Evaluation of similarity measures for rigid registration of multi-modal head images

Darko Škerl¹, Boštjan Likar¹, J. Michael Fitzpatrick², and Franjo Pernuš¹

¹University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
²Vanderbilt University, Department of Electrical Engineering and Computer Science, Nashville, TN, USA

Abstract. Image registrations that are based on similarity measures simply adjust the parameters of an appropriate spatial transformation model until the similarity measure reaches an optimum. The numerous similarity measures that have been proposed in the past are differently sensitive to the imaging modality, image content and differences in image content, selection of the floating and target image, partial image overlap, etc. To be able to study the impact of these factors on the behavior of similarity measures, we have recently developed a protocol for optimization-independent evaluation of similarity measures. In this paper, we use this protocol to evaluate and compare 12 similarity measures for rigid registration. To study the impact of different imaging modalities on the behavior of similarity measures, we have used 16 CT/MR and 6 PET/MR image pairs with known “gold standard” registrations. The results for PET/MR registration and for registration of CT to both rectified and unrectified MR images indicate that mutual information, normalized mutual information, and the entropy correlation coefficient are the most accurate similarity measures and have the smallest risk to be trapped in a local optimum. The results of an experiment on the impact of exchanging the floating and target image indicate that, especially MR/PET registrations, the behavior of some similarity measures, like mutual information, significantly depends on which image is the floating and which is the target.

1 Introduction

Nowadays, medical diagnosis, therapy planning and execution, and monitoring of the progress of disease or effects of treatment heavily depend on medical imaging. Automated analysis of medical images of different modalities and dimensions is a means of fast, accurate, robust, efficient, and objective extraction of crucial information from images. In the past two decades, image registration has rapidly grown into a major area of research in the field of medical image analysis [1-4]. The reasons for the increased interest in image registration derive directly from the purpose of image acquisition and from the acquisition characteristics: (i) in a variety of clinical indications, images are acquired by means of various modalities so that complementary information about patient anatomy and physiology, as well as about pathological processes, is obtained [5-7]; (ii) longitudinal studies in which the state of an organ or a tissue is followed in time so as to monitor progression of disease or efficacy of treatment, are performed with increasing frequency [8-11]; (iii) with the advent of digital image archives, image data bases are constructed, which allow for comparison of the images of individual patients to a digital atlas with known statistical properties [12, 13]; (iv) the rapid growth of minimally invasive interventional procedures has increased the surgeon’s reliance on image guidance – generally by means of two-dimensional (2D) or low quality three-dimensional (3D) images acquired intra-operatively to provide a
correspondence between the actual situation in the treatment room and the treatment planning that was made on the basis of high quality 3D preoperative images [14-17].

Image registration techniques can be classified into two categories: feature-based and voxel-based methods. Feature-based methods require the extraction of features that are visible in both images. Features can be fiducial markers rigidly attached to the patient’s anatomy or distinctive anatomical points or other structures visible in both images to be registered [18]. Accurate automatic extraction of features by image segmentation is by itself a challenging task, while manual segmentation is time consuming and depends on the skills of a human operator. The segmentation error generally propagates to the total error of registration. On the other hand, voxel-based registration methods do not need segmentation and their accuracy is thus not affected by segmentation errors. They optimize a functional measuring the similarity of the images that are registered [18-20]. The main advantage of voxel-based registration techniques is that the calculation of the functional, called the similarity measure (SM), is straightforward.

Registrations based on similarity measures adjust the parameters of an appropriate spatial transformation model until the similarity measure reaches an optimum. Given the images to be registered and the spatial transformation model, the outcome of a registration mainly depends on the similarity measure and the optimization method. The complex interdependence of the similarity measure and optimization makes the assessment of each of them on the registration result difficult even for very specific registration tasks. The numerous similarity measures introduced in the past have different properties and are thus differently sensitive to the imaging modality, image content and differences in image content, selection of floating and target image, sampling, interpolation, histogram binning, partial image overlap, and image degradation, such as noise, intensity inhomogeneities and geometrical distortions. To be able to study the impact of these factors, we have recently developed a protocol for optimization-independent evaluation of similarity measures [21]. In this paper, we use this protocol to evaluate a number of similarity measures, which have been proposed for rigid registration of computed tomography (CT) to magnetic resonance (MR) images and positron emission tomography (PET) to MR images. Besides studying the influence of imaging modalities on the behavior of similarity measures, we have conducted experiments to assess the impact of partial image overlap and of exchanging the floating and target image. In all experiments, we have used the images and the “gold standard” registrations of a large database of multi modal head images from the Vanderbilt University Retrospective Image Registration Evaluation (RIRE) project [22].

2 Material and Methods

2.1 Images
The CT, MR T1 weighted and PET images of the head and their “gold standard” registrations were obtained from the RIRE project, a comparison study of numerous registration methods [22]. The “gold standard” registrations were established via implanted fiducial markers. The estimated accuracy of the “gold standard” was 0.39 mm for CT to MR and 1.6 mm for PET to MR registrations. The CT images were acquired using a Siemens DR-H scanner, the MR images using a Siemens SP 1.5 T scanner, and the PET images with a Siemens/CTI ECAT 933/08 - 16 scanner. The MR T1 image volumes were acquired with an echo time (TE) of 15 ms and a repetition time (TR) of 650 ms (20 slices) or 800 ms (26 slices). For PET, each patient was injected with 10 mCi of $^{18}$F-fluorodeoxyglucose. Scanning was started 40 - 50 minutes after injection and continued for 25 minutes. Image reconstruction was performed using a Hamming reconstruction filter, resulting in images with a full width at half-maximum resolution of 9 mm. Some of the MR images were corrected for static field inhomogeneity using the image rectification technique by Chang and Fitzpatrick [23, 24]. The scale distortions in the MR images were corrected by taking advantage of the attachment to the patients of a COMPASS stereotactic frame as an object of known shape and size. For additional information about the acquisition protocol and image preprocessing see [22].

Image database:

- The CT volumes had a resolution in the x and y directions of 512 pixels, and between 28 and 49 slices. The voxel size was between 0.45 and 0.65 mm in x and y, and between 3.0 and 4.0 mm in z direction.
- The MR volumes had a resolution of 256 pixels in the x and y directions, and between 20 and 52 slices. The voxel size was between 0.86 and 1.28 mm in the x and y directions, and between 3.0 and 4.0 mm in z direction.
- The PET volumes had a resolution of 128 pixels in the x and y directions, and 15 slices. The voxel size was 2.59 mm in x and y, and 8.0 mm in z direction.

Three sets of images were used for evaluation of similarity measures:

- Set 1: 6 corresponding rectified MR T1 and PET images (patients 001, 002, 005, 007, 008 and 009)
- Set 2: 7 corresponding rectified MR T1 and CT images (patients 001 through 007)
- Set 3: 9 corresponding unrectified MR T1 and CT images (patients 101 through 109)

Figure 1 shows image slices from MR T1 weighted, CT, and PET images.

2.2 Similarity measures

We have implemented and evaluated the following similarity measures:

1 http://www.vuse.vanderbilt.edu/~image/registration/
1. symmetric multi-feature mutual information (SMMI) [16]
2. mutual information (MI) [25, 26]
3. normalized mutual information (NMI) [27]
4. entropy correlation coefficient (ECC) [26, 28]
5. joint entropy (H) [29, 30]
6. mutual information based on second order ($\alpha=2$) Havrda-Charvat entropy (HC2) [31-33]
7. mutual information based on third order ($\alpha=3$) Havrda-Charvat entropy (HC3) [31-33]
8. mutual information based on second order ($\alpha=2$) Renyi entropy (RE2) [31, 34]
9. mutual information based on third order ($\alpha=3$) Renyi entropy (RE3) [31, 34]
10. energy of the histogram (E) [35]
11. correlation ratio (COR) [36, 37]
12. Woods criterion or partitioned intensity uniformity (PIU) [38]

All similarity measures were computed from overlapping voxels of the floating and target images and formulated on the 2D joint histogram or joint probability distribution of the intensities and gradients (SMMI) of the two images. Partial volume interpolation was used to obtain the joint histograms [26]. As all similarity measures, except H and PIU, have a maximal value when the images match the best. The H and PIU similarity measure were multiplied by minus one.

2.3 Evaluation protocol

The similarity measure evaluation protocol requires that images, typical for a specific registration task, and “gold standard” registration of these images are given, and that a spatial transformation model (rigid, affine, etc.) is selected [21]. The spatial transformation $T$ that is supposed to bring two images, a floating and a target image, into correspondence was assumed to be rigid and was therefore composed of three translational ($t_x$, $t_y$, $t_z$) and three rotational ($\Theta_x$, $\Theta_y$, $\Theta_z$) parameters. The six-dimensional parametrical space was first normalized so that equal changes of each of the 6 parameters in the normalized parametrical space produced approximately equal voxel shifts when averaged over the whole image volume. This normalization was done so that Euclidean metrics could be used on the parametrical space to determine distances between transformations. We can, for example, measure the distance of a given transformation from the transformation at which the images are “best” aligned (“gold standard”) or from the transformation for which a given similarity measure is maximal. The origin $X_0$ of the 6-dimensional parametrical space was set at the known “gold standard” position. Similarity measure values $SM(X_{n,m})$ were defined for image pairs, with the target image remaining stationary at the origin $X_0$ and the floating image being transformed from the origin to location $X_{n,m}$. Values $SM(X_{n,m})$, $n=1,2,...,N$; $m=-M/2,...,M/2$, of a similarity measure were defined on $N$ lines probing the six dimensional parametrical space and at $M+1$ points evenly spaced along each line. Each of the $N$ lines was defined by a randomly
selected starting position $X_{n,M/2}$ on a hyper sphere of radius $R$, $\|X_{n,M/2}\| = R$, in the normalized 6-dimensional parametrical space and its mirror point $X_{n,M/2}$ on the hyper sphere. Along each line, the position and the value of the global optimum of the similarity measure were denoted by $X_{n,opt}$ and $SM(X_{n,opt})$, respectively.

In this paper, we have evaluated a similarity measure by estimating its accuracy and robustness. These two properties were assessed by the accuracy ($ACC$) (Eq. 1) and the risk of nonconvergence ($RON$) (Eq. 2) features, respectively. Each measure is closely tied to the definition of a “global optimum” on each of the $N$ lines that probe the parameter space. The global optimum is defined to be the position $X_{n,opt}$ (i.e., $m = opt$) for which $SM(X_{n,m})$ is maximal on that line.

- **Accuracy $ACC$** of a similarity measure is defined as the root mean square of distances $||X_{n,opt} - X_{0}||$ between the origin $X_{0}$ and each of the $N$ global optima $X_{n,opt}$, $n = 1, 2, ..., N$

$$ACC = \sqrt{\frac{1}{N} \sum_{n=1}^{N} ||X_{n,opt} - X_{0}||^2} \text{ [mm]}$$  (Eq. 1)

- **Risk of nonconvergence $RON$**, which is related to the smoothness (number and extent of the local minima) of a similarity measure around the $N$ global optima $X_{n,opt}$ and estimates robustness, is defined as the average of all positive gradients $g_{n,m}$ within the probed normalized parametrical space:

$$RON = \frac{1}{2RN} \sum_{n=1}^{N} \sum_{m=-M/2}^{M/2} g_{n,m} \text{ [10^6/mm]}$$  (Eq. 2)

where $g_{n,m}$ was

$$g_{n,m} = \begin{cases} SM(X_{n,m+1}) - SM(X_{n,m}) & \text{if } m < opt \text{ & } SM(X_{n,m+1}) > SM(X_{n,m}) \\ SM(X_{n,m-1}) - SM(X_{n,m}) & \text{if } m > opt \text{ & } SM(X_{n,m+1}) > SM(X_{n,m}) \\ 0 & \text{otherwise} \end{cases}$$  (Eq. 3)

Smaller $RON$ values indicate that a similarity measure is more smooth and therefore more robust. It is expected that the optimization of a similarity measure that is more smooth is more likely to converge to the global optimum than the optimization of a similarity measure whose behavior is less smooth. More details about the evaluation protocol can be found in [21] and in the online version

### 3 Experiments and Results

The number of image intensity bins used to calculate the joint histograms was set to 64. $R$ was set to 35 mm, $N$ to 50, and $M$ to 80. In a previous publication we have shown that the variances of $ACC$ and $RON$ as a function of $N$ were negligible if $N$ had been set to 50 or more. We have chosen

---

2 http://lit.fe.uni-lj.si/Evaluation
such that the distance between two consecutive points on a line was smaller than or equal to the size of the smallest voxel of the two images [21].

### 3.1 The impact of imaging modalities

We first show results of the analysis of similarity measures in PET/MR and CT/MR registrations. In this and all the following experiments, the first modality was the floating image and the second one was the target image. Figure 2 shows box-whiskers diagrams of \( ACC \) and \( RON \) of the 12 similarity measures in PET/MR registrations of images of Set 1. The results indicate that the best similarity measures for PET/MR registrations were MI, NMI, ECC and COR. They all were accurate as their median \( ACC \) was little above 2 mm and they were also robust. The most robust of all were MI, NMI, and ECC. SMMI was the most accurate measure of all but had a high \( RON \) (not shown in Figure 2 because it was out of scale) indicating that it had numerous distinctive local optima within the probed parametrical space and was therefore not smooth. HC2 was rather smooth but not very accurate.

Figure 3 shows box-whiskers diagrams of \( ACC \) and \( RON \) of the 12 similarity measures applied to images of Sets 2 and 3. Results for CT to rectified MR images (top of Figure 3) indicate that MI, NMI, and ECC were again the most accurate and that MI, NMI, ECC, HC2, HC3, and COR were the smoothest and thus potentially the most robust. As expected, the similarity measures were less accurate when applied to Set 3, i.e. to CT and unrectified MR images (bottom of Figure 3). Not only the absolute values but also the variations of \( ACC \) values were larger for image Set 3 than for image Set 2. Again, the measures that performed best were the MI, NMI, and ECC, which behaved almost identically. The comparison of results for image Set 1 and Set 2 (Figure 2 and top of Figure 3) indicate that almost all measures were more accurate when CT and MR images instead of PET and MR images had been registered. Generally, the measures were also smoother.

![Figure 1](image.png)

**Figure 1.** Box-whiskers diagrams, showing the minimum, maximum, median, 1st and 3rd quartile of the distribution of \( ACC \) (left) and \( RON \) (right) properties of 12 similarity measures (sorted according to the median of \( ACC \)) in PET/MR registrations.
As the risk of nonconvergence of the MI, NMI, ECC and COR similarity measures was almost the same and close to zero (Figures 2 and 3), we have conducted a further experiment in which we increased the radius $R$ from 35 mm to 70 mm. $N$ was set to 50 and $M$ to 160. In this way, partial image overlap was reduced and it was expected that NMI would outperform MI as Studholme et al. [27] found that maximization of NMI is superior to MI in case the region of overlap of both images is relatively small. Within the space of larger misregistrations defined by $R$, we measured $RON$ of MI, NMI, ECC and COR. The results are presented in Figure 4 for PET(MR)/MR(PET) and CT(MR)/MR(CT) registrations. The results indicate that MI was the most robust, followed by NMI, ECC and COR. Interestingly, when the MR image had been chosen as the floating image, the four similarity measures were more robust than when the MR had been selected as the target image. This was true for both MR/PET and MR/CT registrations.

Figure 2. Box-whiskers diagrams, showing the minimum, maximum, median, 1st and 3rd quartile of the distribution of ACC and RON properties of 12 similarity measures (sorted according to the median of ACC) in CT to rectified MR (top) and CT to unrectified MR (bottom) image registrations.

Figure 3. Box-whiskers diagrams showing the minimum, maximum, median, 1st and 3rd quartile of the distribution of RON of four similarity measures for which RON was almost zero in Figure 2.
3.2 The impact of exchanging the floating and target image

In the next experiment, we have assessed the impact of exchanging the target and floating images on the behavior of similarity measures. First, we have derived the $ACC$ and $RON$ values of all similarity measures for PET(MR)/MR(PET) and CT(MR)/MR(CT) registrations. For the CT/MR and MR/CT experiments, image Set 2 with rectified MR images had been used. To estimate the impact of exchanging the floating and the target image, two features $N\_ACC$ and $N\_RON$ were derived by normalizing the differences of $ACC$s or $RON$s by corresponding average values, respectively:

$$N\_ACC = 2 \frac{ACC_{I_1,I_2} - ACC_{I_2,I_1}}{ACC_{I_1,I_2} + ACC_{I_2,I_1}} \times 100\%$$

$$N\_RON = 2 \frac{RON_{I_1,I_2} - RON_{I_2,I_1}}{RON_{I_1,I_2} + RON_{I_2,I_1}} \times 100\%$$

(Eq. 4)

Therefore, if the value of $N\_ACC$ or $N\_RON$ is equal to zero, $ACC$ or $RON$ values are the same, regardless which image is the floating and which the target. Negative values of $N\_ACC$ ($N\_RON$), obtained when $ACC_{I_1,I_2}$ ($RON_{I_1,I_2}$) is smaller than $ACC_{I_2,I_1}$ ($RON_{I_2,I_1}$), indicate that it is better to take $I_1$ than $I_2$ for the floating image. Figure 5 shows the minima, maxima, medians, first, and third quartiles of the distribution of $N\_ACC$ and $N\_RON$ values. The values of $N\_ACC$ for MR/PET and PET/MR registrations were distributed around zero, which indicates that exchanging the floating and the target image had little impact on $ACC$ of most of the similarity measures, except for SMMI, which was more accurate if a PET image had been used as the floating and an MR as the target. On the other hand in MR/PET and PET/MR registrations, the distributions of $N\_RON$ for SMMI, MI, NMI, ECC, H, HC2, HC3, COR and PIU indicate that the convergence properties were better (more robust) if the MR rather than the PET image had been used as a floating image. The values of $N\_ACC$ in MR/CT and CT/MR registrations were also distributed around zero, indicating that it did not make a difference whether the MR or the CT image had been used as the floating image. The distributions of $N\_RON$ values in MR/CT registrations indicate that for some measures (distributions below zero) using MR as the floating image would yield slightly better convergence, while for some other measures (distributions above zero) the CT image should be used as the floating image. It seems, that for the MI it did not matter which image had been the floating and which the target, but that when using NMI or ECC slightly better results can be expected if the CT image is used as the floating image.
3.3 Comparison of Similarity Measures

In the next experiment, we assessed the statistical significance of differences between values of ACC and RON obtained by different similarity measures when applied to MR(PET)/PET(MR) and to MR(CT)/CT(MR) registrations. The differences between the values of similarity measures properties were assessed by a paired Student’s t-test. For MR/CT and CT/MR registrations images from Set 2 were used. The p values in the upper right part of Tables 1-4 correspond to MR/PET (Tables 1 and 3) or MR/CT (Tables 2 and 4) registrations, while the p values in the lower left part correspond to PET/MR (Tables 1 and 3) or CT/MR (Tables 2 and 4) registrations. All p values that are larger than a level of significance $\alpha=0.05$ are shown in bold and indicate that similarity measures were not statistically significantly different. In all tables, 0.00 indicate values smaller than 0.005.

Regarding the accuracy of CT/MR registrations (Table 2), the SMMI similarity measure was significantly different from the other measures in that it was more accurate, while MI, NMI and ECC were not significantly different from each other. Tables 1-4 show that there was no significant difference between MI and NMI similarity measures regarding both accuracy and robustness.

**Table 1.** $P$ values for ACC of MR/PET (upper right) and PET/MR registration (lower left)

<table>
<thead>
<tr>
<th></th>
<th>SMMI</th>
<th>MI</th>
<th>NMI</th>
<th>ECC</th>
<th>H</th>
<th>HC2</th>
<th>HC3</th>
<th>RE2</th>
<th>RE3</th>
<th>E</th>
<th>COR</th>
<th>PIU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMMI</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
<td>0.48</td>
<td>0.22</td>
<td>0.08</td>
<td>0.18</td>
<td>0.11</td>
<td>0.09</td>
<td>0.06</td>
<td>0.31</td>
<td></td>
</tr>
</tbody>
</table>
### Table 2. $P$ values for $ACC$ of MR/CT (upper right) and CT/MR registration (lower left)

<table>
<thead>
<tr>
<th>SMMI</th>
<th>MI</th>
<th>NMI</th>
<th>ECC</th>
<th>H</th>
<th>HC2</th>
<th>HC3</th>
<th>RE2</th>
<th>RE3</th>
<th>E</th>
<th>COR</th>
<th>PIU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMMI</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MI</td>
<td>0.00</td>
<td>0.31</td>
<td>0.35</td>
<td>0.10</td>
<td>0.22</td>
<td>0.04</td>
<td>0.01</td>
<td>0.04</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>NMI</td>
<td>0.00</td>
<td>0.38</td>
<td>0.18</td>
<td>0.04</td>
<td>0.19</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ECC</td>
<td>0.00</td>
<td>0.39</td>
<td>0.18</td>
<td>0.04</td>
<td>0.20</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>H</td>
<td>0.00</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.39</td>
<td>0.19</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HC2</td>
<td>0.00</td>
<td>0.22</td>
<td>0.24</td>
<td>0.25</td>
<td>0.43</td>
<td>0.17</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>HC3</td>
<td>0.00</td>
<td>0.07</td>
<td>0.07</td>
<td>0.08</td>
<td>0.42</td>
<td>0.33</td>
<td>0.11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>RE2</td>
<td>0.00</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.12</td>
<td>0.00</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>RE3</td>
<td>0.00</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.12</td>
<td>0.01</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>E</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.21</td>
<td>0.32</td>
<td>0.30</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>COR</td>
<td>0.00</td>
<td>0.25</td>
<td>0.11</td>
<td>0.11</td>
<td>0.17</td>
<td>0.47</td>
<td>0.33</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>PIU</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.45</td>
<td>0.44</td>
<td>0.36</td>
<td>0.07</td>
<td>0.08</td>
<td>0.00</td>
<td>0.34</td>
</tr>
</tbody>
</table>

### Table 3. $P$ values for $RON$ of MR/PET (upper right) and PET/MR registration (lower left)

<table>
<thead>
<tr>
<th>SMMI</th>
<th>MI</th>
<th>NMI</th>
<th>ECC</th>
<th>H</th>
<th>HC2</th>
<th>HC3</th>
<th>RE2</th>
<th>RE3</th>
<th>E</th>
<th>COR</th>
<th>PIU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMMI</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>MI</td>
<td>0.00</td>
<td>0.15</td>
<td>0.16</td>
<td>0.12</td>
<td>0.22</td>
<td>0.43</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>NMI</td>
<td>0.00</td>
<td>0.31</td>
<td>0.17</td>
<td>0.05</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.43</td>
<td>0.02</td>
</tr>
<tr>
<td>ECC</td>
<td>0.00</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>H</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.09</td>
<td>0.03</td>
<td>0.04</td>
<td>0.02</td>
<td>0.07</td>
<td>0.06</td>
<td>0.18</td>
</tr>
<tr>
<td>HC2</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>HC3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
<td>0.03</td>
<td>0.13</td>
<td>0.01</td>
<td>0.37</td>
<td>0.04</td>
</tr>
<tr>
<td>RE2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.01</td>
<td>0.11</td>
<td>0.00</td>
<td>0.40</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>RE3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.03</td>
<td>0.21</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>E</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.36</td>
<td>0.16</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>COR</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>0.04</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>PIU</td>
<td>0.35</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>

### Table 4. $P$ values for $RON$ of MR/CT (upper right) and CT/MR registration (lower left)

<table>
<thead>
<tr>
<th>SMMI</th>
<th>MI</th>
<th>NMI</th>
<th>ECC</th>
<th>H</th>
<th>HC2</th>
<th>HC3</th>
<th>RE2</th>
<th>RE3</th>
<th>E</th>
<th>COR</th>
<th>PIU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMMI</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MI</td>
<td>0.00</td>
<td>0.20</td>
<td>0.44</td>
<td>0.02</td>
<td>0.12</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.17</td>
<td>0.09</td>
</tr>
<tr>
<td>NMI</td>
<td>0.00</td>
<td>0.09</td>
<td>0.14</td>
<td>0.02</td>
<td>0.12</td>
<td>0.08</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.17</td>
<td>0.14</td>
</tr>
<tr>
<td>ECC</td>
<td>0.00</td>
<td>0.09</td>
<td>0.18</td>
<td>0.02</td>
<td>0.12</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.17</td>
<td>0.15</td>
</tr>
<tr>
<td>H</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>HC2</td>
<td>0.00</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.01</td>
<td>0.14</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.25</td>
<td>0.12</td>
</tr>
<tr>
<td>HC3</td>
<td>0.00</td>
<td>0.07</td>
<td>0.05</td>
<td>0.05</td>
<td>0.00</td>
<td>0.18</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.26</td>
<td>0.04</td>
</tr>
<tr>
<td>RE2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.08</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
4 Discussion

A standard method for solving the image registration problem is to treat it as a mathematical optimization, using a similarity measure to quantify the alignment or similarity of the two images for any given spatial transformation. Given the images to be registered and the spatial transformation model, the outcome of a registration mainly depends on the similarity measure and the optimization method. Although in the past most attention has been focused on defining similarity functions, the impact of the optimization method has also been studied [39, 40]. Because of the complex interdependence of the similarity measure and optimization, the impact of each of them on the registration result is difficult even for very specific registration tasks. The numerous similarity measures introduced in the past have different characteristics and are thus differently sensitive to the imaging modality, image content and differences in image content, selection of floating and target image, partial image overlap, sampling, interpolation, histogram binning, and image degradation, such as noise, intensity inhomogeneities and geometrical distortions. To develop a reliable, automatic registration method that is based on optimization of a similarity measure, the characteristics of typical similarity measures need to be understood, i.e. some a priori information on the behavior of the similarity function with regard to the factors above is needed. It is clear that even in the simple case of rigid registration of two 2D images, which requires optimization of only three parameters, the parametrical space is too large to permit a thorough analysis of the similarity function at every point of the space. Limited and indirect information on the similarity measure may, however, be obtained from the accuracy and robustness of the registration itself. A number of studies, comparing several registrations based on different similarity measures have been published in the past [11, 22, 31, 35, 41-45] (see [19, 20] for many more references). Researchers have used different approaches to evaluate such registrations. If the correct registration ("gold standard") is known, registration results of images registered from one or many starting estimates may be analyzed [11, 41, 43]. If the “gold standard” registration is not available or if it is not accurate enough, registration accuracy may be quantified by visual evaluation of registration estimates [25, 41, 46-48] or by assessing the consistency of transformation [11, 39, 49, 50]. More directly, excluding the impact of optimization, similarity measures may be evaluated by drawing plots or traces, showing their behavior when one image is systematically translated from and/or rotated around the “gold standard” registration position [20, 26, 27, 39, 51, 52]. This more direct evaluation also gives only limited information on the behavior of the similarity measure because it is evaluated only at a very small fraction of the parametrical space. To be able to assess the main characteristics of similarity measures we have recently developed a protocol for a more
thorough and optimization-independent evaluation of different characteristics of similarity measures [21]. In this paper we have focused on only two characteristics, i.e. on the accuracy and robustness, because these characteristics are of most interest. Accuracy (i.e., agreement with ground truth) is the distance between the ground-truth transformation to the transformation to which registration converges, while robustness is the ability of a registration method to produce similar results on all trials [53], regardless of starting position, or differences in image contents, which are either due to true variations or a consequence of image degrading effects, or both. Robustness depends on the number and extent of local optima within the space of possible initial misregistrations.

Using the estimates of accuracy ($ACC$) and robustness ($RON$) we have evaluated and compared 12 similarity measures—nine information-theory measures and three others. The most popular information theory measures are the mutual information and normalized mutual information [19, 20, 25, 26, 54]. Other information theory measures have received much less attention in the past. Pluim et al. [45] compared the mutual information with several other $f$-information measures by applying them to rigid registration of clinical MR, CT and PET images from the RIRE image database [22]. Wachowiak et al. [31] also compared information theoretic measures based on Renyi and Havrda-Charvat entropies setting $\alpha$ to 0.25, 0.5, 0.9, 1.1, 1.25, 1.5 and 2. Our results confirm the results of Pluim et al. [45] and Wachowiak et al. [31] that robustness of similarity measures based on Renyi and Havrda-Charvat entropies was influenced by the value of $\alpha$. These similarity measures became less smooth when $\alpha$ was increased from two to three. The symmetric multi-feature mutual information (SMMI), comprised of image intensity and gradient features, has been only recently introduced [16]. Results in Figure 2 show that for PET to MR registrations SMMI was the most accurate but the least robust indicating that this measure would be difficult to optimize. Poor robustness was observed for all other measures except MI, NMI, ECC and COR. As SMMI was the most accurate but not robust, one of the robust similarity measures, for instance MI could be used to start a registration and after converging, registration could be continued with SMMI, which would yield better accuracy. NMI is supposed to be more robust and accurate for registration of images with fewer overlapping voxels [27]. However, the results in Figures 2 and 3 and the results of the paired t-test (Tables 1-4) indicate that there was no significant difference between MI and NMI regarding either accuracy or robustness for either CT/MR or PET/MR registrations. The reason for the similar behavior of MI and NMI was probably the relative large overlapping region of the floating and the target images. It seems that by setting the radius of the hyper sphere $R$, which defines the extent of spatial transformations, to 35mm the overlap was still relatively large. Registrations of PET to MR images were expected to be less accurate and less robust than CT/MR image registrations as the CT images contain more information than PET images. This has been confirmed by both the results in this paper and the results published on the RIRE homepage.

In many papers on image registration, the authors did not explicitly indicate which image
was the floating and which was the target. Besides, different authors used different modalities for the floating (target) image. For instance, Meyer et al. [55] used the CT or MR images as the floating ones and PET or SPECT images as the target (reference) images while Zhu and Cochoff [56] used the SPECT images as the floating ones and the MR images as the target images. Some authors, like Maes et al. [26], have shown that it did matter which image is the floating and which the target. Using the images from the RIRE project image database and mutual information as the similarity measure, they observed that CT to MR registrations had performed clearly worse in terms of accuracy than MR to CT registrations. In terms of robustness, they observed that MR to CT registrations were somewhat more robust than CT to MR registrations. For MR and PET images, they observed that MR/PET registrations where little more accurate than PET/MR registrations. On the other hand, Wells et. al. [25] believed that the assignment of the floating and target volumes in MR to CT and MR to PET registrations should be of little importance. To assess the impact of the floating and target image on the accuracy and robustness indicators of different similarity measures, we have conducted an experiment in which we have exchanged the floating and target image. Results in Figure 5 and Tables 1-4 indicate that some similarity measures behave differently, even if they are mathematically symmetric regarding the floating and target images. The exchange of floating and target image had the highest impact on the robustness (RON) of MR to PET registrations. It seems that when using MI and NMI for MR/PET registrations, much more robust registrations are obtained with MR as the floating and PET as the target image. The difference in behavior can be attributed to the different interpolation conditions, when the floating and the target images are exchanged. These results confirm the observation of Guimond et. al. [57], who non-rigidly registered brain CT, MR and PET images, that for their type of registration the floating image should be the image which has more structure. Generally, the exchange of floating and target images had a small impact on the accuracy of MR/PET and MR/CT registrations and on the robustness of the MR/CT registrations.

5 Conclusion

We hope that the results presented here will help researchers in choosing the right similarity measure, or combination of them, for their registration task. We have compared twelve similarity measures and found that mutual information appears in general to be the most appropriate similarity measure for registration of CT and PET to MR images. We have shown that distortion correction improves the behavior of all similarity measures except SMMI. Finally, the comparison of mutual information to normalized mutual information showed that for this registration task there is very little difference between them.
6 Acknowledgements

The images and »gold standard« transformations were provided as part of the »Retrospective Image Registration Evaluation« project, National Institutes of Health, Project Number 8R01EB002124-03, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN, USA. The authors also wish to thank Ramya Balachandran and Ning Xu from Vanderbilt University, USA for their help with the “gold standard”. This work was supported by the Ministry of Higher Education, Science and Technology, Republic of Slovenia under Grant P2-0232.

7 Literature


