters are then used to obtain a 3-D motion estimate as described in Section II.

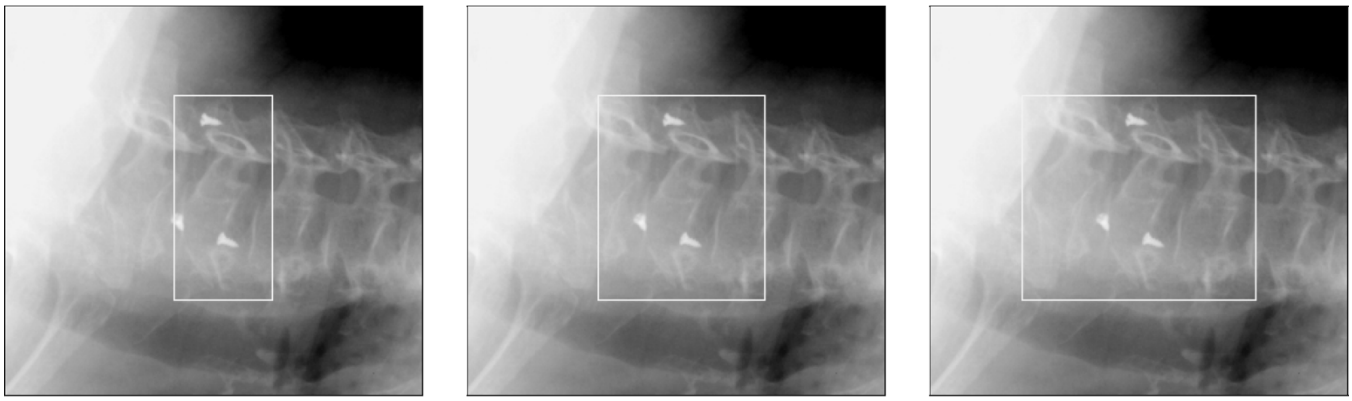
Note that there is a fundamental difference between backprojection of marker-based in-plane motion and 3-D point-based registration of triangulated marker positions. While out-of-plane motion affects the in-plane marker-based motion estimation, it does not interfere with triangulation and point-based registration in 3-D. So although the same fiducial markers are used for the 2-D motion estimation and the TRE computation in 3-D, the TRE will in general not be zero. The actual magnitude of the TRE is a measure of how accurately we can track 3-D target motion, given the exact 2-D in-plane motion from two perpendicular projection images. In other words, the larger the out-of-plane motion that is not captured by the 2-D marker motion, the larger the TRE, and vice versa.

IV. Evaluation of 3-D Target Tracking

A. Image Data and Gold-Standard Transformations

We apply the methods proposed in this paper to image data from 10 patients treated for spinal tumors using the CyberKnife radiosurgery system. All projection images have 512×512 pixels with a pixel size of 0.4 mm (see Fig. 2 for examples). The preoperative CT images have between 180 and 300 slices with a slice thickness of 1.2 mm. The in-plane pixel size of the CT images is between 0.5 mm and 1.0 mm. The true coordinate transformations between physical space and patient (i.e., CT image) coordinates are known from bone-implanted fiducial markers [35].

For spinal applications in particular, there is a potential trade off involving the size of the tracking region. Larger regions likely contain more trackable features and improve the algorithm's capture range. Smaller regions, on the other hand, strengthen the local rigidity assumption and can, therefore, help safeguard against nonrigid motion within the tracked



(a) Small region (b) Medium region (c) Large region

Fig. 8. Comparison of three different tracking region sizes for one projection from Patient #4 (see Table I for details). This is the same patient used in previous work [34]. Recall that, in order to avoid bias, the fiducial markers that are visible in these images are segmented and excluded from tracking for the purpose of this study. (a) Small (target vertebra). (b) Medium (target vertebra plus partial neighbors). (c) Large (target vertebra plus full neighbors). Note that the tracking regions are aligned with the image coordinate to facilitate their manual definition. This is not a limitation of the tracking algorithms.

TABLE I
OVERVIEW OF PATIENT DATA, TARGET VOLUMES, AND TRACKING REGIONS

Patient	Target Region	Size of TRE Volume [mm^3]	Number of X-ray Frames	Tracking Region Sizes Camera A [pixels]			Tracking Region Sizes Camera B [pixels]		
				small	medium	large	small	medium	large
1	C3/C4	$29 \times 22 \times 64$	20	59×151	119×151	168×160	57×133	108×133	161×133
2	C4	$31 \times 32 \times 12$	19	50×133	81×133	127×133	52×129	87×129	131×129
3	C5/C6	$27 \times 22 \times 29$	20	60×152	106×152	156×152	59×139	97×139	153×139
4	C6	$57 \times 53 \times 18$	21	66×137	112×137	157×137	65×146	111×146	156×146
5	C7/T1	$24 \times 36 \times 25$	18	59×158	91×158	150×158	52×140	79×140	106×140
6	C7/T1	$39 \times 37 \times 38$	8	62×146	117×146	184×146	65×152	114×152	172×152
7	C7/T1	$26 \times 24 \times 15$	20	48×124	97×124	150×124	57×120	104×120	148×120
8	T7/T8	$39 \times 58 \times 28$	20	79×165	167×165	248×165	83×168	164×168	248×168
9	T10/T11	$25 \times 58 \times 40$	19	91×203	161×203	261×203	73×209	167×209	236×209
10	pelvis	$81 \times 64 \times 54$	22	—×—	137×198	—×—	—×—	165×175	—×—

The second column lists the target regions, mostly cervical (C) and thoracic (T) vertebrae. The third column gives the sizes of the target volumes for TRE computation. The fourth column gives the number of X-ray frames available for tracking from each patient (including frame 0 for each patient). The remaining columns summarize the sizes in X-ray pixels of the three different tracking regions for each patient. See text for details. Patient #10 was treated in the pelvic region and only a single tracking region was evaluated.

region. We evaluate our motion backprojection method for three different tracking region sizes for each patient.

1) *Small region:* encloses only the target vertebra.
2) *Medium region*: encloses the target vertebra plus approximately half of its neighbors on either side.
3) *Large region:* encloses the target vertebra plus its entire two neighbors.

The three tracking regions defined for one of the patients used in this study are shown in Fig. 8. The sizes in pixels of all three respective regions for each patient are listed in Table I. Note that for one of the patients (patient 10) we only evaluated one tracking region as the target region for this patient was within the pelvis. The tracking regions were the same for all motion tracking methods. For hyperplane tracking, we tracked the respective region as described in Section III-A above. For 2-D to 2-D registration we cropped the reference frame 0 to the respective ROI while keeping the complete floating image. For 2-D to 3-D registration, we computed DRR images for pixels inside the ROI only.

In order to avoid bias of the evaluation due to the presence of markers in the X-ray images, these are excluded from tracking. For each patient, a binary mask for frame 0 from each camera is defined by first segmenting the implanted markers and subsequently applying eight iterations of a morphological dilation operator. The resulting mask marks those areas in the X-ray image that are prohibited from template point assignment for the region tracking algorithm. The margin of eight pixels around each of the markers ensures that template points in the neighborhood cannot access the marker image, as long as the region motion remains under 3.2 mm (8 pixels $\times$ 0.4 mm pixel size). The marked regions are also excluded from the intensity-based registration algorithm.

For the purpose of this evaluation, we assume that the correct transformation (i.e., the gold standard) between patient and CT image coordinates at time $k = 0$ is known. Let this transformation be denoted by a homogeneous matrix $\mathbf{T}_{\text{gold}}^{(0)}$. For the subsequent times $k > 0$ we estimate transformation matrices $\mathbf{T}^{(k)}$. The accuracy of the estimated transformation is then computed as the TRE relative to the respective gold-stan-

TABLE II
SUMMARY OF 3-D PATIENT MOTION AND TARGET TRACKING RESULTS

Tracking Method	Tracking Region	Frames with TRE $>$ 1.2 mm	Frames with TRE $>$ 2.4 mm	TRE (mm)	CPU Time / Frame (s)	
				mean ± std.dev.	mean ± std.dev.	max
No Tracking (Patient Motion)	—	135 (76%)	62 (35%)	1.4 ± 0.5	—	—
Marker-based 2D Motion	—	6 (3%)	1 (1%)	0.6 ± 0.3	—	—
Hyperplane Tracking	small	117 (66%)	59 (33%)	1.3 ± 0.5	$4.7 \times 10^{-3} \pm 0.8 \times 10^{-3}$	5.2*
	medium	102 (58%)	45 (25%)	1.2 ± 0.4	$4.7 \times 10^{-3} \pm 0.8 \times 10^{-3}$	5.2*
	large	78 (44%)	32 (18%)	1.1 ± 0.5	$4.7 \times 10^{-3} \pm 0.8 \times 10^{-3}$	5.2*
2D-2D Registration (NMI)	small	60 (34%)	22 (12%)	0.9 ± 0.6	0.8 ± 0.5	2.5
	medium	40 (23%)	3 (2%)	0.9 ± 0.4	1.2 ± 0.6	2.7
	large	32 (18%)	7 (4%)	0.8 ± 0.4	1.7 ± 0.8	4.7
2D-2D Registration (PI)	small	47 (27%)	22 (12%)	0.8 ± 0.4	2.0 ± 1.2	6.6
	medium	32 (18%)	10 (6%)	0.7 ± 0.4	3.1 ± 1.4	8.2
	large	28 (16%)	5 (3%)	0.7 ± 0.4	5.4 ± 1.9	12.0
2D-3D Registration (PI)	small	113 (64%)	44 (25%)	1.3 ± 0.5	34.9 ± 26.2	222
	medium	106 (60%)	40 (23%)	1.3 ± 0.5	53.9 ± 30.4	190
	large	100 (57%)	39 (22%)	1.3 ± 0.5	94.6 ± 45.2	293
2D-2D Registration (PI) followed by 2D-3D Registration (PI)	small	104 (59%)	58 (33%)	1.1 ± 0.5	52.4 ± 48.8	340
	medium	77 (44%)	36 (20%)	1.0 ± 0.5	63.8 ± 31.6	233
	large	72 (41%)	31 (18%)	0.9 ± 0.5	95.5 ± 44.3	293

*For the hyper-plane tracking algorithm, the training step for the first frame on average required 5.2 s per projection using 200 uniformly distributed random template points.

dard transformation at time k, i.e., $\mathbf{T}_{\text{gold}}^{(k)}$. The TRE itself is computed as the root-mean-square (rms) difference between coordinates in some region V mapped using the estimated transformation versus those mapped using the gold-standard transformation

$$\text{TRE}^{(k)} = \frac{1}{|V|} \sum_{\vec{x} \in V} \left(\mathbf{T}^{(k)} \vec{x} - \mathbf{T}_{\text{gold}}^{(k)} \vec{x} \right)^2 . \tag{16}$$

The region V is the target volume of the surgical procedure. In this study, it is the manually defined bounding box of the vertebra targeted during radiosurgery.

For comparison, we also compute the "uncorrected TRE," that is, the TRE without any motion correction. The uncorrected TRE uses the gold-standard transformation for frame $k = 0$ as the reference, which is based on the assumption that the initial position of the patient is known perfectly. For all subsequent frames $k > 1$, the uncorrected TRE $m^{(k)}$ in the target volume relative to frame 0 is then computed as the rms difference of the gold-standard transformations at time k and time 0

$$m^{(k)} = \frac{1}{|V|} \sum_{\vec{x} \in V} \left(\mathbf{T}_{\text{gold}}^{(k)} \vec{x} - \mathbf{T}_{\text{gold}}^{(0)} \vec{x} \right)^2 . \tag{17}$$

The uncorrected TRE $m^{(k)}$ is identical to the actual patient motion.

Similar to previous evaluation studies on 2-D to 3-D registration [26], [35], we distinguish between successfully and unsuccessfully tracked frames by defining a TRE threshold. For the present study, we use two thresholds, the first at 1.2 mm, which is the slice distance of the CT images that were acquired for our patients. Patient motion estimates in 3-D with TRE values below this threshold have sub-pixel accuracy with respect to the usual 2-D to 3-D registration approach. The second threshold is 2.4 mm.

B. Results

The results of the 3-D target tracking evaluation are summarized in Table II. All computation times refer to processing a single frame from one X-ray camera on a single Pentium 4 Xeon CPU with 3.0 GHz clock speed. The 2-D to 3-D registration algorithm was run on a machine equipped with 2 GB of main memory to allow sufficient space for the PAF-based computation of DRR images. We group the results by different criteria and discuss in detail the accuracy, robustness, and computational performance of the different methods.

Over all ten patients (177 tracked X-ray frames), the actual 3-D patient motion exceeded 1.2 mm in 76% of all frames (135 frames) and 2.4 mm in 35% of all frames (62 frames), with a maximum of 16.4 mm (these are results labeled "No Tracking (Patient Motion)" in Table II). By backprojecting the in-plane fiducial marker motion, we obtain a best-case estimate of the target tracking accuracy for perfect in-plane tracking (these are results labeled "Marker-based 2-D Motion" in Table II). Using this benchmark for 3-D motion estimation, the number of frames with TRE above 1.2 mm was reduced to 6 frames (3%) and the number of frames with TRE above 2.4 mm was 1 (1%). The mean TRE of the transformations in the frames with TRE less than 2.4 mm was 0.6 mm, compared to the mean patient motion of 1.4 mm.

1) Three-Dimensional Motion From 2-D Tracking: The hyperplane tracking algorithm was substantially faster than the 2-D to 2-D registration-based tracking, and it achieved some improvement in TRE. But 2-D to 2-D registration-based tracking produced substantially better improvements in TRE than hyperplane tracking. Between the two intensity-based image similarity measures, 2-D to 2-D registration using the PI metric performed better than the NMI metric, and it did so in particular using smaller regions for registration. The best results were

achieved using the PI metric on the large tracking regions, but at an average computational cost per frame of over 5 s. 2-D to 2-D registration using NMI produced slightly less accurate tracking results (by approximately 0.1 mm TRE and 2–7% more frames with $\mathrm{TRE} > 1.2$ mm), but was several times faster. It also appears that NMI registration benefited more from an increased tracking region size than PI registration did.

All 2-D tracking algorithms performed better with the "large" tracked regions compared to the "medium" and "small" regions, presumably because the "large" regions contain more useful image information. Using even larger tracking regions, tracking accuracy did not increase further compared to the results in Table II. Using hyperplane tracking for example, the number of frames with TRE over 1.2 mm increased from 44% to 55% when adding an additional 10 pixels (corresponding to 4 mm) on all sides of the "large" region. We observed similar behavior for the remaining tracking and registration methods, and the effect increased as more data was added (25 and 50 pixels, corresponding to 10 mm and 20 mm, respectively). Increasing the tracking region sizes also substantially reduced computational efficiency of the registration-based tracking methods.

2) Motion Backprojection Versus 2-D to 3-D Registration: Initializing the 2-D to 3-D registration algorithm with a transformation from 2-D motion backprojection slightly improves the results of the registration algorithm, both in terms of successfully registered frames and TRE values. Because when registration fails it does so rather fast, the increase in successfully registered frames leads to a slight increase in computation time of the 2-D to 3-D registration, although the difference is only 1% for the large tracking region.

The most important result when comparing 2-D to 3-D registration and motion backprojection is that, although motion backprojection from only two projections is lacking one rotational degree of freedom, its accuracy is clearly superior to that of 2-D to 3-D registration. This is true even when 2-D to 3-D registration is initialized with the transformation obtained from motion backprojection. It appears that for motion backprojection the advantage of working on the high-resolution X-ray images rather than DRR images more than makes up for the lack of one degree of freedom. Of course this is possible only as long as there is either no substantial change of this particular parameter due to actual patient motion, or the parameter has limited influence on the TRE, which seems to be the case for our data. We discuss this phenomenon in more detail in Section V.

3) In-Plane Versus Out-of-Plane Rotations: In Fig. 9, the TRE of motion tracking using backprojected 2-D fiducial marker motion is plotted against the actual out-of-plane rotation angle as computed from the triangulated 3-D marker positions. The out-of-plane rotation, which in the case of our data is a rotation around the body axis, appears to provide a linear lower bound for the TRE that corresponds to approximately 0.3 mm TRE for each degree of out-of-plane rotation. This number depends on the relative location of tracked region and surgical target, and it would be larger if the distance between the two was larger.

The TRE results using the best actual 2-D motion tracking (2-D to 2-D registration, PI similarity measure, large region) are plotted versus the in- and out-of-plane rotation angles in Fig. 10.

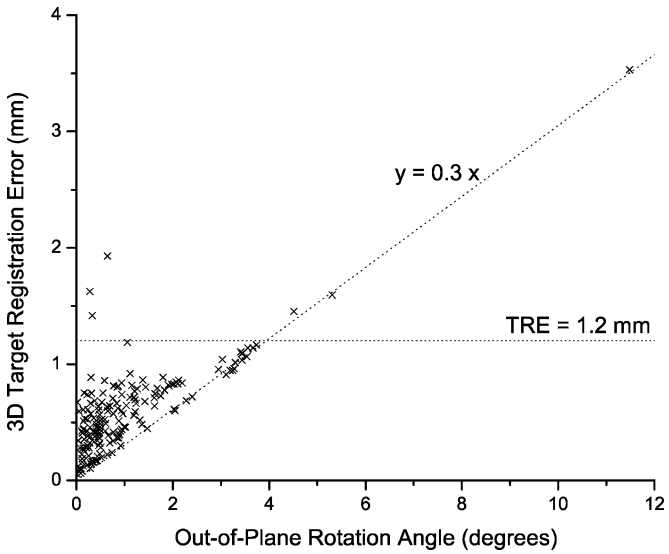


Fig. 9. TRE of 3-D target tracking using backprojected 2-D fiducial marker motion versus absolute out-of-plane rotation angle over 177 X-ray frames from 10 patients.

With few outliers, most frames were tracked with TRE better than 1.2 mm up to out-of-plane rotation angles of 2.5°. Similarly, most frames were tracked with TRE better then 1.2 mm for total in-plane rotations below approximately 5°. We note that the vast majority of all frames had rotation angles below these respective thresholds.

In Fig. 11, we illustrate the mutual effects between in- and out-of-plane rotations, robustness of the tracking algorithm, and resulting 3-D TRE. For each X-ray frame, the tracked region is interpolated based on the 2-D motion parameters estimated by 2-D to 2-D registration using the PI and the NMI similarity metrics. Accurate tracking in the absence of out-of-plane rotations should, therefore, leave the resulting images visually unchanged.

In frames 1 and 2, small to moderate rotations (up to 3.7°) have little effect on the 3-D tracking accuracy (3.4 mm and 3.4 mm patient motion corrected to 0.3 mm–0.7 mm TRE). In frame 3, we observe a combination of out-of-plane rotation by 3.0° and a projection A in-plane rotation by 6.3°, where the latter is out-of-plane for projection B. This causes 2-D tracking using PI to fail in projection B, possibly due to the appearance change of the tracked region. However, the region is still successfully tracked using the NMI similarity measure, resulting in an acceptable TRE of 1.3 mm. This is fairly close to the result using marker-based 2-D motion estimation (1.0 mm) and demonstrates that the failure of tracking using PI is due to a lack of robustness of the 2-D to 2-D registration, rather than due to the fundamental issue of missing one degree of freedom in the 3-D transformation.

4) Execution Times: Execution time is an important consideration for the clinical application of 3-D motion estimation, especially for radiosurgery procedures in which patient motion is repeatedly estimated using periodically acquired X-ray images (e.g., the CyberKnife). The most common way to estimate 3-D motion is repeated 2-D to 3-D image registration. The mean CPU time for 2-D to 3-D registration in our work was between

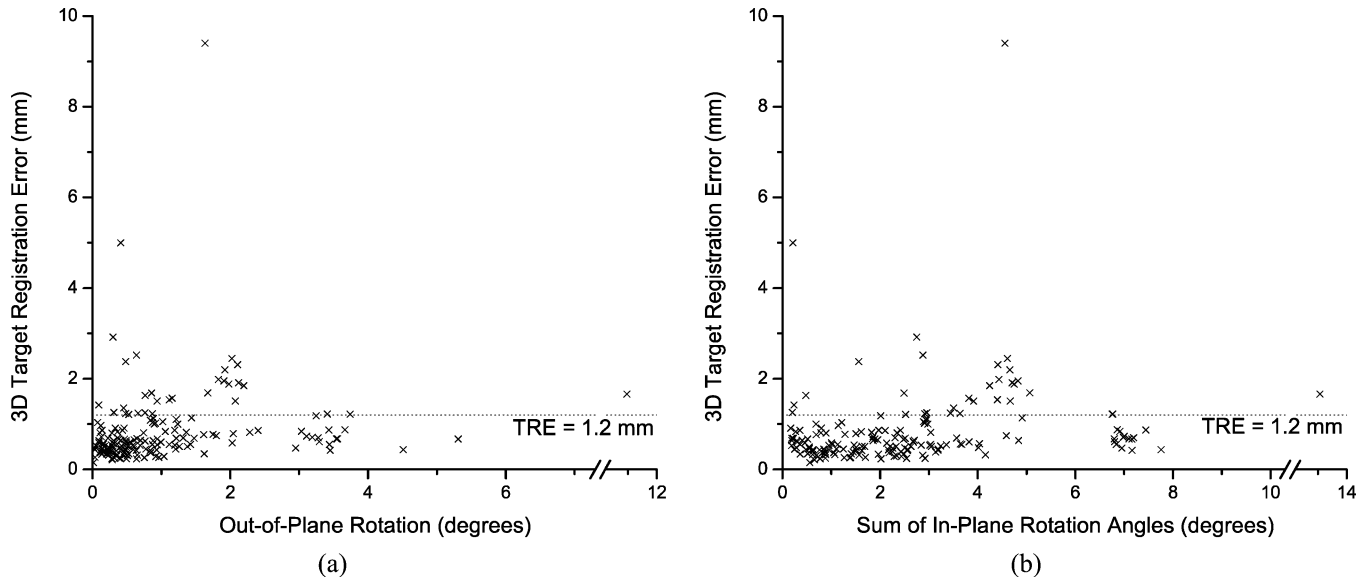
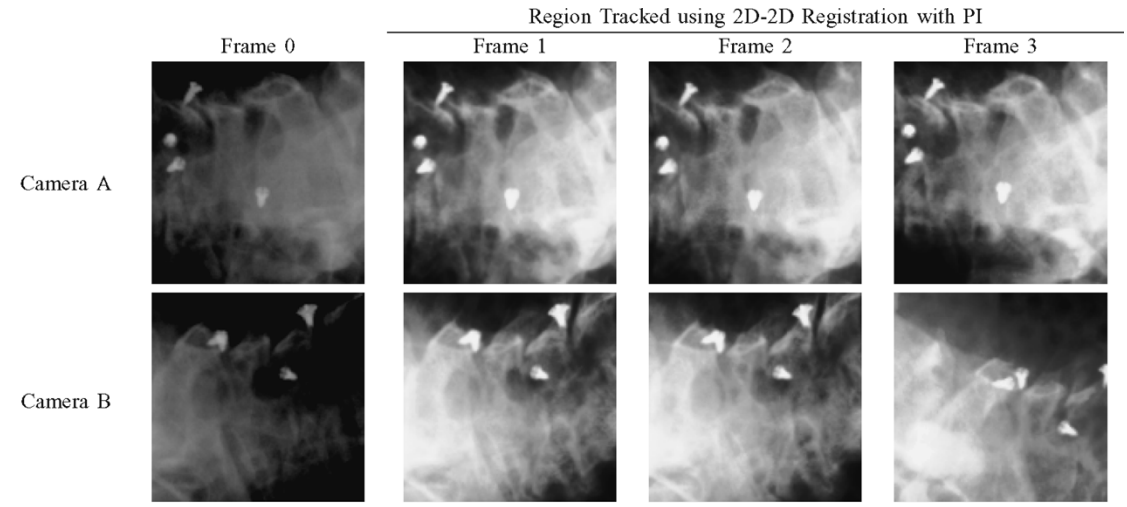


Fig. 10. TRE values of 3-D target tracking using backprojected motion from 2-D registration (PI; large tracking region) versus in-plane and out-of-plane rotation angles. (a) TRE versus absolute out-of-plane rotation angles. (b) TRE versus sum of the two absolute in-plane rotation angles.



In-Plane Rotation Camera A	0.3°	1.1°	6.3°
In-Plane Rotation Camera B	2.1°	3.7°	0.6°
Out-of-Plane Rotation	0.4°	0.5°	3.0°
Patient Motion (TRE without tracking)	3.4 mm	3.3 mm	9.6 mm
TRE using Marker-based 2D Motion	0.3 mm	0.6 mm	1.0 mm
TRE using 2D-2D Registration (PI)	0.3 mm	0.7 mm	9.4 mm
TRE using 2D-2D Registration (NMI)	0.3 mm	0.5 mm	1.3 mm

Fig. 11. Illustration of 3-D TRE versus in-plane and out-of-plane rotation angles (as computed from the triangulated 3-D marker positions) in one case (Patient #2; large region). The two rows of images show the tracking region in five frames of projection images from camera A and B, respectively. The first column, frame 0, is the reference frame. In the remaining frames, the tracking region has been reformatted from the original X-ray image according to the 2-D motion parameters computed by 2-D to 2-D registration using PI. The implanted markers were excluded from computation of the similarity measures to avoid bias. For each frame, the two in-plane rotation angles and the out-of-plane rotation angle are given in degrees. The TRE of the 3-D transformation from motion backprojection is given in mm. See text for additional information.

35 s and 95 s per frame, depending on the size of the X-ray ROI used for registration (Table II). These times are comparable to the times of other reported intensity-based 2-D to 3-D registration algorithms. Generation of DRRs during the optimization search is the primary computational expense in the intensity-based 2-D to 3-D registration process. The fastest generation of a DRR with 200 × 200 pixels (which is a typical size of an ROI in our registration work) that we are aware of requires about 30 ms (e.g., [27]). Each iteration of the parameter search requires the generation of 24 DRRs (two DRRs per transformation parameter per image, six rigid transformation parameters, two orthogonal images), and thus each iterative step requires about 720 ms. Assuming that the number of iterations in the search is 30–40 iterations, which is easily achieved using a gra-

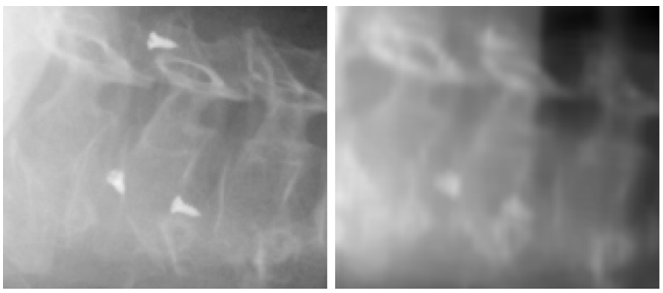


Fig. 12. Comparison of X-ray versus DRR image resolution. (a) X-ray image of the cervical spine. (b) Corresponding DRR computed from the CT image using marker-based gold standard transformation.

dient-based search, the total execution time is about 25 s. The time of 3-D motion estimation using 2-D tracking is 5 s for the first frame, but only 5 ms for subsequent frames (the additional time for the first frame is due to the estimation of the matrix $\mathbf{A}$ in (11)). Even with common 2-D image registration, the average processing time is about 1 s per frame using the MI metric, and about 5 s using PI. This is still substantially faster than 2-D to 3-D image registration. Even if 3-D motion estimation using 2-D tracking turns out in further studies to be less accurate than using repeated 2-D to 3-D image registration, 2-D tracking in this study was quite accurate and thus is potentially useful as an extremely fast indicator of patient motion, which has clinical relevance for radiosurgery.

V. Discussion

This paper has introduced the concept of tracking the motion of a target region in 3-D based on the consistent back-projection of 2-D motion observed in multiple X-ray projection images. To the best of our knowledge, our work is the first to achieve markerless real-time tracking of 3-D patient motion during image-guided procedures. Our initial results on clinical data from a spinal radiosurgery procedure show that our method is accurate and fast. We have also shown that it can be combined with intensity-based 2-D to 3-D registration and improves both accuracy and computational efficiency of the latter.

The 2-D tracking can take advantage of the full resolution of the X-ray projection images (0.4 mm pixel size), while the 2-D to 3-D registration is essentially limited by the resolution of the preoperative CT image (1.2 mm slice thickness; 1 mm in-plane pixel size). The difference between a typical DRR image and the corresponding actual X-ray projection image is illustrated in Fig. 12.

On the other hand, 2-D tracking cannot correctly identify components of the 3-D transformation that are out of plane for the respective projection. In its current form, our method cannot predict all three angles of 3-D rotations, since for two projection images, at least one angle is always entirely out of plane for both. Also, changes in the tracked region due to out-of-plane rotations can potentially interfere with the correct estimation even of the in-plane motion components: because X-ray images are line integrals of attenuation coefficients encountered along rays from the X-ray source to the detector, the image features used by the 2-D tracker can change with rotation.

Note that we evaluate our method using clinical data from an actual treatment and imaging system that was optimized for marker-based target tracking. Two perpendicular projection images are well-suited for triangulation of fiducial marker position, whereas for tracking using 2-D motion a third projection would be needed to capture all 6 degrees of freedom of the 3-D rigid transformation. The design of the imaging system therefore creates a bias against our method that could be avoided by designing an imaging setup specifically with 2-D motion tracking in mind. While adding a third projection would increase the radiation exposure to the patient by 50% per frame, this may be made up for by the advantages of avoiding marker implantation. Ultimately, an active X-ray acquisition system [36] that monitors and adapts to target motion may not only improve tracking accuracy, but also reduce radiation exposure by optimizing the imaging-related X-ray dose.

The intensity-based 2-D to 3-D registration does not suffer from these limitations. Thus, although 3-D motion estimation from 2-D tracking was more accurate than 2-D to 3-D registration for the ten patients in this study, this might not generally be true, especially for more substantial patient motion involving rotation about the axis that is out of plane for both X-ray projection images. Using three perpendicular projection images and occasional reinitialization of the tracking algorithm after rotations have exceeded a maximum threshold, we hope to also make the tracking robust to changes of the tracked features due to out-of-plane rotations.

Unlike marker-based registration, tracking motion from our clinical data is further complicated by the relatively large time between X-ray acquisitions, which is on the order of about 1 min. At video frame rates, moderate object motion leads to relatively small motion from one frame to the next. From a clinical perspective, it may also be advantageous to reduce the time between X-ray acquisition, since patient motion that occurs immediately after acquisition of one set of X-rays will remain undetected until the next set is acquired. So in the worst case, the treatment targeting may be inaccurate for almost 1 min.

In this paper, we have focused on the problem of determining incremental patient motion from successive X-ray projection images. However, all 3-D motion estimates that result from tracking are only relative to the absolute 3-D transformation $\mathbf{T}^{(0)}$ between patient and CT image coordinates at time 0. A possible solution to obtain $\mathbf{T}^{(0)}$ is illustrated in Fig. 13 and makes use of the fact that tracking the relative motion can be performed without knowledge of $\mathbf{T}^{(0)}$. One can therefore initiate a full 2-D to 3-D registration as soon as the X-ray images at time 0 are available. Tracking of the 2-D motion begins with the second frame and proceeds in parallel until the 2-D to 3-D registration completes. All relative transformations computed by tracking so far and in the future relate to the resulting transformation from 2-D to 3-D registration. This implies that before the registration completes, we have to allow for an *initialization phase*, during which we can determine relative patient motion, but not yet the absolute patient position. Because tracking also requires negligible computation time compared to typical X-ray frame rates (the CyberKnife ASDs currently can produce an X-ray image every 4 s), the 2-D to

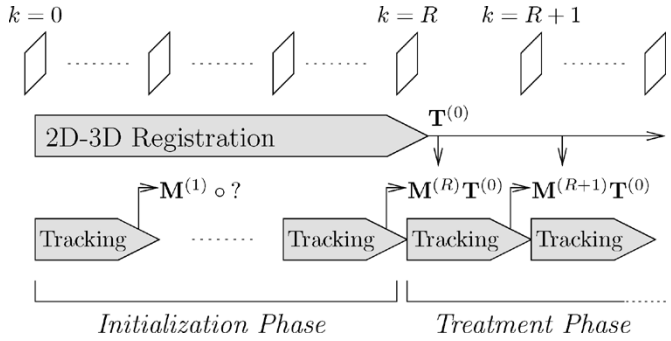


Fig. 13. A possible solution for the first frame registration problem: separation of initialization phase (registration and tracking) and treatment phase (tracking only). Registration requires computation time equal to or less than acquiring and tracking R projection images. Once registration completes, the absolute transformation $\mathbf{T}^{(0)}$ is not known, and treatment commences using the relative patient motion $\mathbf{M}^{(k)}$ for $k \geq R$.

3-D registration can even be performed on the same CPU as the tracking without substantial delay. In order to reduce X-ray exposure during the initialization phase, X-ray acquisition can be suspended, but this may complicate relative motion tracking as patient motion may have accumulated in the meantime.

Another issue we have not addressed in this paper is how the accuracy of the initial transformation effects the accuracy of the 3-D target motion estimate, which is the composition of the initial transformation and the relative motion estimated by tracking. As West & Maurer [37] showed, TREs of independent serial transformations add in quadrature, i.e., the square of the total TRE is the sum of the squares of the individual TREs. This implies that the TRE of the combined transformation in our application is smaller than, or in the worst case equal to, the sum of the individual TREs of initial transformation and relative motion estimate. If, for example, both transformations have the same TRE, then the TRE of their concatenation is only $\sqrt{2} \approx 1.4$ times the individual TRE.

We note that, according to the results presented in this paper, the accuracy of intensity-based 2-D to 3-D registration is less than that achieved by 3-D motion backprojection. However, as Russakoff *et al.* [38] showed recently, incorporating a single fiducial marker in the intensity-based 2-D to 3-D registration can greatly improve its robustness. Using only a single marker substantially reduces patient discomfort and risk of complications during implantation, but it is obviously not sufficient for marker-based registration. It may, however, be a viable strategy when combined with a hybrid marker-based and intensity-based 2-D to 3-D registration for obtaining the initial transformation estimate, and our markerless target tracking algorithm for relative motion tracking. Alternatively, recently presented results by Van der Kraats *et al.* [39] suggest that DRR-based registration methods may not be the best choice for spinal applications, and that gradient-based methods may perform substantially better. Such methods could also easily be used to obtain the initial absolute position estimate needed for motion tracking.

Certainly, since we rely exclusively on clinical data for our quantitative evaluation, data from more patients is needed for reliable results regarding the application accuracy and robustness of our method. Also, like intensity-based 2-D to 3-D registration, our algorithm is limited to tracking bony structures that are clearly visible in projection X-ray images. It cannot be applied to tracking soft-tissue organs that are not attached to bony structures, such as the prostate or the liver. Nevertheless, we conclude from our preliminary analysis that markerless in-plane motion tracking and motion backprojection is a promising technique that is capable of tracking 3-D target region motion in real time with high accuracy.

ACKNOWLEDGMENT

The authors are grateful to K. Mori (University of Nagoya, Japan) for providing his optimized ray casting engine for DRR-based 2-D to 3-D registration. J. Dooley, G. Kuduvalli, and M. Core (Accuray, Inc., Sunnyvale, CA) generously provided technical information, file formats, and utilities that helped us read and use the CyberKnife data. A. Ho (Stanford University, Stanford, CA) retrieved and transferred the archived CyberKnife data. The authors thank the anonymous reviewers for their numerous helpful comments and suggestions that have substantially improved this paper.

REFERENCES

[1] J. R. Adler Jr., M. J. Murphy, S. D. Chang, and S. L. Hancock, "Image-guided robotic radiosurgery," *Neurosurgery*, vol. 44, no. 6, pp. 1299–1306, 1999.

[2] M. Murphy, "Tracking moving organs in real time," *Semin. Radiat. Oncol.*, vol. 14, no. 1, pp. 91–100, 2004.

[3] A. Schweikard, H. Shiomi, and J. Adler, "Respiration tracking in radiosurgery," *Med. Phys.*, vol. 31, no. 10, pp. 2738–2741, 2004.

[4] L. Lemieux, R. Jagoe, D. R. Fish, N. D. Kitchen, and D. G. T. Thomas, "A patient-to-computed-tomography image registration method based on digitally reconstructed radiographs," *Med. Phys.*, vol. 21, no. 11, pp. 1749–1760, 1994.

[5] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. G. Hill, and D. J. Hawkes, "A comparison of similarity measures for use in 2-D to 3-D medical image registration," *IEEE Trans. Med. Imag.*, vol. 17, no. 4, pp. 586–595, Aug. 1998.

[6] G. P. Penney, P. G. Batchelor, D. L. G. Hill, D. J. Hawkes, and J. Weese, "Validation of a two- to three-dimensional registration algorithm for aligning preoperative CT images and intraoperative fluoroscopy images," *Med. Phys.*, vol. 28, no. 6, pp. 1024–1032, 2001.

[7] A. Hamadeh, S. Lavallee, and P. Cinquin, "Automated 3-dimensional computed tomographic and fluoroscopic image registration," *Comput. Aided Surg.*, vol. 3, no. 1, pp. 11–19, 1998.

[8] D. Tomaževič, B. Likar, T. Slivnik, and F. Pernuš, "3-D/2-D registration of CT and MR to X-ray images," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1407–1416, Nov. 2003.

[9] H. Livyatan, Z. Yaniv, and L. Joskowicz, "Gradient-based 2-D/3-D rigid registration of fluoroscopic X-ray to CT," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1395–1406, Nov. 2003.

[10] F. Jurie and M. Dhome, "Hyperplane approximation for template matching," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 7, pp. 996–1000, 2002.

[11] D. Sarrut and S. Clippe, "Geometrical transformation approximation for 2-D/3-D intensity-based registration of portal images and CT scan," in *Lecture Notes in Computer Science*, W. J. Niessen and M. A. Viergever, Eds. Berlin, Germany: Springer-Verlag, 2001, vol. 2208, Proc. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2001: 4th Int. Conf., pp. 532–540.

[12] D. Plattard, M. Soret, J. Troccaz, P. Vassal, J.-Y. Giraud, G. Champleboux, X. Artignan, and M. Bolla, "Patient set-up using portal images: 2-D/2-D image registration using mutual information," *Comput. Aided Surg.*, vol. 5, no. 4, pp. 246–262, 2000.

[13] W. Birkfellner, J. Wirth, W. Burgstaller, B. Baumann, H. Staedele, B. Hammer, N. C. Gellrich, A. L. Jacob, P. Regazzoni, and P. Messmer, "A faster method for 3-D/2-D medical image registration—a simulation study," *Phys. Med. Biol.*, vol. 48, no. 16, pp. 2665–2680, 2003.

[14] J. Brewer, M. Betke, D. P. Gierga, and G. T. Chen, "Real-time 4D tumor tracking and modeling from internal and external fiducials in fluoroscopy," in *Lecture Notes in Computer Science*, C. Barillot, D. P. Haynor, and P. Hellier, Eds. Berlin, Germany, 2004, vol. 3217, Medical Image Computing and Computer-Assisted Intervention—MICCAI 2004, 7th Int. Conf., pp. 594–601.

[15] C. Gräßl, T. Zinßer, and H. Niemann, "Illumination insensitive template matching with hyperplanes," in *Lecture Notes in Computer Science*, B. Michaelis and G. Krell, Eds. Berlin, Germany, 2003, vol. 2781, Proc. Pattern Recognition—25th DAGM Symp., Magdeburg, Germany, pp. 273–280.

[16] C. Gräßl, T. Zinßer, and H. Niemann, "Efficient hyperplane tracking by intelligent region selection," in *Proc. 6th IEEE Southwest Symp. Image Analysis and Interpretation*, Lake Tahoe, NV, Mar. 2004, pp. 51–55.

[17] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 10, pp. 1025–1039, Oct. 1998.

[18] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures," *Med. Phys.*, vol. 24, no. 1, pp. 25–35, 1997.

[19] J. Weese, G. P. Penney, P. Desmedt, T. M. Buzug, D. L. G. Hill, and D. J. Hawkes, "Voxel-based 2-D/3-D registration of fluoroscopy images and CT scans for image-guided surgery," *IEEE Trans. Inform. Technol. Biomed.*, vol. 1, no. 4, pp. 284–293, Dec. 1997.

[20] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3-D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.

[21] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[22] A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *Lecture Notes in Computer Science*, W. M. Wells, III, A. C. F. Colchester, and S. Delp, Eds. Berlin, Germany: Springer-Verlag, 1998, vol. 1496, Proc. Medical Image Computing and Computer-Assisted Intervention—MICCAI'98, 1st Int. Conf., pp. 1115–1124.

[23] P. A. Viola, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[24] J. Weese, T. M. Buzug, C. Lorenz, and C. Fassnacht, "An approach to 2-D/3-D registration of a vertebra in 2-D X-ray fluoroscopies with 3-D CT images," in *Lecture Notes in Computer Science*, J. Troccaz, W. E. L. Grimson, and R. Mösges, Eds. Heidelberg, Germany, 1997, vol. 1205, Proc. CVRMed-MRCAS'97, 1st Joint Conf. Computer Vision, Virtual Reality and Robotics in Medicine and Medial Robotics and Computer-Assisted Surgery, pp. 119–128.

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[26] T. Rohlfing, D. B. Russakoff, J. Denzler, K. Mori, and C. R. Maurer Jr., "Progressive attenuation fields: fast 2-D to 3-D registration without precomputation," *Med. Phys.*, vol. 32, no. 9, pp. 2870–2880, Sep. 2005.

[27] D. B. Russakoff, T. Rohlfing, D. Rueckert, and C. R. Maurer Jr., "Fast calculation of digitally reconstructed radiographs using light fields," *Proc. SPIE—Medical Imaging: Image Processing*, vol. 5032, pp. 684–695, 2003.

[28] D. B. Russakoff, T. Rohlfing, K. Mori, D. Rueckert, J. R. Adler Jr., and C. R. Maurer Jr., "Fast generation of digitally reconstructed radiographs using attenuation fields with application to 2-D to 3-D image registration," *IEEE Trans. Med. Imag.*, vol. 24, no. 11, pp. 1441–1454, Nov. 2005.

[29] D. LaRose, "Iterative X-ray/CT registration using accelerated volume rendering," Ph.D. thesis, Carnegie Mellon Univ., Pittsburgh, PA, 2001.

[30] K. Mori, Y. Suenaga, and J.-i. Toriwaki, "Fast software-based volume rendering using multimedia instructions on PC platforms and its application to virtual endoscopy," *Proc. SPIE—Medical Imaging 2003: Physiology and Function: Methods, Systems, and Applications*, vol. 5031, pp. 111–122, 2003.

[31] M. Sramek and A. Kaufman, "Fast ray-tracing of rectilinear volume data using distance transforms," *IEEE Trans. Visual. Comput. Graphics*, vol. 6, no. 3, pp. 236–252, Jul.-Sep. 2000.

[32] C. R. Maurer Jr., R. Qi, and V. Raghavan, "A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 265–270, Feb. 2003.

[33] J. M. Fitzpatrick, J. B. West, and C. R. Maurer Jr., "Predicting error in rigid-body, point-based registration," *IEEE Trans. Med. Imag.*, vol. 17, no. 5, pp. 694–702, Oct. 1998.

[34] T. Rohlfing, J. Denzler, D. B. Russakoff, C. Gräßl, and C. R. Maurer, Jr., "Markerless real-time target region tracking: application to frameless stereotactic radiosurgery," in *Proc. 9th Fall Workshop Vision, Modeling, and Visualization,* , B. Girod, M. Magnor, and H.-P. Seidel, Eds., 2004, pp. 5–12.

[35] D. B. Russakoff, T. Rohlfing, A. Ho, D. H. Kim, R. Shahidi, J. R. Adler Jr., and C. R. Maurer Jr., "Evaluation of intensity-based 2-D to 3-D spine image registration using clinical gold-standard data," in *Lecture Notes in Computer Science*, J. C. Gee, J. B. A. Maintz, and M. W. Vannier, Eds. Berlin, Germany: Springer-Verlag, 2003, vol. 2717, Proc. Biomedical Image Registration—2nd Int. Workshop, WBIR 2003, pp. 151–160.

[36] J. Denzler and C. M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 145–157, Feb. 2002.

[37] J. B. West and C. R. Maurer Jr., "Designing optically tracked instruments for image-guided surgery," *IEEE Trans. Med. Imag.*, vol. 23, no. 5, pp. 533–545, May 2004.

[38] D. B. Russakoff, T. Rohlfing, J. R. Adler Jr., and C. R. Maurer Jr., "Intensity-based 2-D to 3-D spine image registration incorporating a single fiducial marker," *Acad. Radiol.*, vol. 12, no. 1, pp. 37–50, 2005.

[39] E. B. van de Kraats, G. P. Penney, D. Tomaževič, T. van Walsum, and W. J. Niessen, "Standardized evaluation of 2-D to 3-D registration," in *Lecture Notes in Computer Science*, vol. 3216, Medical Image Computing and Computer-Assisted Intervention—MICCAI 2004, 7th Int. Conf., Pt. I, C. Barillot, D. P. Haynor, and P. Hellier, Eds. Berlin, Germany, 2004, pp. 574–581.