Published in final edited form as:
Phys Med Biol. 2016 April 21; 61(8): 3009–3025. doi:10.1088/0031-9155/61/8/3009.

3D–2D image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch

T De Silva[1], **A Uneri**[2], **M D Ketcha**[1], **S Reaungamornrat**[2], **G Kleinszig**[3], **S Vogt**[3], **N Aygun**[4], **S-F Lo**[5], **J-P Wolinsky**[5], and **J H Siewerdsen**[1,2,4,5]

[1] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

[2] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

[3] Siemens Healthcare XP Division, Erlangen, Germany

[4] Department of Radiology and Radiological Science, Johns Hopkins Medical Institute, Baltimore, MD 21287, USA

[5] Department of Neurological Surgery, Johns Hopkins Medical Institute, Baltimore, MD 21287, USA

Abstract

In image-guided spine surgery, robust three-dimensional to two-dimensional (3D–2D) registration of preoperative computed tomography (CT) and intraoperative radiographs can be challenged by the image content mismatch associated with the presence of surgical instrumentation and implants as well as soft-tissue resection or deformation. This work investigates image similarity metrics in 3D–2D registration offering improved robustness against mismatch, thereby improving performance and reducing or eliminating the need for manual masking.

The performance of four gradient-based image similarity metrics (gradient information (GI), gradient correlation (GC), gradient information with linear scaling (GS), and gradient orientation (GO)) with a multi-start optimization strategy was evaluated in an institutional review board-approved retrospective clinical study using 51 preoperative CT images and 115 intraoperative mobile radiographs. Registrations were tested with and without polygonal masks as a function of the number of multistarts employed during optimization. Registration accuracy was evaluated in terms of the projection distance error (PDE) and assessment of failure modes (PDE > 30 mm) that could impede reliable vertebral level localization.

With manual polygonal masking and 200 multistarts, the GC and GO metrics exhibited robust performance with 0% gross failures and median PDE < 6.4 mm (±4.4 mm interquartile range (IQR)) and a median runtime of 84 s (plus upwards of 1–2 min for manual masking). Excluding manual polygonal masks and decreasing the number of multistarts to 50 caused the GC-based registration to fail at a rate of >14%; however, GO maintained robustness with a 0% gross failure rate. Overall, the GI, GC, and GS metrics were susceptible to registration errors associated with

jeff.siewerdsen@jhu.edu.

content mismatch, but GO provided robust registration (median PDE = 5.5 mm, 2.6 mm IQR) without manual masking and with an improved runtime (29.3 s).

The GO metric improved the registration accuracy and robustness in the presence of strong image content mismatch. This capability could offer valuable assistance and decision support in spine level localization in a manner consistent with clinical workflow.

Keywords

3D–2D registration; spine surgery; image similarity metrics; image-guided surgery; intraoperative imaging

1. Introduction

In image-guided spine surgery, integrating three-dimensional (3D) preoperative images and two-dimensional (2D) intraoperative images via 3D–2D registration can provide valuable assistance during target localization. For instance, the overlay of vertebral labels identified in preoperative computed tomography (CT) images onto the intraoperative radiographs can provide decision support system for surgeons in accurate vertebral level identification (Otake *et al* 2012b, 2013, 2015, Lo *et al* 2015). This could potentially decrease wrong-level errors, reported to occur in approximately 1 in 3110 spine surgery procedures (Mody *et al* 2008), and reduce the time and stress associated with resolving anatomical uncertainty. This registration requires a high degree of robustness in registration performance. However, the presence of surgical instrumentation, hardware implants, and soft-tissue resection/displacement is common in intraoperative radiographs relative to preoperative CT images, causing image content mismatch and potentially confounding image-based registration. Manual/semi-automatic methods to mask out such extraneous (or missing) content are time consuming, user dependent, error prone, and disruptive to workflow. The development and validation of image registration methods (Nithiananthan *et al* 2011, Uneri *et al* 2013) that are robust in the presence of content mismatch are important to achieving reliable registration in a manner consistent with the needs of intraoperative workflow and clinical integration.

Several image similarity metrics have been proposed in 3D–2D registration involving x-ray and CT images. Table 1 lists a number of previous studies comparing the performance of various image similarity metrics in 3D–2D registration for clinical applications ranging from image guidance in the spine (Penney *et al* 1998) and intracranial interventions (Hipwell *et al* 2003, McLaughlin *et al* 2005) to patient positioning/target alignment in radiotherapy (Khamene *et al* 2006, Kim *et al* 2007, Wu *et al* 2009, Gendrin *et al* 2011). While mismatch in image content due to changes in anatomy and/or instrumentation between images acquired at two different time points is commonly encountered in many clinical procedures, the degree of mismatch and the extent to which it challenges image-based registration depend strongly on the specific clinical procedure under consideration. In spine surgery, for example, the following sources of mismatch are common occurrences (as shown in figure 1) and present particular challenges to 3D–2D registration: (a) The presence of surgical instrumentation in the intraoperative radiograph (e.g. retractors, screws, implants, clips, etc); and (b) non-rigid anatomical deformation (e.g. soft tissues, gas inside the colon, and skin lines) and anatomical positioning (e.g. arms up, or at the sides), especially when the intraoperative patient setup does not match the position in preoperative scanning (e.g. preoperative supine CT versus intraoperative prone radiograph).

Penney *et al* (1998) presented related work in the context of spine surgery, where the performance of multiple image similarity metrics was evaluated in 3D–2D registration in the presence of partial mismatch. However, the mismatch in that study was simulated using phantom data with clinical image features overlaid. While the amount of clinical data used for validation in the studies listed in table 1 is limited, clinical images often pose additional challenges for similarity metrics during registration (Penney *et al* 1998, McLaughlin *et al* 2005, Wu *et al* 2009). Given the high degree of mismatch common in spine surgery procedures and the results of previous studies showing variability in similarity metric behavior in clinical images, validation in a large clinical data set is essential to the assessment of accuracy and robustness in registration performance under mismatch. Moreover, such a data set captures the broad variety of realistic anatomical features and pathologies that challenge image registration.

From the similarity metrics that have been previously investigated, gradient-based metrics have shown promise in 3D–2D image registration problems involving intraoperative radiographs and preoperative CT images (Penney *et al* 1998, Gendrin *et al* 2011, Otake *et al* 2012b, 2015). Image gradients can be used to filter out low spatial frequency differences such as soft tissue (Penney *et al* 1998, Pluim *et al* 2000) and act as a high-pass filter to focus the registration on the bony anatomy clearly visible in x-ray images. Gradient information (GI), gradient correlation (GC), and gradient difference (GD) have shown superior performance in 3D–2D registration applications in previous studies, while GI and GC metrics have been developed with the inherent capacity to handle registrations with partial mismatch. For example, previous work on vertebral level localization in spine surgery (Lo *et al* 2015, Otake *et al* 2015) showed promising results with GC similarity. However, the challenge due to content mismatch was mitigated by manually defined polygonal masks that excluded extraneous regions during similarity metric evaluation. The definition of polygons to mask extraneous objects is a slow, subjective step that is error prone and limits the practical utility of image-based registration within streamlined workflow. Alternatively, automatic/semi-automatic segmentation methods (Linguraru *et al* 2007) could be implemented as a pre-processing step to mask extraneous objects within the field-of-view of the radiograph. However, such an approach increases the computational burden and makes the registration algorithm susceptible to segmentation errors.

In order to achieve robust 3D–2D registration, the registration framework (and specifically the similarity metric) should inherently handle content mismatch, rather than rely upon user input that is subject to variability. In this work, we evaluate the performance of gradient-based image similarity metrics that have shown promise in previous work, and we develop modifications that yield further improvement. Registration performance was evaluated in a fairly large clinical data set (51 patients and 115 radiographs) that consisted of preoperative CT images and intraoperative radiographs acquired during thoracolumbar spine surgery procedures. The accuracy and robustness of vertebral level localization were quantified in terms of geometric accuracy (projection distance error) and failure rate (attributed to previously identified failure modes (Lo *et al* 2015)) in the context of vertebral level localization. We also evaluated the properties of the objective function search space and compared the capture ranges for each metric. The system is envisioned as a means of decision support and an 'independent check' in level localization by projecting preoperative CT vertebral labels onto the intraoperative radiographs via 3D–2D registration.

2. Methods and materials

2.1. Clinical study

The clinical data were collected retrospectively in an institutional review board (IRB) approved study from patients undergoing thoracolumbar spine surgery within the past four years at our institution. Basic inclusion criteria were the availability of a preoperative CT image acquired within one year prior to the surgery date and at least one radiograph acquired during the procedure. From an initial data set comprising 60 patients and containing 154 intraoperative radiographs, some data were excluded according to the following criteria: (1) Inappropriate field of view (12 radiographs demonstrated inadequate positioning of the detector and/or collimators with respect to vertebral levels of interest); (2) poor image quality (9 radiographs exhibited poor image contrast likely due to suboptimal imaging techniques); and (3) limited overlap/correspondence between the CT and radiograph (8 cases presented preoperative CT that appear to have been acquired for purposes other than the spinal pathology or surgical planning and contained little or no overlap with the intraoperative radiograph).

The remaining data set comprised 51 patients and 115 radiographs for the study reported below. These data exhibited a substantial degree of realistic variation in imaging protocols/techniques, image quality, and imaging systems used for preoperative and intraoperative imaging. For example, the 51 preoperative CT images included three scanner manufacturers (Toshiba Corporation, Tokyo, Japan; Siemens Healthcare, Erlangen, Germany; and GE Healthcare, Little Chalfont, UK), and scan techniques varied from 120–140 kVp and 80–660 mAs with slice thickness ranging from 0.24–3.00 mm. All of the intraoperative radiographs were acquired using a common model of mobile radiography system (DRX-1, Carestream Health, Rochester, NY) with 0.14 × 0.14 mm pixel dimensions and included 105 lateral (LAT) and 10 anterior–posterior (AP) projections. All the CT images were acquired with the patient supine, whereas intraoperative radiographs were typically in prone position.

2.2. Image pre-processing and target definition

2.2.1. Spine level definition—Preoperative CT images were de-identified and prepared for registration by first defining the vertebral levels. Labeling in CT can potentially be done automatically (Klinder *et al* 2009, Ma and Lu 2013), but in this study, a board-certified neuroradiologist annotated the vertebral labels at the approximate centroid of each vertebral body. For each CT scan, all the vertebrae within the scan volume were labeled according to standard designations (T1, T2, etc), including normal anatomical variations (e.g. the existence of T13 and/or L6).

2.2.2. 2D radiographic masking and 3D volumetric masking—To constrain the registration to regions proximal to the spine and to test the registration robustness in the presence of image content mismatch, the images were masked using two different approaches: (1) Manual 2D radiographic masking; and (2) automatic 3D CT masking. Manual 2D radiographic masking was performed as in (Otake *et al* 2015) to exclude regions associated with content mismatch. Previous radiographic masking (Lo *et al* 2015, Otake *et al* 2015) involved manual definition of polygonal exclusion masks (denoted *p*) to exclude regions with instrumentation, surgical implants, and strong gradients produced by non-rigid anatomy; however, manual masking is a time-consuming and subjective process. A faster, less labor-intensive approach would be advantageous to clinical implementation considering the time sensitive workflow requirements in the operating room. For example, we implemented a simple rectangle (denoted *r*) manually defined by two points capturing the region containing vertebrae and possibly avoiding areas due to collimation/text annotations.

On the other hand, automatic 3D volumetric masking applies a binary spatial mask in the CT image around the labeled vertebrae and requires no additional effort in the workflow, since the label positions have already been identified. From the label coordinates in the CT image, a 3D mask was automatically generated as an elliptical cylinder (minor axis 50 mm and major axis 25 mm) encompassing the region about the vertebrae and applied to all the registrations performed in this study.

In all cases, to further decrease the effect of deformable soft tissue gradients on the registration performance, a simple intensity threshold (50 HU) was applied to the CT image as in (Otake *et al* 2012b) to exclude anatomical regions of low intensity.

2.3. 3D–2D registration framework

2.3.1. Basic framework—The registration is performed by optimizing the image similarity between the preoperative CT image and the intraoperative radiograph $\left(I_2 : \mathbb{R}^2 \to \mathbb{R}\right)$ in a rigid 6D transformation space. Figure 2 illustrates the overall 3D–2D registration process and variable parameters relating to masking (*p*, *r*, or *v*) and similarity metrics (section 2.3.2).

During registration, the image similarity is computed between the intraoperative radiograph I_2 and a digitally reconstructed radiograph (DRR) ($\left(I_1 : \mathbb{R}^2 \to \mathbb{R}\right)$) generated from the CT image. A parallelized GPU implementation is utilized to efficiently generate DRRs (Otake *et al* 2012a). The patient and imaging system positioning and intrinsic parameters of system geometry were initialized from a set of pre-defined discrete configurations. While this initialization approximated the patient and imaging setup, simple additional manual initialization was performed by translating the CT image along the superior–inferior (S–I) direction to ensure a reasonable initial overlap between the DRR and the radiograph. The PDE following initialization ranged from 11.3 to 111.1 mm (median PDE = 43.1 mm, interquartile range 20.9 mm).

The image similarity metric is an important component within the registration framework that plays an important role in the robustness of registration in the presence of content mismatch. The GI and GC metrics have been shown previously (Penney *et al* 1998, Pluim *et al* 2000, Otake *et al* 2012b) to be useful when handling partial content mismatch. We introduce modified versions of these metrics with the aim of improving registration robustness. The details of the modified similarity metrics are described in section (2.3.2).

Content mismatch caused by extraneous objects and tissue deformation can create a challenging search space with multiple local optima. To achieve robust performance under these conditions, the registration process incorporates a multistart optimization with the covariance matrix adaptation; evolution strategy (CMA-ES) reported by (Hansen 2006). In CMA-ES, population samples (with nominal population size $\lambda = 100$) were randomly drawn according to a multivariate normal distribution during optimization. At each iteration, the mean, covariance matrix, step-size, and evolution paths are updated deriving from the currently sampled population. For the multistart strategy, the six-dimensional (6D) search space is partitioned to multiple, equal-size subspaces following a k-d tree partitioning algorithm, and a separate CMA-ES search was performed for each subspace initialized at their respective centers (Otake *et al* 2013). The partitioned search range was defined as a region centered on the initialization pose and spanning ±200 mm along the superior–inferior direction, ±100 mm along the anterior–posterior direction, ±100 mm along the left–right direction, and ±10 degrees along each of the three rotational directions in the search space. The solution at the subspace search achieving the highest similarity metric at convergence is selected as the initialization to a subsequent second-level CMA-ES search. Increasing the number of multistarts (denoted S in figure 2) boosts the density of seed points positioned within the multidimensional space and therefore improves the robustness of finding the global optima at the cost of additional computation in performing parallel searches (roughly proportional to the number of multi-starts). The similarity metric evaluations for all the multistarts as well as population sample evaluations within a single start were parallelized in our implementation to improve the computation time, as characterized below.

2.3.2. Similarity metrics

2.3.2.1. Gradient information (GI): When calculating GI, the min operator applied to gradient magnitudes of the two images [$|\nabla_u I_1$ and $|\nabla_u I_2|$ where $u = (x, y)$] intends to exclude strong extraneous gradients. The metric is computed according to equation (1) by combining gradient magnitude and orientation (w) terms where the gradient orientation is computed as the cosine angle between two gradient vectors (cos θ) using the vector dot product.

$$GI=\frac{1}{N}\sum_{i\in\Omega}^{N} w(i)\ min\left(\left|\nabla_u I_1(i)\right|, \quad \left|\nabla_u I_2(i)\right|\right) \qquad (1)$$

where $w(i)=\frac{cos\ \theta_i+1}{2}$ and $cos\ \theta_i=\frac{\nabla_u I_1(i).\nabla_u I_2(i)}{\|\nabla_u I_1(i)\|\|\nabla_u I_2(i)\|}$

2.3.2.2. Gradient correlation (GC): GC is calculated as the normalized cross correlation (NCC) between two gradient images according to equation (2). The corresponding gradient images in intraoperative radiographs and DRRs can vary due to the differences in imaging techniques used or due to image content mismatch. GC can accommodate such differences up to a linear factor and therefore offers some degree of robustness against mismatch:

$$GC=\frac{1}{2}\left\{\sum_{i\in\Omega}^{N}\frac{\nabla_x I_1(i).\nabla_x I_2(i)}{\|\nabla_x I_1(i)\|\|\nabla_x I_2(i)\|}+\sum_{i\in\Omega}^{N}\frac{\nabla_y I_1(i).\nabla_y I_2(i)}{\|\nabla_y I_1(i)\|\|\nabla_y I_2(i)\|}\right\} \qquad (2)$$

2.3.2.3. Gradient information with linear scaling (GS): The min operator in GI assumes that anatomical structures produce a constant gradient magnitude in both DRRs and radiographs. However, this assumption is often not valid, since extraneous objects, differences in real/simulated projections may contribute to the similarity metric value. To mitigate this effect, we introduced a linear scaling factor (a) to match different levels of corresponding gradient magnitudes. This scaling factor (a) was determined dynamically as $a = \gamma \cdot \max(\nabla_u I_1)/\max(\nabla_u I_2)$. The parameter γ is a constant that accounts for the presence of strong extraneous gradient magnitudes in I_2. With the scaling factor (a), GS is calculated as:

$$GS=\frac{1}{N}\sum_{i\in\Omega}^{N} w(i)\ min\left(|\nabla_u I_1(i)|,\quad \alpha|\nabla_u I_2(i)|\right) \qquad (3)$$

2.3.2.4. Gradient orientation (GO): While all of the above metrics rely upon gradient magnitudes, gradient orientation on its own has been widely used for feature matching (Lowe 2004), object detection (Dalal and Triggs 2005), and image registration (De Nigris *et al* 2012), and shown promise in robust matching under partial occlusion and view point changes. In this study, we introduce a GO metric to achieve robust registration performance under mismatch. We hypothesized that GO could help to mitigate the effect of strong gradient magnitudes produced by extraneous instrumentation in the registration solution. However, gradient orientation could be susceptible to noise in regions that produce low gradient magnitudes. Therefore, in computing GO, we only consider pixels with gradient magnitudes exceeding the threshold t_1 or t_2 in the images I_1 or I_2, respectively. These thresholds were defined as the median gradient magnitude of each image. This eliminated 50% of the image pixels containing low gradient magnitudes from the GO computation. Then we perform an intersection operation between the two thresholded gradient magnitudes according to equation (4) and the resulting number of overlapping pixels contributing to the overall GO value can be variable. To prevent degenerate solutions with minimal overlap between the images during metric calculation, we introduce a lower bound (N_{LB}) along with the total number of evaluated pixels N and normalized the metric by $\max(N,N_{\mathrm{LB}})$ to penalize when the number of contributing pixels falls below N_{LB}. In addition, the natural log was used instead of cosine weighting to introduce a more sharply decaying penalty to gradient orientation mismatch:

$$GO=\frac{1}{max\left(N,N_{LB}\right)}\sum_{i\in\{\Omega:|\nabla_u I_1(i)|>t_1\cap|\nabla_u I_2(i)|>t_2\}} w'(i)$$

$$\text{where}\quad w'(i)=\frac{2-\left(ln\left(|cos^{-1}(cos\,\theta_i)|+1\right)\right)}{2} \tag{4}$$

Figure 3 provides a simple illustration of the four similarity metrics defined above. In each case, GI, GC, GS, and GO ‘maps’ (i.e. $M(u, v)$ the metric depicted in the projection domain prior to summation over pixels) are shown for a simple simulation of content mismatch between the DRR I_1 and radiograph I_2. The top row in figure 3 simulates the desired solution (i.e. images aligned) in the presence of content mismatch (the addition of a triangle in I_2) and intensity mismatch (hypodensity of the circle and square in I_2) and the bottom row simulates a degenerate solution (i.e. images misaligned due to additional content) with the same mismatch in content and intensity. While all metrics align corresponding square and circle elements at the desired solution, the mismatching triangle at the degenerate solution causes the GI, GC metrics to yield a larger relative contribution to the overall metric value.

2.4. Performance evaluation

2.4.1. Accuracy, robustness, and run time—

To provide a reference/truth definition in registration, a board-certified neuroradiologist annotated the approximate centroid of each vertebral body in the 2D radiographs. The geometric accuracy of registration was measured using projection distance error (PDE) calculated as the distance (in the projection domain at the detector) between registered CT labels in DRRs and (‘true’) vertebrae centroids in radiographs. For the acquired radiographs, the magnification factor at the detector typically ranged from 1.4–1.6. For each similarity metric, the performance was evaluated as a function of the degree of masking (manual polygonal p or automatic rectangular r) and registration runtime (determined primarily by S, set to 200 or 50 multistarts), with three nominal configurations ($R(p, 200)$, $R(p, 50)$, and $R(r, 50)$) defined as follows:

***R*(*p*, 200)** denoted the configuration in which polygonal masks (p) were manually defined in 2D radiographs, and the number of multistarts in the optimization was $S = 200$. This configuration corresponds to that demonstrated in clinical studies by (Lo *et al* 2015). It has the advantages of excluding extraneous gradients via manual masking and increased robustness due to the high number of multi-starts. However, manual masking presents an obvious challenge to workflow, and the large number of multistarts carries a longer run time.

***R*(*p*, 50)** denoted the configuration in which polygonal masks were manually defined as in the previous case, but the number of multistarts was reduced to $S = 50$. This configuration was intended as a stress test to improve the run time and to examine potential deterioration in registration robustness.

***R*(*r*, 50)** denoted the configuration in which a simple rectangular mask (r) was applied to the 2D x-ray images (simply masks the collimator edges, and could be automated), and the number of multistarts was set to $S = 50$. This configuration exposes all the extraneous gradients within the rectangle and poses a challenging registration problem with content mismatch and relatively few multistarts.

> However, it involves minimal manual intervention and a faster run time, each advantageous to clinical workflow.

In addition to geometric accuracy (PDE), we evaluated the overall robustness of registration in terms of the ‘failed’ registrations in which the registered labels were positioned outside the correct vertebral body. Considering an average thoracolumbar vertebral height (superior–inferior direction) of 22 mm (Busscher *et al* 2010), we defined failure as any registration with PDE > 30 mm (equal to 22 mm (in the object) times an approximate typical magnification factor of ~1.4). Registration was performed on a desktop Windows 7 64-bit workstation with an Intel Xeon 2 processor (2.4 GHz) and GeForce TITAN Black GPU (nVidia, Santa Clara CA), with GPU implementations of the forward projection (DRR) calculation, and similarity metric calculation.

2.4.2. Quality of the objective function search space—To evaluate the effect of the various similarity metrics in improving robustness to image content mismatch, we assessed the quality of the search space for the most challenging configuration $R(r, 50)$. A one-dimensional (1D) surrogate for visualization of the 6D search space was formed from line profiles sampled across different directions centered around a known *ground truth*, obtained from a successful registration (low PDE) using the $R(p, 200)$ configuration. The profiles were normalized such that a unit change in all the transform directions (i.e. both rotations and translations) cause the same mean shift in physical dimensions in the image (Škerl *et al* 2006). This provided a highly simplified visualization of the 6D search space and elucidated how various similarity metrics determined the quality/condition of the search (e.g. more or fewer false local maxima, etc).

2.4.3. Sensitivity to initialization—The simple manual initialization in the S–I direction described above is intended to provide coarse, nominal overlap between the radiographs and the DRR at the start of the registration. For example, if the preoperative CT covers the entire spinal column, but the intraoperative radiograph covers only a portion of the thoracic spine, then the user simply slides a 2D rectangle window on the DRR to roughly demark the thoracic spine. Other forms of initialization variability exist as well, e.g. magnification associated with variable source-to-image distance, but disparity in the S–I position is believed to be the most susceptible for level misidentification (e.g. registration errors resulting in label positions away from the true level by one or more levels in the S–I direction). We evaluated the capture range over which registration was robust against S–I initialization error. For each intraoperative lateral radiograph, we varied the initialization pose (p_0) along the S–I direction from −300 to +300 mm, and PDE was evaluated to quantify the capture range (i.e. extent in the S–I direction for which the registration was robust against failure).

3. Results

3.1. Registration accuracy and robustness

Figure 4 compares the registration performance for each similarity metric under various parameter configurations. Table 1 shows the median ± interquartile range (IQR) in PDE along with failure rates (fraction of cases for which PDE > 30 mm). For the $R(p, 200)$

configuration (i.e. 2D polygonal masks and 200 multistarts), both the GC and GO metrics exhibited an accurate and robust performance. The GI and GS metrics failed to achieve a clinically acceptable level of accuracy, suffering a >16% failure rate, although GS improved the gross failure rate compared to GI by approximately 50%. The performance exhibited by the GC metric is consistent with that in (Lo *et al* 2015).

Reducing to 50 multistarts in the $R(p, 50)$ configuration (as shown in figure 4(b)) improves the runtime (detailed below and in table 2) but compromises the robustness of the registration and results in an increasing failure rate for all of the similarity metrics. The GC and GO metrics maintained superior performance in comparison to GI and GS.

As shown in figure 4(c), configuration $R(r, 50)$ further challenged the algorithm by exposing all the extraneous gradients within the simple rectangular field of view in the radiograph. The performance of the GI, GC, and GS metrics degraded significantly, e.g. the gross failure rate for GC increasing to 15%. The GO metric, on the other hand, maintained geometric accuracy and robustness. A slight improvement in the GO registration when compared with $R(p, 50)$ is attributed to the additional benefit of relevant anatomical information included with rectangle masking but masked out with polygonal masking.

Since our PDE distributions were not normally distributed, we tested the statistical significance of differences between the PDE distributions for each configuration using a non-parametric pairwise Wilcoxon signed rank test. Among distributions $R_{\mathrm{GO}}(p, 200)$, $R_{\mathrm{GO}}(p, 50)$, and $R_{\mathrm{GO}}(r, 50)$, we found p-value >0.94 failing to reject the null hypothesis that the PDE distributions have the same median. This suggests that the challenging configuration $R_{\mathrm{GO}}(r, 50)$ produces a similar PDE distribution to the more manually intensive $R_{\mathrm{GO}}(p, 200)$ and $R_{\mathrm{GO}}(p, 50)$ configurations. Comparing the GI, GS, and GC metrics to the GO metric, table 2 marks the cases (*) for which the PDE distribution was statistically significantly different ($p < 0.05$).

The computation time was primarily a function of the number of multistarts used during the optimization. Owing to GPU implementations for each metric, the average computation time ranged from 23–29 s for the $S = 50$ configurations, compared to 65–85 s for the $S = 200$ configuration. Thus, the improvements in robustness observed with 200 multistarts come with approximately three-fold increase in computation time. Overall, the $R_{\mathrm{GO}}(r, 50)$ configuration outperformed all of the GI, GS, and GC configurations in terms of geometric accuracy, robustness, and run time and presents an implementation that is advantageous with respect to clinical workflow (via a simple, potentially automatic definition of a rectangular mask).

The similarity metric maps shown in figure 5 depict the metric value at each pixel location (prior to summation, as in the toy illustrations in figure 3). In this example, the gradients produced by instrumentation appearing in the radiograph (but not the CT) tend to dominate the similarity metric value for the GI, GC, and GS metrics. For the GI and GC metrics, this tends to direct the registration to a false (erroneous) solution. For the GS metric, this effect is mitigated partly by the scaling parameter a; however, the exclusion of extraneous gradients requires estimating a to accurately match the corresponding gradients in the two images. For the GO metric, the gradient magnitude values do not dominate the overall metric value, and the desired solution is reached wherein corresponding vertebrae gradients align in a manner that is robust against extraneous gradients and occlusions from instrumentation.

3.2. Quality of the objective function search space

The quality of the search space is illustrated in the spaghetti plots of figure 6, each plot showing 1000 1D line profiles centered about the true solution in the 6D objective function search space. The plots correspond to a single case, similar to that in figure 5. For all the metrics, the space exhibits numerous false local optima associated with surgical instruments present in the radiograph (but not the CT) or the semiperiodic structure of the spine itself in the S–I direction. For GI and GS, the peak at the solution is less distinct from surrounding local maxima, although GS appears improved over GI (consistent with improved robustness noted above). The min operator applied in these metrics imparts a flattening effect in the search space and can challenge the optimizer to reach the solution. On the other hand, GC shows a more distinct optimum at the solution, but it also exhibits sharp local maxima attributed to strong gradient magnitudes far from the solution (surgical instruments). Therefore, unless these extraneous gradients are masked from the image (as with the manual *p* mask as in previous work (Lo *et al* 2015), the registration is susceptible to failure. The GO metric exhibits the most distinct global optimum at the solution and suppresses false maxima far from the solution.

3.3. Sensitivity to initialization/capture range

Figure 7 shows the median and third quartile of the distribution in PDE over all cases for each similarity metric as a function of displacement in the S–I direction from the manual initialization. For each case, the capture range was computed as the displacement in the S–I direction that resulted in PDE within 5 mm of that at the true S–I position. Consistent with the results above, GO exhibited the best overall performance with a median capture range of 375 mm ± 122 mm IQR measured at the detector. This capture range spans approximately 12 vertebra in the S–I direction. The other metrics demonstrated a fairly broad capture range as well: GI capture range = 192 mm ± 214 mm IQR; GC capture range = 345 mm ± 185 mm IQR; and GS capture range = 293 mm ± 180 mm IQR. Such broad capture ranges are partly attributable to the similarity metric and perhaps more so to the robust optimization strategy involving multiple parallel searches from multiple initial seed points. The result is encouraging in that it suggests a high level of robustness to S–I initialization error, particularly considering the ability to move through multiple false local maxima associated with one-level error, two-level error, and so on.

4. Discussion and conclusions

Image content mismatch caused by surgical instruments introduced between the preoperative CT and intraoperative radiographs can be problematic in achieving robust registration. Our results indicate that similarity metrics that rely upon gradient magnitudes tend to be more susceptible to registration failure under such conditions. This may be attributed to the fact that the extraneous objects produce stronger gradient magnitudes than those of the vertebrae and therefore can dominate the search space and drive the solution to false local maxima.

The main advantage of GO is that both the instrumentation and vertebrae gradients contribute similar weights to the overall metric value. As long as vertebral bodies produce a larger spatial distribution of gradients in comparison to those produced by the extraneous objects, the desired solution of accurately aligning vertebrae is more favorable within the GO search space. The GO similarity metric thereby exhibits robust performance against content mismatch and enables the definition of a simple rectangular mask of the radiograph collimators to reduce or eliminate the need for polygonal masking.

In addition to the time required during surgery, the definition of manual polygonal masks is subject to user variability and tends to exclude regions that may be disadvantageous to registration. For example, complex polygonal masks about surgical implants within vertebrae can partially occlude vertebral boundaries (cortical margins) that are salient features for registration. The slight improvement in registration accuracy for $R_{\mathrm{GO}}(r, 50)$ compared to $R_{\mathrm{GO}}(p, 50)$ could be attributable to this additional image content available with the simple rectangular mask. Another drawback of the polygonal masking approach is that it requires the user to anticipate and identify problematic gradients in a radiograph with many subtleties. Considering the multiple sources of image mismatch, ranging from skinline deformation to surgical instrumentation, the manual delineation of such masks could be cumbersome and susceptible to user variability and repeat registration with mask refinement.

Although the GS metric somewhat improved the registration performance compared to GI, it did not exhibit particularly strong robustness to failure (>16%) over all the tested configurations. One challenge in the design of the GS metric is the selection of the parameter that attempts to match the gradient magnitudes of similar anatomical objects in the DRR and radiograph. We evaluated this parameter by computing the ratio of maximum gradient magnitudes observed between the two images; however, the parameter should be adjusted according to mismatch caused by strong gradient magnitudes. Automatic determination of the parameter for any given pair of images is non-trivial. With the objective of improving performance in GC under mismatch we experimented with a modified version according to equation (5):

$$GC_{\beta}=\frac{1}{2}\left\{\sum_{i=\beta\%}^{(100-\beta)\%}\frac{\nabla_x I_1(i)\,.\nabla_x I_2(i)}{\|\nabla_x I_1(i)\|\|\nabla_x I_2(i)\|}+\sum_{i=\beta\%}^{(100-\beta)\%}\frac{\nabla_y I_1(i)\,.\nabla_y I_2(i)}{\|\nabla_y I_1(i)\|\|\nabla_y I_2(i)\|}\right\} \quad (5)$$

by truncating β percentiles from the top-most and the bottom-most values prior to GC summation. The top β percentile was excluded to eliminate extraneous strong gradients due to instrumentation, whereas the bottom β percentile was excluded to mitigate the contribution from soft tissues. While the modified metric showed an improved overall performance compared to conventional (untruncated) GC, it suffered from the same limitation of automatically determining the parameter β that is dependent on the degree of image content mismatch.

The challenging search space illustrated in figure 6 shows multiple local maxima resulting from extraneous objects and stands to benefit from the multistart optimization strategy to improve robustness. The benefit of increasing the number of multistarts is evident in table 2 (column $R(p, 200)$ relative to $R(p, 50)$), showing a reduction in failure rate (i.e. PDE > 30 mm). However, increasing the number of multistarts did not necessarily improve the precision of the registration, as observed in the median ± IQR PDE, but such mm-scale variations in PDE in table 2 were not statistically significant ($p = 0.52, 0.72, 0.10$, and 0.07 for GI, GC, GS, and GO, respectively) and are likely not clinically significant with respect to localizing a particular vertebral body. The small variations in PDE could be attributed to errors in the localization of the vertebrae centroid (expert reader truth definition), the stochasticity of the CMA-ES optimizer, and/or fundamental limitations arising from the use of similarity metrics as surrogate measures of registration accuracy during the optimization process. Moreover, the corresponding increase in robustness has diminishing return due to the 'curse of dimensionality' in the 6D search space, whereas the computation time increases approximately linearly with the number of multistarts. Due to its ability to handle mismatch, GO yielded a comparable performance when the number of multistarts was decreased from 200 to 50. A smaller number of multistarts is desirable in decreasing the computation time. Therefore, it is desirable for the registration framework to rely less upon the number of multistarts, an attribute exhibited with the incorporation of the GO similarity metric.

Deformation of the spine is not specifically addressed in the current (rigid) registration approach, but the large clinical dataset did present realistic levels of anatomical deformation. The main objective in this work was to achieve the best rigid alignment even in the presence of realistic deformation. Given the clinical need to identify the correct vertebral level within approximately 30 mm PDE (i.e. within the correct vertebral body), the rigid registration framework may be sufficient if the labels can be projected in close proximity to the accurate vertebral levels when deformation occurs. This assertion is the subject of ongoing clinical evaluation of the algorithm.

In conclusion, we have extended the LevelCheck 3D–2D registration algorithm to incorporate an improved objective function that exhibits stronger robustness to the image content mismatch observed in real clinical images. These improvements better enable clinical translation of the registration algorithm, and we are currently evaluating registration performance in a prospective clinical study.

Acknowledgments

This work was supported by NIH Grant No. R01-EB-017226 and academic-industry collaboration with Siemens Healthcare (XP Division, Erlangen Germany).

References

Busscher I, Ploegmakers JJW, Verkerke GJ, Veldhuizen AG. Comparative anatomical dimensions of the complete human and porcine spine. Eur. Spine J. 2010:191104–14.

Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2005; 1:886–93.

De Nigris D, Collins DL, Arbel T. Multi-modal image registration based on gradient orientations of minimal uncertainty. IEEE Trans. Med. Imaging. 2012; 31:2343–54. [PubMed: 22987509]

Gendrin C, et al. Validation for 2D/3D registration. II: the comparison of intensity- and gradient-based merit functions using a new gold standard data set. Med. Phys. 2011; 38:1491–502. [PubMed: 21520861]

Hansen, N. Towards A New Evolutionary Computation. Vol. 192. Springer; Berlin: 2006.

Hipwell JH, Penney GP, McLaughlin RA, Rhode K, Summers P, Cox TC, Byrne JV, Noble JA, Hawkes DJ. Intensity-based 2D–3D registration of cerebral angiograms. IEEE Trans. Med. Imaging. 2003; 22:1417–26. [PubMed: 14606675]

Khamene A, Bloch P, Wein W, Svatos M, Sauer F. Automatic registration of portal images and volumetric CT for patient positioning in radiation therapy. Med. Image Anal. 2006; 10:96–112. [PubMed: 16150629]

Kim J, Li S, Pradhan D, Hammoud R, Chen Q, Yin F-F, Zhao Y, Kim JH, Movsas B. Comparison of similarity measures for rigid-body CT/dual x-ray image registrations. Technol. Cancer Res. Treat. 2007; 6:337–46. [PubMed: 17668942]

Klinder T, Ostermann J, Ehm M, Franz A, Kneser R, Lorenz C. Automated model-based vertebra detection, identification, and segmentation in CT images. Med. Image Anal. 2009; 13:471–82. [PubMed: 19285910]

Linguraru MG, Vasilyev NV, Del Nido PJ, Howe RD. Statistical segmentation of surgical instruments in 3D ultrasound images. Ultrasound Med. Biol. 2007; 33:1428–37. [PubMed: 17521802]

Lo S-FL, et al. Automatic localization of target vertebrae in spine surgery: clinical evaluation of the LevelCheck registration algorithm. Spine. 2015; 40:E476–83. [PubMed: 25646750]

Lowe DG. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 2004; 60:91–110.

Ma J, Lu L. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. Comput. Vis. Image Underst. 2013; 117:1072–83.

McLaughlin RA, Hipwell J, Hawkes DJ, Noble JA, Byrne JV, Cox TC. A comparison of a similarity-based and a feature-based 2D–3D registration method for neurointerventional use. IEEE Trans. Med. Imaging. 2005; 24:1058–66. [PubMed: 16092337]

Mody MG, Nourbakhsh A, Stahl DL, Gibbs M, Alfawareh M, Garges KJ. The prevalence of wrong level surgery among spine surgeons. Spine. 2008; 33:194–8. [PubMed: 18197106]

Nithiananthan S, et al. Demons deformable registration of CT and cone-beam CT using an iterative intensity matching approach. Med. Phys. 2011; 38:1785–98. [PubMed: 21626913]

Otake Y, Schafer S, Stayman JW, Zbijewski W, Kleinszig G, Graumann R, Khanna AJ, Siewerdsen JH. Automatic localization of target vertebrae in spine surgery using fast CT-tofluoroscopy (3D–2D) image registration. Proc. SPIE. 2012a; 8316:83160N.

Otake Y, Schafer S, Stayman JW, Zbijewski W, Kleinszig G, Graumann R, Khanna AJ, Siewerdsen JH. Automatic localization of vertebral levels in x-ray fluoroscopy using 3D–2D registration: a tool to reduce wrong-site surgery. Phys. Med. Biol. 2012b; 57:5485–508. [PubMed: 22864366]

Otake Y, Wang AS, Uneri A, Kleinszig G, Vogt S, Aygun N, Lo SL, Wolinsky J-P, Gokaslan ZL, Siewerdsen JH. 3D–2D registration in mobile radiographs: algorithm development and preliminary clinical evaluation. Phys. Med. Biol. 2015; 60:2075–90. [PubMed: 25674851]

Otake Y, Wang AS, Webster Stayman J, Uneri A, Kleinszig G, Vogt S, Khanna AJ, Gokaslan ZL, Siewerdsen JH. Robust 3D–2D image registration: application to spine interventions and vertebral labeling in the presence of anatomical deformation. Phys. Med. Biol. 2013; 58:8535–53. [PubMed: 24246386]

Penney GP, Weese J, Little JA, Desmedt P, Hill DL, Hawkes DJ. A comparison of similarity measures for use in 2D–3D medical image registration. IEEE Trans. Med. Imaging. 1998; 17:586–95. [PubMed: 9845314]

Pluim JPW, Maintz JBA, Viergever MA. Image registration by maximization of combined mutual information and gradient information. IEEE Trans. Med. Imaging. 2000; 19:809–14. [PubMed: 11055805]

Škerl D, Likar B, Pernuš F. A protocol for evaluation of similarity measures for rigid registration. IEEE Trans. Med. Imaging. 2006; 25:779–91. [PubMed: 16768242]

Uneri A, Nithiananthan S, Schafer S, Otake Y, Stayman JW, Kleinszig G, Sussman MS, Prince JL, Siewerdsen JH. Deformable registration of the inflated and deflated lung in cone-beam CT-guided thoracic surgery: initial investigation of a combined model- and image-driven approach. Med. Phys. 2013; 40:017501. [PubMed: 23298134]

Wu J, Kim M, Peters J, Chung H, Samant SS. Evaluation of similarity measures for use in the intensity-based rigid 2D–3D registration for patient positioning in radiotherapy. Med. Phys. 2009; 36:5391–403. [PubMed: 20095251]

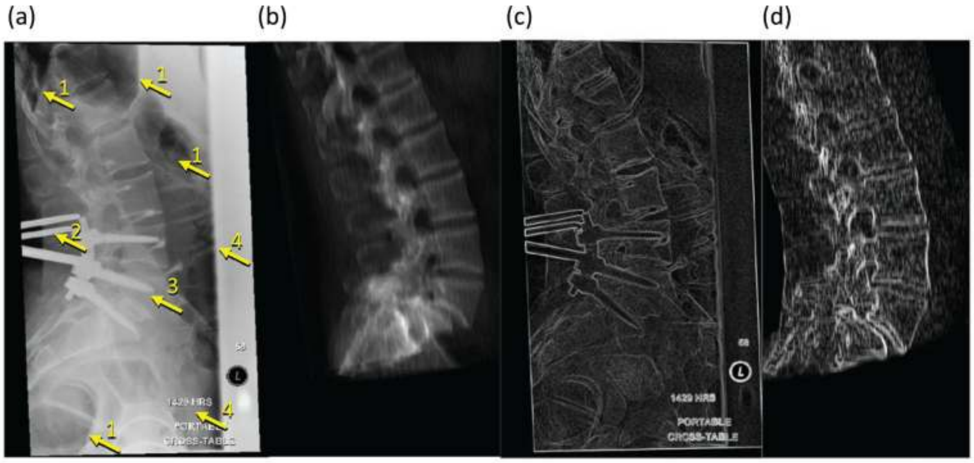

Figure 1.

Illustration of image content mismatch present in the (a) intraoperative radiograph in comparison to (b) the projection image from CT. The extraneous gradients caused by content mismatch are conspicuous in (c) the radiograph gradient image in comparison to (d) CT projection gradients. Such gradients are caused by, for example: (1) anatomical deformation associated with the lung, colon, gas bubbles, etc; (2) interventional instruments; (3) surgical implants; and (4) collimators and ‘burnt-in’ text annotations.

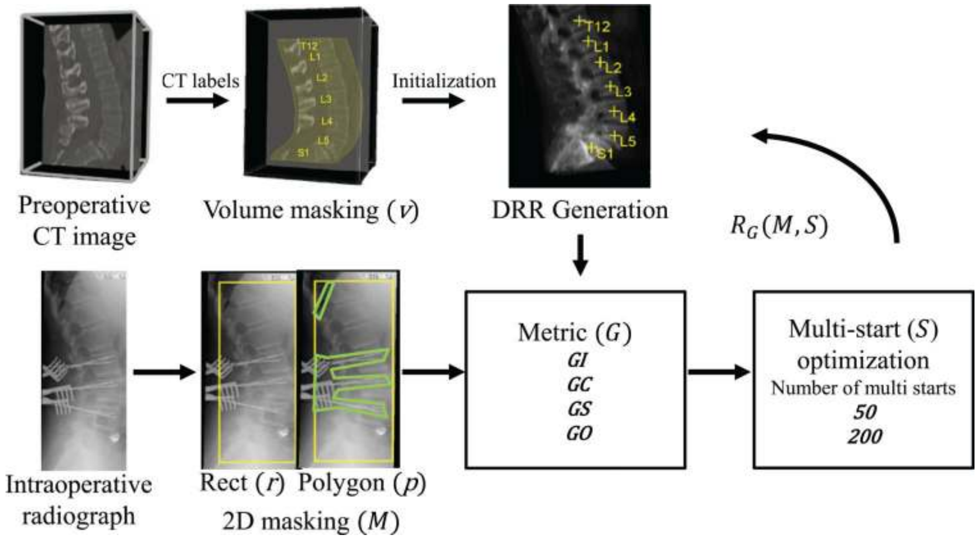

Figure 2.

Overall 3D–2D image registration process. The experimental variables in the study are shown in italics. Three masking methods (denoted by *M*) include 2D radiographic rectangular (*r*) or polygonal (*p*) masks or 3D volumetric (*v*) masks. Four similarity metrics (denoted by *G*) are GI, GC, GS, or GO. The number of multistarts (*S*) varied from 50 to 200. $R_G(M, S)$ represents a registration performed using similarity metric *G*, masking *M*, and multistart *S*.

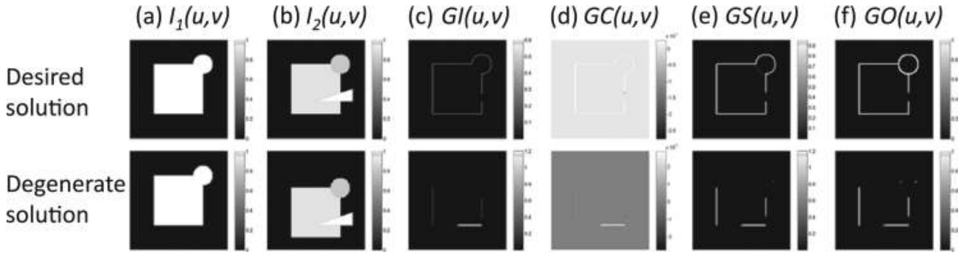

Figure 3.

Illustration of the four similarity metrics evaluated in this work for various simple simulations of content mismatch between the (a) I_1 (DRR) and (b) I_2 (radiograph). In each case, the similarity map shows the metric prior to summation over pixels. (c)–(f) All the metrics, at the desired solution, align the corresponding square and circle elements overcoming mismatch due to the triangle. However, the triangle mismatch at the degenerate solution, causes a larger relative contribution in (c) and (d) GI, GC metrics when compared with (e) and (f) GS, GO metrics.

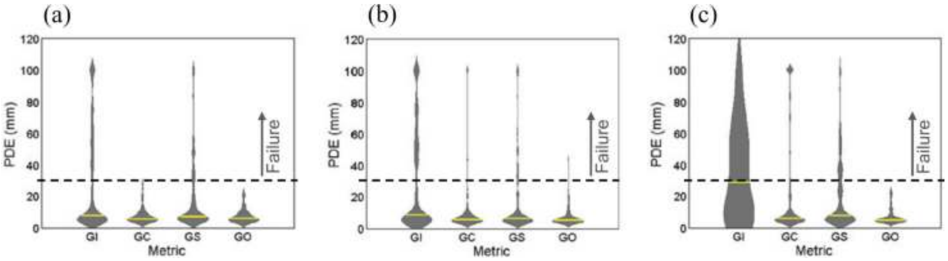

Figure 4.

Violin plots showing the distribution in PDE for registration using the four similarity metrics considered in this work. The three configurations of parameter settings ((a) *R*(*p*, 200), (b) *R*(*p*, 50), and (c) *R*(*r*, 50)) represent increasingly challenging conditions but with an improved run time. The horizontal line of each distribution marks the median PDE, and the ‘Failure’ line at PDE > 30 mm demarks the threshold for which the registered label is likely outside the true vertebra.

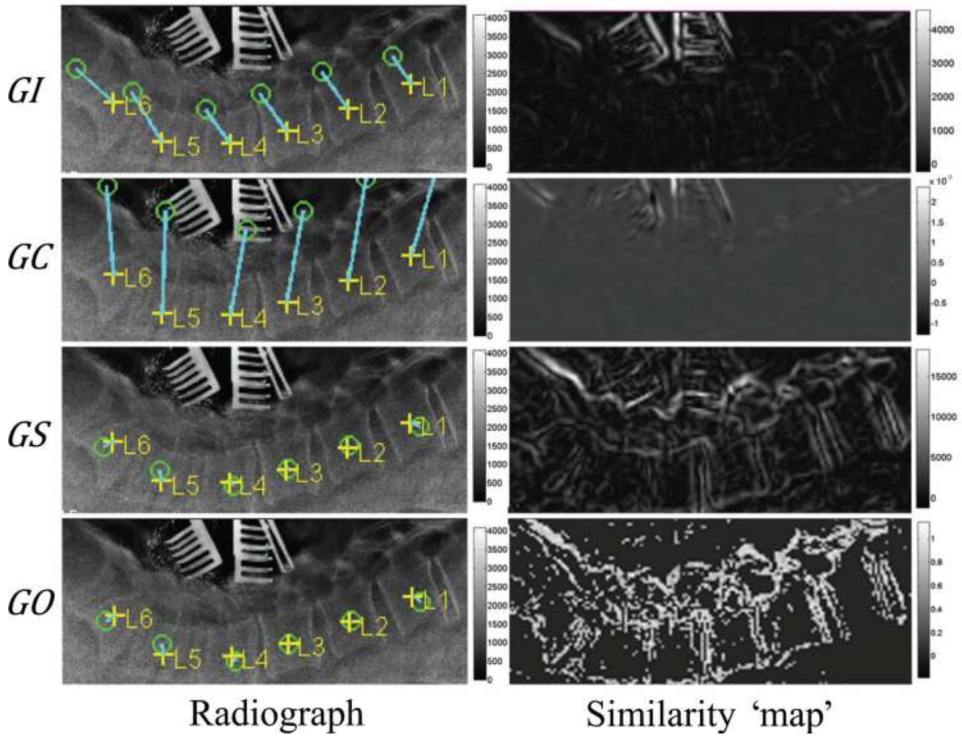

Figure 5.

Illustration of image content mismatch in a typical case presenting surgical instrumentation in the intraoperative radiograph (but not the preoperative CT). The robustness of each similarity metric is evident in the similarity 'maps'. The images on the left show the intraoperative radiograph overlaid by ground truth labels (crosses), registered labels (circles), and the distance between the two (PDE, marked as a line segment).

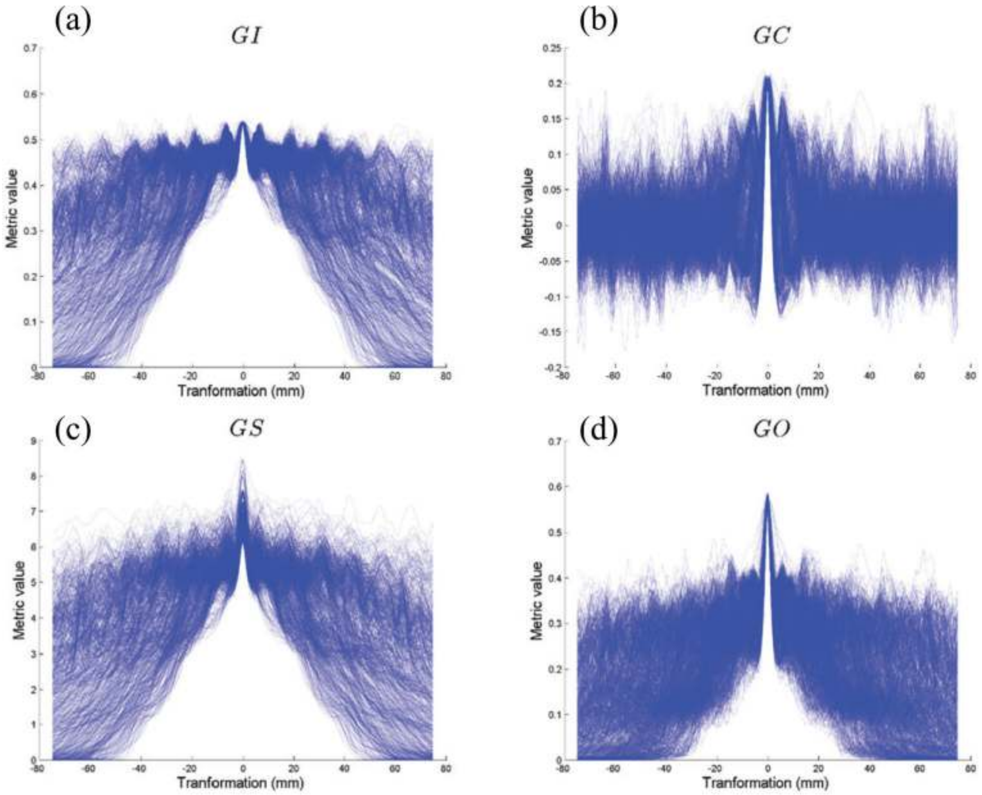

Figure 6.

Profiles through the objective function search space centered on the true registration solution. Each plot is for a given similarity metric—(a) GI, (b) GC, (c) GS, and (d) GO—and corresponds to a single case in which surgical instruments are present in the intraoperative radiograph but not the preoperative CT. The central peak corresponds to the true registration solution, and other peaks correspond to false local maxima. The GO metric exhibits the highest quality search space in presenting the most distinct global optimum.

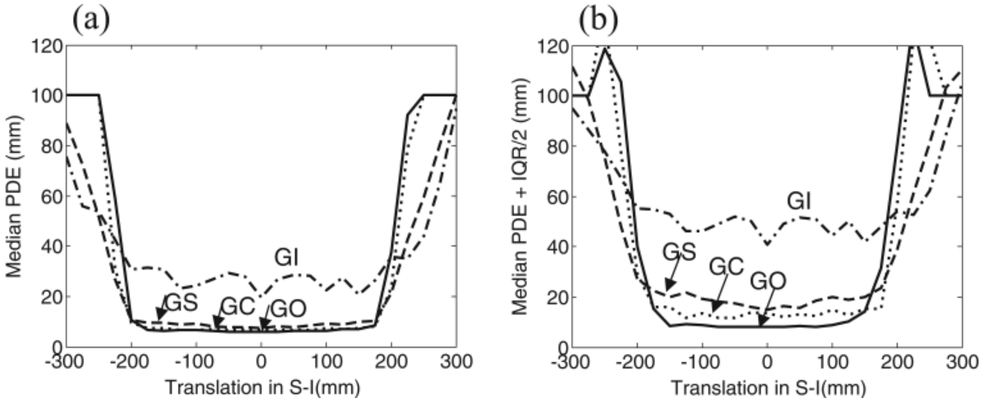

Figure 7.

Capture range measurements analyzing the sensitivity to initialization errors in the S–I direction for each similarity metric. (a) Median PDE. (b) The third quartile in the PDE distribution.

Table 1

Summary of an example previous work comparing the performance of similarity metrics in 3D-2D registration.

Publication	3D-2D Registration application	Metrics	Validation
Penney *et al* (1998)	CT to projection x-ray in spine surgery	NCC, entropy, MI, GC, PI, GD	Phantom + simulated clinical images
Hipwell *et al* (2003)	Magnetic resonance angiography (MRA) to digital subtraction angiography (DSA) in neurointerventions	NCC, GC, entropy, MI, PI, GD	Phantom + 4 patients
McLaughlin *et al* (2005)	MRA to projection x-ray in neurointerventions	GD	Phantom + 4 patients
Khamene *et al* (2006)	CT to projection x-ray in radiotherapy	LNC, GC, PI, GD, VWC, MI, CR, NCC	Phantom
Kim *et al* (2007)	CT to dual projection x-ray in radiotherapy	NCC, entropy, GC, GD, PI, MI	3 phantoms
Wu *et al* (2009)	CT to dual projection x-ray in radiotherapy	PIU, NMI, NCC, PI, GC, GD	2 phantom studies + 1 patient
Gendrin *et al* (2011)	CT to projection x-ray in radiotherapy	NCC, RC, CR, MI	Phantom
Otake *et al* (2012, 2013, 2015)	CT to projection x-ray in spine surgery	GI, GC	Simulation, phantom, and cadaver

Note: Metrics include: normalized cross correlation (NCC), entropy of the difference image, mutual information (MI), gradient correlation (GC), pattern intensity (PI), gradient difference (GD), correlation ratio (CR), local normalized correlation (LNC), variance weighted correlation (VWC), partitioned intensity uniformity (PIU), and rank correlation (RC).

Table 2

Median PDE, interquartile range, failure rate (Failure), and median run time (Run) for the four similarity metrics and three parameter configurations.

	***R*(*p*, 200)**				***R*(*p*, 50)**				***R*(r, 50)**			
Metric	**PDE (mm)**	**IQR (mm)**	**Failure (%)**	**Run (s)**	**PDE (mm)**	**IQR (mm)**	**Failure (%)**	**Run (s)**	**PDE (mm)**	**IQR (mm)**	**Failure (%)**	**Run (s)**
GI	8.2*	38.9	31.3	65.1	8.9*	46.7	36.5	23.7	29.1*	48.1	49.5	23.7
GC	6.2	3.7	0	84.5	6.1	3.5	2.6	29.0	6.4*	6.9	14.7	28.9
GS	7.7*	10.3	16.5	60.5	6.8*	7.5	16.5	22.7	8.2*	26.1	26.9	22.2
GO	6.4	4.4	0	84.1	5.6	2.9	0.8	28.8	5.5	2.6	0	29.3

Note: The asterisks mark cases for which the distribution in PDE was statistically significantly different ($p < 0.05$) from that of the GO case.