Available online at www.sciencedirect.com
SCIENCE DIRECT
Image and Vision Computing 23 (2005) 956–971
image AND vision COMPUTING
www.elsevier.com/locate/imavis

Development and validation of a multi-step approach to improved detection of 3D point landmarks in tomographic images

Sönke Frantz[a,*], Karl Rohr[b], H. Siegfried Stiehl[a]

[a]Department of Computer Science, University of Hamburg, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
[b]University of Heidelberg, IPMB, DKFZ Heidelberg, Dept. Intelligent Bioinformatics Systems, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

Received 17 May 2004; received in revised form 20 May 2005; accepted 20 May 2005

Abstract

We introduce a novel multi-step approach to improved detection of 3D anatomical point landmarks in tomographic images. Such landmarks serve as important image features for a variety of 3D medical image analysis tasks (e.g. image registration). Existing approaches to landmark detection, however, often suffer from a rather large number of false detections. Our multi-step approach combines an existing robust 3D detection operator with two different novel approaches to the reduction of false detections, and is applied within a semi-automatic procedure allowing for interactive control by the user. Experimental results obtained for a number of different anatomical landmarks of the human head in 3D CT and MR images demonstrate that both automatic ROI size selection and incorporation of a priori knowledge of the intensity structure at a landmark significantly improve the detection performance. The applicability of semi-automatic landmark extraction is thus considerably improved. We also summarize the results of a validation study in which we compare the performance of semi-automatic landmark extraction with that of a (standard) manual procedure for landmark extraction. As an exemplary application, we consider rigid MR/CT registration. The main result of our study is that compared to a purely manual procedure, semi-automatic landmark extraction (a) significantly reduces the elapsed time for landmark extraction, (b) generally yields registration results of comparable quality, and (c) increases the reproducibility of the results.

Keywords: Anatomical point landmarks; Differential approaches; Landmark detection; Validation

1. Introduction

Anatomical point landmarks are useful features for a variety of medical image analysis tasks. If selected suitably, such landmarks may represent substantial image information very concisely, which is important in medical applications with regard to the vast amount of data one has to deal with. For example, anatomical point landmarks are often used in image registration (e.g. [1–5]), which is a key issue in clinical diagnosis and the planning of surgical interventions. Examples of other applications utilizing anatomical point landmarks are the generation of shape models or morphometrics. The main problem with anatomical point landmarks is, however, their reliable and accurate extraction from 3D tomographic images. In practice, landmarks are usually manually extracted from 3D images, which is tedious and time-consuming and also difficult to reproduce. Consequently, automated procedures for landmark extraction are of central interest. In particular, our emphasis is on semi-automatic procedures allowing for interactive control by the user, which we consider important in clinical applications ('keep-the-user-in-the-loop'; see, e.g. [6]). Such a semi-automatic procedure, generally, comprises three steps: (i) the user interactively determines a coarse position of the landmark of interest, (ii) a computational approach to landmark extraction is applied within a region-of-interest (ROI) around that position to provide potential landmark candidates, and (iii) the user selects the most promising candidate.

When developing computational approaches to landmark extraction, key issues are reliable and robust landmark detection as well as accurate localization. This contribution is concerned with the crucial step of landmark detection.

* Corresponding author. Present address: YXLON International X-Ray GmbH, Essener Bogen 15, 22419 Hamburg, Germany.
E-mail addresses: frantz@informatik.uni-hamburg.de (S. Frantz), k.rohr@dkfz.de (K. Rohr), stiehl@informatik.uni-hamburg.de (H.S. Stiehl).

0262-8856/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2005.05.019

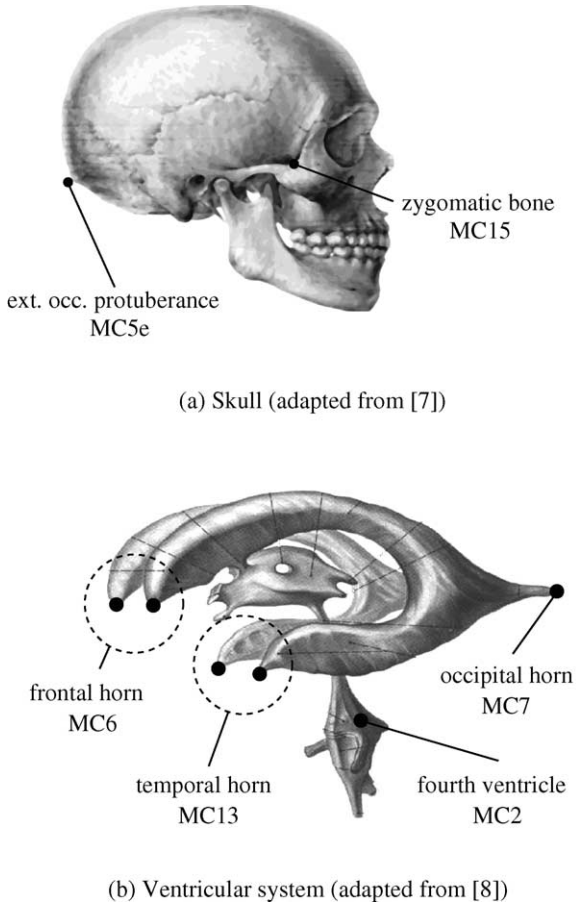

Fig. 1. Examples of anatomical point landmarks of the human head. (a) Skull (adapted from [30]), (b) ventricular system (adapted from [31]).

Among the different types of point landmarks, we here focus on the (generic) geometric classes of tips and saddle points. Examples of landmarks of such types in the case of the human head, in which we are primarily interested here, are shown in Fig. 1.

Existing work on the detection of such types of landmarks has mainly concentrated on differential approaches (e.g. [7–9]). Differential approaches exploit only local image information and, therefore, the application of such approaches is, generally, associated with rather low computational costs. A general problem is, however, that often a rather large number of false detections are obtained. A large number of false detections does not only affect in a negative way the effort required for selecting the correct candidate, but also the reliability, and thus reduces the confidence of a user in the results. One reason for false detections is image noise. In particular, differential approaches that employ image derivatives of higher order are affected by this (e.g. [7,9], where second or even third order derivatives are utilized); see also [10] for a comparative study of the detection performance of various 3D differential operators for landmark detection. Second, a considerable number of false detections are caused by neighboring anatomical structures that are captured by the region-of-interest (ROI) for landmark detection. One main contribution of this paper is a novel multi-step approach to improved landmark detection that significantly reduces the number of false detections and thus increases the applicability of semi-automatic landmark extraction. The second main contribution describes the results of a validation study in which different observers participated and where we analyzed the performance of semi-automatic landmark extraction based on our new multi-step approach in comparison to that of a (standard) manual procedure. Preliminary results of the material presented in this paper have been published in [11,12].

The paper is structured into two parts. In the first part, we introduce the multi-step approach to landmark detection, which features the combination of two different novel approaches to the reduction of false detections, while landmark detection is performed by applying an existing robust 3D differential operator that utilizes only first order image derivatives ([8]). The first approach to the reduction of false detections, which is described in Section 2, addresses the problem of selecting an optimal ROI size for a landmark. We present a novel statistical approach to automatic ROI size selection. Second, in Section 3 we take advantage of additional a priori knowledge of the intensity structure at a landmark to impose further constraints on detected candidates. The multi-step approach to landmark detection is finally described in Section 4. The second part of the paper is devoted to the experimental evaluation of the multi-step approach to landmark detection. In Section 5, we first analyze in detail the detection performance of our new multi-step approach for different anatomical point landmarks in 3D MR and CT images of the human head. In Section 6, we are finally concerned with the validation of semi-automatic landmark extraction. As a standard application, we consider landmark-based rigid image registration. The performance is compared with that of a purely manual procedure for landmark extraction. As performance criteria being important for practice, we consider (i) the elapsed time for landmark extraction, (ii) the results of MR/CT registration based on the extracted landmarks, as well as (iii) the inter-observer variability of the localized landmark positions as an indicator for reproducibility.

2. Statistical approach to automatic ROI size selection

Within a semi-automatic procedure for landmark extraction, the user interactively determines a coarse 3D position of a landmark. At this position, a 3D ROI is placed in which a differential operator is applied to detect landmark candidates. However, a general problem is to select a suitable ROI size. On the one hand, the ROI should capture a sufficient amount of the intensity structure to enable

reliable landmark detection. For that reason, the ROI is required to be sufficiently large. On the other hand, the ROI should be sufficiently small so that just the landmark alone is captured by the ROI, while interfering neighboring structures are excluded. In practice, an ROI of fixed size is usually applied. Because of this, however, neighboring anatomical structures are often captured by the ROI, which gives rise to additional, unwanted detections. In this section, we develop a statistical approach to the automatic selection of a suitable ROI size. Our approach is based on a 3D differential edge intersection approach, which was originally developed for refined landmark localization [13]. With the edge intersection approach, tangent planes that are defined based on the intensity gradient are used to locally approximate the surface at a landmark. The landmark position is estimated by intersecting the tangent planes using the least-squares method. Applying this approach in the context of ROI size selection, we exploit the statistical localization uncertainty of the edge intersection position estimate as a criterion for isolating an anatomical landmark (e.g. a tip) within the ROI.

2.1. Related work on ROI size selection

The choice of a suitable ROI size or the choice of an optimal observation window size for an operator is a general problem in computer vision, and some work addressing this issue does exist. However, ROI size selection in the context of 3D landmark extraction has not been considered so far. Related work regarding the selection of suitable windows was done, for example, in the context of matching stereo images (e.g. [15]), optical flow measurements (e.g. [16]), or 2D junction localization (e.g. [17]). In view of our work, the approach presented in [17] has the closest relation. It has to be stressed, however, that in [17] 2D junction localization in 2D images using windows of optimal sizes was considered, while in our case we consider the selection of 3D ROIs around prominent 3D anatomical structures as depicted in 3D medical images. Also, the criterion that was used in [17] for selecting an optimal window size differs from ours. In [17], optimal window sizes were determined based on so-called scale-space maxima of the 2D differential operator in [18] (i.e. the respective scale at which a junction is detected is assumed to be an appropriate scale for the observation window at the structure of interest), while we here exploit the statistical uncertainty of the position estimate resulting from the 3D edge intersection approach.

2.2. Description of our approach

Let us consider, for simplicity, a cubic ROI of width w centered at the interactively determined position (i.e. the ROI size is controlled by one parameter w). Let $\hat{\mathbf{x}}_w$ denote the position estimate resulting from the 3D edge intersection approach [13]. Then, the statistical localization uncertainty of $\hat{\mathbf{x}}_w$ is given by the covariance matrix

$$\sum\nolimits_{\mathrm{w}} = \sigma_\varepsilon^2 \left(\sum_{i=1}^{n} \nabla g(\mathbf{x}_i) \nabla g(\mathbf{x}_i)^{\mathrm{T}} \right)^{-1}, \quad (1)$$

where σ_ε^2 is a data-dependent noise term, $\nabla g(\mathbf{x}_i)$ is the intensity gradient at $\mathbf{x}_i$ (the index i spans over all voxels within the ROI), and $n = w^3$ denotes the number of voxels within the ROI (see Appendix A for details). The matrix Σ_w reflects the consistency of the data observed within the ROI with the assumed polyhedral model of the surface at the landmark. The idea behind ROI size selection is then to vary the ROI width and to select the optimal ROI width based on minimal uncertainty.

Our algorithm for ROI size selection works as follows: we start with a user-specified minimal ROI width (e.g. $w_{\min} = 7$ voxels). As a scalar measure for the uncertainty, we consider the determinant of the covariance matrix, $U_w = \det(\Sigma_w)$, which is also referred to as the generalized variance (e.g. [14]). When the ROI does not capture a sufficient amount of the intensity structure to reliably estimate the landmark position, U_w is large. Hence, by enlarging the ROI and thus by capturing a larger amount of the intensity structure around the landmark, U_w can be expected to decrease. However, when neighboring structures are eventually captured by the ROI, U_w significantly increases, which suggests that further enlargement of the ROI is not adequate. In our implementation, we detect such a change of U_w at w_{increase}, say, by requiring that (a) U_w increases and (b) the relative spatial variation of the position estimate resulting from the edge intersection approach exceeds a threshold t_V. For criterion (a), we check whether $U_w > U_{w-2}$, while for criterion (b), we test if $||\hat{\mathbf{x}}_w - \hat{\mathbf{x}}_{w-2}|| \geq t_V$ ($\hat{\mathbf{x}}_w$ and $\hat{\mathbf{x}}_{w-2}$ denote the position estimates based on ROI widths of w and $w-2$, resp.). Experimentally, we found that the combination of these two criteria turned out to be very robust. Additionally, a maximal ROI width $w_{\max}$ is prescribed. Finally, the optimal ROI width w_{opt} is selected in between, that is, $w_{\text{opt}} = \arg\min U_w$ with $w_{\min} \leq w \leq \min\{w_{\text{increase}}, w_{\max}\}$.

3. Incorporation of additional a priori knowledge of the intensity structure at a landmark

In Section 2, we have described an approach to the automatic selection of a suitable ROI at a landmark to avoid the inclusion of neighboring structures in the ROI and thus to reduce the number of false detections. In this section, we follow a different approach to reducing the number of false detections, namely, by taking advantage of additional a priori knowledge of the intensity structure at a landmark. For example, the user generally knows the type of the landmark of interest (e.g. a tip or a saddle point) as well as the modality and the imaging parameters of the image used (e.g. CT or T1-, T2-, or PD-weighted MR). This knowledge

can be used to impose additional constraints on the detected candidates. We here suggest using differential measures to classify the intensity structure at the detected candidates. Then, candidates can be rejected which have an intensity structure being inconsistent with the a priori knowledge of the intensity structure at the landmark. For classification, we exploit curvature properties of the isointensity surfaces at the detected candidates.

3.1. Related work on the utilization of curvature measures for landmark extraction

Measures based on the curvature of isointensity surfaces have previously been utilized for the detection of 3D point landmarks in [7,9], for example. However, using such measures requires computing image derivatives up to the second order or even up to the third order. Note that the estimation of high order derivatives is, generally, very sensitive to noise, which gives rise to detections due to noise only (see also [10]). For reasons of robustness, we here apply a 3D differential operator for landmark detection that requires only first order image derivatives ([8]). Afterwards, we additionally take into account second order image derivative information and utilize curvature measures to classify the local intensity structure at the detected candidates.

3.2. Description of our approach

Suppose we have detected a point $\mathbf{x}_d$ (with $\nabla g(\mathbf{x}_d) \neq 0$) on the surface of an anatomical structure and assume that the isointensity surface at $\mathbf{x}_d$, which is implicitly defined by $g(\mathbf{x}) - g(\mathbf{x}_d) = 0$, well approximates the surface of the anatomical structure. The type of the surface (e.g. a tip or a saddle) can then be determined based on the signs of the principal curvatures of the isointensity surface or, alternatively, based on the signs of the Gaussian curvature and the mean curvature of the isointensity surface (for closed formulae of the Gaussian and the mean curvature, see, e.g. [7,19]). Additionally, we can distinguish different landmarks w.r.t. their appearance as compared to the surrounding intensity (i.e. a dark or a bright tip).

In the case of a tip, both principal curvatures have the same sign, which implies that the Gaussian curvature, which is the product of both principal curvatures, is positive. In the case of a saddle point, the principal curvatures have opposite signs and the Gaussian curvature is negative. Finally, it is possible to distinguish objects w.r.t. their appearance as compared to the surrounding intensity by exploiting the sign of the mean curvature (recall that the mean curvature is the mean value of both principal curvatures). This is possible because the sign of the mean curvature depends on the orientation of the gradient of the intensity function, which is normal to the surface. Note, however, that this additional discrimination is only useful in the case of tips. In the case of a dark tip as compared to a bright surrounding, the gradient of the intensity function points out of the structure and, consequently, both principal curvatures are negative. Therefore, the mean curvature is negative, too. In the case of a bright tip, the gradient of the intensity function points into the structure and, therefore, both principal curvatures are positive. Consequently, the mean curvature is also positive.

4. A novel multi-step approach to 3D point landmark detection

In summary, our multi-step approach to landmark detection comprises three steps:

1. A cubic ROI is centered at the interactively determined position, and then an optimal size for the ROI is automatically selected (see Section 2).
2. Landmark candidates are detected by applying a computationally efficient 3D differential operator [8]. The operator is applicable to different types of landmarks and is relatively robust w.r.t. noise since only first order image derivatives are used (cf. [10]). The operator reads $\mathrm{Op3} = \det(\mathbf{C})/\mathrm{trace}(\mathbf{C})$, where $\mathbf{C}$ denotes the averaged dyadic product of the intensity gradient, $\mathbf{C} = \overline{\nabla g \nabla g^{\mathrm{T}}}$. Landmark candidates are usually determined by searching for local maxima of the operator responses.
3. Detected candidates with an intensity structure being inconsistent with the a priori knowledge of the intensity structure at the landmark at hand are automatically rejected (see Section 3).

In contrast, in previous work (e.g. [8,20]) the 3D detection operator Op3 was applied alone, that is, an ROI of fixed size was used and additional a priori knowledge of the intensity structure at a landmark in terms of the surface curvature was not incorporated.

5. Evaluation of the detection performance using 3D MR and CT images

In this section, we analyze in detail the detection performance of the new multi-step approach for different anatomical point landmarks in 3D MR and CT images of the human head. In particular, we compare the detection performance with that of an existing 3D detection operator.

5.1. Performance evaluation

To evaluate the detection performance, we use a certain type of performance visualization as well as a scalar quantity measuring the detection capability (cf. [21]). To visualize the detection performance, the operator responses at the detected candidates are plotted as a function of the distance to the ROI center. Thus, these plots reflect

the spatial scatter of the detected positions within the ROI, the number of detections, as well as the significance of the different detections in terms of the strength of the operator response. To evaluate the detection performance quantitatively, we use a performance measure that takes into account both the number of detections and the significance of the different detections. This measure reflects the separability of the detections w.r.t. the operator response. Suppose we have obtained n detections for a landmark. Let $R_i \geq 0$, $i=1,\ldots,n$, denote the operator responses at the detected positions and let $R_{\max}$ denote the maximum of these values. Then, the performance measure denoted by ψ reads:

$$\psi = \begin{cases} 0 & n = 0, \\ \dfrac{\sum_{i=1}^{n} R_i}{R_{\max}} & n \geq 1. \end{cases} \tag{2}$$

When there is only one correct detection, we obviously have $\psi=1$. Additional false detections with small operator responses yield $\psi \approx 1$. In this case, the correct detection can clearly be distinguished from the false detections. On the other hand, when there are detections with operator responses similar to the maximal operator response, ψ is much larger than 1. In the case of equal responses $R_1=\cdots=R_n$, we have $\psi=n$. In sum, the closer ψ is to 1, the better is the detection performance. Note, however, that these considerations assume that the detection with the strongest operator response is also a correct detection (we speak of a correct detection if the detection is associated with the landmark at hand).

5.2. Experimental set-up

As landmarks, we consider visually salient features located at the skull and within the brain: the saddle points at the zygomatic bones (MC15), the tip of the external occipital protuberance (MC5e), the tip of the vermex of the cerebellum at the fourth ventricle (MC2), the junction at the top of the pons (MC18), and the tips of the frontal (MC6) and occipital (MC7) ventricular horns (see Fig. 1). We use five T1-weighted 3D MR/CT image pairs from different patients: one (C06) acquired at Utrecht University Hospital, The Netherlands, and four (V101, V104, V107, V109) acquired at Vanderbilt University, USA. The voxel sizes of the C06 data are $0.86\times0.86\times1.2\ \text{mm}^3$ (MR image) and $0.63\times0.63\times1.0\ \text{mm}^3$ (CT image). The voxel sizes of the original data provided by Vanderbilt University are about $0.85\times0.85\times3.0\ \text{mm}^3$ (MR images) and $0.42\times0.42\times3.0\ \text{mm}^3$ (CT images). Here, we used up-sampled data based on cubic B-spline image interpolation ([22]) with a slice thickness of 1.0 mm.

For automatic ROI size selection, the minimal and maximal width of the cubic ROI was set to $w_{\min}=7$ voxels and $w_{\max}=21$ voxels. The threshold for the spatial variation of the position estimate was set to $t_V=0.5$ mm. For estimating the image derivatives, we applied a scheme based on cubic B-spline interpolation and Gaussian smoothing. With this scheme, the derivatives are calculated from the reconstructed continuous signal, while the anisotropic image resolution is taken into account (see [23] for details). The standard deviation σ of the Gaussian filters (given in mm units) was thereby coarsely adapted to the scale of the landmark at hand: a smaller scale ($\sigma=1.0$ mm) was used for the zygomatic bone and the top of the pons, and a somewhat larger scale ($\sigma=1.5$ mm) was used for the external occipital protuberance, the fourth ventricle, and the frontal and occipital ventricular horns. The components of the averaged dyadic product of the gradient of the intensity function (matrix $\mathbf{C}$ utilized by the detection operator Op3) were computed within $5\times5\times5$ neighborhoods. Landmark candidates were determined by searching for local maxima of the operator responses within $3\times3\times3$ neighborhoods.

5.3. Experimental results

In the following, we first present in detail the experimental results obtained for one MR/CT image pair. Then, we summarize the results obtained for all five MR/CT image pairs from above.

5.3.1. Analysis of the detection performance using one MR/CT image pair

Exemplarily, we here consider six landmarks located within the mid-sagittal plane and within the left hemisphere. In these first experiments, the respective ROI center was the position resulting from best possible manual landmark localization in the images (in agreement with two persons). Thus, using this ground truth, we can also study the impact of our approach on the localization accuracy. No thresholds were applied to the operator responses. In Figs. 2 and 3, the results obtained for one MR image and one CT image (V109 image pair) are shown, respectively. Thick bars indicate those detections obtained with the multi-step approach, whereas narrow bars indicate the detections that would additionally occur by applying the 3D detection operator Op3 alone, that is, by using an ROI of fixed size and without the incorporation of a priori knowledge of the intensity structure at a landmark (for comparison, we here set the ROI width to $w=w_{\max}=21$ voxels). Additionally, the number of detections n as well as the value of the detection performance measure ψ are given.

From Figs. 2 and 3, one can conclude that the detection performance of the multi-step approach is significantly better than that of applying the 3D detection operator alone. A rather large number of false detections with significant operator responses is suppressed, for example, in the MR image in the case of the external protuberance and the fourth ventricle (Figs. 2b and c) and in the CT image in the case of the fourth ventricle and the top of the pons (Figs. 3c and d).

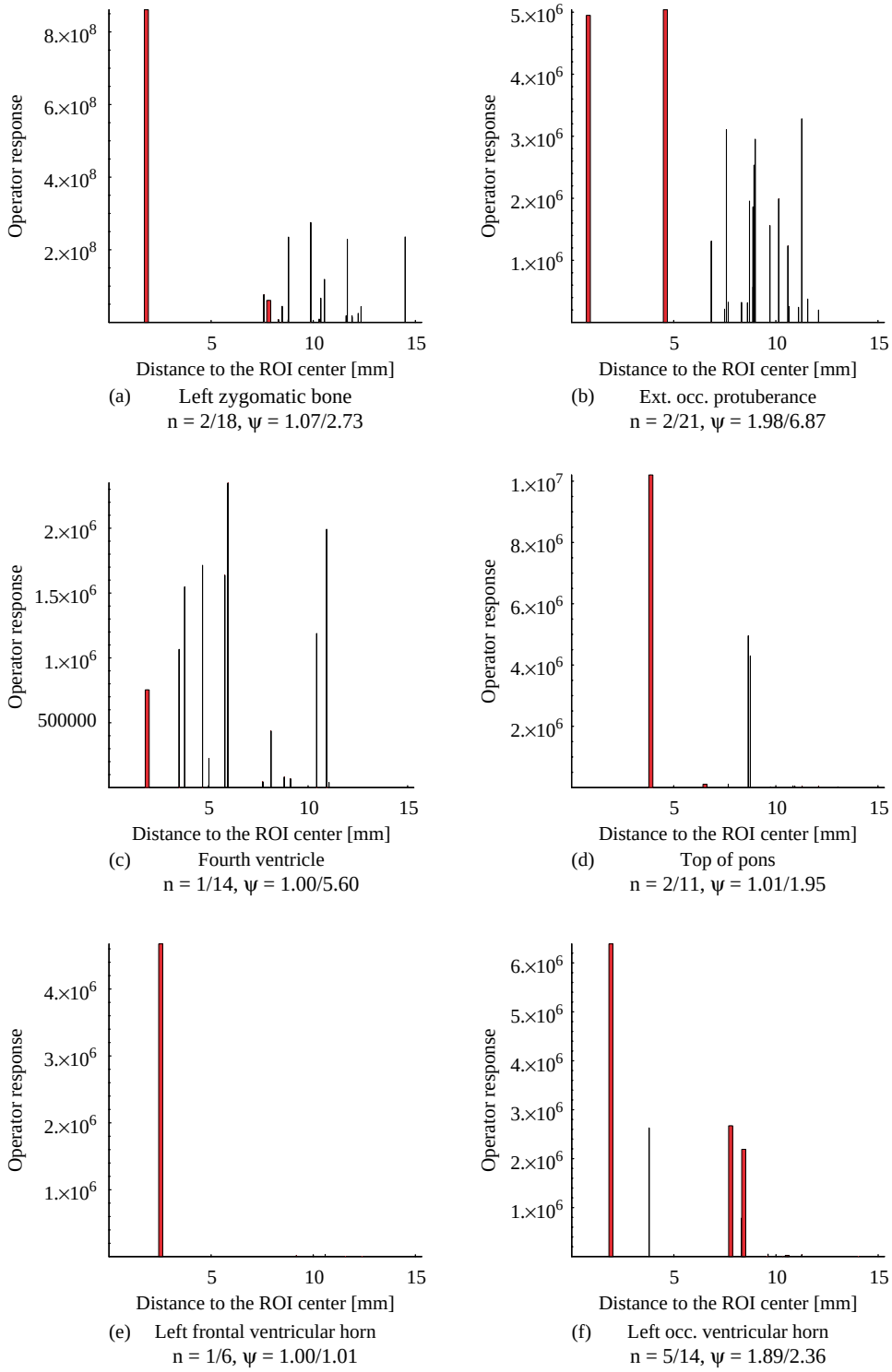

Fig. 2. Detection performance of the multi-step approach for different landmarks in a T1-weighted 3D MR image. The detection operator responses at detected positions are drawn as a function of the distance to the ROI center. Thick bars indicate those detections obtained with the multi-step approach, whereas narrow bars indicate the detections that would additionally occur when applying an existing 3D detection operator alone. In addition, the number of detections n as well as the value of the detection performance measure ψ are given (e.g. in (a) $n=2/18$ means two detections for the multi-step approach and 18 detections as the result of applying the 3D detection operator alone).

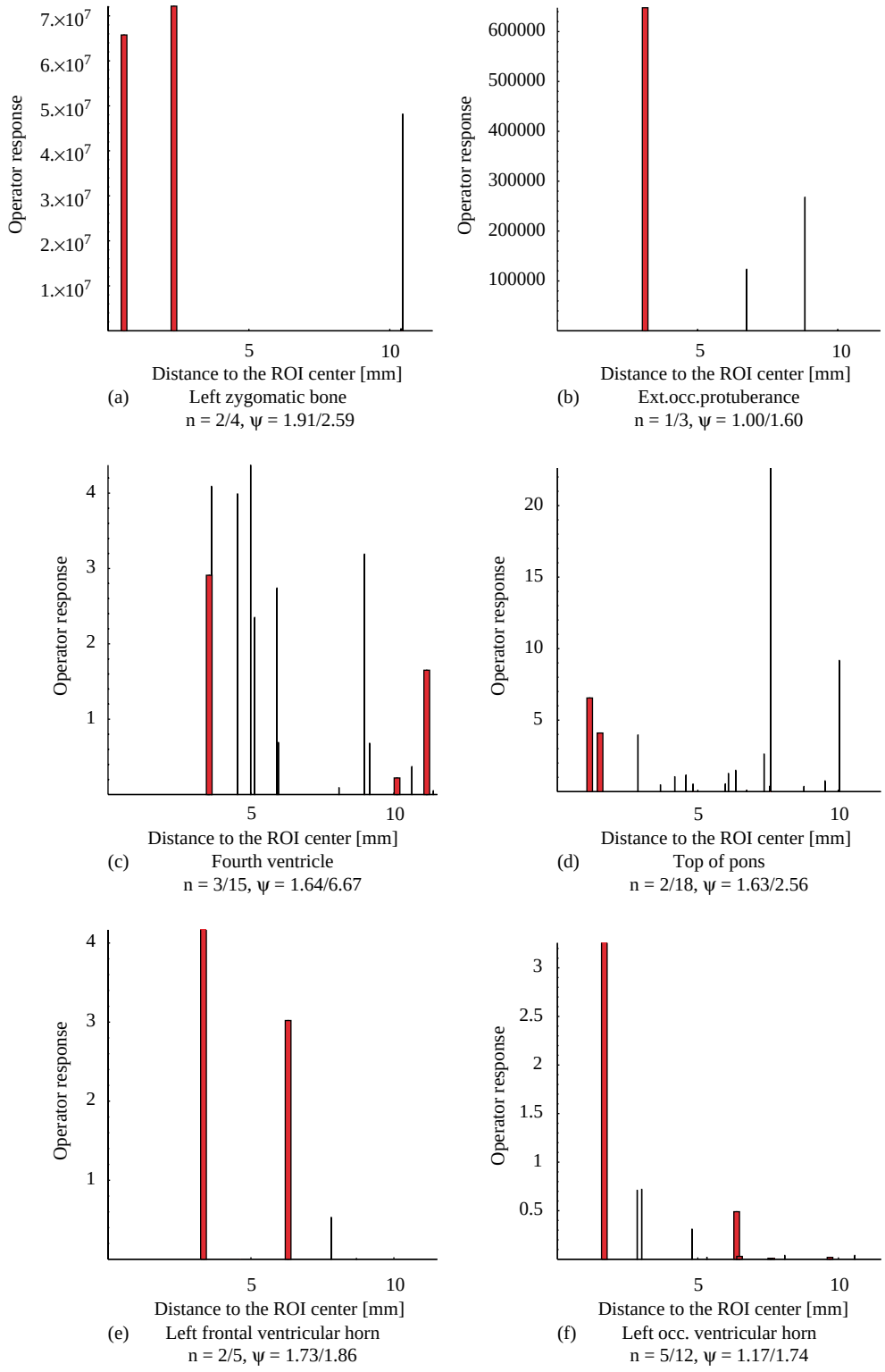

Fig. 3. Same as Fig. 2 but for a 3D CT image.

Note that in the MR image in the case of the fourth ventricle and in the CT image in the case of the fourth ventricle and the top of the pons, a number of significant detections with larger operator responses than the respective candidate with minimal distance to the manually specified position were automatically rejected. Moreover, among the remaining detections, the candidate with minimal distance to the manually specified position (i.e. the correct candidate) is

Table 1
Detection performance of the multi-step approach for different landmarks in five T1-weighted 3D MR images

	MC151	MC15r	MC5e	MC2	MC18	MC61	MC6r	MC71	MC7r
$\bar{n}$	2.0/4.8	3.3/8.3	1.8/8.3	1.0/12.0	1.0/5.0	2.5/3.0	2.4/3.4	1.8/2.8	2.0/3.4
$n_{\max}$	3/8	5/12	2/13	1/15	1/7	3/4	3/5	3/4	3/5
$\Delta\bar{n}$	59%	61%	79%	92%	80%	17%	30%	36%	42%
$\bar{\psi}$	1.5/2.6	1.8/3.6	1.5/4.0	1.0/4.9	1.0/2.6	1.7/1.8	1.7/2.2	1.3/1.5	1.4/1.8
$\psi_{\max}$	2.1/4.5	2.8/4.9	2.0/6.0	1.0/4.9	1.0/3.5	2.3/2.7	2.4/3.6	1.9/2.0	1.6/2.5

For comparison, the results of applying an existing 3D detection operator alone are given. For each landmark, the mean $\bar{n}$ and the maximal $n_{\max}$ number of detections, the relative difference $\Delta\bar{n}$ between the mean values, as well as the mean $\bar{\psi}$ and the maximal value $\psi_{\max}$ of the detection performance measure ψ are listed (e.g., in the case of MC151 $\bar{n} = 2.0/4.8$ denotes in the mean 2.0 detections for the multi-step approach and 4.8 detections as the result of applying the 3D detection operator alone).

distinguished by the largest operator response. Thus, for a user the ease of the selection of the correct candidate is considerably increased. The fact that the remaining detections are better distinguishable w.r.t. the operator response is also reflected by smaller values of the performance measure ψ. Also, it has to be pointed out that both approaches to the reduction of false detections, that is, automatic ROI size selection as well as incorporation of a priori knowledge of the intensity structure at a landmark, complementarily contribute to the improvement of the detection performance. For example, in the MR image in the case of the fourth ventricle (Fig. 2c), we found that automatic ROI size selection alone yielded two detections, while the incorporation of a priori knowledge of the intensity structure at that landmark yielded four detections. The combination of both approaches actually yielded the optimal result of one detection. In all cases, the localization performance was not affected, that is, the detection with minimal distance to the manually specified position was retained. For the other four MR/CT image pairs, we obtained similar results.

5.3.2. Summary of results obtained for five MR/CT image pairs

The results presented in the following were obtained within a validation study of the applicability of semi-automatic landmark extraction (see Section 6 for more details). In this study, five observers used a semi-automatic procedure for landmark extraction based on the new multi-step approach to landmark detection from Section 4 to determine seven up to nine corresponding landmarks in the five MR/CT image pairs from above. The landmarks were those from above plus the respective landmarks in the right hemisphere.

In Tables 1 and 2, the results of one observer are shown. We compare the performance of the multi-step approach with that of applying the 3D detection operator Op3 alone. This was possible since the user inputs in the validation study (e.g. the interactively determined positions) were automatically recorded by the computer system so that, for comparison, an alternative algorithm for landmark detection could be applied retrospectively. In the experiments, we applied a dynamic threshold to the operator responses to suppress insignificant detections (i.e. detections with less than 10% of the maximal operator response in an ROI). Table 1 shows the results obtained for the five MR images, while Table 2 shows the results obtained for the five CT images. For abbreviation, we used the landmark symbols introduced in Section 5.2. A suffix added to the landmark symbols indicates the respective hemisphere, for example. MC151 refers to the saddle point at the left zygomatic bone. For each landmark, the mean $\bar{n}$ and the maximal number $n_{\max}$ of detections, the relative difference $\Delta\bar{n}$ between the mean values ($\Delta\bar{n} = (\bar{n}_{\text{Op3}} - \bar{n}_{\text{multi-step}})/\bar{n}_{\text{Op3}}100\%$), as well as the mean $\bar{\psi}$ and the maximal value $\psi_{\max}$ of the detection performance measure ψ are given.

For all landmarks, the mean number of detections was reduced using the multi-step approach. In most cases, we obtained significantly fewer detections, as compared to applying the 3D detection operator alone. The values of the detection performance measure ψ are much smaller and closer to 1, which again indicates that the remaining detections are better distinguishable w.r.t. the operator response. In the MR image, especially in the case of the saddle points at the zygomatic bones (MC15), the external occipital protuberance (MC5e), the fourth ventricle (MC2), and the top of the pons (MC18) the results were significantly improved. In the CT image, in the case of MC2 and MC18

Table 2
Same as Fig. 1 but for five 3D CT images

	MC151	MC15r	MC5e	MC2	MC18	MC6 1	MC6r	MC7 1	MC7r
$\bar{n}$	1.6/2.6	1.8/3.0	1.5/2.3	1.0/8.0	3.0/6.0	1.3/2.0	2.8/3.6	1.6/3.8	2.0/3.8
$n_{\max}$	2/3	2/4	2/4	1/10	5/7	2/2	5/6	2/7	3/7
$\Delta\bar{n}$	39%	40%	35%	88%	50%	35%	23%	58%	48%
$\bar{\psi}$	1.3/1.9	1.6/2.1	1.2/1.4	1.0/3.5	1.7/2.7	1.0/1.2	1.9/2.2	1.3/2.1	1.5/2.0
$\psi_{\max}$	1.8/2.3	2.0/2.8	1.5/2.0	1.0/3.9	2.6/3.3	1.2/1.2	3.6/3.9	1.9/2.6	1.8/2.9

the results were also considerably improved. In general, the detection performance was better in the case of the CT images, that is, the number of detections was smaller and also the value of ψ was smaller. In most cases, we obtained only one or two detections with the multi-step approach (in the case of the MR images in 72% of all cases and in the case of the CT images in 83% of all cases). However, in either case, we obtained not more than five detections for a landmark, while applying the 3D detection operator alone yielded 15 detections in the worst case.

6. Validation study of semi-automatic landmark extraction

The validation of methods intended for clinical routine use is imperative. In this section, we present the results of a first, rather extensive validation study of the applicability of semi-automatic landmark extraction. We consider the extraction of corresponding anatomical point landmarks from multi-modality images for the purpose of landmark-based image registration. In comparison to a purely manual procedure for landmark extraction, which is commonly used in practice but which is generally time-consuming and lacks reproducibility, an automated procedure offers the possibility to reduce the elapsed time for landmark extraction and to increase the reproducibility of the results. The aim of our study was to investigate to what extent this can be achieved.

6.1. Validation strategy

Exemplarily, we here consider the extraction of anatomical landmarks from MR and CT images of the same patient. To assess the performance of semi-automatic landmark extraction, we consider the following criteria: (a) the elapsed time spent for landmark extraction, (b) the accuracy of the results of rigid MR/CT registration based on the extracted landmarks, and (c) the reproducibility of the results for different observers (for a detailed description of these criteria, see Section 6.2). The performance of semi-automatic landmark extraction is compared with that of a (standard) manual procedure. Five observers participated in our study (three computer scientists, one physicist, and one physician), and five MR/CT image pairs from different patients were used. Although none of the participants can be considered a clinical expert in radiology or neurosurgery, all persons were sufficiently familiar with 3D brain anatomy and the task at hand.

6.1.1. Scenario for landmark extraction and registration

Since we compare the performance of two procedures for landmark extraction that are basically different and which require both user interaction with the data, it is important to ensure independence of the experiments. In particular, a bias towards one procedure due to the results of the respective other procedure must be avoided. To diminish a bias towards the (new) semi-automatic procedure, each observer started with this procedure and afterwards used the manual procedure. Thus, familiarity with the data would not have been to the advantage of the semi-automatic procedure. In addition, to diminish adaption effects to the data and to ensure independence of the experiments, the second run-through of each observer with the manual procedure was scheduled at least four weeks later. Also, to enable a fair comparison of the results, the user-interfaces for semi-automatic and manual landmark extraction were almost identical.

Landmarks were simultaneously extracted from both modalities through simultaneous display on two monitors that were placed in a darkened room to guarantee high-contrast image display. The simultaneous display helps the user to orient within the 3D images as well as to establish corresponding landmark positions. The extracted landmarks were then used for MR/CT registration. We assume that the multi-modal images were acquired approximately at the same time. In such a case, a global rigid transformation comprising translation and rotation is generally sufficient for image registration. Local deformations, for example, due to scanner-induced distortions or patient movements, are not considered. Although a rigid transformation has only a few degrees of freedom, in practice a rather large number of landmarks should be used for registration for reasons of robustness. However, in our case only seven up to nine landmarks were used owing to the relatively poor quality of the used CT data. Consequently, the registration results can be expected to be sensitive to errors in landmark localization. Therefore, the observers were allowed to reject landmarks if the registration results were not satisfactory. To support the identification of possibly erroneous correspondences, each observer was asked to coarsely rank his confidence in having established corresponding landmark positions after extracting a landmark from both modalities. After inspecting the registration result, each observer was allowed to reject up to two landmarks according to his confidence in the results and to redo registration using the reduced set of landmarks. The basis for this decision were the visual impression of the registration result, the landmark rankings, and the computed registration error based on the landmarks used for registration.

6.1.2. Related work on experimental evaluation of landmark extraction

Only a few experimental studies on point landmark extraction have been published up to now. A detailed study of the performance of an automated procedure for landmark extraction as compared to the performance of a (standard) manual procedure does not exist. In [7], 3D differential operators using image derivatives up to the third order were applied within a fully automatic procedure to extract general feature points for the special case of rigid registration of monomodality images. In [24], image registration based on 3D anatomical point landmarks was validated. However, in contrast to our validation study, only manual landmark extraction was considered. Moreover, the elapsed time for landmark extraction was not

reported. In [9], the performance of various 3D differential operators for landmark detection was studied. The operators involve image derivatives up to the second or third order and were applied within a semi-automatic procedure to extract 3D anatomical point landmarks from MR and CT images. However, only one person was involved in the experiments, and the elapsed time for landmark extraction was not recorded. In [10], the detection performance of various 3D differential operators using first order image derivatives as well as first and second order derivatives was investigated. However, the extraction of corresponding landmarks from images of different modalities within a semi-automatic procedure involving different observers as well as the use of the landmarks for image registration were not considered.

6.2. Performance criteria

In this section, we describe in detail the performance criteria used in our study.

6.2.1. Elapsed time for landmark extraction

To measure the effort required for landmark extraction, we consider the time interval between the time at which the system is ready for user interaction and the time at which the last landmark has been successfully extracted. Thus, the time for loading of the data and for initialization, which may considerably vary depending on the actual net load in our workstation cluster and which thus would make the time measurements hardly comparable, is not considered. The elapsed time was automatically recorded by the computer system. We then compare the elapsed time for semi-automatic (t_{sem}) and manual (t_{man}) landmark extraction. However, drawing conclusions which are based only on the absolute elapsed time can be problematic. One reason is that in our study no clinical expert was involved. Moreover, both the quality of the data and the number of landmarks varied. Therefore, we primarily analyze the relative differences $\Delta t_{\text{man,sem}}$ between the elapsed time for manual and semi-automatic landmark extraction:

$$\Delta t_{\text{man,sem}} = \frac{t_{\text{man}} - t_{\text{sem}}}{t_{\text{man}}} 100\%. \tag{3}$$

Our assumption is that the relative difference is comparable with that observed in the case of clinical experts.

6.2.2. Accuracy of the registration results

The extracted landmarks are used for rigid registration. To evaluate the registration results, we here follow different ways to assess the results both qualitatively and quantitatively:

1 First, we visually assess the registration results based on the overlay of the transformed MR images with automatically extracted edges from the CT images. In addition, we consider the root-mean-squared (RMS) sum of the residuals of the positions of the used landmarks after registration

$$e_{\text{RMS}} = \sqrt{\frac{1}{n_l} \sum_{l=1}^{n_l} \|\mathbf{e}_l\|^2}, \tag{4}$$

where n_l is the number of landmarks used for registration and $\mathrm{e}_l = \hat{\mathbf{R}}\mathbf{x}_{l,\text{MR}} + \hat{\mathbf{t}} - \mathbf{x}_{l,\text{CT}}$ denotes the residual error at the lth landmark with $\hat{\mathbf{R}}$ and $\hat{\mathbf{t}}$ being the estimated rotation matrix and translation vector, respectively. We are well aware that visual assessment is highly subjective, but unfortunately standardized assessment methods do not exist yet. Note also that drawing conclusions regarding the registration accuracy based on the measure in (4) is generally critical (cf. [25]).

2 In the case of one MR/CT image pair, we estimate the registration error using a number of spatially scattered landmarks that were independently manually specified by an expert participating in another study ([24]). Note, however, that the positions of these landmarks (voxel positions) may be prone to error and, therefore, conclusions based on these error measurements must be drawn carefully.

3 For some part of the data, we estimate the registration error using a marker-based gold standard transformation. That data were provided from the registration evaluation project carried out at Vanderbilt University, USA ([26]). Note, however, that also the gold standard carries some inherent error (see [26] for details).

6.2.3. Reproducibility of the results for different observers

A general requirement in practical applications is the reproducibility of a method. To assess the reproducibility of the results based on semi-automatic and manual landmark extraction for different observers, we study the inter-observer variability of the localized landmark positions. For each procedure, we analyze the variability of the localized landmark positions separately for each landmark in each image. Let $\mathbf{x}_o$ denote the position of a certain landmark in a certain image localized by the oth observer. The mean position of the landmark is then given by $\bar{\mathbf{x}} = 1/n_o \sum_{o=1}^{n_o} \mathbf{x}_o$, where n_o denotes the number of observers who localized the landmark. As a variability measure for the spatial scatter of the landmark positions obtained with one procedure, we consider the RMS distance d_{RMS} from the mean landmark position:

$$d_{\text{RMS}} = \sqrt{\frac{1}{n_o} \sum_{o=1}^{n_o} \|\mathbf{x}_o - \bar{\mathbf{x}}\|^2}. \tag{5}$$

6.3. Experimental set-up

As mentioned above, we used five MR/CT image pairs from different patients (see Section 5.2 for a description of

the image data). Depending on the image quality, the field-of-view of the images, and the presence of lesions, we selected between seven and nine landmarks for each image pair that had to be extracted using both a semi-automatic and a manual procedure. In the case of semi-automatic landmark extraction, the parameter setting was the same as above (see Section 5.2). However, to suppress insignificant detections, we compared the operator responses at the detected positions with a dynamic threshold based on the respective maximal operator response. Note that dynamic here means that detections with operator responses less than 10% w.r.t. the respective maximal operator response were suppressed. To facilitate candidate selection in the case of several detections, the detected candidates were presented to the user in the order of decreasing strength of the operator responses. To estimate the parameters of the rigid transformation, we used the closed-form solution in [27].

6.4. Results of the validation study and discussion

6.4.1. Elapsed time for landmark extraction

In Table 3, the elapsed time for semi-automatic (t_{sem}) and manual (t_{man}) landmark extraction is given for each observer and each image pair. In addition, the relative difference $\Delta t_{man,sem}$ between the elapsed time for manual and semi-automatic landmark extraction is listed. Note that these values refer to the extraction of all landmarks from one image pair.

From Table 3, one can see that in all cases, the elapsed time for semi-automatic landmark extraction was shorter (in part significantly shorter) than the elapsed time for manual landmark extraction. The achieved relative reduction of the elapsed time for semi-automatic landmark extraction as compared to manual landmark extraction ranged between 3 and 66%. One reason for the variation of the elapsed time among the observers is that the strategies for exploring the 3D data as well as the degree of mastering the user-interface were different for the different observers. Note that these differences could be diminished by appropriate training. Also note that the number and types of landmarks to be extracted varied among the different image pairs. Another reason for the variation of the elapsed time is that the quality of the data varied noticeably in terms of contrast and noise. Finally, the spatial pose of the patients relative to the scanner (image) coordinate system varied, which in part made it difficult for the observers to orient within the 3D images since the appearance of anatomy in consecutive slices differed significantly among the different data sets. In the mean, semi-automatic extraction of nine corresponding landmarks from both modalities took 11′30 min, while manual landmark extraction took 18′28 min. Thus, the achieved mean relative reduction of the elapsed time for semi-automatic landmark extraction compared to manual landmark extraction was about 38%.

6.4.2. Accuracy of the registration results

In most cases, the different observers used all landmarks that were extracted from the MR/CT image pairs for registration. By assessing the registration results of both procedures through visual inspection, we found that the results generally showed comparable quality. This is also reflected by Table 4, which gives the computed registration errors based on the landmarks used for registration. From this table, one can conclude that the landmark registration error in general lies between 1 and 3 mm, which indicates reasonable registration results in the case of both procedures. However, one can also see that in most cases, the landmark registration error is somewhat larger in the case of the semi-automatic procedure. Also, the mean landmark registration error is, with the exception of the results obtained for the C06 data, slightly larger in the case of this procedure. Note, however, that drawing conclusions regarding the registration accuracy based on the landmark registration error is generally critical (cf. [25]). Figs. 4 and 5 exemplarily depict the registration results of one observer (observer 1), which represent typical results. Here, the transformed MR data were overlaid with the computed edges of the CT data, which were extracted using a 3D extension of the approach in [28]. In the overlaid data, those voxels lying on a detected edge in the CT image were assigned a high intensity.

To analyze the registration results in more detail, in the following we consider error measures that are independent

Table 3
Elapsed time for semi-automatic (t_{sem}) and manual (t_{man}) landmark extraction (in minutes)

MR/CT data	Obs. 1 t_{sem}/t_{man} $\Delta t_{man,sem}$	Obs. 2 t_{sem}/t_{man} $\Delta t_{man,sem}$	Obs. 3 t_{sem}/t_{man} $\Delta t_{man,sem}$	Obs. 4 t_{sem}/t_{man} $\Delta t_{man,sem}$	Obs. 5 t_{sem}/t_{man} $\Delta t_{man,sem}$
C06 7 landmarks	6′20/18′37 66%	6′13/14′21 57%	14′39/20′25 28%	8′38/9′23 8%	12′27/16′38 25%
V101 7 landmarks	9′17/18′35 50%	12′03/18′12 34%	16′24/25′40 36%	11′20/13′59 19%	17′42/18′13 3%
V104 7 landmarks	8′23/20′17 59%	9′54/13′56 29%	12′04/21′36 44%	8′26/12′05 30%	11′16/25′40 56%
V107 8 landmarks	7′32/18′14 59%	11′09/17′01 34%	13′23/16′54 21%	10′07/15′30 35%	12′10/25′50 53%
V109 9 landmarks	13′38/20′36 34%	16′01/21′46 26%	16′28/24′47 34%	11′56/15′42 24%	12′38/20′08 37%

$\Delta t_{man,sem}$ denotes the relative reduction between t_{man} and t_{sem}.

Table 4
Root-mean-squared (RMS) registration error e_{RMS} and respective maximal registration error e_{max} based on the landmarks used for registration

MR/CT data	Procedure	Obs. 1 e_{RMS}/e_{max}	Obs. 2 e_{RMS}/e_{max}	Obs. 3 e_{RMS}/e_{max}	Obs. 4 e_{RMS}/e_{max}	Obs. 5 e_{RMS}/e_{max}	$\bar{e}_{RMS}$
C06 7 landmarks	Semi-automatic	1.05/1.53	1.36/1.99	1.05/1.53	1.05/1.53	3.02/4.49	1.51
	Manual	1.19/1.56	1.77/2.58	1.93/2.78	1.29/1.56	2.08/2.70	1.65
V101 7 landmarks	Semi-automatic	2.45/3.45	3.24/3.96	2.80/4.54	2.73/3.70	3.40/6.51	2.92
	Manual	1.44/2.04	3.61/6.30	2.59/4.53	2.21/2.96	1.55/2.74	2.28
V104 7 landmarks	Semi-automatic	1.69/3.46	1.99/3.70	3.54/5.40	1.99/3.70	1.68/2.93	2.17
	Manual	1.66/2.39	1.20/1.94	3.04/4.75	1.75/2.39	1.51/2.34	1.83
V107 8 landmarks	Semi-automatic	2.81/4.78	1.97/2.78	2.92/5.09	1.54/2.19	2.81/4.41	2.41
	Manual	1.96/2.72	2.22/3.19	2.44/3.57	2.13/2.87	1.81/2.76	2.11
V109 9 landmarks	Semi-automatic	2.67/3.64	3.42/5.74	2.34/3.97	2.56/3.68	2.99/5.16	2.79
	Manual	1.81/2.66	1.97/2.87	2.83/4.24	1.78/2.85	2.10/2.96	2.09

All values are given in mm units.

of the landmarks used for registration. In the case of the C06 data, we estimate the registration error based on 21 landmarks that were independently determined by an external expert ([24]). In Table 5, the computed registration errors are summarized. The subtle differences in the RMS registration errors, which are in most cases well below 1 mm, confirm our visual impression. The registration errors at the considered landmarks generally ranged between 1 and

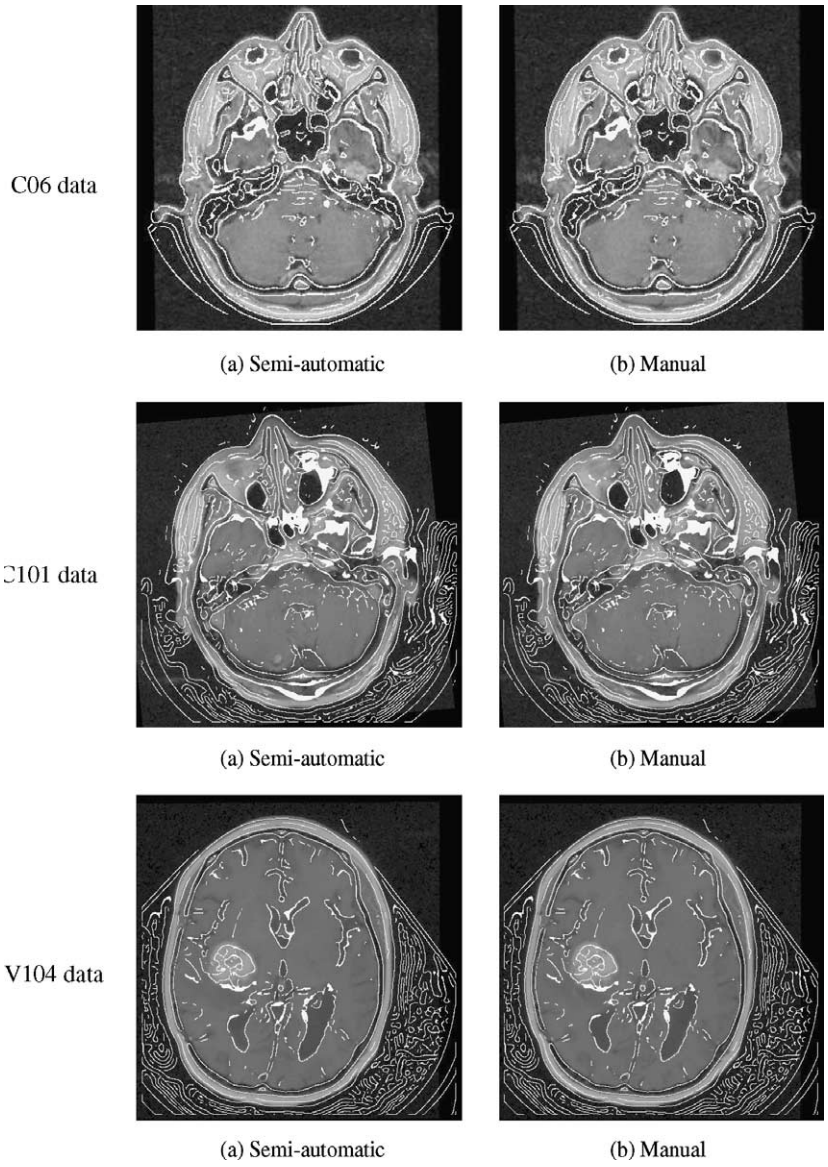

Fig. 4. MR/CT rigid registration results based on semi-automatic and manual landmark extraction (for C06, V101, and V104 data and observer 1).

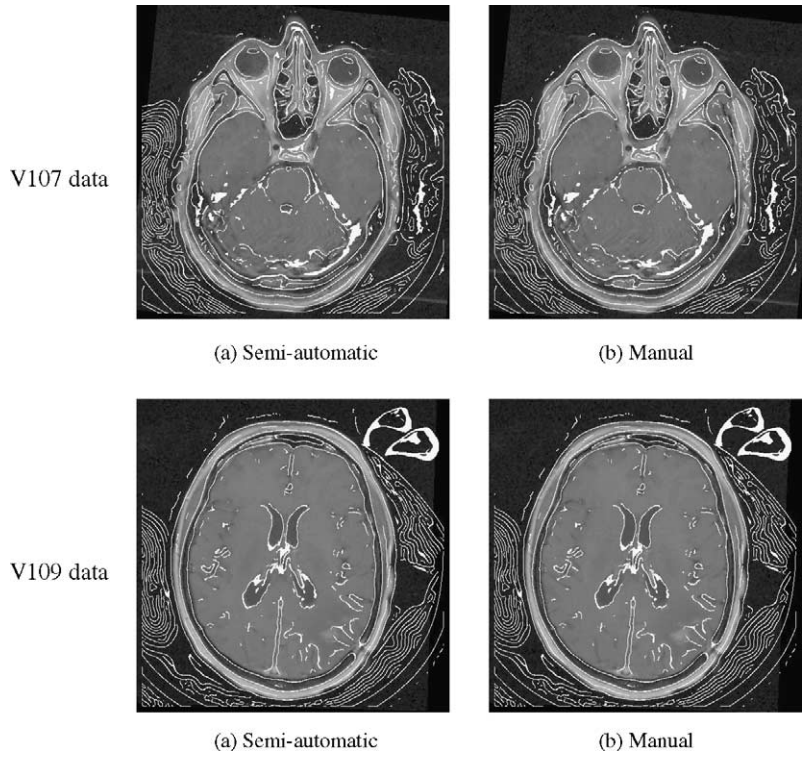

Fig. 5. MR/CT rigid registration results based on semi-automatic and manual landmark extraction (for V107 and V109 data and observer 1).

3 mm. The relatively large maximal error, which occurred in all cases at the same landmark, indicates that the positions for this landmark, which were determined by the expert, are possibly erroneous. In the mean, the semi-automatic procedure gave a registration error of $\bar{e}_{\mathrm{RMS,sem}} = 2.45$ mm, whereas the manual procedure yielded a slightly smaller error of $\bar{e}_{\mathrm{RMS,man}} = 2.28$ mm

In addition, part of the registration results obtained for the V101, V104, V107, and V109 data were evaluated using a marker-based gold standard transformation, which was used to determine corresponding positions as well as the registration error ([26]). In Table 6, the estimated registration errors obtained in the case of the results of observer 1 are given. From this table, one can again conclude that the registration results based on the manual procedure are slightly better. In the mean, the semi-automatic procedure yielded here a registration error of $\bar{e}_{\mathrm{RMS,sem}} = 2.47$ mm, whereas the manual procedure gave an error of $\bar{e}_{\mathrm{RMS,man}} = 1.88$ mm. Note, however, that the results refer to only one observer and recall also that while semi-automatic and manual landmark extraction actually yielded similar registration results, while the elapsed time for landmark extraction was significantly reduced with the semi-automatic procedure.

6.4.3. Reproducibility of the results for different observers

In total, each observer extracted up to 76 landmarks (38 landmarks in five MR images and 38 landmarks in five CT images). We found that for 58 out of the 76 landmarks (i.e. in 76% of all cases) the variability of the localized landmark positions was smaller in the case of the semi-automatic procedure. For 47 out of these 58 landmarks all observers even selected the same candidate out of the set of detected candidates for a landmark (i.e. in 62% of all cases we have $d_{\mathrm{RMS,sem}} = 0$ mm for the semi-automatic procedure). For seven out of the remaining 11 landmarks we found that only one observer selected a different (outlier) candidate. Only for 18 out of the 76 landmarks the semi-automatic procedure showed a larger variability of the localized landmark positions than the manual procedure.

Table 5
Root-mean-squared (RMS) registration error e_{RMS} and respective maximal registration error $e_{\max}$ based on 21 manually specified landmarks that were independently determined by an expert (for C06 data and all observers)

Procedure	Obs. 1 $e_{\mathrm{RMS}}/e_{\max}$	Obs. 2 $e_{\mathrm{RMS}}/e_{\max}$	Obs. 3 $e_{\mathrm{RMS}}/e_{\max}$	Obs. 4 $e_{\mathrm{RMS}}/e_{\max}$	Obs. 5 $e_{\mathrm{RMS}}/e_{\max}$	$\bar{e}_{\mathrm{RMS}}$
Semi-automatic	2.09/4.57	2.41/5.20	2.09/4.57	2.09/4.57	3.56/6.44	2.45
Manual	2.26/4.70	2.09/4.79	2.14/5.13	2.26/5.07	2.67/5.80	2.28

All values are given in mm units.

Table 6
Root-mean-squared (RMS) registration error e_{RMS} and respective maximal registration error e_{max} using a marker-based transformation (for V101, V104, V107, and V109 data and observer 1)

Procedure	V101 e_{RMS}/e_{max}	V104 e_{RMS}/e_{max}	V107 e_{RMS}/e_{max}	V109 e_{RMS}/e_{max}	$\bar{e}_{RMS}$
Semi-automatic	3.38/4.78	2.29/3.02	2.21/2.93	1.97/2.28	2.47
Manual	2.56/2.62	1.58/2.39	1.72/2.41	1.66/2.08	1.88

All values are given in mm units.

By inspecting the results further, we found that for 10 out of these 18 landmarks only one observer selected a different (outlier) candidate. In summary, in the case of the semi-automatic procedure the mean variability of the localized landmark positions was $\bar{d}_{RMS,sem} = 1.06$ mm while in the case of the manual procedure the mean variability was $\bar{d}_{RMS,man} = 2.22$ mm. Given the relatively small sample size, however, it is hard to draw final conclusions w.r.t. the variability of the localized positions of single landmarks. The experimental results revealed, however, that the variability of the localized positions was similar for both modalities: We found that in the case of the MR images for 30 out of 38 landmarks the variability was smaller in the case of the semi-automatic procedure ($\bar{d}_{RMS,sem,MR} = 0.94$ mm), while in the case of the CT images for 28 out of 38 landmarks the variability was smaller in the case of the semi-automatic procedure ($\bar{d}_{RMS,sem,CT} = 1.19$ mm).

7. Conclusion

The detection performance is crucial for an automated procedure for landmark extraction. Ideally, such a procedure should yield only one correct detection for the landmark of interest. Existing approaches to landmark detection, however, often suffer from a rather large number of false detections, which in turn would affect in a negative way the acceptance by a user. One main contribution of this paper was a novel multi-step approach to improved landmark detection, which combines an existing robust 3D differential operator for landmark detection with two different approaches to the reduction of false detections. We analysed in detail the detection performance of the multi-step approach for a number of different anatomical point landmarks of the human head in 3D MR and CT images. Our experimental evaluation showed that the combination of automatic ROI size selection with the incorporation of a priori knowledge of the intensity structure at a landmark is very effective in reducing the number of false detections. In comparison to applying a 3D detection operator alone, as has been done in previous work, the detection performance is considerably improved, which thus significantly improves the applicability of semi-automatic landmark extraction.

As a second main contribution, we then presented the results of a validation study in which we compared the performance of semi-automatic landmark extraction based on the new multi-step approach to landmark detection with that of a purely manual (standard) procedure for landmark extraction. As application, we exemplarily considered rigid registration of 3D MR and CT images. The main result of our study was that compared to a purely manual procedure, (a) the elapsed time for landmark extraction can be significantly reduced with the semi-automatic procedure, (b) the registration results based on semi-automatic landmark extraction generally show comparable quality, and (c) the reproducibility of the results in terms of the inter-observer variability of the localized landmark positions is significantly smaller with the semi-automatic procedure. The results of the validation study demonstrated the applicability of semi-automatic landmark extraction based on the multi-step approach to landmark detection for the purpose of image registration. However, the quantitative evaluation of the registration results revealed also that compared to the results based on manual landmark extraction, the semi automatic procedure showed a slightly larger registration error. Thus, additional steps for the purpose of refined localization of 3D anatomical landmarks (e.g. using the differential approach in [13] and also a recent more global approach in [29] based on deformable models) are assumed to improve the results.

Acknowledgements

This work was supported by Philips Research Hamburg, Germany, within the project IMAGINE (IMage-and Atlas-Guided Interventions in NEurosurgery). Images were provided through the E.U. AIM project COVIRA and as part of the project 'Evaluation of Retrospective Image Registration', National Institutes of Health, Project Number 1 RO1 NS33926-01, Principal Investigator, Dr. J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN, USA. We thank Dr Derek L.G. Hill, Guy's Hospital, London, England, for providing the 'ground truth' positions for the C06 data.

Appendix A. Statistical localization uncertainty of the edge intersection position estimate

The 3D edge intersection approach relies on a simple polyhedral model of the surface at a landmark. The idea is to

locally approximate the landmark's surface through tangent planes, which are defined based on the intensity gradient, and then to intersect these tangent planes based on a distance minimization to obtain a position estimate for the landmark. To emphasize edge points, the distances are here weighted by the gradient magnitude.

The gradient-weighted perpendicular distance from an arbitrary point $\mathbf{x}$ to the tangent plane at a point $\mathbf{x}_i$ within the considered ROI then reads $\varepsilon_i(\mathbf{x}) = \langle \nabla g(\mathbf{x}_i),\ \mathbf{x} - \mathbf{x}_i \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product. The position estimate is obtained by minimizing the residual distances using the least-squares method

$$E(\mathbf{x}) = \sum_{i=1}^{n} \varepsilon_i(\mathbf{x})^2 \rightarrow \min., \tag{A.1}$$

where n denotes the number of voxels within the ROI. Minimizing (A.1), we obtain as position estimate:

$$\hat{\mathbf{x}} = \left(\sum_{i=1}^{n} \nabla g(\mathbf{x}_i) \nabla g(\mathbf{x}_i)^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} \nabla g(\mathbf{x}_i) \nabla g(\mathbf{x}_i)^{\mathrm{T}} \mathbf{x}_i \right). \tag{A.2}$$

Now let us assume that the residuals ε_i, which (in the case of an isolated structure) in general characterize the noise, are statistically independent and normally distributed with zero mean and variance σ_ε^2. This implies that the position estimate $\hat{\mathbf{x}}$ resulting from the edge intersection approach is also normally distributed, and its covariance matrix reflecting the statistical localization uncertainty can be stated as

$$\sum_{\mathrm{w}} = \sigma_\varepsilon^2 \left(\sum_{i=1}^{n} \nabla g(\mathbf{x}_i) \nabla g(\mathbf{x}_i)^{\mathrm{T}} \right)^{-1}, \tag{A.3}$$

where an estimate of the variance is given by $\hat{\sigma}_\varepsilon^2 = E(\hat{\mathbf{x}})/(n-3)$ (e.g. [14]).

References

[1] F.L. Bookstein, Principal warps: thin-plate splines and the decomposition of deformations, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (6) (1989) 567–585.

[2] D.L.G. Hill, D.J. Hawkes, J.E. Crossman, M.J. Gleeson, T.C.S. Cox, E.E.C.M.L. Bracey, A.J. Strong, P. Graves, Registration of MR and CT images for skull base surgery using point-like anatomical features, The British Journal of Radiology 64 (1991) 1030–1035.

[3] A.C. Evans, W. Dai, L. Collins, P. Neelin, S. Marrett, in: M.H. Loew (Ed.), Warping of a Computerized 3-D Atlas to Match Brain Image Volumes for Quantitative Neuroanatomical and Functional Analysis Proceedings of the SPIE's International Symposium on Medical Imaging 1991: Image Processing, San Jose, CA, USA vol. 1445, SPIE Press, Bellingham, WA, 1991, pp. 236–246.

[4] D.L.G. Hill, D.J. Hawkes, M.J. Gleeson, T.C.S. Cox, A.J. Strong, W.-L. Wong, C.F. Ruff, N. Kitchen, D.G.T. Thomas, J.E. Crossman, C. Studholme, A.J. Gandhe, S.E.M. Green, P. Robinson, Accurate frameless registration of MR and CT images of the head: applications in surgery and radiotherapy planning, Radiology 191 (1994) 447–454.

[5] K. Rohr, H.S. Stiehl, R. Sprengel, T.M. Buzug, J. Weese, M.H. Kuhn, Landmark-based elastic registration using approximating thin-plate splines, IEEE Transactions on Medical Imaging 20 (6) (2001) 526–534.

[6] F.A. Gerritsen, C.W.M. van Veelen, W.P.Th. Mali, A.J.M. Bart, H.L. T. de Bliek, J. Buurman, A.H.W. van Eeuwijk, M.J. Hartkamp, S. Lobgret, L. Moreira Pereira Ramos, L.J. Polman, P.C. van Rijen, C.P. Visser, Some requirements for and experience with COVIRA algorithms for registration and segmentation in: L. Beolchi, M.H. Kuhn (Eds.), Medical Imaging-Analysis of Multimodality 2D/3D Images, IOS Press, Amsterdam, 1994, pp. 4–28.

[7] J.-P. Thirion, New feature points based on geometric invariants for 3D image registration, International Journal of Computer Vision 18 (2) (1996) 121–137.

[8] K. Rohr, On 3D differential operators for detecting point landmarks, Image and Vision Computing 15 (3) (1997) 219–233.

[9] W. Beil, K. Rohr, H.S. Stiehl, Investigation of approaches for the localization of anatomical landmarks in 3D medical images in: H.U. Lemke, M.W. Vannier, K. Inamura (Eds.), Proceedings Computer Assisted Radiology and Surgery 1997 (CAR'97), Berlin, Germany, Elsevier Science, Amsterdam, 1997, pp. 265–270.

[10] T. Hartkens, K. Rohr, H.S. Stiehl, Performance of 3D differential operators for the detection of anatomical point landmarks in MR and CT images in: K.M. Hanson (Ed.), Proceedings of the SPIE's International Symposium on Medical Imaging 1999: Image Processing, San Diego, CA, USA vol. 3661, SPIE Press, Bellingham, WA, 1999, pp. 32–43.

[11] S. Frantz, K. Rohr, H.S. Stiehl, in: C.J. Taylor, A.C.F. Colchester (Eds.), Improving the detection performance in semi-automatic landmark extraction, Proceedings of the Second International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'99), Cambridge, England Lecture Notes in Computer Science vol. 1679, Springer-Verlag, Berlin, 1999, pp. 253–262.

[12] S. Frantz, K. Rohr, H.S. Stiehl, S.-I. Kim, J. Weese, in: H.U. Lemke, M.W. Vannier, K. Inamura, A.G. Farman (Eds.), Validating point-based MR/CT registration based on semi-automatic landmark extraction Proceedings of the Computer Assisted Radiology and Surgery 1999 (CARS'99), Paris, France, Elsevier Science, Amsterdam, 1999, pp. 233–237.

[13] S. Frantz, K. Rohr, H.S. Stiehl, in: K.M. Hanson (Ed.), Refined localization of three-dimensional anatomical point landmarks using multi-step differential approaches Proceedings of the SPIE's International Symposium on Medical Imaging 1998: Image Processing, San Diego, CA, USA vol. 3338, SPIE Press, Bellingham, WA, 1998, pp. 28–38.

[14] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, fourth ed., Prentice-Hall, Upper Saddle River, NJ, 1998.

[15] M. Okutomi, T. Kanade, A locally adaptive window for signal matching, International Journal of Computer Vision 7 (2) (1992) 143–162.

[16] W.J. Niessen, J.S. Duncan, M. Nielsen, L.M.J. Florack, B.M. ter Haar Romeny, M.A. Viergever, A multiscale approach to image sequence analysis, Computer Vision and Image Understanding 65 (2) (1997) 259–268.

[17] T. Lindeberg, Junction Detection with Automatic Selection of Detection Scales and Localization Scales Proceedings of the First International Conference on Image Processing (ICIP'94), Austin, TX, USA, vol. I, IEEE Computer Society Press, Los Alamitos, CA, 1994. pp. 924–928.

[18] L. Kitchen, A. Rosenfeld, Gray-level corner detection, Pattern Recognition Letters 1 (2) (1982) 95–102.

[19] L.M.J. Florack, B.M. ter Romeny, J.J. Koenderink, M.A. Viergever, General Intensity Transformations and Differential Invariants, Journal of Mathematical Imaging and Vision 4 (2) (1994) 171–187.

[20] K. Rohr, in: W.M. Wells, A. Colchester, S. Delp (Eds.), Image Registration Based on Thin-Plate Splines and Local Estimates of

Anisotropic Landmark Localization Uncertainties, Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'98), Cambridge, MA, USA Lecture Notes in Computer Science vol. 1496, Springer-Verlag, Berlin, 1998, pp. 1174–1183.

[21] K. Rohr, Extraction of 3d anatomical point landmarks based on invariance principles, Pattern Recognition 32 (1) (1999) 3–15.

[22] M. Unser, A. Aldroubi, M. Eden, B-Spline signal processing: part I-theory, IEEE Transactions on Signal Processing 41 (2) (1993) 821–833.

[23] S. Frantz, Local and Semi-Global Approaches to the Extraction of 3D Anatomical Landmarks from 3D Tomographic Images, PhD thesis, Fachbereich Informatik, Universität Hamburg, Germany, 2001, Published in Dissertationen zur Künstlichen Intelligenz, Band 253. Akademische Verlagsgesellschaft, Berlin, 2001.

[24] K.C. Strasters, J.A. Little, J. Buurman, D.L.G. Hill, D.J. Hawkes, in: J. Troccaz, E. Grimson, R. Mösges (Eds.), Anatomical landmark image registration: validation and comparison, Proceedings of the First Joint Conference Computer Vision, Virtual Reality and Robotics in Medicine and Medical Robotics and Computer-Assisted Surgery (CVRMedMRCAS'97), Grenoble, France Lecture Notes in Computer Science vol. 1205, Springer-Verlag, Berlin, 1997, pp. 161–170.

[25] J.M. Fitzpatrick, J.B. West, C.R. Maurer Jr., Predicting Error in Rigid-Body Point-Based Registration, IEEE Transactions on Medical Imaging 17 (5) (1998) 694–702.

[26] J. West, J.M. Fitzpatrick, M.Y. Wang, B.M. Dawant, C.R. Maurer Jr., R.M. Kessler, R.J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P.A. van den Elsen, S. Napel, T.S. Sumanaweera, B. Harkness, P.F. Hemler, D.L.G. Hill, D.J. Hawkes, C. Studholme, J.B.A. Maintz, M.A. Viergever, G. Malandain, X. Pennec, M.E. Noz, G.Q. Maguire Jr., M. Pollack, C. A. Pelizzari, R.A. Robb, D. Hanson, R.P. Woods, Comparison and evaluation of retrospective intermodality brain image registration techniques, Journal of Computer Assisted Tomography 21 (4) (1997) 554–566.

[27] K.S. Arun, T.S. Huang, S.D. Blostein, Least-squares fitting of two 3-D point sets, IEEE Transactions on Pattern Analysis and Machine Intelligence 9 (5) (1987) 698–700.

[28] J.F. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (6) (1986) 679–698.

[29] S. Frantz, K. Rohr, H.S. Stiehl, in: S.L. Delp, A.M. DiGioia, B. Jaramaz (Eds.), Localization of 3D Anatomical Point Landmarks in 3D Tomographic Images Using Deformable Models, Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000), Pittsburgh, PA, USA Lecture Notes in Computer Science vol. 1935, -Verlag, Berlin, 2000, pp. 492–501.

[30] R. Bertolini, G. Leutert, Atlas der Anatomie des Menschen, Band 3: Kopf, Hals, Gehirn, Rückenmark und Sinnesorgane, vol. 3, Springer-Verlag, Berlin, 1982.

[31] J. Sobotta, Atlas der Anatomie des Menschen Band 1: Kopf, Hals, obere Extremität Haut, München, 19th edition, Urban and Schwarzenberg, 1988.