Automatic accuracy measurement for multi-modal rigid registration using feature descriptors

Frida Hauler[1⋆], Miro Jurisic[1], Hugo Furtado[1,2], Umberto Sabatini[3], Anne Laprei[4], Ursula Nestle[5], and Wolfgang Birkfellner[1,2]

[1] Medical University Vienna, Center for Medical Physics and Biomedical Engineering, Waeringer Guertel 18-20., 1090 Vienna, Austria
[2] Medical University Vienna, Christian Doppler Laboratory for Medical Radiation Research for Radiation Oncology, Vienna, Austria,
[3] Fondazione Santa Lucia, Rome, Italy
[4] Institute Claudius Regaud, Toulouse, France
[5] Universitaet Klinikum Freiburg, Germany
{frida.hauler,
wolfgang.birkfellner}@meduniwien.ac.at
http://summer-project.eu

Abstract. In radiotherapy (RT) for tumor delineation and diagnostics, complementary information of multi-modal images is used. Using high ionizing radiation, the accuracy of registered volume data is crucial; therefore a reliable and robust evaluation method for registered images is needed in clinical practice. Multi-modal image registration aligns images from different modalities like computed tomography (CT) and magnetic resonance imaging (MRI) or cone beam computed tomography (CBCT) into one common frame of reference. The gold standard validation methods are visual inspection by radiation oncology experts and fiducial-based evaluation. However, visual inspection is a qualitative measure with a range of 2-6 mm inaccuracy, it is time consuming and prone to errors. The fiducial-based evaluation is an invasive method when fiducial markers are fixated to bone or implanted in organs. Therefore, in clinical practice a robust non-invasive automated method is needed to validate registration of multi-modal images.
The aim of this study is to introduce and validate an automatic landmark-based accuracy measure for multi-modal image rigid registration using feature descriptors. A porcine dataset with fixed fiducial markers was used to compare our accuracy measure with the target registration error of fiducial markers.In addition, the robustness of our evaluation method

⋆ This work is part of the SUMMER Marie Curie Research Training Network (PITN-GA-2011-290148), which is funded by the 7th Framework Programme of the European Comission (FP7-PEOPLE-2011-ITN). 10 lung dataset used in this work were obtained from Christian Doppler Laboratory for Medical Radiation Research for Radiation Oncology, Vienna, Austria; the 10 brain and 5 lung dataset, from SUMMER project partners Institute Claudius Regaud, Toulouse, France and Universitaet Klinikum Freiburg, Germany

was tested on multi-vendor database consisted of 10 brain and 20 lung cases comparing the automatic landmark accuracy measure based on feature descriptors with manual landmark based evaluation.
An automatic, non-invasive method based on feature descriptors for accuracy evaluation of multi-modal rigid registration was introduced. The method can be used to provide accuracy information slice-by slice on CT, CBCT and CT, MR-T1, -T2 weighted, MR-T1 contrast enhanced (ce) multi-modal images.

Keywords: accuracy of multi-modal rigid registration, accuracy measure based on feature descriptors, automatic landmark based evaluation

1 Introduction

In cancer treatment, RT is one of the main therapeutic measures next to surgery and chemotherapy. The goal of RT is to give a high dose of ionizing radiation to target volumes while sparing the surrounding healthy tissue. To successfully destroy the tumor cells a dose of 50 to 90 Gy is necessary, delivered in a cycle of up to 30 daily fractions. To spare the surrounding healthy tissue and organs at risk (OAR), utmost precision is necessary to define the tumor structures (clinical target volume - CTV) and exact beam control is needed to deliver the high dose to the planned target volume (PTV).
Different image modalities help the diagnosis and target volume definition providing different diagnostic information in RT. These can be anatomical images as CT, MRI, X-rays and ultrasound (US) or functional images such as Positron Emission Tomography (PET), functional MRI (fMRI) and single-photon emission computerized tomography (SPECT). However, no single modality can contain all the diagnostic information for reliable determination and delineation of malignant tissues.
To obtain better tumor targeting during RT treatment using complementary information from multi-modal images, the volumes need to be aligned into same coordinate system using 3D-3D registration algorithms. Registration is the determination of an optimal geometrical transformation which aligns one dataset (moving image) with corresponding areas in an other dataset (fixed image) taken at various points in time or by different scanners. Registration is a wide field with an arsenal of proven algorithms[1], but still there is a gap in defining the accuracy of the registration which for tumor delineation is crucial.
The current gold standard validation methods of registration are visual inspection by a radiation oncologist expert and fiducial-based evaluation [2]. The visual inspection is a qualitative measure depending on inter-observer variability between the experts and lack a standard grading of registration accuracy in clinical practice. However, registration will be used by experts and their opinion has crucial importance for validation of any evaluation method.
A more quantitative validation is based on fiducial markers applied on surface or inside of the body. Fixed fiducials require an invasive intervention while fiducial markers positioned on the skin can move. In any case the fiducial-based valida-

tion is considered as a gold standard evaluation for a quantitative measure of registration error; nevertheless in absence of datasets containing fiducial markers a manual landmark based evaluation can be used manually annotating anatomical features by radiation oncologists. This is a time consuming process prone to intra- and inter observer variability error.
We propose a reliable, automatic and non-invasive method for measuring the accuracy and robustness of the outcome of registration using feature descriptors. For validating this method accuracy, a porcine dataset with fixed fiducial markers is used comparing our method with the target registration error of fiducial markers. The robustness of the method was tested on 10 brain and 20 lung clinical dataset. In absence of fiducial markers in patient dataset, a landmark-based accuracy value was compared with our automated landmark based accuracy measure.

2 Methods

2.1 Phantom Dataset

The accuracy of the automatic evaluation method was performed on a phantom dataset of a porcine head, consisting seven fixed bone fiducial markers and a known registration gold standard [3]. The dataset is available to the public on `http://midas3.kitware.com/midas/community/3`.
The reference dataset consists of CT, CBCT with big and small field of view (FOV) and MR-T1, MR-T2 images. The CT volumes were scanned with a Spiral Philips CT scan consisting of 825 slices with 0.8 slice thickness, each slice containing 512 x 512 pixel of 0.63 x 063 x 0.40 mm^3 voxel size. The CBCT images were acquired by Elekta Synergy linear accelerator (LINAC) with two different field of view (FOV), with 540 x 540 x 520 voxels of 0.5 mm^3 size and a larger FOV of 410 x 410 x 264 voxels of 1.0 mm^3 size. MR volumes were acquired by 3 T Philips MR Scan, the T1 weighted images with 240 x 240 x 150 mm^3 image resolution and 1 x 1 x 1.99 mm voxel size and the T2 weighted images with 192 x 256 x 39 image resolution and 0.89 x 0.89 x 3 mm^3 voxel size.

2.2 Patient datasets

The robustness of the method have been tested on multi-center clinical datasets including 10 brain and 20 lung patients. The brain dataset consists of CT, MR-T1, MR-T1 CE with Gd and MR-T2 weighted images. CT volume scanned by GE Medical System with 512 x 512 x 95 image resolution, 2.5 mm slice thickness and 0.97 x 0.97 x 2.5 mm^3 voxel size. The MR images were acquired by 1.5 T Siemens scanner; T1-weighted and T1 Gd CE volumes with 224 x 256 FOV, image matrix of 224 x 256 x 160 and 1 x 1 x 1 mm^3 voxel size with 1 mm slice thickness. The MR-T2 weighted images of 192 x 256 x 39 image resolution, 0.89 x 0.89 x 3 mm3 voxel size and 3 mm slice thickness.
The 20 lung dataset are coming from two different centers, each patient consisting

of CT and CBCT images. 5 patient data were acquired on Siemens Medcom Scanner, with 512 x 512 x 143 mm^3 image resolution and 0.97 x 0.97 x 2 mm^3 voxel size; Varian Medical Cone-beam CT with 1.17 x 1.17 x 2.96 mm^3 voxel size and 384 x 384 x 52 mm^3 image resolution. Rest of the 15 patient data have been scanned on Siemens Medcom Scanner, with 0.97 x 0.97 x 2 mm^3 voxel size and 512 x 512 x 143 mm^3 image resolution and the CBCT on Elekta LINAC kv image during the treatment with 1 x 1 x 4 mm^3 voxel size and 410 x 410 x 66 image resolution.

2.3 Rigid registration and pre-processing

Due to different slice thickness of multi-modal images, rigid registration is sensitive to uncertainties. To avoid these uncertainties, all images were preprocessed by re-sampling to an 1 mm^3 isotropic voxel size, using cubic spline interpolation.

For registration, we used a commercial software Analyze 11.0 (AnalyzeDirect Inc., Visualization and Analysis Software) based on mutual information (MI) metric; therefore the fiducial markers and our method were only used for evaluation purposes of the registration accuracy. During registration, the CT volume was always considered the fixed image and the CBCT or MR images the moving image.

2.4 Evaluation of rigid registration

To validate the accuracy of registration we adapted error measures from Maurer et al.[4].

- Fiducial registration error (FRE), the root-mean-square distance between corresponding fiducial points after registration.
- Target registration error (TRE), the distance between corresponding points other than fiducial points after registration.

In order to define corresponding points of the anatomical target points for TRE calculation after registration, landmarks have been defined in two ways, found automatically using feature descriptors and annotated manually. For porcine phantom data, the manually annotated points on CT, CBCT and MR-T1, -T2 images using Analyze 11.0 have been chosen based on the fact they are not deform with the soft tissue and are visible on all used image modalities. Manual landmarks have been annotated on reference slices which has been chosen to contain the fiducial markers for later comparison of FRE to TRE for both automatic and manual landmarks. (Fig. 1).

Same procedure has been applied for patient datasets, after rigidly registering the brain (CT with MR-T1, T2 weighted,T1-Gd contrast enhanced) and lung (CT with CBCT) images, features on both fixed and moving images are located using the SURF algorithm from Matlab OpenSURF Computer Vision Library [5]. The interest points as distinctive locations like corners, blobs, T-junctions are detected by Hessian detector (Fig.1 f.-j.). The neighbourhood of every interest point is represented by a feature vector. The calculation time is directly

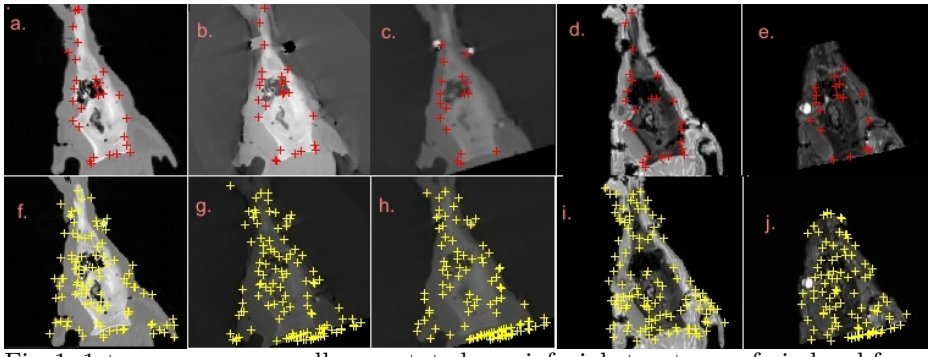

Fig. 1: 1st row: a.-e. manually annotated craniofacial structures of pig head from coronal view; 2nd row: automatic landmarks found on same slices by SURF. Fig. a.,f.-CT, b.,g.-CBCT sFOV, c.,h.-CBCT MFOV, d.,i.-MR-T1, e.j.-MR-T2.

proportional to the dimension of the descriptor, so SURF detector relies on integral images [6] and only 64 dimensions are used to reduce the computational time. The size of the filters is set by the octave parameter. Higher octaves use larger filters and sub-sample the image data to find larger size blobs. Increasing the number of scale levels to compute per octave detect more blobs at finer scale. Indifferent if the landmark have been found automatically by SURF or manually (Fig. 2), the feature vector elements or the coordinates of manual annotation, are matched between both images by building a correlation matrix from pair of points which correlate in both directions inside of a maximum search radius. The outliers from the hypothetical matches have been eliminated by RANdom SAmple Consensus (RANSAC)[8].

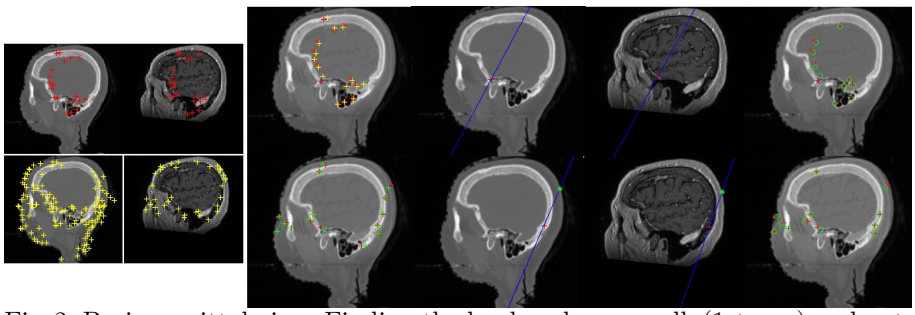

Fig. 2: Brain, sagittal view. Finding the landmarks manually(1st row) and automatically (2nd row), using the same matching algorithm for finding the corresponding point-pairs

2.5 Validation of accuracy method

Manually annotated brain features were chosen based on features investigated during visual inspection by radiation oncologist (Table 1). The anatomical landmarks on lung cases have been chosen based on Grgic et al. [9], a complete list presented in Table 2. The registered volumes have been also checked by visual inspection by a radiologist expert. For cross-validation of the quantitative

accuracy measure obtained by automated feature descriptor based algorithm a manual landmark based evaluation was applied on 10 brain and 20 lung clinical patient data set. Each 10th slice was manually annotated by an expert and approved by two independent radiologist. Annotation of a lung patient on CT and CBCT image modality from three view took 2 hour and 6-8 hours for annotating a brain patient's CT, MR-T1, -T2 weighted and MR-T1 contrance enhanced images.

Table 1: Anatomical features annotated on brain

Visual inspection	**Brain manual landmarks**
Skull, near to lesion	Frontal process, Parietal bone, Sphenoidal sinus, Sphenoidal bone
Eye balls	Orbital Surface, Superior orbital fissure
Sella Turcica	Tuberculum Sellae, Dorsum Sellae, Anterior/Posterior clinoid process
	Clivus, Zygomatic process, Articular tubercle
	Head of mandible, Mandibular fossa, Squamous part, Tympanic part
Chiasma optica	Intratemporal surfaces, Ramus mandible, Pterygoid fovea
	Occipital candyle, Styloid process, Mastoid process

Table 2: Manually annotated features on lung

Lung anatomical landmarks		
Lung apices	Spine	Sternum, Base of ribs
Aortic arch	Heart	Carina (Bronchus bifurcation)
Diaphragm	Tumor	

3 Results

In the pig case, the difference between the fiducial registration error (FRE) and our accuracy measurement is in range of 1.0 ± 0.5 mm on sagittal view, 1.1 ± 0.5 mm on coronal and 0.4 ± 0.2 mm on axial view.

For all patient cases, the SURF method presents higher errors. However, the mean of target registration error (TRE) is less than 0.5 mm compared to manual annotated cases, even in lung cases (Fig.3).

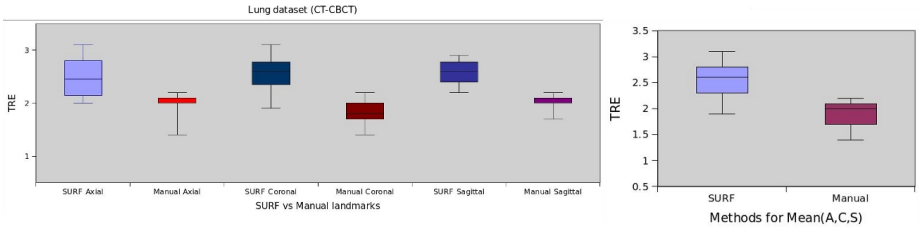

Fig. 3: TRE comparison for SURF and manual landmarks. Left figure: TRE calculated on each view; right figure: a cumulated value is shown between TRE of SURF and landmark evaluation

The robustness of the SURF method is lower than the manual annotation method (Fig. 4). Nevertheless, the mean difference of TRE between SURF and

manual methods for the brain dataset is in range of 1.4 mm and 1.8 mm. The highest difference between SURF and Manual method was observed in MR-T2 image cases.

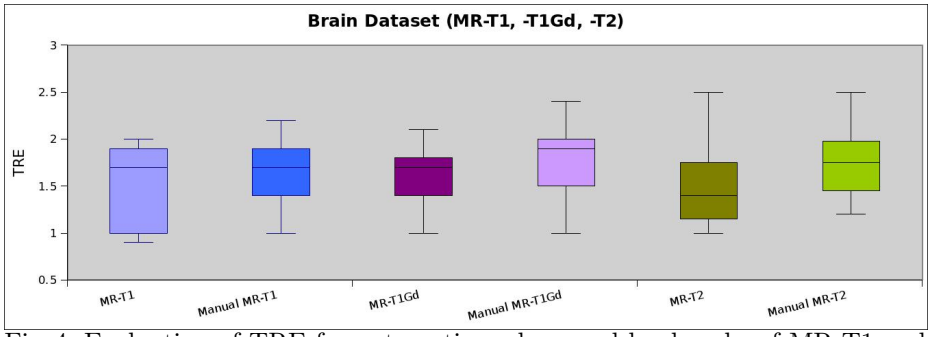

Fig. 4: Evaluation of TRE for automatic and manual landmarks of MR-T1 and CT rigid registration on each view

Comparing the accuracy measures from different image modalities of brain dataset, no significant differences have been observed. All but one was below 1.5 mm in manual method case and all was below 2.0 mm for SURF.(Table 3)

Table 3: TRE of SURF vs. Manual annotation accuracy methods for all multi-modal image modalities used for brain registration

	SURF	**Manual**
	Axial – Sagittal – Coronal	**Axial – Sagittal – Coronal**
MR-T1 –CT	1.9 mm — 1.7 mm — 1.7 mm	1.3 mm — 1.2 mm — 1.2 mm
MR-T1Gd CE –CT	1.7 mm — 1.6 mm — 1.8 mm	1.0 mm — 1.0 mm — 1.0 mm
MR-T2 –CT	1.6 mm — 2 mm — 1.2 mm	1.2 mm — 1.9 mm — 1.5 mm

4 Discussion

The aim of this study was to define the registration accuracy by an automatic non-invasive method which is reliable and robust enough to be suitable in the future use in radiation therapy independent of registration metric applied before. With multi-modal image data, a huge number of surface descriptors might appear that have no direct counterpart in the other modality, which will render reliable point-to-point registration difficult and will also impose robustness issues. Intensity - based registration, on the other hand, is known to work reliable in many cases [3], but the lack of a quantitative measure for registration is a known issue [10]. It therefore makes sense to combine two different methods of aligning multi-modal image data, where one is only used to validate the other. A future possibility might be the validation of the reverse approach, but this is complicated by the fact that most intensity -based measures like MI do not provide absolute values as a result of the optimization procedure which would allow for a quantitative assessment of the registration effort. Our results showed, SURF accuracy measure has approximately 1 mm higher errors in comparison

with FRE and TRE of landmark based evaluation. This can be considered still acceptable, taking in consideration the non-invasiveness of the method and the gained time. As the SURF descriptor can be considered robust, this bias can be originated from parameter tuning of maximum search distance or the window size of the correlation matrix during the matching of corresponding landmark pairs. Nevertheless, the mean error found by SURF based evaluation is close to the manual and the fiducial based accuracy value. The high standard deviation differences between the accuracy measures gives cause for further investigation of robustness of our method, to rank the manual error with different sigma(=1,2,3 mm) using Gaussian random distributed error and investigate if it follows the trend of introduced error values. It would be also interesting to see how big miss-registration can be detected by our method for registration of other anatomical parts which can be more deformed as the head or the lung.

5 Conclusion

Based on the results obtained, we can say that an automatic accuracy measure using feature descriptors is a promising method. Comparing the accuracy values, measured by different methods, we can say, our approach can be considered adequate for a non-invasive, automatic measure. In the future, we need to further assess the robustness of our method. Also, the 2D feature based detection could be extended to 3D features.

References

1. Hajnal, J., Hill D. and Hawkes, D.J.: Medical Image Registration. Boca Raton London New York Washington, D.C. (2001)
2. Fitzpatrick, J.M. and West, J.B.: The distribution of target registration error in rigid-body point-based registration. IEEE Transactions on Medical Imaging. 20,9, 917–927 (2001)
3. Pawiro, S.A. et. al and Wolfgang Birkfellner: Validation for 2D/3D registration I: A new gold standard data set. Medical Physics. 38, 1481–1490 (2011)
4. Maurer Jr., C.R. and Michael Fitzpatrick et al.: Registration of head volume images using implantable fiducial markers. IEEE Transactions on Medical Imaging. 16,4. 447–462 (1997)
5. Christopher Evans: Notes on the OpenSURF Library. Lecture Notes, University of Bristol (2009)
6. Bay, H. and Ess, A. and Tuytelaars, T. and Van Gool, L.: Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding. 110,3. 346–359 (2008)
7. Lowe, D. G.: Object recognition from local scale-invariant features, Proceedings of the IEEE International Conference on Computer Vision 2 : 1150-1157. (1999)
8. Zuliani, M. and Kenney, C.S. and Manjunath, B.S.: The multiransac algorithm and its application to detect planar homographies. Proceedings - International Conference on Image Processing, ICIP. 3. 153–156 (2005)
9. Grgic, A. and Nestle, U. and Schaefer-Schuler, A. and Kremp, S. and Kirsch, C.-M. and Hellwig, D., FDG-PET-Based Radiotherapy Planning in Lung Cancer: Optimum Breathing Protocol and Patient Positioning-An Intraindividual Comparison, International Journal of Radiation Oncology Biology Physics, 73.1: 103-111.(2009)

10. Ulin K, Urie MM, Cherlow JM. Results of a multi-institutional benchmark test for cranial CT/MR image registration. Int J Radiation Oncology Biol Phys. 2010 Aug 1;77(5):1584-9.