2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

The truth is hard to make: Validation of medical image registration

Josien P.W. Pluim[1,2], Sascha E.A. Muenzing[3], Koen A.J. Eppenhof[1], Keelin Murphy[4]
[1] Eindhoven University of Technology, Eindhoven, The Netherlands
[2] University Medical Center Utrecht, Utrecht, The Netherlands
[3] RWTH Aachen University, Aachen, Germany
[4] University College Cork, Cork, Ireland
j.pluim@tue.nl, muenzing@lfb.rwth-aachen.de, k.a.j.eppenhof@tue.nl, keelinm@gmail.com

***Abstract*— An unsolved problem in medical image analysis is validation of methods. In this paper we will focus on image registration and in particular on nonlinear image registration, which is one of the hardest analysis problems to validate. The paper covers currently used methods of validation, comparative challenges and public datasets, as well as some of our own work in this area.**

Keywords- image registration; validation; evaluation

I. INTRODUCTION

Since some of the very early works in the 1970s (e.g. [1], [2]), the field of medical image registration has progressed tremendously. A plethora of new developments has been proposed for similarity metrics, deformation models, optimization and regularization. One aspect that has not quite developed at the same pace is the *validation* of registration methods. This became obvious when we recently looked back at a review of image registration methods from 20 years ago [3]. Clearly, defining a proper ground truth or even a gold standard is a challenging task, especially for nonlinear registration.

In this paper we will give an overview of validation approaches for medical image registration, followed by a number of public challenges on the topic and freely accessible datasets for validation. We will also describe some of our own contributions to this field.

II. VALIDATION OF MEDICAL IMAGE REGISTRATION

A. Proposed methods

Validation methods can be divided into those that qualify and those that quantify registration error. Visual inspection is an example of the former category, using, for instance, side-by-side viewing, overlays of structures or checkerboard displays [4], [5]. Besides the disadvantage that the evaluation is qualitative and often subjective, these methods cannot be used for large quantities of data. In the remainder we will focus on quantitative approaches, which are preferred.

A second subdivision of validation methods is into those that employ a ground truth and those that use a gold standard. The former contain an exact definition of the correct transformation whereas the latter employ an approximation of it.

There are few methods that can truly provide a ground truth for validation. One is the application of artificial transformations to images [6]. When an image is deformed with a known transformation and the result is registered to the original, the exact error can be computed at every position in the image. Unfortunately, it is extremely hard to define a realistically deformed image, including proper noise and imaging variations. Generally, artificially deformed images are a simplified representation of the true registration problem and they do not yield a reliable evaluation. This was, for instance, demonstrated in an evaluation study of registration methods we conducted [7], in which the difference in performance of the various methods was almost indistinguishable on artificially deformed images (though not on other data).

The other type of validation approach that can potentially define a ground truth is that based on simulated images [8] and phantoms [9]. Physical models of known proportions and known deformations can quantitatively evaluate true registration error, except for possible small deviations as a result of acquisition error. Simulations and phantoms, however, also suffer from lack of realism, making them unsuitable methods of validation for clinical performance of registration methods.

Obviously, the truth is hard to make. All other methods of validation rely on a gold standard rather than a ground truth.

One of the most frequently used gold standard approaches to measuring registration accuracy is the alignment of corresponding anatomical structures after registration. This can be defined by their overlap, by the distance of surfaces or by similar quantitative measures [10]. The quality of such validation depends on the feasibility of defining structures and on the accuracy of the segmentations. If these conditions are met, the overlap measure can be a useful method of validation. However, Rohlfing warns against possible traps in such approaches [11]. He convincingly shows that overlap only takes into account whether corresponding structures are aligned, not how this alignment is achieved. Both in between and within structures large misregistrations are possible, even when the overlap is considered high. Only with relatively small, local structures, can overlap be considered a reliable measure of registration quality.

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

Sometimes other similarity metrics than the one used in the registration method are computed to evaluate the results, e.g. the quality of a registration based on mutual information is evaluated by computation of the sum of squared differences after registration. Such approaches are doomed to fail. If the evaluation metric can accurately measure registration error, the metric in the registration method is not an optimal choice. If on the other hand the registration metric is the superior one, the evaluation by another metric is not valid.

An interesting approach is one based on consistency of transformations. Inverse consistency measures whether the concatenation of the transformations from image *A* to *B* and back from *B* to *A* is equal to the identity transformation [12], [13]. This is a measure of consistency and not of accuracy, because errors in the two individual transformations may cancel out to some degree. Datteri and colleagues showed that when consistency between all possible triplets of a large number of images is calculated, the errors of the separate transformations between the image pairs can be found [14]. The transformation between an image pair is included in several triplets and consequently a sufficient number of equations are defined to solve for all separate transformation errors. This is possible for both rigid and nonlinear transformations [15].

In general, the preferred measure of validation is the distance between corresponding points in the two images. These can be anatomical landmarks or tailor-made markers. Accurate annotation of these points is required, which depends on many factors, such as how the points are indicated, how well they can be imaged (size and contrast) and how they are attached to the patient (in case of markers) [16] [17]. Indicating landmarks for evaluation is generally feasible for rigid transformations, which require only a few points. For nonlinear transformations, on the other hand, the number necessary for a proper evaluation makes this option close to impossible [18].

In recent years, the attention for the topic of uncertainty evaluation has grown. Such approaches attempt to provide a local measure of the reliability of the registration result, for instance, by quantifying the shape of the search space locally [19], by bootstrapping [20] or by a probabilistic approach [21]. Although these methods do not truly measure registration accuracy, they do provide a related evaluation that deserves a mention.

B. *Public challenges*

In the past ten years, so-called *challenges* have become increasingly popular in the field of medical image analysis. The core idea of a challenge is that researchers apply their methods to the same data (made available by the organizers), but that they are blinded to the gold standard. Results are to be submitted to the challenge organizers who evaluate them and return the outcomes. This is an excellent way to compare methods on the same data using the same measures, in a fair and blinded manner.

Many of the challenges are collected on the Grand Challenges site: http://grand-challenge.org/. The number of challenges per year shows how fast this concept has caught on. It is also striking how few of the challenges are on image registration, most likely because of the difficulty in creating a gold standard for this type of problem. Interestingly, the very first challenge in medical imaging, dating back as far as 1996, was one on image registration. The *Retrospective Registration Evaluation Project* by Vanderbilt University challenged researchers to rigidly register multimodal brain images [22]. It ran until very recently, but the data is now freely available (via http://www.insight-journal.org/rire/). Since then, two projects on nonlinear registration were initiated. The first is the Non-rigid Image Registration Evaluation Project (NIREP) by the University of Iowa, to be found at http://www.nirep.org/ [23]. It is not truly a challenge, but it provides a platform for sharing databases and metrics for evaluation. The most recent one is the Evaluation of Methods for Pulmonary Image Registration (EMPIRE10) challenge by the University Medical Center Utrecht: http://empire10.isi.uu.nl/ [7].

C. *Available datasets*

Besides the challenges that offer data and an evaluation mechanism for the registration results, there are also a number of public datasets with gold standard available. The ones below are created for image registration:

- Popi-model (Université Lyon 1) [24]: 4D lung CT, with manually identified landmarks, https://www.creatis.insa-lyon.fr/rio/popi-model
- Dir-lab (the University of Texas Medical Branch) [18]: lung CT, both 4D and inspiration-expiration, with manually identified landmarks, http://www.dir-lab.com/
- Vienna registration phantom (Medical University Vienna) [25]: CT, MRI and x-rays of a porcine head, with marker-based gold standard, http://midas3.kitware.com/midas/community/3
- In addition, datasets with segmented structures exist. Though these were usually created for evaluation of segmentation rather than registration, some can nonetheless be employed for the latter (bearing in mind the work of Rohlfing), see for example the evaluation study by Klein et al. and the datasets therein [26].

III. CONTRIBUTIONS TO VALIDATION OF IMAGE REGISTRATION ACCURACY

Next we treat a few examples of our own work in the area of validation of image registration accuracy.

A. *Semi-automatic construction of reference standards for evaluation*

Corresponding pairs of landmarks are considered one of the most reliable ways of quantifying registration accuracy. Obtaining a sufficiently large number of them to evaluate nonlinear registration of 3D medical data is a daunting task. We proposed a framework that involves the human in a limited way, making the creation of large sets feasible [27], [28].

The basic principle of the framework consists of three phases:

1. Landmarks are automatically defined in one of the two images to be registered,
2. A human observer starts annotations of the corresponding landmarks in the other image, while the computer simultaneously makes estimates,
3. The computer estimates are of sufficient quality for the computer to complete the annotations autonomously.

In the first phase, points p are selected that would be suitable for annotation. This selection is based on the points' *distinctiveness value D(p),* which combines a number of image characteristics. First of all, good landmarks are assumed to lie on edges of structures and the gradient magnitude of the image at a landmark should therefore be high. Secondly, landmarks are expected to stand out from their surroundings. This is measured by the difference between the intensities in a region of interest (ROI) around point p and a set of neighbouring points q_i:

$$Diff(ROI(p), ROI(q_i)) = \frac{1}{N}\sum_{k=1}^{N}\left|ROI(p)_k - ROI(q_i)_k\right|$$

with N the number of voxels in the ROI and q_i a neighbouring point on a circle around p with user-defined radius, see Figure 1. The distinctiveness value of a point p, $D(p)$, is defined

$$D(p) = \frac{G(p)}{\max G}\frac{1}{m}\sum_{i=1}^{m} Diff(ROI(p), ROI(q_i))$$

with $G(p)$ the gradient magnitude at p, normalized by $maxG$, the maximum gradient magnitude of all points considered. Subsequently, n landmarks are selected, based on their distinctiveness value and given some restrictions to have a reasonable spread of landmarks across the image volume.

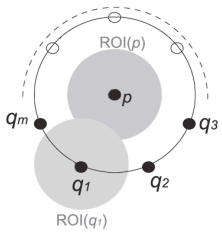

Figure 1 Definition of a region of interest (grey areas) around a point p and its neighbouring points q_i

In the second phase, an observer is shown individual landmarks in one image and asked to indicate the corresponding point in the other image. Examples of landmarks are shown in Figure 2. A bespoke user interface for this process was created. Alongside the user, the computer estimates corresponding points in the following manner: a thin-plate spline deformation through the current human-defined landmark pairs is computed, the position of the corresponding point is extracted and subsequently refined by finding the optimal position in a small neighbourhood using block-matching. Every computer estimate is compared with the point selected by the observer. Once x out of the previous y estimates are within a distance d of the observer-selected point (with x, y and d user-defined variables), the computer system is deemed capable of performing the annotations for the remaining landmarks (phase 3). When in doubt, e.g. the block matching fails, the system can pass a landmark back to the observer to indicate the corresponding point.

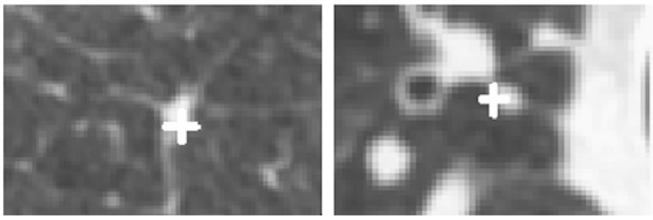

Figure 2 Examples of automatically selected landmarks in lung CT.

The system for reference standard creation was demonstrated on low-dose lung CT data (baseline and follow-up) of 47 subjects, with the aid of three observers (one expert, two medical students). We set the system to select 300 landmarks per image. The first estimates of corresponding points by the computer were poor, as is to be expected. However, the distance of the automatically estimated corresponding points to the observer annotations fell rapidly and generally was smaller than 1 mm after about 15 landmarks (see Figure 3). Moreover, the study showed that for evaluation of registration methods, manually and automatically defined landmark pairs are equally suitable. Furthermore, we demonstrated that the annotations of non-expert observers with this system do not differ significantly from those of an expert. All in all, the system produces a large number of high-quality landmark pairs for evaluation of image registration with minimal user effort. The system is freely available from http://isimatch.isi.uu.nl/.

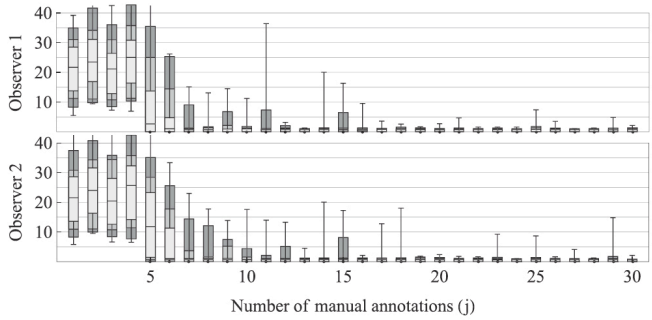

Figure 3 Box-and-whisker plots of the distances between automatic and manual annotations for all image pairs against the number of manual annotations performed thus far. The top and bottom plot show the result for two different observers.

B. Automatic quality assessment of image registration

Ultimately, evaluation of registration accuracy results in a map indicating the quality of the registration locally throughout the image volume. We have developed an approach that learns local registration error from prior data in a supervised manner [29]. To learn local registration quality, a

reference set of alignment patterns was created. Using the previously described landmark method, sets of 30 corresponding point pairs were defined on lung CT images of 51 subjects. For each subject a baseline and follow-up scan were available, acquired 3-15 months apart. The image pairs were registered using a number of different transformation models, from rigid to fully deformable. Using the known correspondences of landmark pairs, we could compute for each landmark and each registration result the error of that registration at that position. We deliberately included transformation models that would not be able to register the images perfectly, to obtain training examples of both good and poor alignment.

The classification problem was initially trained to distinguish three categories of registration error: Correct Alignment (CA, a registration error at that point of less than 2mm), Poor Alignment (PA, 2-5mm) and Wrong Alignment (WA, > 5mm). In order for the classifier to learn the characteristics of registration results of different quality, features were computed on subvolumes around the registered landmarks. These features included intensity-based features, such as intensity differences, correlation and entropy, as well as features on the deformation field (e.g. Jacobian determinant). Features were computed at various scales, leading to a total of 66 features. Classification was performed in a two-stage manner: first a distinction between correct and incorrect alignment (CA vs. ¬CA) was made, followed by a further subdivision of the latter class into poor and wrong alignment (PA vs. WA). The system was trained using cross-validation. Feature selection and various classifiers were included to find an optimal approach.

Evaluation of the system was performed on a separate set of corresponding landmarks, not used for training. Accuracy of classification is given in Table 1. The results are good for the extreme classes, but poorer for the middle class. Many of the misclassifications occur at the boundaries between classes.

Table 1 Confusion matrix of estimated versus true registration error class

True subsets	Estimated subsets						Totals
	CA′		PA′		WA′		
CA	2708	(97%)	73	(3%)	0	(0%)	2781
PA	196	(23%)	570	(67%)	81	(10%)	847
WA	0	(0%)	55	(11%)	439	(89%)	494
Totals	2904		698		520		4122

Once the quality of a registration result can be determined, it cannot only be used to inform the user, but also to improve that result. We developed approaches for boosting registration algorithms based on the theory of hypothesis boosting. We consider the displacement field u_n of registration n the hypothesis h_n. In order to boost a registration algorithm similarly to classifier boosting, we require an estimate of local registration accuracy. Accordingly, each hypothesis u_n is weighted by a weight based on the local error estimate. One approach to boosting is to iteratively employ a registration method and adaptively focus the registration on remaining errors [30]. In this manner we aim to reduce non-systematic registration errors, thereby obtaining more robust registrations and overall improved registration results. We validated the approach on three different deformable registration algorithms (ANTS gSyN, NiftyReg, and DROP) on three independent reference datasets of pulmonary intra-subject images. It consistently and significantly reduced registration errors yielding an improvement of the registration accuracy by about 5%–30% depending on the dataset and the registration algorithm employed.

A second boosting approach combined deformation fields from a number of different registration methods based on estimated local error [31]. After registration of an image pair with a number of registration methods, the results were classified per small regions using the automatic error estimation scheme and the boosted deformation field consisted of the combination of the best result per region.

The classification was later replaced by a regression scheme, which produces a scalar value of local registration error [32]. These values were used as weights in the boosting framework. The various registration results were combined locally by weighing them with the inverse of their estimated registration error. A number of state-of-the-art registration methods for lung CT data were employed (ANTS, NiftyReg and Elastix). Table 2 shows that the boosting results based on the automatic quality assessment outperform each of the separate methods, as demonstrated on the data of the EMPIRE10 challenge.

Table 2 Results of the boosting approach and the separate registration methods on the EMPIRE10 challenge data

DIR	Boundary [%]		Fissure [%]		Landmark [mm]		Singularity [%]		Final Rank
	avg	max	avg	max	avg	max	avg	max	
ANTS-NiftyReg-Elastix	1.E-03	1.E-02	0.18	1.98	0.64	1.38	0	0	3
ANTS	1.E-02	2.E-01	0.12	1.57	0.75	1.99	6.E-06	2.E-04	5
NiftyReg-Elastix	8.E-04	9.E-03	0.33	3.27	0.73	1.82	0	0	6
NiftyReg	2.E-02	5.E-01	0.38	4.00	0.90	3.19	0	0	11
Elastix	9.E-04	5.E-03	0.46	4.57	0.82	2.23	1.E-02	2.E-01	13

C. Supervised local error estimation using convolutional neural networks

Training systems to estimate registration error locally can be based on classifiers (as above) or on convolutional neural networks. We describe here a proof of principle and first results of registration validation using CNNs.

To estimate the error at any position in the images, a patch of 32 by 32 pixels around that position is defined in the registered images. These two patches are fed to the CNN, which returns the norm of the local deformation. The network is a rather standard setup of two convolutional layers with 3x3 kernels followed by a 2x2 pooling layer. This combination is repeated one time and completed by three fully connected layers, see Figure 4.

The method is applied to 2D Digital Subtraction Angiography images of the head and neck. Training data were created by randomly deforming images. The original images and their deformed versions were used as input pairs with

known deformations to train the CNN to estimate local deformation. Data augmentation was included to achieve a feasible number of input pairs.

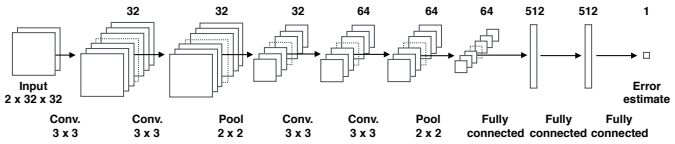

Figure 4 Schematic overview of convolutional neural network

Validation was on another set of DSA image pairs, of 11 patients. These consisted of genuine pre- and post-contrast images, for which the true deformation is unknown. To create a gold standard, image pairs were nonlinearly registered in a multiresolution approach of four levels, which we assumed would yield a good alignment. To validate the neural network, we used 'misregistered' pairs. These were formed by the registered images after three levels, i.e. excluding the finest resolution. For evaluation, the error maps produced by the neural network were compared with the deformation field of the four-resolution registration (our gold standard). Figure 5 contains an example of the estimated and gold standard error.

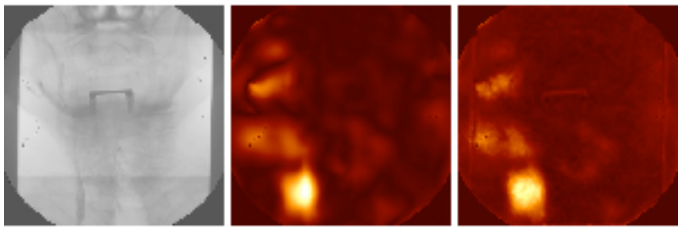

Figure 5 Example of, from left to right, registered DSA image, gold standard error map, estimated error map

Figure 6 shows the correlation of the estimated and gold standard errors for all pixels in the 11 image pairs. Good correlation between true and estimated errors is found, up to roughly eight pixels. Larger errors are underestimated, because errors of that size are not included in the training set.

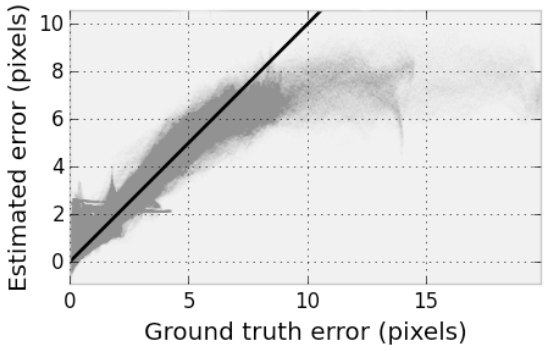

Figure 6 Estimated error (vertical axis) versus gold standard error (horizontal axis)

The current system is being expanded to 3D images and evaluated more extensively.

IV. FUTURE

Validation of nonlinear image registration is a complicated problem. Techniques for validation do exist, but they all have their shortcomings. Overall, it should be considered an unsolved problem, yet proper validation is essential for translation of automatic registration methods into clinical practice.

Further progress is sorely needed, on the one hand, in the area of validation measures and frameworks for automatic quantification of error. Future developments in pattern recognition, classification and deep learning may play a vital role.

On the other hand, feasible solutions for obtaining reference standards (whether ground truth or gold standard) are lacking. A potential approach that has been gaining attention for medical image applications, is crowd-sourcing. The idea originated as far back as 1907, when sir Francis Galton published some remarkable findings in Nature [33]. He had witnessed a competition to guess the weight of an ox at a local market, a quite famous anecdote by now. Nearly 800 visitors, from complete laymen to farmers and butchers, made an estimate of the weight. Not a single person predicted the correct amount, yet the median of all their estimates was extremely close to the truth: 1207 vs. 1198 lb. (note: the mean turned out to be even closer to the truth 1197 lb., but Galton firmly believed the median to be the correct measure to use, as he explained in a letter to the editor three weeks later [34]). Would it be possible to use large numbers of non-experts to produce reliable gold standards for medical imaging problems? Some first studies into this question indicate that it may very well be a potential solution, although it is currently not clear what the requirements for such approaches are and what type of clinical problems they would be suitable for. Nguyen *et al.* employed the general public to classify potential polyps in CT images as actual or false polyps. With minimal training, laymen achieved a performance similar to an automatic CAD system [35]. Maier-Hein and colleagues [36] enlisted the crowd to segment instruments in endoscopy images and found the quality of the results to be similar to those of experts. Games are now being developed to entice the public to contribute to our field in a fun, yet valuable way (see, for instance, http://biogames.ee.ucla.edu/ [37], [38]).

Still, such approaches do not solve one dilemma: whether or not the human should be the standard. Is it not the aim of many of the automatic solutions to improve upon human performance .... ?

V. REFERENCES

[1] D. J. Dowsett and B. J. Perry, "A comparative statistical analysis of brain scans using a digital computer," *Brit J Radiol*, vol. 43, pp. 617–628, 1970.

[2] J. H. Kinsey and B. D. Vannelli, "Application of digital image change detection to diagnosis and follow-up of cancer involving the lungs," presented

at the SPIE Application of Optical Instrumentation in Medicine IV, 1975, vol. 70, pp. 99–112.

[3] M. A. Viergever, J. Maintz, S. Klein, and K. Murphy, "A survey of medical image registration–under review," *Med Image Anal*, vol. 33, pp. 140–144, 2016.

[4] J. M. Fitzpatrick, D. L. G. Hill, Y. Shyr, J. West, C. Studholme, and J. C R Maurer, "Visual assessment of the accuracy of retrospective registration of MR and CT images of the brain," *IEEE Trans Med Imaging*, vol. 17, no. 4, pp. 571–585, 1998.

[5] E. R. E. Denton, L. I. Sonoda, D. Rueckert, S. C. Rankin, C. Hayes, M. O. Leach, D. L. G. Hill, and D. J. Hawkes, "Comparison and evaluation of rigid, affine, and nonrigid registration of breast MR images," *J Comput Assist Tomo*, vol. 23, no. 5, 1999.

[6] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of nonrigid image registration using finite-element methods: application to breast MR images," *IEEE Trans Med Imaging*, vol. 22, no. 2, pp. 238–247, 2003.

[7] K. Murphy, B. van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. L. Cao, K. F. Du, G. E. Christensen, V. Garcia, T. Vercauteren, N. Ayache, O. Commowick, G. Malandain, B. Glocker, N. Paragios, N. Navab, V. Gorbunova, J. Sporring, M. de Bruijne, X. Han, M. P. Heinrich, J. A. Schnabel, M. Jenkinson, C. Lorenz, M. Modat, J. R. McClelland, S. Ourselin, S. E. A. Muenzing, M. A. Viergever, D. De Nigris, D. L. Collins, T. Arbel, M. Peroni, R. Li, G. C. Sharp, A. Schmidt-Richberg, J. Ehrhardt, R. Werner, D. Smeets, D. Loeckx, G. Song, N. Tustison, B. Avants, J. C. Gee, M. Staring, S. Klein, B. C. Stoel, M. Urschler, M. Werlberger, J. Vandemeulebroucke, S. Rit, D. Sarrut, and J. P. W. Pluim, "Evaluation of registration methods on thoracic CT: The EMPIRE10 Challenge," *IEEE Trans Med Imaging*, vol. 30, no. 11, pp. 1901–1920, 2011.

[8] K. Rohr, M. Fornefett, and H. S. Stiehl, "Spline-based elastic image registration: integration of landmark errors and orientation attributes," *Computer Vision and Image Understanding*, vol. 90, no. 2, pp. 153–168, 2003.

[9] A. Roche, X. Pennec, G. Malandain, and N. Ayache, "Rigid registration of 3-D ultrasound with MR images: a new approach combining intensity and gradient information," *IEEE Trans Med Imaging*, vol. 20, no. 10, pp. 1038–1049, 2001.

[10] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans Med Imaging*, vol. 25, no. 11, pp. 1451–1461, 2006.

[11] T. Rohlfing, "Image smilarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable," *IEEE Trans Med Imaging*, vol. 31, no. 2, pp. 153–163, 2012.

[12] M. Holden, D. L. G. Hill, E. R. E. Denton, J. M. Jarosz, T. C. S. Cox, T. Rohlfing, J. Goodey, and D. J. Hawkes, "Voxel similarity measures for 3-D serial MR brain image registration," *IEEE Trans Med Imaging*, vol. 19, no. 2, pp. 94–102, 2000.

[13] T. Netsch, P. Rsch, J. Weese, A. van Muiswinkel, and P. Desmedt, "Grey value-based 3-D registration of functional MRI time-series: comparison of interpolation order and similarity measure," presented at the Medical Image Computing and Computer Assisted Intervention - MICCAI, 2000, vol. 3979, pp. 1148–1159.

[14] R. Datteri and B. M. Dawant, "Estimation of rigid-body registration quality using registration networks," presented at the SPIE Medical Imaging, 2012, vol. 8314, pp. 831419–831419–12.

[15] R. D. Datteri, Y. Liu, P. F. D'Haese, and B. M. Dawant, "Validation of a nonrigid registration error detection algorithm using clinical MRI brain data," *IEEE Trans Med Imaging*, vol. 34, no. 1, pp. 86–96, 2015.

[16] J. M. Fitzpatrick, D. L. G. Hill, and J. C R Maurer, "Handbook of Medical Imaging, volume 2. Medical Image Processing and Image Analysis," in *Image Registration*, no. 8, J. M. Fitzpatrick and S. Milan, Eds. SPIE Press, 2000.

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

[18] R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero, "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets," *Phys Med Biol*, vol. 54, no. 7, pp. 1849–1870, 2009.

[19] G. Saygili, M. Staring, and E. A. Hendriks, "Confidence estimation for medical image registration based on stereo confidences," *IEEE T Med Imaging*, vol. 35, no. 2, pp. 539–549, 2016.

[20] J. Kybic, "Bootstrap resampling for image registration uncertainty estimation without ground truth," *IEEE T Image Process*, vol. 19, no. 1, pp. 64–73, Jan. 2010.

[21] I. J. A. Simpson, J. A. Schnabel, A. R. Groves, J. L. R. Andersson, and M. W. Woolrich, "Probabilistic inference of regularisation in non-rigid registration," *NeuroImage*, vol. 59, no. 3, pp. 2438–2451, 2012.

[22] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, J. C R Maurer, 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, J. G Q Maguire, M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J Comput Assist Tomo*, vol. 21, no. 4, pp. 554–566, 1997.

[23] G. E. Christensen, X. Geng, J. G. Kuhl, J. Bruss, T. J. Grabowski, I. A. Pirwani, M. W. Vannier, J. S. Allen, and H. Damasio, "Introduction to the Non-rigid Image Registration Evaluation Project (NIREP)," presented at the International Workshop on Biomedical Image Registration, Berlin, Heidelberg, 2006, pp. 128–135.

[24] J. Vandemeulebroucke, S. Rit, J. Kybic, P. Clarysse, and D. Sarrut, "Spatiotemporal motion estimation for respiratory-correlated imaging of the lungs," *Med Phys*, vol. 38, no. 1, pp. 166–178, 2011.

[25] S. A. Pawiro, P. Markelj, F. Pernus, C. Gendrin, M. Figl, C. Weber, F. Kainberger, I. Nobauer-Huhmann, H. Bergmeister, M. Stock, D. Georg, H. Bergmann, and W. Birkfellner, "Validation for 2D/3D registration I: A new gold standard data set," *Med Phys*, vol. 38, no. 3, pp. 1481–1490, 2011.

[26] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.

[27] K. Murphy, B. van Ginneken, S. Klein, M. Staring, B. J. de Hoop, M. A. Viergever, and J. P. W. Pluim, "Semi-automatic construction of reference standards for evaluation of image registration," *Med Image Anal*, vol. 15, no. 1, pp. 71–84, 2011.

[28] S. Kabus, T. Klinder, K. Murphy, B. van Ginneken, C. Lorenz, and J. P. W. Pluim, "Evaluation of 4D-CT lung registration," presented at the Medical Image Computing and Computer Assisted Intervention, 2009, vol. 5761, pp. 747–754.

[29] S. E. A. Muenzing, B. van Ginneken, K. Murphy, and J. P. W. Pluim, "Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration," *Med Image Anal*, vol. 16, no. 8, pp. 1521–1531, 2012.

[30] S. E. A. Muenzing, B. van Ginneken, M. A. Viergever, and J. P. W. Pluim, "DIRBoost–An algorithm for boosting deformable image registration: Application to lung CT intra-subject registration," *Med Image Anal*, vol. 18, no. 3, pp. 449–459, 2014.

[31] S. E. A. Muenzing, B. van Ginneken, and J. P. W. Pluim, "On combining algorithms for deformable image registration," presented at the International Workshop on Biomedical Image Registration, Berlin, Heidelberg, 2012, pp. 256–265.

[32] S. E. A. Muenzing, "Learning-based approaches to deformable image registration," Utrecht University, 2014.

[33] F. Galton, "Vox populi," *Nature*, vol. 75, p. 450, 1907.

[34] F. Galton, "The ballot box," *Nature*, vol. 75, pp. 509–510, 1907.

[35] T. B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J. E. Burns, and R. M. Summers, "Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography," *Radiology*, vol. 262, no. 3, pp. 824–833, Mar. 2012.

[36] L. Maier-Hein, S. Mersmann, D. Kondermann, S. Bodenstedt, A. Sanchez, C. Stock, H. G. Kenngott, M. Eisenmann, and S. Speidel, "Can masses of non-experts train highly accurate image classifiers?," presented at the Medical Image Computing and Computer Assisted Intervention, 2014, vol. 8674, pp. 438–445.

[37] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, and A. Ozcan, "Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study," *PLoS ONE*, vol. 7, no. 5, p. e37245, 2012.

[38] S. Mavandadi, S. Feng, F. Yu, S. Dimitrov, R. Yu, and A. Ozcan, "BioGames: A platform for crowd-sourced biomedical image analysis and telediagnosis," *Games for Health Journal*, vol. 1, no. 5, pp. 373–376, Oct. 2012.