Medical Image Analysis 16 (2012) 1521–1531

Contents lists available at SciVerse ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration

Sascha E.A. Muenzing [a,*], Bram van Ginneken [b,a], Keelin Murphy [a], Josien P.W. Pluim [a]

[a] Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100, Room Q0S.459, 3584 CX Utrecht, The Netherlands
[b] Diagnostic Image Analysis Group, Anthonie van Leeuwenhoeklaan 5, 6525 HK Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 6 October 2011
Received in revised form 13 April 2012
Accepted 25 June 2012
Available online 24 July 2012

Keywords:
Non-rigid registration
Registration error
Image registration quality
Pattern recognition
Supervised learning

ABSTRACT

A novel method for automatic quality assessment of medical image registration is presented. The method is based on supervised learning of local alignment patterns, which are captured by statistical image features at distinctive landmark points. A two-stage classifier cascade, employing an optimal multi-feature model, classifies local alignments into three quality categories: correct, poor or wrong alignment. We establish a reference registration error set as basis for training and testing of the method. It consists of image registrations obtained from different non-rigid registration algorithms and manually established point correspondences of automatically determined landmarks. We employ a set of different classifiers and evaluate the performance of the proposed image features based on the classification performance of corresponding single-feature classifiers. Feature selection is conducted to find an optimal subset of image features and the resulting multi-feature model is validated against the set of single-feature classifiers. We consider the setup generic, however, its application is demonstrated on 51 CT follow-up scan pairs of the lung. On this data, the proposed method performs with an overall classification accuracy of 90%.

1. Introduction

Image registration is a crucial step in many image analysis tasks (Hill et al., 2001; Lester and Arridge, 1999; Maintz and Viergever, 1998). A large number of registration algorithms have been developed but for each application a specific solution or parameter setting has to be found to guarantee optimal registration results. Even then, poor registration performance for certain image pairs, or for regions within image pairs is common. Most registration algorithms do not provide information about whether results are good enough for subsequent processing.

Several methods are reported in the literature to address this issue. Crum et al. (2004) presented a method for automatic estimation of errors in voxel-based registration. The Gaussian scale-space of the post-registration residual image is used to establish an approximate scale of image differences. Spatial localization of estimated errors is not possible with this approach and applicability is limited to rigid image transformations. Fedorov et al. (2008) described a method for evaluation of brain MRI alignment. The proposed error estimation is based on a robust version of the Hausdorff distance metric applied to recovered image edges. Park et al. (2004) used local information measures for an adaptive registration approach. A general mismatch measure based on local estimates of mutual information and entropy was proposed to identify misaligned image areas. Sofka and Stewart (2008) proposed a location registration and recognition approach for longitudinal evaluation of corresponding regions in CT volumes. It is a feature-based method for separate matching of regions at pre-selected locations (e.g. nodules). A verification algorithm – using accuracy and stability measures combined in a support vector classifier – decides whether a region of one scan has been correctly recognized and aligned. Möller and Posch (2008) proposed a hierarchical analysis scheme that distinguishes between various underlying registration error sources. In the final stage a machine learning approach based on principal component analysis and support vector regression is adopted to automatically analyze patterns from so-called quality maps and to estimate the amount of radial lens distortions in 2-D camera images. Castillo et al. (2009) evaluated spatial accuracy of lung CT image registrations using large sets of expert-determined landmark point pairs. They investigate the uncertainty of spatial error estimates related to different registration algorithms. Further, sample size calculations were conducted to estimate the average spatial accuracy of an algorithm within certain confidence intervals. The study demonstrated that landmark pairs can be used to assess spatial registration accuracy within a narrow uncertainty range.

To tackle the need for objective and efficient registration assessment, we propose a general framework for quality assessment in

* Corresponding author.
E-mail address: sascha@isi.uu.nl (S.E.A. Muenzing).

1361-8415/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.media.2012.06.010

medical image registration. Its core is a method for automatic classification of local image alignments into three quality categories: correct, poor and wrong alignment. The method is supervised, that is, it can distinguish registration qualities because it has learned characteristics of different alignment categories in a training procedure.

This paper is based on an earlier publication (Muenzing et al., 2009), in which we presented a proof of concept and preliminary results.

2. Materials

The data in this work consists of a set of low-dose thoracic computed tomography (CT) scans which form part of a lung cancer screening trial (Xu et al., 2006). Fifty-one subjects (47 male, 4 female, ages 51–74 yrs), each with a baseline and a follow-up scan (3–15 months apart) were chosen randomly from the screening trial database. All scans were obtained at full inspiration and without contrast injection on a 16 detector-row scanner (Mx8000 IDT or Brilliance 16P, Philips Medical Systems, Cleveland, OH). Exposure settings were 30 mAs at 120 kVp for subjects weighing up to 80 kg or 30 mAs at 140 kVp for those weighing over 80 kg. A soft reconstruction filter (Philips "B") was used. The scans have a per-slice resolution of 512×512 voxels, and the number of slices per scan ranges from 374 to 579 (on average 462). Slice thickness is 1 mm with slice-spacing of 0.7 mm. Pixel spacing in the X and Y directions varied from 0.61 mm to 0.89 mm with an average spacing of 0.73 mm.

3. Methods

The proposed method for automated quality assessment of image registration consists of four main components which are described in the following subsections. Section 3.1 describes the employed automatic landmark detection method which provides a consistent way to determine a set of well-distributed landmarks for each inspected image. Based on this landmark detection scheme we create a reference set of registration errors (3.2). A set of image features (3.3) is extracted for every landmark, and the proposed classification system (3.4) is trained and validated. The main components of the proposed framework are depicted in the flow chart in Fig. 1. Application of the proposed system to new registrations, i.e. registrations not included in the training set, starts with the automatic landmark landmark detection (3.1). These landmark locations are inspected (3.3) and registration quality is assessed by the supervised learning system (3.4).

3.1. Automatic landmark detection

Our proposed method for evaluation of image registration employs automatic landmark detection to define regions of interest to be inspected. We do not use explicitly defined anatomical landmarks but instead define landmarks automatically by statistical properties. Automatic landmark determination provides a consistent way of generating landmarks. Moreover it allows the composition of landmark sets of adjustable cardinality. Note that we do not aim to conduct image registration by automated matching of corresponding landmarks. Here, landmarks are only determined in the fixed image and serve as reference locations to which the corresponding registered image is compared, in order to assess registration quality.

For our experiments on lung CT scans we use a landmark detection scheme which proved reliable in covering the anatomy of lungs in CT scans. First, the global search area is constrained to the volume within the lungs by a lung mask (van Rikxoort et al., 2009). Second, from the remaining lung volume a set of initial landmark points is determined by a distinctiveness measure which estimates the dissimilarity of a voxel with its surrounding region. All voxels with low intensity gradient are excluded from processing as they are likely to be extremely difficult to match reliably in the follow-up image. Last, in addition to choosing the most distinctive points as landmarks, an even distribution of the landmarks throughout the lungs is required. Generated landmarks are typically located at vessel bifurcations. The technique is based on Murphy et al. (2011).

3.2. Reference registration error set

The Reference Registration Error Set (*RRES*) forms the basis dataset for the intended pattern recognition approach. It consists of alignment samples, i.e. landmark correspondences between fixed, moving and registered images. Image registrations are acquired by applying several different transformation models. That way we have for each object (patient) two original scans (baseline&follow-up) and a set of registered scans. Knowing the position of the landmark in the fixed image, one can use the computed image transformations to determine the registered landmark position in the moving image. The establishment of landmark correspondences is described in 3.2.1, the automatic registration procedure is described in 3.2.2 and the composition of a reference registration error set is described in 3.2.3.

3.2.1. Landmark correspondence

Since our experiments are completely based on real clinical data, no ground truth of the deformation field is available which would represent the underlying mapping of points in the fixed image onto points in the moving image. We resolve this issue by employing manual landmark correspondences.

We have available a set of landmark correspondences which were established with a system for semi-automatic construction of reference standards (Murphy et al., 2011). This system uses an automatic landmark detection method as described in Section 3.1. For each of the 51 scan pairs a set of 100 landmarks l_{Fi} were automatically generated for each baseline scan I_{Fi}. From those 100 landmarks, the first 30 landmarks were matched manually. The ordering of the detected landmarks is such that each landmark is as far away as possible from preceding selected landmarks (Murphy et al., 2011). However, we opt to define the detected landmarks as a set of points, because the underlying ordering is of no further relevance to the subsequently performed processing. In the following we use this landmark subset $L_F \doteq \{l_{Fip} | i = 1 \dots 51, p = 1 \dots 30\}$ for which a corresponding point l_{Mip} in the moving image I_{Mi} was obtained manually. The manual matching is conducted by means of a software application with a graphical user interface (Murphy et al., 2011). It shows sagittal, coronal and transverse plane views with adjustable window levels for both baseline and follow-up scan. The system denotes a landmark l_{Fip} in the baseline scan views and asks a human observer o to manually find the corresponding anatomic position l_{Mip} in the follow-up scan. Two human observers (medical students) independently established manual landmark matchings

$$L_M \doteq \left\{ \left\langle l_{Fip}, l_{Mipo} \right\rangle, i = 1 \dots 51, p = 1 \dots 30, o = 1 \dots 2 \right\}. \quad (1)$$

The accuracy of physical position measurement has been limited to whole voxels (voxel $\approx 0.7^3$ mm^3), for both the automatic landmark detection and the annotation of the manual matching.

3.2.2. Automatic registration

Prior to registration the baseline and follow-up scans are downsampled in order to improve speed and reduce memory consump-

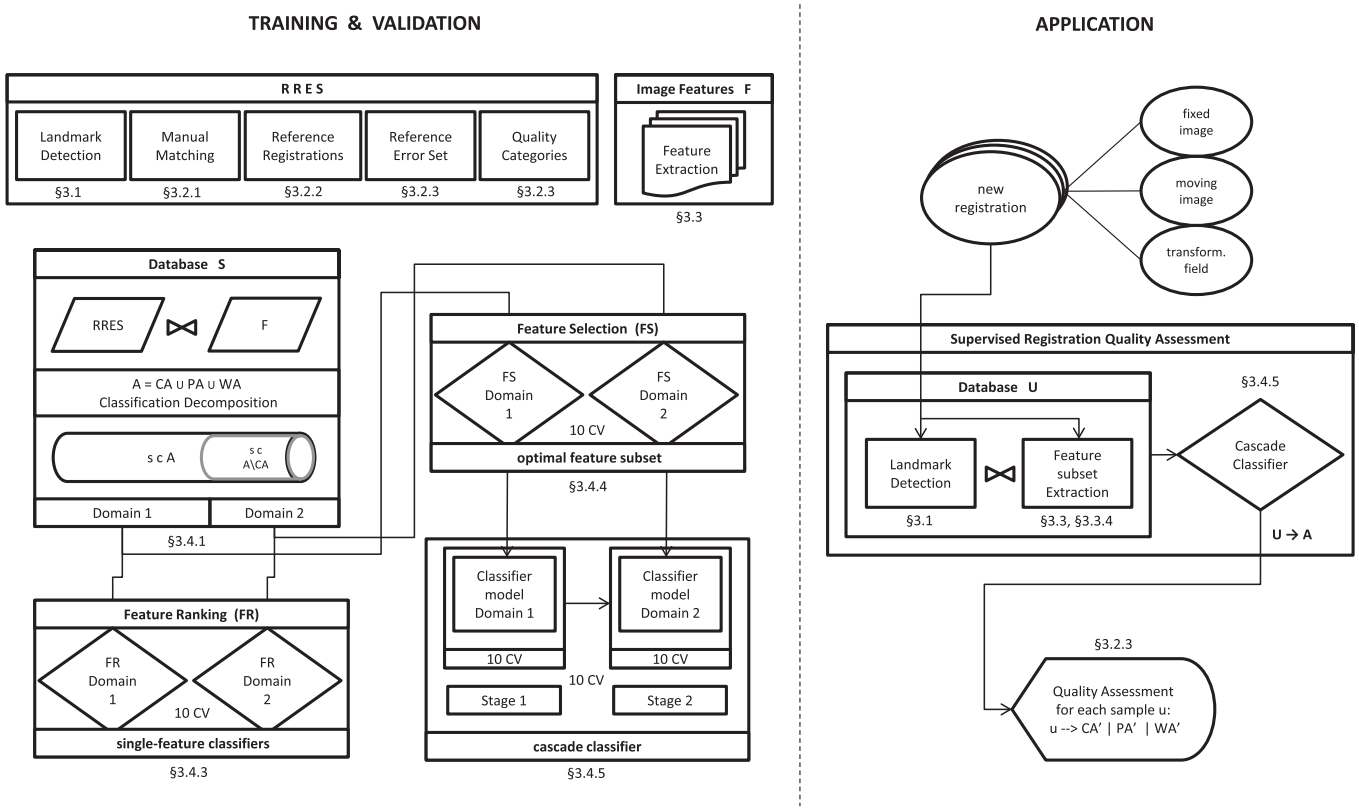

Fig. 1. Flow chart with the main components of proposed framework. Each component is accompanied by a section number where more information can be found.

tion. The downsampling is by means of block-averaging such that the matrix size of 512×512 in the original images is reduced to 256×256, with the number of slices reduced to form isotropically sampled data. The downsampled follow-up scan (moving image) is registered to the downsampled baseline scan (fixed image) and the resulting transformation is subsequently applied to the full resolution follow-up scan. Lung mask images (van Rikxoort et al., 2009) are used to exclude confounding structures like the ribs and heart.

Registrations are carried out using elastix (ver. 3.9) which is a registration toolkit for intensity-based medical image registration (Klein et al., 2010). A multi-resolution strategy is taken to avoid local minima in the cost function. A Gaussian pyramid is employed, using a sub-sampling factor of two in each dimension. Also, a multi-grid approach is used for the B-spline registration: the registration starts with a coarse B-spline control point grid, which is refined in subsequent resolutions. A mutual information cost function is used along with an iterative stochastic gradient descent optimizer (Klein et al., 2007). The derivative of the mutual information is calculated based on a small subset of images samples, randomly chosen every iteration. For all registration methods 32 histogram bins were used. Termination of the optimization procedure occurred in each resolution after a fixed number of iterations. Three types of transformation were applied:

Similarity Registration[1] ($\mathbf{T_S}$) settings: 5 resolution levels, 512 iterations, 4096 spatial samples.

Affine Registration ($\mathbf{T_A}$) settings: 5 resolution levels, 512 iterations, 4096 spatial samples.

Freeform Registration ($\mathbf{T_F}$) We use a B-spline based registration preceded by an affine registration. Settings affine registration: 4 resolution levels, 256 iterations, 2048 spatial samples. Settings B-Spline registration: 5 resolution levels, 256 iterations, 4096 spatial samples, final grid-size spacing of 8 voxels in each dimension. The listed settings have been experimentally determined to be relatively fast and accurate on this type of data. We refer to the entity of fixed and moving image, and corresponding similarity, affine and freeform transformations as:

$$R_i \doteq \{I_{Fi}, I_{Mi}, \mathbf{T_{Si}}, \mathbf{T_{Ai}}, \mathbf{T_{Fi}}\}, \quad i = 1 \ldots 51. \tag{2}$$

3.2.3. Reference set construction

For the Reference Registration Error Set we consider only those $L_M(i,p)$ valid correspondences that are within a certain interobserver deviation

$$\delta_o(i,p) \doteq \left\| L_{M(o=1)}(i,p) - L_{M(o=2)}(i,p) \right\|_2. \tag{3}$$

We compute the average position of corresponding observer matchings $L_{M\bar{o}} \doteq \overline{\langle L_{M(o=1)}, L_{M(o=2)} \rangle}$ and establish the Reference Landmark Correspondence Set:

$$L_C \doteq \{\langle L_F(i,p), L_{M\bar{o}}(i,p) \rangle, i = 1 \ldots 51, p \in N_i\}, \tag{4}$$

with $N_i \doteq \{p : \delta_o(i,p) \leqslant 2 \text{ mm}\}$. This results in $|L_C| = 1374$ landmark correspondences from originally 1530 manual matchings, rejecting about 10% (approximately normally distributed with a median of 3 and maximum of 8 rejections per scan pair).

Next, several alignment categories are distinguished. We define the landmark registration error

$$LRE(i,p) \doteq \|\mathbf{T}(L_F(i,p)) - L_{M\bar{o}}(i,p)\|_2, \tag{5}$$

as alignment difference between manual correspondence and automatic registration, where $\mathbf{T}(L_F)$ are the registered positions of L_F in the corresponding image I_M. We categorize registration errors into three alignment classes: *Correct Alignment* (CA), *Poor Alignment* (PA) and *Wrong Alignment* (WA), and define the following disjunct subsets:

[1] A similarity transformation involves translation, rotation and isotropic scaling (7 dof in 3D).

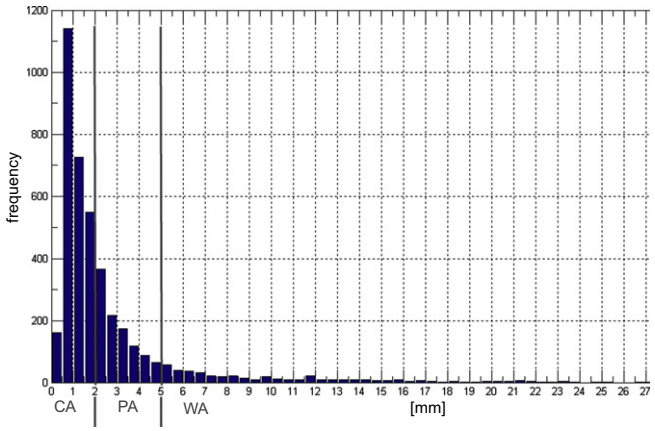

Fig. 2. Data distribution of reference registration error set. *X*-axis shows registration errors (bin size 0.5 mm) and corresponding frequencies on *Y*-axis. Largest misalignment is at 54 mm with 77 samples being sparsely distributed in the range from 27 to 54 mm (not displayed). Also shown are categories (CA, PA, WA) of alignment quality (3.2.3).

$$\begin{aligned} \mathrm{CA} &\doteq \{\mathbf{T}(L_F) : LRE \leqslant 2\ \mathrm{mm}\} \\ \mathrm{PA} &\doteq \{\mathbf{T}(L_F) : 2 < LRE < 5\ \mathrm{mm}\} \\ \mathrm{WA} &\doteq \{\mathbf{T}(L_F) : LRE \geqslant 5\ \mathrm{mm}\} \end{aligned} \tag{6}$$

so that registration quality $A = \mathrm{CA} \cup \mathrm{PA} \cup \mathrm{WA}$. The assignment of the category boundaries is based on considerations of the landmark annotation precision and the interobserver variability. In L_C the average and standard deviation (SD) of δ_o are 0.51 mm and 0.52 mm respectively. Based on these scores and the maximum $\delta_o = 2$ mm, we choose the interval $[0 \ldots 2]$ for CA. Further, from Fig. 2 it becomes clear that a binary split with a decision border at $LRE = 2$ mm would yield a category of wrong alignment samples of which the majority would be located close to the border of CA. Such binary classification would be highly sensitive to the placement of this single decision border. Therefore we introduce the poor alignment category of size 3 mm ($\pm 3 \times SD$) and consequently assign all other samples to WA. We compute *LRE* for all $\mathbf{T}$ of R_i and join[2] it with the previously derived information:

$$RRES \doteq L_C \bowtie R \bowtie LRE \bowtie A, \tag{7}$$

establishing the Reference Registration Error Set (*RRES*).

3.3. Image features

The aim of this work is automatic quality assessment of image registration by means of analysis of the alignment of distinct image structures. To capture the characteristic appearance of such alignment patterns we extract a specific set of local image features for all landmarks, sampled at different spatial scales. The feature set consists of features related to Gaussian features, correlation, entropy and deformation. All features are calculated locally either from a cubic subvolume or by employing Gaussian scale space. Further, each feature sample distribution is transformed by a Box-Cox power transformation, and then normalized to zero mean and unit standard deviation (*z*-score normalization). A power transformation stabilizes variance and thereby transforms skewed, non-symmetric distributions into symmetric, standard Gaussian like distributions (Sakia, 1992; Osborne and Carolina, 2010).

We define the entity of all features as feature set *F*, which yields for each alignment sample a vector $\mathbf{f}$ combining 66 feature values obtained from nine image feature types at different spatial scales (see Table 1).

[2] $\bowtie$ denotes the natural join operator in relational algebra (Maier, 1983).

Table 1
Overview of image features. *vol* denotes cubic subvolumes with edge lengths given in brackets. σ denotes the scales used within the Gaussian scale space paradigm. In the following the acronym of feature type plus scale value is used to refer to a particular feature (e.g. PMI-20).

#	Feature		Scale (mm)
1:7	*SID*	σ	{0, 0.5, 1, 2, 4, 8, 16}
8:13	*GSID*	σ	{0.5, 1, 2, 4, 8, 16}
14:19	*GID*	σ	{0.5, 1, 2, 4, 8, 16}
20:27	*NC*	*vol*	{5, 10, 15, 20, 25, 30, 35, 40}
28:35	*JM*	σ	{0, 0.5, 1, 2, 4, 8, 16}
36:42	*NMI*	*vol*	{5, 10, 15, 20, 25, 30, 35, 40}
43:50	*PMI*	*vol*	{5, 10, 15, 20, 25, 30, 35, 40}
51:58	*NMIS*	*vol*	{5, 10, 15, 20, 25, 30, 35, 40}
59:66	*PMIS*	*vol*	{5, 10, 15, 20, 25, 30, 35, 40}

3.3.1. Gaussian intensity features

A natural framework for examining image structures at different scales is provided by scale-space theory (Koenderink, 1984; ter Haar Romeny, 2009). We use a discrete analogue of the Gaussian kernel $G(\mathbf{x};\sigma)$ and accordingly a discrete Gaussian derivative kernel $\nabla G(\mathbf{x};\sigma)$ with normalization across scale-space (Lindeberg, 1993). In image registration it is common to visually assess registration results by examining the residual image, which we define as $I_\Delta(\mathbf{x}) \doteq I_F(\mathbf{x}) - I_M(\mathbf{x}) \circ \mathbf{T}(\mathbf{x})$. We extract intensity-related features as follows:

The squared intensity difference

$$SID(\mathbf{x};\sigma) \doteq I_\Delta^2(\mathbf{x}) * G(\mathbf{x};\sigma) \tag{8}$$

at $\sigma = \{0, 0.5, 1, 2, 4, 8, 16\}$ mm, and the gradient magnitude *GSID* of *SID* at $\sigma = \{0.5, 1, 2, 4, 8, 16\}$ mm.

The gradient of intensity difference

$$GID(\mathbf{x};\sigma) \doteq \|I_\Delta(\mathbf{x}) * \nabla G(\mathbf{x};\sigma)\|_2 \tag{9}$$

at $\sigma = \{0.5, 1, 2, 4, 8, 16\}$ mm.

3.3.2. Correlation

The normalized correlation (*NC*) between two signals is a standard approach in pattern recognition (Duda et al., 2001). In comparison to the proposed Gaussian features, the *NC* is insensitive to multiplicative factors between two images. We compute *NC* locally on cubic image subvolumes *V* with edge length *vol* centered at $\mathbf{x}$ for $vol = \{5, 10, 15, 20, 25, 30, 35, 40\}$ mm.

3.3.3. Entropy

As entropy-based feature, different modifications of the mutual information (*MI*) metric are employed. The major advantage of *MI* over the aforementioned *NC* and Gaussian features, is its ability to account for non-linear dependencies (Pluim et al., 2003). We compute the *normalized mutual information NMI* using 64 bins for the histogram-based density estimation.

A *modification of NMI* which normalizes the *MI* with the local entropy *H* of the corresponding subvolumes as proposed in Park et al. (2004) is

$$PMI(\mathbf{x};vol) \doteq \frac{MI(I_F(\mathbf{x}), I_M(\mathbf{x}) \circ \mathbf{T}(\mathbf{x}))}{min(H(I_F(\mathbf{x})), H(I_M(\mathbf{x}) \circ \mathbf{T}(\mathbf{x})))} \tag{10}$$

In addition we compute probability density functions (PDFs) based on histograms with dynamically determined *numbers of bins*. There is no reason to assume that a fixed bin size is best when calculating *MI* locally. Excessive quantization caused by too large bin sizes may cause a loss of important information (“oversmoothing”), and too small bin sizes may result in many bins becoming

sparsely populated, and consequently making the PDF estimates unreliable (“undersmoothing”). We opt for an adaptive histogram binning approach according to Sturges’ rule (Sturges, 1926), where the number of classes of a histogram is the closest integer to

$$|B| = log_2(n) + 1, \tag{11}$$

with n being the number of observations. We included adaptive histogram binning separately for both aforementioned normalized *MI* metrics yielding two additional features which we denote by $NMIS(\mathbf{x}; vol, |B|)$ and $PMIS(\mathbf{x}; vol, |B|)$. All *MI* metrics are locally calculated on cubic image subvolumes V with edge length vol centered at $\mathbf{x}$ for $vol = \{5, 10, 15, 20, 25, 30, 35, 40\}$ mm.

3.3.4. Deformation

For analysis of the image deformations $\mathbf{d}(\mathbf{x})$ produced by the registration algorithm we examine the so called Jacobian map

$$JM(\mathbf{x}; \sigma) \doteq |J(\mathbf{u}(\mathbf{x}) * G(\mathbf{x}; \sigma))| \tag{12}$$

which is the determinant of the Jacobian matrix J of a displacement field $\mathbf{u}(\mathbf{x})$ (Leow et al., 2007). *JM* encodes the local volume change between I_F and I_M. We derive deformation-based features by calculating *JM* within the Gaussian scale-space of $\mathbf{u}(\mathbf{x})$ at $\sigma = \{0, 0.5, 1, 2, 4, 8, 16\}$ mm.

3.4. Supervised learning

We employ supervised learning for automated classification of registration errors into three classes (CA, PA, WA) which were introduced in Section 3.2.3. In the following subsections we describe the classification setup (3.4.1) employed, which includes a decomposition of the multi-class problem and a cross-validation procedure for evaluation of classification performance. A set of standard classifiers (3.4.2) is investigated along with a feature selection procedure (3.4.4) in order to find the best suited multi-feature-classifier model for the underlying problem domain. In addition we conduct feature rankings (3.4.3) to investigate the performance of single features, and to establish reference performances comparable to manual thresholding of e.g. the residual image. The final classification model, a multi-feature two-stage classifier cascade is described in Section 3.4.5.

3.4.1. Classification setup

A supervised binary classification method is used, which means that a classifier is first trained on labeled samples from a positive and a negative class (Duda et al., 2001).

Database. The aim is creating a database comprising a rich variety of different alignment patterns after automatic registration due to unrecovered deformations between supposedly registered images. We build such a database $S \doteq RRES \bowtie F$, based on the entire Reference Registration Error Set and correspondingly extracted features. S comprises $3 \cdot |L_C| = 4,122$ alignment samples s, each of which is assigned a label of alignment category A, and each of which is represented by a corresponding feature vector $\mathbf{f}$, described in Section 3.3.

Validation. We keep all data in one set S on which we apply k-fold cross-validation (CV) to estimate classification performances (Toussaint, 1974). In k-fold CV the dataset S is randomly split into k mutually exclusive subsets (the folds) $S_1, S_2, \ldots, S_k$ of approximately equal size. A classifier is trained and tested k times; each time $t \in \{1, 2, \ldots, k\}$ it is trained on $S \setminus S_t$ and tested on S_t. In order to optimize algorithm parameters we use another nested loop of cross-validation by further splitting each of the $S \setminus S_t$ training sets into smaller training sets and validation sets (internal CV). For each combination of the classifier parameters, we compute cross-validation performance and select the best performing parameters based on this internal CV. Next, we build a classification model with the best parameters on the training sets from the outer folds and apply this model to the corresponding testing sets from the outer folds (external CV). It is indicated in Kohavi (1995) that 10-fold stratified CV is a good strategy for model selection. In stratified cross-validation, the folds are stratified so that they contain approximately the same proportions of labels as the original dataset. We use 10-fold CV in the feature ranking (3.4.3), feature selection (3.4.4) and validation of the classifier cascade (3.4.5), and internal 5-fold CV for parameter estimations. The cross-validation sampling technique used is random but ensures preservation of patient entities, meaning that all alignment samples originating from one registration scan pair (Eqn. (2)) are only present in one fold S_t. Moreover, the sampling technique involves rejection of a folding set and retry of the sampling until stratification is achieved. For consistency, training and testing of all investigated classification schemes is based on one fixed 10-fold set (paired experimental design) (Hanley and McNeil, 1983).

We evaluate all classification performances by means of the area under the ROC (Receiver Operating Characteristics) curve (AUC) (Bradley, 1997). We calculate the AUC by trapezoidal integration and estimate corresponding standard deviations from the cross-validation results, i.e. the AUC is calculated for the k ROC curves and then averaged, giving an estimate of the true area and an estimate of its standard error, calculated from the standard deviation of the k areas.

Domains. The multi-class classification problem is decomposed into two binary classification problems where alignment samples s are assigned to domains as follows:

$$\begin{aligned} \text{Domain1} &\doteq \forall s \in A \quad \mapsto \quad \text{CA}' \vee \text{MA}' \\ \text{Domain2} &\doteq \forall s \in A \setminus \text{CA} \quad \mapsto \quad \text{PA}' \vee \text{WA}' \end{aligned} \tag{13}$$

with the quotation mark denoting a prediction of a classifier, i.e. the classifier function assigns to each sample s a registration quality label of A and where $\text{MA} \doteq \text{PA} \cup \text{WA}$. We opt for this decomposition in order to obtain specific classifiers and features for each quality of alignment.

3.4.2. Classifiers

There is a wide choice of classifiers available in the literature (Duda et al., 2001; Jain et al., 2000) with no superior learning method overall, in fact the underlying problem determines which classifier provides a better performance. We employ a set of standard classifiers from statistical pattern recognition, which we group into the following categories:

Linear discriminant classifiers – LBN, LFD. The linear Bayes–Normal classifier (LBN) assumes Gaussian distributions of the samples of each class, and equal covariance matrices for each distribution (Duda et al., 2001). Fisher’s linear discriminant (Fisher, 1936) classifier (LFD) is a non-density based classifier. It computes a linear discriminant using a minimum least squares method.

Non-linear classifiers – kNN, RBFSVM. The k-nearest-neighbor (kNN) classifier is a non-parametric classifier, with a free parameter k. In the kNN rule, the posterior probability for each of the classes is estimated by the fraction of training samples among the k nearest neighbors that belongs to that class. In our implementation k is automatically optimized over the training set using a leave-one-out error estimation (Duin et al., 2007). A useful feature of the support vector machine (SVM) is that this method can be kernelized if linear discriminants are not appropriate for a given dataset. By mapping original feature vectors into a higher-dimensional feature space and solving an SVM optimization problem there, a highly non-linear classification function can be obtained in the original feature space. The performance of the SVM for a particular classification problem depends on a correctly estimated penalty parameter and a suitable kernel function. We em-

ploy an SVM using the L_2 norm with radial basis function (RBF) kernel (Chang and Lin, 2001), and refer to this configuration as RBFSVM. The penalty and kernel parameters (C, γ) are optimized using grid search with an internal 5-fold CV.

Ensemble learning – DS, AB, RF. An ensemble of learners is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples. Ensembles are often much more accurate than the individual learners that make up the ensemble. The main reason is that many learning algorithms apply local optimization techniques, which may get stuck in local optima (Bauer and Kohavi, 1999; Kotsiantis, 2011).

Boosting (Schapire, 2002) is a meta-algorithm that obtains a strong classifier from weak classifiers, combining each one of these individual classifiers with different weights to obtain a consensus decision that minimizes the error rate. Adaptive boosting (AdaBoost) (Freund and Schapire, 1997) is an iterative approach to improve the performance of a weak classifier by assigning weights to training samples, and where incorrectly classified training samples will gain a larger weight in the process. In Tieu and Viola (2000); Viola and Jones (2004) a feature selective AdaBoost (AB) is proposed to establish ensembles that rely only on a subset of features. The weak classifiers used (thresholded single features) can be viewed as single node decision trees, so-called decision stumps (DS). For each feature, the weak learner determines the optimal threshold classification function, such that the minimum number of examples is misclassified. Thus, in every iteration the weak classifiers selects the single highly selective feature along which the positive examples are most distinct from the negative examples. We employ this feature selective AdaBoost scheme (AB) and establish ensembles of single-feature classifiers (DS).

Random forests (RF) (Breiman, 2001) involves the idea of Bagging (Breiman, 1996) and the random selection of features (Ho, 1995; Amit and Geman, 1997; Ho, 1998). It is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Random forests generates decision trees by randomly selecting a limited number of features from all available features for node splitting, and each tree casts a vote for the final decision. RF contains internal estimates to monitor error, strength, and correlation, which can be used to measure variable importance. We employ Random forests with default settings (Liaw and Wiener, 2002), i.e. 500 trees are generated and the number of randomly selected features is given by the square root of the number of features in the dataset.

3.4.3. Feature ranking

We conduct feature ranking on each classifier,[3] i.e. evaluating classification performance of every single-feature classifier for Domain 1 and Domain 2, respectively. Here, the purpose of the feature ranking is threefold. First, it enables the comparison of features w.r.t. classifier performances, i.e. it gives insight into the importance of a feature for the particular classification problem. Second, it allows us to evaluate the performance gain due to a multi-feature model approach. Third, we consider single-feature classification performances of simple classifiers (DS, LBN) as reference for best performances achievable by manual thresholding.

3.4.4. Feature selection

The goal of feature selection is to find a subset of features such that the classification performance is maximal and the number of selected features minimal. A classifier which depends only on a subset of features is more efficient since fewer features have to be computed, and in addition, a classifier which depends on few features is more likely to generalize well. We limit the maximum subset size to 20 features and employ different feature selection strategies depending on the classifier. The AdaBoost (AB) classifier performs inherent feature selection. We run the AB algorithm for 20 iterations, yielding a strong classifier which depends upon 20 features. The Random Forests (RF) classifier provides an internal estimate of feature importance. Based on this importance measure (accuracy loss (Breiman, 2002)), we subsequently select up to the 20 most important features. The RF classifier is then retrained for every feature subset and tested using internal CV to obtain unbiased estimates of feature subset classification performance. The best subset found is evaluated on the external CV test set S_t. For all other classifiers we employ sequential forward floating search (SFFS) (Jain and Zongker, 1997) with a wrapper setup (Kohavi and John, 1997) where the classifier performance is used as decision criterion. The maximum feature subset size is set to 20, equivalently to above, and SFFS is conducted on each training set $S \setminus S_t$, whereas within the SFFS procedure internal CV is conducted, and the subset found is evaluated on the test set S_t of the external CV (3.4.1). Due to this validation scheme, the risk of overfitting in feature selection can be avoided (Reunanen, 2003). That way, for each classifier (3.4.2) an optimal feature subset is established (model selection).

3.4.5. Classifier cascade

A classifier cascade is a sequential combination of classifiers (Viola and Jones, 2001). We propose a classification scheme employing multi-stage classification and model selection, involving both feature selection and classifier selection. In the context of this work, the major advantage of cascade classification over multi-class classification (one-against-all scheme) is by design the better discrimination between non-adjacent classes. In Section 3.4.1 we already introduced different classification domains by decomposing the initial three-class problem into two individual classification problems. For the given registration quality assessment problem, we opt for following classification scheme:

1. Feature selection conducted as described in Section 3.4.4 provides feature models optimized separately for all investigated classifier types (first model selection).
2. Evaluation of classification performance of each classifier using the corresponding optimal feature subset. Based on these results – classification performance as well as feature subset size – for each domain a best classifier is selected (second model selection).
3. Establishment of classifier cascade by connecting best Domain 1 classifier to best Domain 2 classifier. Cross-validation of Stage 2 as described in Section 3.4.1. In addition, the Domain 2 classifier is optimized in favor of detection sensitivity of class WA by choosing an operating point on the ROC curve.

In contrast to individual domain classifiers, the training procedure of the classifier cascade consists of two stages with alignment samples s as follows:

$$\begin{aligned} \text{Stage 1} &\equiv \text{Domain1} \\ \text{Stage 2} &\doteq \forall s \in \mathrm{MA}' \mapsto \mathrm{PA}' \vee \mathrm{WA}', \end{aligned} \tag{14}$$

with $\mathrm{MA}' \doteq A \setminus \mathrm{CA}'$. The proposed classifier cascade yields a combined classification function that assigns to each alignment sample exactly one label and thereby eventually estimates registration quality locally, for each alignment sample according to A into: correct (CA), poor (PA), or wrong (WA) alignment.

[3] For AB and RF only the base classifier (DS) is ranked.

4. Results

4.1. Reference registration error set

The reference registration error set (*RRES*) is constructed as described in Section 3.2.3. Table 2 summarizes the results. In addition to the merged dataset (*RRES*), also subsets are listed for each type of registration and for each of the three alignment categories of correct, poor and wrong alignment (CA, PA, WA).

Fig. 2 displays a histogram of the frequency distribution of the merged dataset. The x-coordinate shows the registration error *LRE* (Eqn. (5)) with the resolution determined by a fixed bin size of 0.5 mm. The histogram shape depicts a pronounced right skewness of the alignment distribution.

4.2. Feature ranking

The results of the feature ranking are shown in Table 4, separately for each domain. To ease the comparison, only the five best ranked features are listed per classifier, with feature ID (see Table 1), average AUC and standard deviation of 10-fold CV. Best performing models (single-feature classifiers) are in bold font, as well as the performances that are not significantly worse than the best (tested over folds using a two-sided paired t-test with a confidence of $\alpha = 0.05$). In Domain 1 the best classification performances are obtained with mutual information features (PMIS, NMIS), whereas Gaussian features (GID, SID) prevail in Domain 2. In Domain 1 there are good classification performances achievable with single features, whereas the classification task in Domain 2 appears more difficult with the best AUC only 0.82, compared to best AUC of 0.96 in Domain 1.

4.3. Feature selection

Table 5 shows the results of the feature selection (model selection) for Domain 1 and Domain 2. Feature selection was conducted as described in Section 3.4.4. Average feature subset size and corresponding standard deviation are listed next to AUC values. The right part of the table, separated by double lines, shows feature subsets found by a feature selection conducted on the entire learning set *S*, guided by a 10-fold CV. We show these feature subsets to ease the insight into which feature combination is important – i.e. would be chosen if trained on the entire *S* in order to classify further unseen image registrations – rather than listing per classifier 10 different feature subsets or all chosen features separately along with their frequency. All classifiers perform better using an optimal multi-feature model than corresponding best single-feature model. In Domain 1, the difference between best single-feature model (kNN) and best multi-feature model is relatively small. In Domain 2, however, classification performances (AUC) increased on average by about 14%, calculated based on best single-feature classifier vs. classifier using optimal multi-feature subset.

Table 2
Composition of reference registration error set: alignment samples per registration method (in rows), and (in columns) per alignment category (CA, PA, WA). Ratios in parentheses are calculated w.r.t. totals of each dataset.

	CA		PA		WA		Totals
Freeform	1300	(94.6%)	34	(2.5%)	40	(2.9%)	1374
Affine	875	(63.7%)	324	(23.6%)	175	(12.7%)	1374
Similarity	606	(44.1%)	489	(35.6%)	279	(20.3%)	1374
Merged	2781	(67.5%)	847	(20.5%)	494	(12.0%)	4122

4.4. Classifier cascade

We opt in Stage 1 for the RF model because of its superior consistency (lower variation), and in Stage 2 we opt for the RBFSVM model because of its superior feature space reduction (subset of 11 features). Stage 2 of the classifier cascade is validated based on the posteriori probabilities, obtained from 10-CV of Stage 1 classifier. In a final step, in Stage 2 the point on the ROC curve that is closest to the north-west corner ($FP = 0, TP = 1$) is chosen as operating point. This results in an increased WA sensitivity compared to the default operating point.

The performance of the classifier setup is evaluated by means of a confusion matrix. Table 3 presents the confusion matrix for the entire three-class problem. Categories predicted by the classification cascade are listed in the matrix columns. The diagonal shows the number of true positives for each category, respectively. The combined classification accuracy is 90%, with the following sensitivities and specificities CA: (97%, 93%), PA: (67%, 82%) and WA: (89%, 84%). No misclassifications occur between CA and WA.

5. Discussion

A method for automatic quality assessment of medical image registration has been presented. Its core is a supervised two-stage multi-feature classifier cascade. The automatic classification provides for each inspected landmark location a categorical quality assessment. The method is demonstrated on intra-patient lung CT registrations. For each scan pair a set of registrations are obtained using three different transformation models. On this pooled registration data, the proposed method performs with an overall classification accuracy of 90%. In the following subsections we first discuss the employed image features (5.1) and classification scheme (5.2). Second, we briefly present two possible application settings (5.3) for which the automatic quality assessment can be directly utilized. Third, the limitations (5.4) of the proposed method are discussed, including general limitations as well as issues that have to be addressed when applying the method for other applications, and image data. Last, we discuss the generic framework (5.5) of the proposed method, that is the generic setup based on pattern recognition approaches which allows automated adaptation to different data.

5.1. Image features

The results in Tables 4 and 5 show that there are different features and feature combinations, respectively, that perform best for each domain. This is a rather general statement, however it is a common finding in pattern recognition which motivates the use of machine learning methods in order to automatically determine optimal models for a particular application. In our experiments on lung CT registrations, the most important single features for Domain 1 are mutual information features, and Gaussian intensity features for Domain 2. The proposed derivations of mutual infor-

Table 3
Confusion matrix of final classification results, i.e. assessment of registration by alignment categories. Ratios in parentheses are calculated w.r.t. corresponding row totals.

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

mation (*NMIS*, *PMIS*) outperformed standard *MI* (Table 4). In the following we discuss more generic findings regarding entropy, correlation and deformation features.

Mutual information. Although the employed *MI* features seem to be strongly correlated, during feature selection more than one *MI* feature type was chosen. That means, discriminative power must have been added in order to improve classification performance. Providing a pool of similar features might enable the feature selection process to conduct a sort of feature weighting and thereby improving discriminative power. Either way, results indicate that it is beneficial to consider different histogram bin sizes and *MI* normalization approaches.

Correlation. Given the above findings of employing different *MI* features, also the correlation feature *NC* could be combined with a scale-space approach, either directly based on Gaussian smoothed subvolumes or analogous to *MI* by calculating the correlation measure based on histograms using different bin sizes.

Jacobian map. The impact of the deformation-based features (*JM*) was expected to be low since landmark locations are determined based on intensity and intensity gradient, and therefore it is plausible that intensity based features are more prominent. Another reason is in the registration algorithms applied. Similarity and affine registrations produce only global deformations. Yet, in freeform registrations, a local deformation measure (from different scales) can be advantageous in combination with other feature types (cp. Table 5 Domain 1: AB, Domain 2: LFD, AB, RF).

5.2. Classification scheme

As mentioned in Section 3.4.3, we consider single-feature classification performances of simple classifiers (DS, LBN) as representative for the performances that can be achieved by manual thresholding, which is currently common practice (cp. e.g. Isgum et al. (2009)). From Table 4 it is apparent that more complex classification approaches such as LFD, kNN, RBFSVM clearly outperform simple classifiers. Therefore we conclude that the application of state-of-the-art classification approaches are significantly superior to manual thresholding approaches on empirical features such as residual images. The merit of proposed pattern recognition methods becomes even more evident when multi-feature models are considered (cp. performances of DS and LBN in Table 4 with best classification performance in Table 5). We include the LBN classifier in Table 5 to allow a complete comparison to Table 4. Although a multi-feature model improves the classification performance of the LBN, the overall performance is clearly inferior to the other classifiers employed.

From the multi-feature models listed in Table 5, it appears that Domain 1 is linearly separable (LFD) whereas Domain 2 seems better fitted by a non-linear model (RBFSVM). Inspection of the RBFSVM parameters however revealed that the automatic parameter optimization resulted in a linear SVM (RBF radius $\gamma \approx 0$). Thus a linear discriminant function fitted best on this data.

The Domain 2 problem is obviously more difficult to discriminate than Domain 1. Because the a priori probabilities of both do-

Table 4
Feature ranking results. For each classifier the best five features are listed, ranked by AUC. AUC performances (×100) are averaged over folds and the corresponding standard deviation is given in parentheses. Results in bold are the best performances and the performances that are not significantly worse than the best (tested using a two-sided paired *t*-test with a confidence $\alpha = 0.05$).

DS			LBN			LFD			*k*NN			RBFSVM		
Domain 1														
NC-5	86.3	(3.2)	NC-10	89.9	(3.7)	SID-2	94.3	(1.0)	**PMIS-15**	**96.1**	**(0.5)**	NMIS-5	95.2	(0.7)
NC-10	85.9	(2.3)	NC-15	89.9	(3.8)	NC-10	93.9	(2.4)	**NMIS-5**	**96.0**	**(0.7)**	PMIS-15	95.1	(0.6)
NC-15	85.7	(2.4)	NC-5	89.4	(3.5)	NC-15	93.8	(2.5)	PMI-15	95.3	(0.8)	**NMI-10**	**94.3**	**(2.4)**
SID-2	85.4	(1.8)	NC-20	88.5	(3.8)	SID-4	93.6	(1.2)	NMI-5	95.3	(0.6)	**PMI-20**	**94.3**	**(2.4)**
NC-20	85.0	(2.5)	SID-2	88.4	(1.8)	GSID-4	93.6	(1.2)	PMIS-10	95.2	(1.4)	NMI-5	94.3	(2.1)
Domain 2														
GID-4	73.9	(8.1)	SID-16	71.4	(8.3)	**SID-16**	**81.5**	**(7.9)**	**GID-4**	**81.3**	**(6.9)**	**GID-4**	**82.2**	**(6.1)**
SID-16	72.0	(9.8)	GID-4	69.3	(8.3)	**GID-4**	**78.7**	**(10.0)**	**SID-16**	**81.0**	**(7.9)**	**SID-16**	**81.3**	**(7.7)**
NC-15	71.3	(12.7)	NC-20	65.2	(10.8)	**SID-8**	**77.1**	**(7.7)**	**NMIS-5**	**78.5**	**(5.2)**	**GID-8**	**80.1**	**(9.7)**
GID-8	69.9	(9.6)	NC-25	64.7	(10.4)	**NC-20**	**76.1**	**(14.1)**	**GID-8**	**78.1**	**(10.9)**	**SID-8**	**77.0**	**(7.6)**
NC-20	69.9	(12.2)	NC-15	64.6	(11.3)	**NC-15**	**75.5**	**(13.5)**	**NC-15**	**77.7**	**(12.3)**	**NC-20**	**76.4**	**(11.3)**

Table 5
Feature selection results of tested classifiers in both domains. AUC performances (×100) are averaged over folds and the corresponding standard deviation in parentheses. Bold: best performances and performances that are not significantly worse than the best (two-sided paired *t*-test $\alpha = 0.05$).

Classifier	AUC		#	Feature subset
Domain 1				
LBN	93.7	(2.5)	3	NC-10, PMI-10, PMIS-15
LFD	**97.7**	**(0.7)**	20	SID-4, SID-8, SID-16, GSID-16, NC-15, NMI-10, NMI-15, NMI-20, NMI-35, NMI-40, PMI-20, NMIS-5, NMIS-10, NMIS-15, NMIS-30, NMIS-35, PMIS-5, PMIS-20, PMIS-25, PMIS-30
kNN	97.3	(1.1)	15	SID-8, NMI-5, NMI-10, NMI-20, NMI-25, PMI-15, PMI-20, PMI-5, PMI-10, NMIS-5, NMIS-10, NMIS-20, PMIS-10, PMIS-20, PMIS-30
RBFSVM	**97.7**	**(0.9)**	17	GID-8, NMI-10, NMI-30, PMI-10, PMI-15, PMI-20, PMI-25, PMI-35, NMIS-5, NMIS-10, NMIS-25, NMIS-30, NMIS-40, PMIS-15, PMIS-20, PMIS-25, PMIS-30
AB	96.8	(0.6)	15	SID-4, SID-8, SID-16, GID-2, NC-10, NC-15, NC-40, JM-0, JM-16, NMI-20, PMI-40, PMIS-5, PMIS-15, NMIS-30, NMIS-40
RF	**97.7**	**(0.5)**	17	SID-1, SID-2, SID-4, SID-8, SID-16, GSID-4, GSID-8, NC-5, NC-10, NC-15, NC-20, NC-25, NMI-20, NMIS-5, NMIS-10, PMIS-10, PMIS-15
Domain 2				
LBN	78.2	(8.9)	5	SID-16, GID-4, NC-20, NC-35, NMIS-10
LFD	**94.1**	**(4.1)**	20	SID-0.5, SID-2, SID-16, GSID-1, GSID-2, GSID-8, GID-2, GID-4, NC-10, NC-15, NC-40, JM-8, JM-16, NMI-20, PMI-5, PMI-10, NMIS-15, NMIS-20, NMIS-30, PMIS-10
kNN	92.8	(4.7)	12	SID-16, GID-8, NC-10, NC-15, NC-25, NMI-25, PMI-5, PMI-30, PMI-40, PMIS-5, PMIS-10, PMIS-25
RBFSVM	**95.4**	**(4.2)**	11	SID-1, SID-16, GID-4, NC-15, NC-25, NC-35, NC-40, PMI-5, PMI-10, PMI-15, NMIS-5
AB	93.5	(4.0)	15	SID-4, SID-16, GSID-2, GSID-8, GID-1, GID-4, GID-8, NC-15, NC-25, NC-40, JM-4, PMI-5, NMIS-5, NMIS-10, NMIS-25
RF	**95.3**	**(3.4)**	14	SID-8, SID-16, GSID-16, GID-4, GID-8, NC-15, NC-20, NC-25, NC-30, NC-35, NC-40, JM-0, NMIS-5, PMIS-15

mains are similar, it can be concluded that the classification challenge in Domain 2 is not due to interclass imbalances. However, a striking difference between both domains is visible in the data distribution of the alignment space (Fig. 2). Class WA samples are spread over a relatively large range compared to class CA and PA. If similar distributions are present in the feature space then class WA is sparsely populated and probably not well represented close to the class border.

5.3. Application example

In this subsection we briefly describe two possible applications for which the proposed system can be directly utilized.

Computer-aided inspection of medical image data from image registration: For every unseen image registration a set of landmarks is determined, image features are extracted and alignment quality is assessed by the proposed multi-feature two-stage classification. Employing the physical coordinates of the landmark correspondences $(L_F, \mathbf{T}(L_F))$ and the results of the classification, registration quality can be visualized as shown in Fig. 3, which contains a few exemplary cases to illustrate the application. Having this application available, the manual inspection of clinical images can be aided, as the application can directly point to wrongly aligned locations.

Registration quality assurance: In this application scenario the general idea is to conduct local quality assessment by means of the proposed method. The classification results – predicted frequencies per alignment category – can be compared, e.g. against a previously established reference of the applied registration algorithm to assure normal registration results (comparison by means of statistical summary, such as mean and standard deviation of spatial accuracy). If abnormal registration is detected then an automatic re-registration with different settings could be triggered, a manual matching could be performed, or at least a warning could be issued. In the clinical environment, automatic quality estimation could be used within a clinical reference setting to assure proper registration quality for certain clinical tasks.

5.4. Limitations

Image information. The proposed method for registration quality assessment is based on assessment of the spatial mapping accuracy by means of landmark samples. Hence we analyze structures that are visible in an intensity image. First, there are regions between landmarks that are not sampled and not explicitly assessed. Second, most of the CT lung volume appears homogeneous and structureless but actually consists of small anatomical structures (lobules, bronchioles, alveoli). Third, on top of technical constraints imaging quality is diminished by practical clinical constraints, such as low-dose scanning, acquisition time and patient movement. The first issue can be alleviated: given a sufficiently large number of well-distributed landmarks, registration accuracy between landmarks can be statistically interpolated (cp. Muenzing et al., 2012). Remaining issues are subject of image acquisition and are not explicitly solvable in the scope of image registration.

Alignment measure. In the proposed framework the registration error is defined as landmark translation error, i.e. it does not account for local orientation or scale. The employed translational error measure works well for lung CT and brain MRI registrations (Murphy et al., 2011). If one assumes some sort of local shape regularization inherent in medical structures, then a local orientation error will also result in a translational error, if not at the same position then at least in the close vicinity. However, there might be applications where this assumption does not hold and in such cases the reference landmark matching needs to be extended to include orientation measures, and also additional image features might be necessary to better capture local orientation patterns.

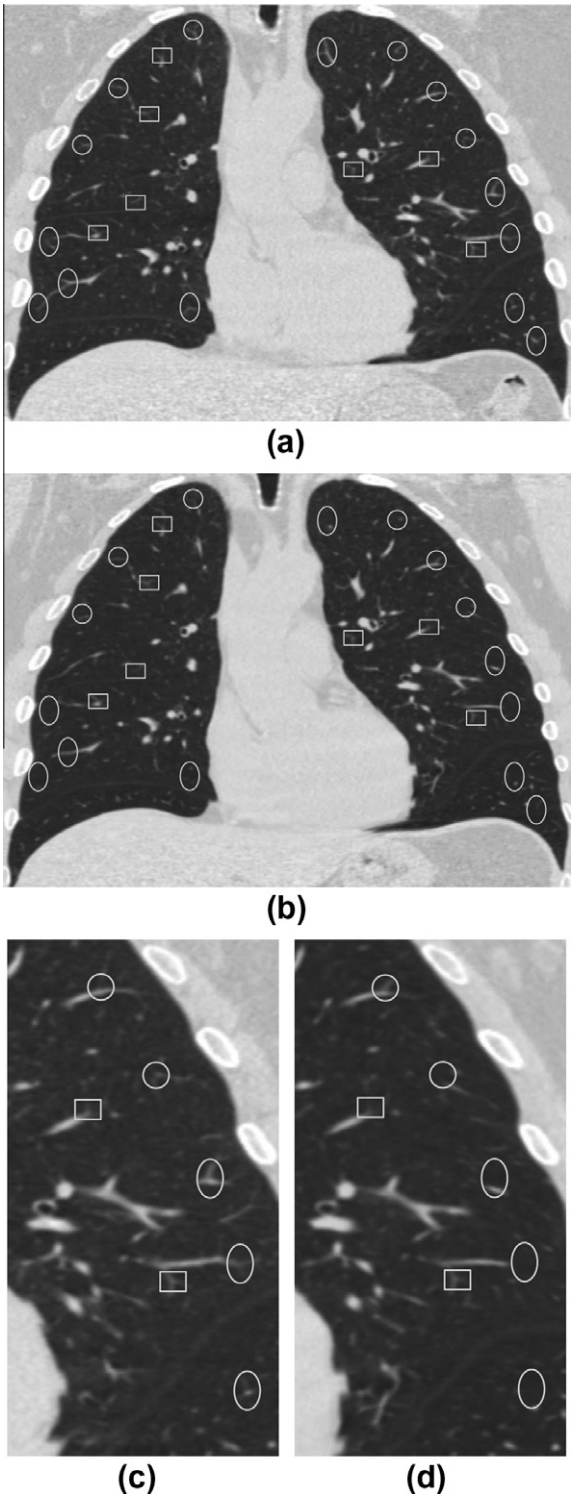

Fig. 3. Examples of supervised quality assessment of lung CT follow-up registrations. The image pair (a), (b) depicts a baseline scan and the corresponding registered follow-up scan. The image pair (c), (d) displays a zoomed view of the scans shown in (a), (b). Alignment categories are marked as follows: CA: rectangle, PA: circle, WA: oval.

Quality assessment. The proposed quality assessment is based on a classification approach, and consequently it defines registration quality on a categorical scale. However, the underlying alignment

problem is measured on a ratio scale, and therefore the introduced quality categories cause a quantization error. In particular there might be a bias towards registration errors just below the quantification threshold. A regression approach appears therefore more suited. This was initially investigated, however experiments with non-linear regressors showed overall unsatisfactory results. In particular alignment samples in the range of WA have pronounced inaccurate estimations (large variance), probably due to the more complex image patterns for larger misalignments.

Genuine change. The proposed supervised learning approach does not differentiate between misalignments that are purely the result of errors produced by the registration algorithm and those misalignments that are caused by alterations in morphology or anatomy, so-called genuine changes. Geometrically speaking, genuine changes constitute inconsistencies in topology. To correctly match genuine changes an image registration algorithm would need to know how to match those particular changes onto unaltered image structures. Given a proper set of alignment samples, a learning-based approach might be used to aid automatic image registration so that areas of genuine change are transformed more appropriately. We therefore consider both, the detection of genuine change and the physiological valid matching of those areas an interesting topic for future work.

5.5. Generic framework

The method described in this paper is optimized for CT lung scans but we expect the proposed framework is sufficiently generic to be adapted to other anatomical regions and to other imaging modalities. In general this requires only a re-training on the new application data. In cases where there is no ground truth data available for a set of sample registrations, a reference standard needs to be established first. This might seem cumbersome but can be done efficiently, aided by a semi-automatic software tool such as described in Murphy et al. (2011). However, it is not certain that it will be possible to detect misregistrations for every single application. For another application it might be necessary to adapt the automatic landmark generation to achieve a sufficient coverage of the anatomy of interest (Murphy et al., 2011). And other image features might be necessary (or most effective) for other applications.

Image features. All scales and volume sizes are based on the image resolution and the ultimate parameter value is automatically determined within the optimization procedure (feature selection). The range of these scales should be adapted, in general it is just necessary to have a sufficient collection. In multi-modality registration the images are acquired from different modalities, and therefore particular structures in one image are usually described by other intensity values in the other image. That means the residual image of two correctly registered images is unequal zero. Still, the residual intensities at anatomical structures might be distinctive from their surroundings and therefore could be described well by the proposed intensity-based features. However, adaptation of the proposed framework to specific multi-modality registrations is subject to further investigation.

Meta-learning. The alignment samples of RRES resemble a unimodal skew distribution which consequently resulted in imbalanced class priors. Generally, in binary classification problems where only few examples of one class exist, the classification problem becomes similar to a so-called one-class problem. Such circumstances give rise to a more difficult classification problem which is susceptible to prediction bias of classification accuracy. For that reason we used the AUC measure for classification validation which is insensitive to class imbalances. In general, the classification accuracy on unseen data is influenced by the class priors of the training set, that is, the best classification performance is to be expected if the training set represents the class distribution of the unseen data. We evaluated predicted class labels of the classifier cascade separately for each registration subset (cf Table 3) and report the following classification accuracies: Freeform 98%, Affine 88%, Similarity 85%. The average of 90% corresponds to the reported overall accuracy calculated from the confusion matrix in Table 3. In conclusion it can be stated that the reported overall classification accuracy likely underestimates the performance of the developed classifier when a freeform registration algorithm is applied on unseen scan pairs (assuming that the set of 51 scan pairs represents the problem domain adequately). However, registration accuracy may vary between different algorithms and scan pairs. A comprehensive database (RRES) is therefore indispensable for the development of a robust classification system. Interestingly, further optimization of the proposed system is possible by automatic adaptation of the trained classifier to unseen data, e.g. by incorporation of predicted class priors into the classification method (Jin and Liu, 2005). However, we believe that the performance of proposed system is in the first place limited by the underlying accuracy of manual landmark annotations, which in itself is limited by the image resolution and image quality. Further, we assume that such uncertainties are not limited to lung CT imaging but rather common to clinical reference standard data. In this work the aim has been to show the principle of employing pattern recognition to the problem domain of quality assessment in medical image registration.

6. Conclusion

The paper introduces a novel method for automatic assessment of registration quality in medical images. Supervised learning is employed to distinguish local alignment patterns, which are captured by statistical image features at distinctive landmark points. The method is trained and tested on 51 CT follow-up scan pairs of the lung by patient-entity preserving 10-fold cross-validation. A two-stage classifier cascade, employing an optimal multi-feature model, classifies registrations locally into three quality categories (correct, poor or wrong alignment). Feature selection is conducted and the established optimal multi-feature model is validated against simple single-feature classifier models to simulate a thresholding approach which is current state-of-the-art practice (Isgum et al., 2009). The proposed two-stage multi-feature classifier yields an accuracy of 90% combined over all alignment categories, and thereby outperforms common approaches, comprising high detection accuracy combined with categorized registration assessment. Although the method is only demonstrated on CT lung images, we expect that the generic setup allows adaptation to other applications.

References

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Computation 9, 1545–1588.
Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning 36, 105–139.
Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145–1159.
Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.
Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.
Breiman, L., 2002. Manual on setting up, using, and understanding random forests v3.1.
Castillo, R., Castillo, E., Guerra, R., Johnson, V.E., McPhail, T., Garg, A.K., Guerrero, T., 2009. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Physics in Medicine and Biology 54, 1849–1870.
Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/cjlin/libsvm>.
Crum, W.R., Griffin, L.D., Hawkes, D.J., 2004. Automatic estimation of error in voxel-based registration. In: Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 821–828.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. John Wiley and Sons, New York.
Duin, R., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D., Verzakov, S., 2007. PRTools4.1, A Matlab Toolbox for Pattern Recognition, Delft University of Technology.
Fedorov, A., Billet, E., Prastawa, M., Gerig, G., Radmanesh, A., Warfield, S.K., Kikinis, R., Chrisochoides, N., 2008. Evaluation of brain MRI alignment with the robust hausdorff distance measures. In: International Symposium on Visual Computing. LNCS, vol. 5358. Springer, pp. 594–603.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.
ter Haar Romeny, B., 2009. Front-End Vision and Multi-Scale Image Analysis: Multi-scale Computer Vision Theory and Applications. Springer, written in Mathematica.
Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843.
Hill, D.L., Batchelor, P.G., Holden, M., Hawkes, D.J., 2001. Medical image registration. Physics in Medicine and Biology 46, R1–45.
Ho, T.K., 1995. Random decision forest. In: 3rd International Conference on Document Analysis and Recognition, pp. 278–282.
Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 832–844.
Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion – application to cardiac and aortic segmentation in CT scans. IEEE Transactions on Medical Imaging 28, 1000–1010.
Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 153–158.
Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 4–37.
Jin, R., Liu, Y., 2005. A framework for incorporating class priors into discriminative classification. In: Ho, T., Cheung, D., Liu, H. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer, pp. 401–412.
Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J., 2010. elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29, 196–205.
Klein, S., Staring, M., Pluim, J.P.W., 2007. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. Transactions on Medical Imaging 16, 2879–2890.
Koenderink, J.J., 1984. The structure of images. Biological Cybernetics 50, 363–370.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: 14th International Joint Conference on Artificial Intelligence. Morgan Kaufman, n, pp. 1137–1143.
Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.
Kotsiantis, S., 2011. Combining bagging, boosting, rotation forest and random subspace methods. Artificial Intelligence Review 35, 223–240.
Leow, A.D., Yanovsky, I., Chiang, M.C., Lee, A.D., Klunder, A.D., Lu, A., Becker, J.T., Davis, S.W., Toga, A.W., Thompson, P.M., 2007. Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration. IEEE Transactions on Medical Imaging 26, 822–832.
Lester, H., Arridge, S., 1999. A survey of hierarchical non-linear medical image registration. Pattern Recognition 32, 129–149.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.
Lindeberg, T., 1993. Discrete derivative approximations with scale-space properties: a basis for low-level feature extraction. Journal of Mathematical Imaging and Vision 3, 349–376.
Maier, D., 1983. The Theory of Relational Databases. Computer Science Press.
Maintz, J.B.A., Viergever, M.A., 1998. A survey of medical image registration. Medical Image Analysis 2, 1–36.
Möller, B., Posch, S., 2008. An integrated analysis concept for errors in image registration. Pattern Recognition and Image Analysis 18, 201–206.
Muenzing, S.E.A., Murphy, K., van Ginneken, B., Pluim, J.P.W., 2009. Automatic detection of registration errors for quality assessment in medical image registration. In: Proceedings of the SPIE, pp. 72590K–72590K–9.
Muenzing, S.E.A., van Ginneken, B., Pluim, J.P.W., 2012. On combining algorithms for deformable image registration. In: Proceedings of the WBIR, vol. 7359, pp. 256-265.
Murphy, K., van Ginneken, B., Klein, S., Staring, M., de Hoop, B., Viergever, M., Pluim, J.P., 2011. Semi-automatic construction of reference standards for evaluation of image registration. Medical Image Analysis 15, 71–84.
Osborne, J.W., Carolina, N., 2010. Improving your data transformations: applying the box-cox transformation. Practical Assessment Research Evaluation 15, 2.
Park, H., Bland, P.H., Brock, K.K., Meyer, C.R., 2004. Adaptive registration using local information measures. Medical Image Analysis 8, 465–473.
Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A., 2003. Mutual-information-based registration of medical images: a survey. IEEE Transactions on Medical Imaging 22, 986–1004.
Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research 3, 1371–1382.
van Rikxoort, E., de Hoop, B., Viergever, M., Prokop, M., van Ginneken, B., 2009. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. Medical Physics 36, 2934–2947.
Sakia, R.M., 1992. The box-cox transformation technique: a review. The Statistician 41, 169.
Schapire, R.E., 2002. The boosting approach to machine learning: an overview. In: Nonlinear Estimation and Classification. Springer.
Sofka, M., Stewart, C.V., 2008. Location registration and recognition (LRR) for longitudinal evaluation of corresponding regions in CT volumes. In: Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 989–997.
Sturges, H.A., 1926. The choice of a class interval. Journal of the American Statistical Association 21, 65–66.
Tieu, K., Viola, P., 2000. Boosting image retrieval. International Journal of Computer Vision, 228–235.
Toussaint, G., 1974. Bibliography on estimation of misclassification. IEEE Transactions on Information Theory 20, 472–479.
Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii.
Viola, P., Jones, M., 2004. Robust real-time face detection. International Journal of Computer Vision 57, 137–154.
Xu, D., Gietema, H., de Koning, H., Vernhout, R., Nackaerts, K., Prokop, M., Weenink, C., Lammers, J., Groen, H., Oudkerk, M., van Klaveren, R., 2006. Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 54, 177–184.