Accuracy Estimation for Medical Image Registration Using Regression Forests

Hessam Sokooti[1(✉)], Gorkem Saygili[1], Ben Glocker[2], Boudewijn P. F. Lelieveldt[1,3], and Marius Staring[1,3]

[1] Leiden University Medical Center, Leiden, The Netherlands
h.sokooti_oskooyi@lumc.nl
[2] Imperial College, London, UK
[3] Delft University of Technology, Delft, The Netherlands

Abstract. This paper reports a new automatic algorithm to estimate the misregistration in a quantitative manner. A random regression forest is constructed, predicting the local registration error. The forest is built using local and modality independent features related to the registration precision, the transformation model and intensity-based similarity after registration. The forest is trained and tested using manually annotated corresponding points between pairs of chest CT scans. The results show that the mean absolute error of regression is 0.72 ± 0.96 mm and the accuracy of classification in three classes (correct, poor and wrong registration) is 93.4 %, comparing favorably to a competing method. In conclusion, a method was proposed that for the first time shows the feasibility of automatic registration assessment by means of regression, and promising results were obtained.

Keywords: Image registration · Registration accuracy · Uncertainty estimation · Regression forests

1 Introduction

Most image registration methods do not provide insights about the quality of their results and devolve this difficult task to human experts, which is very time-consuming. Automatic evaluation of registration reduces the time of manual assessment and can provide information about the registration uncertainty. Having the error of registration is useful to refine the registration, either automatically or with the feedback of human experts. Even if refinement is not possible, information about the registration quality can help decide if subsequent processing is meaningful, and visualizing the error can be helpful in medical applications before making a clinical decision.

Several methods have been suggested to estimate the registration accuracy, such as exploitation of the Bayesian posterior distribution [1] or based on the consistency of multiple registrations [2]. In the stochastic approaches Kybic [3] computed the registration uncertainty by performing multiple registrations with

S. Ourselin et al. (Eds.): MICCAI 2016, Part III, LNCS 9902, pp. 107–115, 2016.
DOI: 10.1007/978-3-319-46726-9_13

bootstrapping on the cost function samples to generate a set of registration solutions. He found a correlation between the variation of the 2D translational parameters and the true registration error but the method is not tested for 3D non-rigid registration with much more transform parameters. Hub *et al.* [4] estimated the uncertainty by perturbing the B-spline grid with random values and check whether or not the local SSD changed. The drawback of this approach is that it is not efficient in homogeneous areas. In 2013, they applied the same perturbation for the Demons algorithm and showed that the variance of the deformation vector field (DVF) is related to the registration error [5]. However, an exhaustive experiment is needed to find large registration errors.

In this paper we turn our attention to methods capable of learning the registration error. This has the advantage that multiple features related to registration uncertainty can be exploited and combined in a single framework. Muenzing *et al.* [6] classified the registration quality into three categories (wrong, poor and correct), and reported that it was not possible to successfully build a regressor. All their features were intensity-based, except for the Jacobian of the transform parameters. In this paper, instead of formulating uncertainty estimation as a classification problem, we formulate it as a regression problem, enabling a continuous prediction of registration accuracy. To the best of our knowledge, there is only one paper that takes a similar approach [7], but it was only tested on synthetically deformed images. We explore several modality independent features (some of them new) related to registration precision, the estimated transformation and the image similarity after registration, and their contribution to the regression performance. The proposed framework can be used in combination with any registration paradigm, i.e. does not depend on specifics such as a Bayesian formulation, and can already be used for pairwise registration.

2 Methods

2.1 System Overview

A block diagram of the proposed algorithm is shown in Fig. 1. The inputs of the system are a fixed I_F and a moving image I_M. We use a limited number of so-called *mother features*, from which much more features are generated using a pooling technique. A regression forest (RF) is then trained from the feature pool to predict local registration error. One class of features is derived from the registration, or from a set of sub-registrations. The other class is derived from the intensities of the fixed and deformed moving images. Details are given in Sect. 2.2.

Mathematically, the registration problems is formulated as an optimization problem in which the cost function $\mathcal{C}$ is minimized with respect to $\boldsymbol{T}$:

$$\widehat{\boldsymbol{T}} = \arg\min_{\boldsymbol{T}} \mathcal{C}\big(\boldsymbol{T}; I_F, I_M\big), \tag{1}$$

where $\boldsymbol{T}$ denotes the transformation. The minimization is solved by an iterative optimization method embedded in a multi-resolution setting. A registration can be initialized by an initial transform $\boldsymbol{T}^{\text{ini}}$.

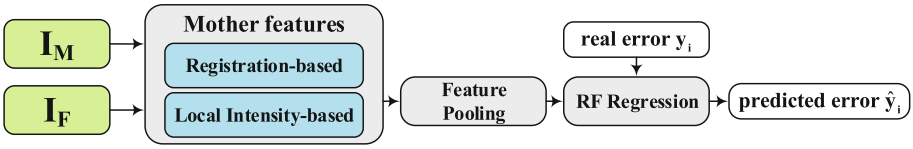


Fig. 1. A block diagram of the proposed algorithm.

2.2 Features and Pooling

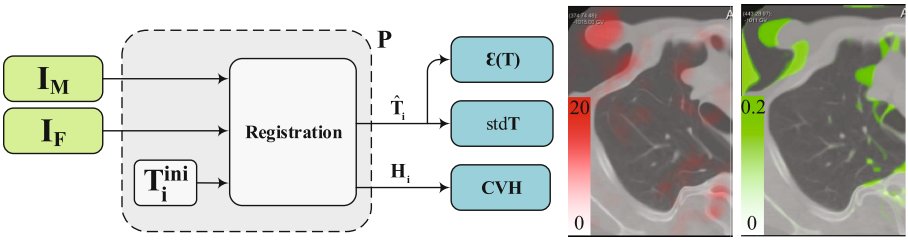
Variation of deformation vector field (std $\boldsymbol{T}$**)** The initial parameters of an optimization problem can affect the final solution for many registration paradigms, especially in areas where the cost function has multiple local minima or is semi-flat. On the other hand, in cases where the cost function is well-defined, variations in the initial transformation are expected to have much less effect on the final registration result. The variation in the final transformation result is then an intuitive measure for the local registration uncertainty, which is a surrogate for the correctness or at least the precision of the registration. A flow chart of the described feature is given in Fig. 2(a). Consider P randomly generated transformations $\boldsymbol{T}_i^{\text{ini}}$ that are used as initializations of the registration algorithm from Eq. (1), resulting in P final transformations $\widehat{\boldsymbol{T}}_i$. The standard deviation of those transformations std $\boldsymbol{T}$ is then used as a mother feature:

$$\overline{\boldsymbol{T}} = \tfrac{1}{P}\sum \widehat{\boldsymbol{T}}_i, \qquad \text{std}\,\boldsymbol{T} = \tfrac{1}{P}\sqrt{\sum \|\widehat{\boldsymbol{T}}_i - \overline{\boldsymbol{T}}\|^2}. \tag{2}$$

The random initializations are generated in this work by adding a uniformly distributed offset to the B-spline coefficients. An example of std $\boldsymbol{T}$ in a manually deformed image is available in Fig. 2(b), for illustration purposes we magnified the imposed deformation field. It is also possible to first perform a registration, resulting in a transformation $\boldsymbol{T}^{\text{base}}$, and then add random offsets to that ($\boldsymbol{T}^{\text{base}} + \boldsymbol{T}_i^{\text{offset}}$), which is approximately similar to Hub's work [5]. Akin to Eq. (2) a mother feature std $\boldsymbol{T}^{\text{Hub}}$ is then derived.

Areas with a small std $\boldsymbol{T}$ are still potentially areas of low registration quality, if the difference between $\overline{\boldsymbol{T}}$ and $\boldsymbol{T}^{\text{base}}$ is too large. We then consider the bias $\mathcal{E}(\boldsymbol{T})$ as a complementary feature to std $\boldsymbol{T}$ computed by $\mathcal{E}(\boldsymbol{T}) = \|\boldsymbol{T}^{\text{base}} - \overline{\boldsymbol{T}}\|$. The mother feature $\mathcal{E}(\boldsymbol{T}^{\text{Hub}})$ is computed similarly.

Coefficient of variation of joint histograms (CVH): Based on the multiple registration results we can additionally extract information about the matched intensity patterns of the images. The first step is to calculate the joint histograms $\text{H}_i, \forall i$ of the fixed image I_F and the deformed moving image $I_M(\boldsymbol{T})$. A large variation in the joint histograms implies a large registration error. The scalar CVH is defined as: CVH $=$ std H$/(\overline{\text{H}} + \epsilon)$. The coefficient of variation is used to compensate for large differences between the elements of $\overline{\text{H}}$, and the constant ϵ is used to ignore small numbers in the joint histogram. Note that this feature



(a) Constructing registration-based features (b) std **T** (c) CVH

Fig. 2. The registration-based features require multiple registrations

can also be used in a multi-modality setting, like all our features. An example of the CVH on a manually deformed image is shown in Fig. 2(c).

Determinant of Jacobian (Jac): In addition to previous registration-based features, we also use Jac. Local changes in volume can point to poor registration quality or discontinuous transformations.

Difference of MIND: Heinrich *et al.* [8] introduced the Modality Independent Neighborhood Descriptor (MIND) to register multimodal images by comparing similarities between same patches in the fixed and moving image. The output of this local self-similarity has n features for each voxel, where n is the size of the search region. The n features is aggregated in a single mother feature by using the Euclidean distance between MIND of I_F and that of $I_M(\boldsymbol{T})$. We calculate MIND with two different search regions, see Sect. 3 for details.

Feature pooling: All features are calculated in a voxel-based fashion. Incorporating local information of each feature can reduce discontinuity and improve interaction with other features. For instance, it is possible to have a high std $\boldsymbol{T}$ in homogeneous regions while the difference of MIND is almost zero. On the other hand, when we have misregistration on the boundaries, the difference of MIND indicates high dissimilarity while std $\boldsymbol{T}$ can have a high value only in the nearby voxels but not exactly on the border. To overcome these problems, the total set of features is largely increased by generating a pool from the mother features by calculating averages and maxima over them using differently sized boxes.

2.3 Regression Forests

Breiman [9] introduced the random forest by extending bagging and making more clever averaging of trees. The general idea is to use some weak learners (trees) and make an efficient combination of them. In contrast to bagging, splitting of each node is done with a random subset of features which speeds up the training phase and reduces correlation between trees in the forest, accordingly decreasing the forest error rate. The reason that we chose the random forest is that it has the ability to handle data without preprocessing such as rescaling data, removing outliers and selecting features. Feature importance is measured

over the out-of-bootstrap samples Ω by permuting features, and computing the difference between the mean square error (MSE) before and after permutation:

$$\mathrm{Imp}(x_i) = \frac{1}{N_t} \sum_{t=1}^{N_t} \left(\underset{j \in \Omega}{\mathrm{MSE}}\left(\hat{y}_{\pi_i j}, y_j\right) - \underset{j \in \Omega}{\mathrm{MSE}}\left(\hat{y}_j, y_j\right) \right), \tag{3}$$

where y_j is the real value, $\hat{y_j}$ the predicted value after the regression, $\hat{y}_{\pi_{ij}}$ the predicted value when permuting feature i, and N_t the number of trees.

3 Experiments and Results

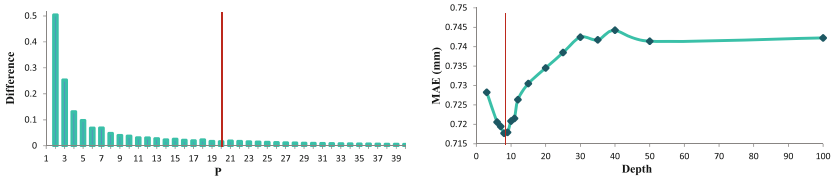
Materials and ground truth: In this study, the SPREAD database [10] has been used, which has 21 pairs of 3D lung CT images. The dimension of the images is about $446 \times 315 \times 129$ with an average voxel size of $0.781 \times 0.781 \times 2.5$ mm. Patients are within the range of 49 to 78 years old and for each patient a baseline image and a follow-up image (after 30 months) are available in which 100 well-distributed corresponding landmarks are selected semi-automatically on distinctive locations [11]. The residual Euclidean distance after registration between the corresponding points can be seen as the accuracy of the registration.

However, 100 training samples for each pair are not enough to reliably train the regression forest. To obtain more training samples, we include voxels in a small local neighborhood of the annotated points. We assume that the registration error is equal to the error at the landmark, which seems reasonable for smooth transformations and within a small region. The neighborhood size is chosen as $10.153 \times 10.153 \times 7.5$ mm, which is approximately equivalent to the final grid space of the B-spline registration.

The main programming language is MATLAB 2015a, while feature pooling is implemented in C++ and the regression forest is computed using the scikit-learn package of Python. All registrations are performed by `elastix` [12].

Evaluation and experimental setup: To evaluate the proposed algorithm, the mean absolute error (MAE) between the real registration error y_i and estimated one $\hat{y}_i$ is calculated by $\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$. We also reported the $\mathrm{MAE_i}$ in three bins with respect to y_i $[0, 3)$, $[3, 6)$ and $[6, \infty)$ mm, corresponding to correct, poor and wrong registration [6]. It is possible to classify the $\hat{y}_i$ based on these bins and calculate the total accuracy (Acc) and accuracy in each bin ($\mathrm{Acc_i}$). We employ k-fold cross validation, using $k = 10$, splitting the data in 15 image pairs for training and the remaining 6 pairs for testing.

Parameters of features and pooling: The feature $\mathrm{std}\,\boldsymbol{T}$ is computed using $P = 20$ initializations $\boldsymbol{T}_i^{\mathrm{ini}}$, which are constructed randomly using a uniform distribution in the range $[-2, 2]$ mm. P was chosen sufficiently large, such that the overall standard deviation of the resulting transformations did not change considerably, as shown in Fig. 3(a). For the registrations, we used three resolutions of 500 iterations each, with a final B-spline grid spacing of $[10, 10, 10]$ mm. The cost function is mutual information, which is optimized by adaptive stochastic



(a) The number of registrations P (b) The tree depth D

Fig. 3. Tuning some of the parameters. The selected ones are indicated by red. (Color figure online)

Table 1. Regression results for the several feature pools

	MAE	MAE_1	MAE_2	MAE_3	Acc	Acc_1	Acc_2	Acc_3
std $\boldsymbol{T}$	0.76 ± 1.03	0.56 ± 0.47	2.29 ± 1.31	4.29 ± 2.84	93.0	95.1	33.2	65.8
std $\boldsymbol{T}^{\text{Hub}}$	0.84 ± 1.25	0.59 ± 0.53	2.15 ± 1.14	6.28 ± 2.61	91.5	94.0	29.4	54.1
CVH	0.90 ± 1.42	0.61 ± 0.67	2.42 ± 0.94	7.05 ± 2.81	90.8	92.5	21.0	29.4
$MIND_{sp}$	0.73 ± 1.05	0.54 ± 0.50	1.81 ± 0.99	4.83 ± 2.67	93.0	95.6	40.0	66.0
$MIND_3$	0.74 ± 1.06	0.53 ± 0.43	2.08 ± 1.20	4.83 ± 2.78	93.0	95.5	36.2	62.9
$\mathcal{E}(\boldsymbol{T})$	0.85 ± 1.25	0.63 ± 0.70	2.13 ± 0.98	5.36 ± 3.08	91.4	94.3	27.7	48.9
$\mathcal{E}(\boldsymbol{T}^{\text{Hub}})$	0.82 ± 1.17	0.58 ± 0.47	2.22 ± 1.13	5.72 ± 2.76	91.9	94.2	29.1	68.2
Jac	0.91 ± 1.43	0.62 ± 0.61	2.26 ± 0.86	7.37 ± 2.86	90.4	92.4	13.8	24.8
All	0.74 ± 1.00	0.55 ± 0.45	2.03 ± 0.98	4.46 ± 2.69	93.1	95.5	34.6	57.1
All-Pooled	0.72 ± 0.96	0.54 ± 0.46	2.00 ± 1.08	4.01 ± 2.66	93.4	95.8	38.9	69.7

gradient descent [12]. In CVH Eq. we set ϵ to 100 in order to ignore small set of voxels. std $\boldsymbol{T}^{\text{Hub}}$ is calculated with the same settings except that one resolution is used. The MIND feature is calculated using a $[3 \times 3 \times 3]$ region as suggested by [8] and also compared with a sparse patch including 82 voxels inside a $[7 \times 7 \times 3]$ box, which is physically more isotropic for our data.

After computing the mother features, average and maximum pooling is performed with box sizes of $[2, 4, 6, ..., 60]$ mm. As a result, for each mother feature we obtain a pool of 60 features: 30 from box averages and 30 from box maxima.

Parameters of the regression forest: The RF is trained on 50 trees with a maximum depth of D, while at least 5 samples remain in the leaf nodes. At each splitting node, f features are randomly selected from the pool ($f = 10$ for each single feature; $f = 2$ for 'All'; $f = 20$ for 'All-pooled'). The parameter D is optimized within the range of $[3, 100]$ by comparing the MAE. From the results in Fig. 3(b), we selected $D = 9$ for the remainder of this paper.

Results: RFs are trained for each single mother feature independently and for the combination of all features with or without feature pooling. Table 1 gives the results in terms of regression MAE and classification accuracy. The two MIND-based features have similar regression performance, but the sparse patch shows better classification accuracy in especially the second and third bin. We therefore

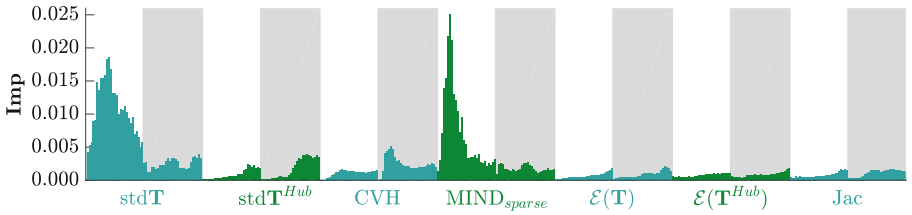


Fig. 4. Feature importance. White areas correspond to box averages, while shaded areas correspond to box maxima.

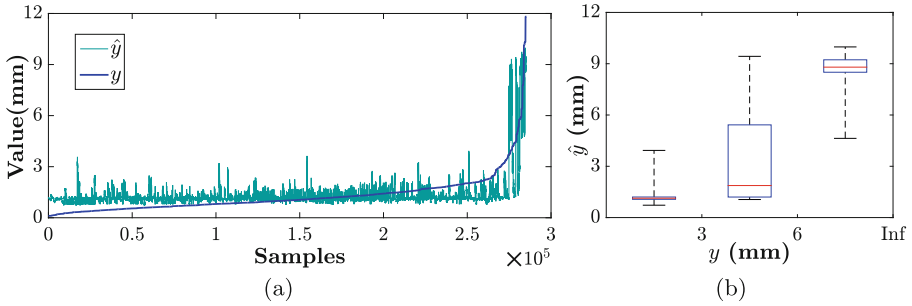


Fig. 5. Real (y) vs predicted ($\hat{y}$) registration error for the combined feature pool.

included the sparse patch MIND in the total feature pool. From Table 1 it can be seen that for the intensity-based features the best performance is obtained from MIND, for the registration-based features from std $\boldsymbol{T}$, and that the joint feature pool performs better than any single feature. The feature importance, see Eq. (3), is displayed in Fig. 4. It confirms that std $\boldsymbol{T}$ and MIND are the features contributing most to the RF performance, followed by CVH. The result of the complete pool is detailed in Fig. 5(a) which shows the real against the predicted error, sorted from small to large. In Fig. 5(b) we grouped the real errors in the three bins, each showing a box plot of the predicted errors. Intuitively, a smaller overlap between the boxes represents a better regression.

4 Conclusion and Discussion

In this paper we proposed a method based on random forests to regress registration accuracy from registration-based as well as intensity-based features. We introduced the variation in registration result from differences in initialization (std $\boldsymbol{T}$) as a feature, which showed higher feature importance and regression and classification performance than an existing variant of it. The proposed feature CVH measuring joint intensity variation also contributed to the regression performance, and can be calculated from the std $\boldsymbol{T}$ results without much additional computation. The combination of those features with several others, using a box-based pooling technique, yielded best overall performance. With a mean overall

regression error of 0.72 ± 0.96 mm and a classification accuracy of 93.4 % we conclude that the proposed method is very promising for the a posteriori assessment of local registration error. In future work, we will include additional information such as the variation of $\boldsymbol{T}$ in $[x, y, z]$ separately and estimate the error in each direction. In the current experiment, the number of samples in the second and third bin (poor and wrong) is considerably less than the number of samples in the first bin. We will therefore add poor registration results to the training set, thereby hopefully improving the regression results, especially in the second bin. A post-processing technique such as smoothing or majority voting over a neighborhood potentially also improves regression accuracy. One of the advantages of the proposed method is that all employed features are modality independent, and allow for parallel (GPU) computation. In the future, we will therefore test the algorithm on multi-modality data. Extra advantages are that additional features can be trivially included in the framework, that our method is compatible with any registration method, can already work in pairwise registration.

References

1. Risholm, P., Janoos, F., Norton, I., Golby, A.J., Wells, W.M.: Bayesian characterization of uncertainty in intra-subject non-rigid registration. Med. Image Anal. **17**(5), 538–555 (2013)
2. Datteri, R.D., Dawant, B.M.: Automatic detection of the magnitude and spatial location of error in non-rigid registration. In: Dawant, B.M., Christensen, G.E., Fitzpatrick, J.M., Rueckert, D. (eds.) WBIR 2012. LNCS, vol. 7359, pp. 21–30. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31340-0_3
3. Kybic, J.: Bootstrap resampling for image registration uncertainty estimation without ground truth. IEEE Trans. Image Process. **19**, 64–73 (2010)
4. Hub, M., Kessler, M.L., Karger, C.P.: A stochastic approach to estimate the uncertaintyinvolved in B-spline image registration. IEEE Trans. Med. Imaging **28**(11), 1708–1716 (2009)
5. Hub, M., Karger, C.: Estimation of the uncertainty of elastic image registration with the Demons algorithm. Phys. Med. Biol. **58**(9), 3023 (2013)
6. Muenzing, S.E., van Ginneken, B., Murphy, K., Pluim, J.P.: Supervised quality assessment of medical image registration: application to intra-patient CT lung registration. Med. Image Anal. **16**(8), 1521–1531 (2012)
7. Lotfi, T., Tang, L., Andrews, S., Hamarneh, G.: Improving probabilistic image registration via reinforcement learning and uncertainty evaluation. In: Wu, G., Zhang, D., Shen, D., Yan, P., Suzuki, K., Wang, F. (eds.) MLMI 2013. LNCS, vol. 8184, pp. 187–194. Springer, Heidelberg (2013). doi:10.1007/978-3-319-02267-3_24
8. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: Mind: modality independent neighbourhood descriptor for multimodal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)
9. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
10. Stolk, J., Putter, H., Bakker, E.M., Shaker, S.B., Parr, D.G., Piitulainen, E., Russi, E.W., Grebski, E., Dirksen, A., Stockley, R.A., Reiber, J.H.C., Stoel, B.C.: Progression parameters for emphysema: a clinical investigation. Respir. Med. **101**(9), 1924–1930 (2007)

11. Murphy, K., van Ginneken, B., Klein, S., Staring, M., de Hoop, B.J., Viergever, M.A., Pluim, J.P.: Semi-automatic construction of reference standards for evaluation of image registration. Med. Image Anal. **15**(1), 71–84 (2011)
12. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)