Artificial Intelligence in Medicine 60 (2014) 151–163

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Intensity-based image registration using scatter search

Andrea Valsecchi [a,*], Sergio Damas [a], José Santamaría [b], Linda Marrakchi-Kacem [c,d]

[a] Applications of Fuzzy Logic and Evolutionary Algorithms Research Unit, European Centre for Soft Computing, Calle Gonzalo Gutiérrez Quirós S/N, 33600 Mieres, Spain
[b] Department of Computer Science, University of Jaén, Edificio Tecnológico y de Ingenierías A-3, Paraje Las Lagunillas S/N, 23071 Jaén, Spain
[c] NeuroSpin, French Alternative Energies and Atomic Energy Commission, Bâtiment 145, Centre d'études de Saclay, 91191 Gif-sur-Yvette, France
[d] Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière, Hôpital Pitié Salpêtrière, Boulevard de l'Hôpital 47, 75013 Paris, France

ARTICLE INFO

Article history:
Received 6 November 2012
Received in revised form 13 January 2014
Accepted 28 January 2014

Keywords:
Global optimization
Heuristics
Scatter search
Image registration
Atlas-based segmentation
Magnetic resonance imaging

ABSTRACT

Objective: We present a novel intensity-based algorithm for medical image registration (IR).
Methods and materials: The IR problem is formulated as a continuous optimization task, and our work focuses on the development of the optimization component. Our method is designed over an advanced scatter search template, and it uses a combination of restart and dynamic boundary mechanisms integrated within a multi-resolution strategy.
Results: The experimental validation is performed over two datasets of human brain magnetic resonance imaging. The algorithm is evaluated in both a stand-alone registration application and an atlas-based segmentation process targeted to the deep brain structures, considering a total of 16 and 18 scenarios, respectively. Five established IR techniques, both feature- and intensity-based, are considered for comparison purposes, and ground-truth data is used to quantitatively assess the quality of the results. Our approach ranked first in both studies and it is able to outperform all competitors in 12 of 16 registration scenarios and in 14 of 18 registration-based segmentation tasks. A statistical analysis confirms with high confidence ($p < 0.014$) the accuracy and applicability of our method.
Conclusions: With a proper, problem-specific design, scatter search is able to provide a robust, global optimization. The accuracy and reliability of the registration process are superior to those of classic gradient-based techniques.

1. Introduction

In its most general formulation, image registration (IR) [1] is the task of aligning two or more images in order to establish a spatial correspondence of their common content. Such images usually have the same or a similar subject but have been acquired under different conditions, such as time and viewpoint, or by multiple sensors. In medical image analysis, IR is a key technology that allows to "fuse" visual information from different sources [2]. Applications include combining images of the same subject from different modalities, aligning temporal sequences of images to compensate for motion between scans, image guidance during interventions and aligning images from multiple subjects in cohort studies. The remarkable developments in medical imaging technology over the last decades determine a constant demand for better image processing and analysis techniques. Dealing with novel, more diverse, and increasingly accurate sources of imaging data is the main challenge in IR and it explains why it is still a very active research field.

The alignment between two images is specified as a spatial transformation, mapping the content of one image to the corresponding area of the other. A popular strategy among IR methods is to perform the alignment by considering only salient and distinctive parts of the image, such as lines, corners and contours, called *features*. This strategy has the advantage of greatly reducing the complexity of the problem, but relies on the ability to detect the features correctly. However, this approach is limited to the cases in which features alone are able to characterize the image content. IR methods following this approach are called *feature-based* [2,1], while the term *intensity-based* (or voxel-based) names the methods in which the whole image data is used.

Regardless of this division, the core of every IR technique is an *optimization process* that explores the space of geometrical transformations. Two strategies are available. In *parameters*-based approaches the search is directly performed in the space of the

* Corresponding author. Tel.: +34 644301318.
E-mail addresses: andrea.valsecchi@softcomputing.es, valsecchi.andrea@gmail.com (A. Valsecchi), sergio.damas@softcomputing.es (S. Damas), jslopez@ujaen.es (J. Santamaría), linda.marrakchi@gmail.com (L. Marrakchi-Kacem).

http://dx.doi.org/10.1016/j.artmed.2014.01.006
0933-3657/

transformation parameters. Hence, a solution is a vector of values for the parameters of the registration transformation. In *matching*-based approaches, features are matched through a search in the space of feature correspondences; once a suitable matching has been found, the transformation parameters are derived accordingly by numerical methods. In both cases the search is guided by a *similarity metric*, a function that measures the degree of resemblance between the input images. This can be done either by comparing the whole images or just their corresponding features. Traditional parameters-based methods use classic gradient-based optimization algorithms, while matching-based methods use matching algorithms like iterative closest point (ICP) [3].

Many features of IR problems, such as noise, discretization and differences in the order of magnitude of the transformation parameters still pose a challenge to traditional optimization methods. A number of alternative approaches are based on *metaheuristics* [4], which have proven their ability to deal with complex real-world problems in a large number of fields, including computer vision and image processing. In particular, metaheuristic-based registration approaches have demonstrated to be a promising solution to overcome the drawbacks of traditional optimization algorithms [5,6]. Scatter search (SS) [7–9] is a prominent example of such techniques. It has already been applied to image registration problems as optimizer of feature-based approaches. In IR, SS has been successful both when used to find matchings among features [10] as well as in searching for the transformation parameters directly [11].

In this work, SS is used as base for a novel intensity-based IR method. The algorithm is specifically designed to take advantage of the characteristics of the IR process to improve the optimization. To evaluate its effectiveness, our method is compared with an heterogeneous group of competitors in two experimental studies involving simulated and real medical images. A thorough analysis of the results is performed, and their significance is assessed by means of different statistical tests.

The paper is structured as follows. In Section 2, we review the image registration problem and present several techniques to solve it. Section 3 introduces the basics of SS and the design of our IR method. In Section 4, we present the experimental studies along with the analysis of their results. Finally, conclusions are provided in Section 5.

2. Image registration

A typical IR problem involves two images, conventionally called *model* (I_M) and *scene* (I_S), with different roles in the registration process. The model is the reference (or target) image, while the scene is the image that is transformed to reach the geometry of the other. The registration aims to find a geometric transformation f that aligns the scene to the model; in other words, f is such that the model I_M and the transformed scene $f(I_S)$ are as similar as possible.

Several components characterize an IR method. First we have the *transformation model*, that determines which kind of transformation can be used to align the images. This choice depends entirely on the concrete application; very simple models such as translation transform can be enough in certain contexts such as remote sensing [12]. At the other end of the spectrum there are *non-rigid* (also called elastic) transformations, such as B-spline and thin-plate splines transformations, able to represent local deformations (warpings) using hundreds or even thousands of parameters. Other common choices include rigid transform, which allows translation and rotation, similarity transform, which also admits scaling, and affine transformation, which can also represent shearing. These are examples of global transformations having respectively 6, 7 and 12 degrees of freedom for 3D images.

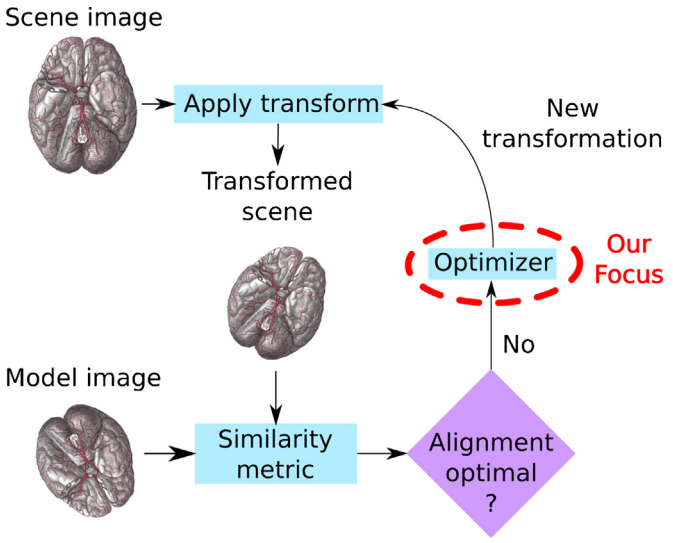


Fig. 1. The interactions among the components of a registration technique.

The second component of any IR method is the *similarity metric*, a function $F(I_1, I_2)$ that measures the degree of resemblance between two images. As the final performance of any IR method depends on the accurate estimation of the alignment of the images, this is a crucial IR component [13]. The quality of a transformation f is obtained by computing the similarity metric over the model I_M and the transformed scene $f(I_S)$. The actual evaluation mechanism depends on the nature of the registration approach. In feature-based methods the similarity metric usually measures the distance between corresponding features [14]. For instance, the alignment can be evaluated using mean square error (MSE) between the points of the model and those of the transformed scene. If the model has r feature points, each point x_i is assigned to the closest point in the transformed scene c_i, and the MSE is given by:

$$\mathrm{MSE} = \frac{1}{r}\sum_{i=1}^{r}\|x_i - c_i\|^2$$

In intensity-based approaches, instead, the resemblance of the intensity values in the two images are considered. Sum of squared differences (SSD), normalized correlation (NC) and mutual information (MI) [15,16] are typically used. The subject of the images and their acquisition technique determine the kind of the relationship between the intensity distributions, which in turn decides what similarity metrics are appropriate. For instance, when two images have been acquired using different sensors, a scenario called *multi-modal* registration, the relationship between the intensity values in the images can be strongly non-linear. While NC can handle a linear relationship, metrics based on information theory, such as MI, are better suited for this scenario. MI is defined as

$$\mathrm{MI} = \sum_{s\in L_S, m\in L_M} p(m,s,f)\log_2\frac{p(m,s,f)}{p_M(m)\ \ p_S(s,f)}$$

where L_M and L_S are sets of regularly spaced intensity bins centers, p is the discrete joint probability and p_M, p_S are the marginal discrete probabilities of the model and scene images.

The third main component of an IR method is the *optimizer*. It is responsible for finding the best transformation, in terms of similarity metric, among the transformations in our transformation model. Fig. 1 shows the flow chart of the whole registration process.

Each optimizer has a different search strategy, which depends also on the nature of the algorithm. One approach is to perform the search directly in the space of the transformation parameters.

Table 1
The IR algorithms included in our experimental study.

	Nature	Strategy	Optimizer
I-ICP	Feature	Matching	Gradient descent
Dyn-GA	Feature	Parameters	Genetic algorithms
SS*	Feature	Matching	Scatter search
ASGD	Intensity	Parameters	Gradient descent
GA+	Intensity	Parameters	Genetic algorithms
SS+	Intensity	Parameters	Scatter search

This turns the registration in a continuous optimization problem, therefore classic numerical optimization algorithms can be used. Gradient descent, Newton's method, Powell's method and discrete optimization [17] are among the most common choices along with approaches based on evolutionary computation (EC) and other metaheuristics [18–26]. IR algorithms that follow this approach are then called *parameter-based*. An alternative approach consists in searching for a matching between features or areas of the image. From the match one can derive the parameters of the corresponding transformation using numerical methods. This class of algorithms is called *matching-based*. The iterative closest point algorithm is a famous example following the latter approach [27–29]. ICP uses the closest-point assignment rule (i.e. a model point is assigned to the closest transformed scene point) to perform the match. Once a match is established, the corresponding transformation is computed using least squares estimation or other more robust model fitting techniques [30].

Finally, we mention two minor components that play a role in the registration. In intensity-based methods, computing the similarity metric on the whole images is usually unfeasible and unnecessary, therefore a *sampling strategy* determines how many and which voxels are actually used. Those are usually selected at random with uniform probability or sampled along a regular grid.

Second, it is common to perform the registration in multiple stages. Increasingly larger and more detailed versions of the input images are used at each stage. In the first stage, the optimizer finds a solution to a coarse version of the registration problem. In each of the further resolutions, part of the details of the input images is restored and the optimizer aims to adapt the solution of the previous phase to fit the new, more detailed data. The *multi-resolution strategy* determines which kind of processing is performed on the images in each different stage of the registration; usually the procedure includes down-sampling and smoothing. The sequence of images used during the registration is called *pyramid*.

In the rest of this section we present a selection of remarkable IR algorithms that will be later compared in an experimental study. The group is quite heterogeneous in terms of nature of the approaches, search strategies and optimizers (see Table 1), encompassing most of the combinations of approaches presented in the previous section. We begin with two modern versions of classic IR methods and then move to algorithms based on EC and other metaheuristics.

I-ICP. This extension of the original ICP proposal has been introduced in Liu [28]. In comparison with its predecessor, I-ICP introduces the use of collinearity to assess the quality of a matching. Consider two points x, y of the model and their corresponding points x', y' in the scene. If the matching is good, x' and y' should both be close to the straight line l passing through x, y. The distances $d(x', l)$ and $d(y', l)$ provide additional information on the matching that I-ICP exploits to find better solutions. Moreover, the algorithm has a more complex termination mechanism. Once the algorithm has converged to a solution, a tiny, random perturbation is applied to it, then the algorithm continues. This step is repeated until the algorithm converges to the same solution found before the perturbation. This helps I-ICP to traverse local optima, increasing its robustness.

Adaptive stochastic gradient descent. Adaptive stochastic gradient descent (ASGD) [31] is an optimization method designed for intensity-based IR. The algorithm is an extension of the Robbins–Monro stochastic gradient descent method in which some of the parameters are automatically computed, in particular the step size. In the comparison of intensity-based IR methods presented in [32], ASGD outperformed the other algorithms in both affine and elastic registration problems. The algorithm uses a random image sampler and a multi-resolution strategy in which the images in the pyramids are obtained by applying both downsampling and Gaussian smoothing.

Dyn-GA. Dyn-GA [22] is a parameter-based IR technique based on a real-coded Genetic Algorithm (GA) [33]. The algorithm uses fitness-proportionate selection and a novel crossover operator that swaps a number of genes between the two individuals. The mutation replaces the value of a randomly selected gene; the new value is drawn from a range that depends on the fitness of the individual: the larger the fitness, the larger the range. Also, as the optimization progresses, the ranges of transformation parameters are restricted around those of the individuals in the population, so that the search is focused near the current solutions.

GA+. GA+ has been proposed in [34] as an intensity-based method whose optimizer is a GA. The registration is performed through a search in the space of transformation parameters, therefore the GA is real-coded and uses real-coded operators such as BLX-α crossover. GA+ can handle a number of transformation models and similarity metrics. It also support multiple resolutions. A restart mechanism is used at the end of the first resolution. If the fitness of the best individual is greater than a fixed threshold, the registration moves to the second resolution, otherwise the first resolution is performed again. Experiments showed this mechanism helps the algorithm to achieve a better performance at a low extra computational cost.

SS.* SS* [10] is a matching-based approach using the SS optimization algorithm (see Section 3). The proposal is specifically designed to work with a class of features, crest lines points, and exploits the knowledge of the local curvature of the points to perform the matching. In addition, the authors proposed an advanced coding scheme, in which a matching is represented as a permutation of points, and a novel design for some of the components of SS.

3. Scatter search design for intensity-based IR

SS was originally proposed by Glover [7] in the context of integer programming. The main idea behind SS is to recombine systematically a *reference set* of high-quality solutions, possibly obtained using different techniques. Most SS implementations to date have been based on the SS *templates* presented in [8,9]; our exposition follows the well-known five-methods template introduced in the latter. These methods are:

- a *diversification generation method* that generates a set of diverse trial solutions;
- an *improvement method*, used to enhance a solution, usually applying an heuristic method;
- a *reference set update method* to build and maintain a reference set of solutions selected for their quality or diversity;
- a *subset generation method*, by which sets of solutions from the reference set are created;
- a *solution combination method* that combines a set of solutions into one or more new solutions.

An outline of the complete SS procedure [9] is shown in Algorithm 1. At the beginning, a number of solutions are created,

improved and stored in a temporary container P of size $PSize$. The reference set is then updated by selecting the best solutions between P and the reference set itself. Next, the process enters a loop (line 8) that iterates the core SS procedure until a stopping condition is met, controlling the duration of the optimization. Then, an inner loop begins. First, subsets of solutions from the reference set are created. Each subset is combined into a new solution, which is then improved and stored in another container, *Pool*. Next, the reference set is updated with the best solutions among those in *Pool* and the current reference sets. If no new solution has entered the reference set, the inner loop ends. Finally, a new set P is created as in the beginning of the algorithm.

A popular variant of the canonical SS design is the *2-tier* design [9], in which the reference set is divided in two tiers. One tier, called the *quality* reference set, stores the b_1 most high quality solutions, while the other, the *diversity* reference set, contains the b_2 solutions having high diversity with respect to those in the first tier. Each tier is ordered according to either quality or diversity, and during the reference set update procedure the new solutions having the highest values of quality or diversity are considered for inclusion into the reference set. The advantage of this design is that it adds diversity to the search instead of focusing only around the best solutions found, avoiding premature convergence.

Algorithm 1 *(A generic optimization procedure based on SS.).*

```
1   ReferenceSet ←∅;
2   P ←∅;
3   While |P| < PSize do
4     x ← DiversificationGeneration();
5     x′ ← ImprovementMethod(x);
6     if x′ ∉ P then add x′ to P
7   end
8   While ¬ stop condition do
9     Update the ReferenceSet by selecting the best b solutions in
      ReferenceSet ∪ P;
10    NewElements ← true;
11    Pool ←∅;
12    While NewElements do
13      Subsets ← SubsetGeneration();
14      NewElements ← false;
15      While Subsets ≠ ∅ do
16        S ← pop(Subsets);
17        x ← SolutionCombination(S);
18        x′ ← ImprovementMethod(x);
19        If x′ ∉ Pool bf then Add x′ to Pool
20      end
21      Update the ReferenceSet by selecting the best b solutions in
      ReferenceSet ∪ Pool;
22      IfReferenceSet has new elements then
23        NewElements ← true
24      end
25      Build a new set P using the diversification generation and
      improvement methods;
26    end
27  end
```

In what follows we present the design of the proposed IR method, called SS$^+$. The effectiveness of any SS implementation depends on a proper design of the previous five methods that must be specific to the optimization problem at hand. As we reviewed in Section 2 the main components of an IR approach are the transformation model, the similarity metric and the optimizer. In general, the choice of the transformation model and the similarity metric are highly problem-dependent. Our method has been designed to support multiple transformation models. SS$^+$ is able to handle all *parametric* transformation models, from the simple Euler transform up to B- and thin-plate splines. Regardless of the concrete transformation model, a solution is stored as a vector of real values, corresponding to its parameters. In medical imaging the images also store the physical size of the voxels, so we can encode some of the transformation parameters in physical units, e.g. translations can be specified in millimeters rather than in number of voxels. Analogously, different similarity metrics are supported and its choice is left to the user. In the algorithm, the actual similarity metric value is directly used as fitness.

Our efforts have been devoted to the design of the optimizer component. Our proposal is based on a SS algorithm with 2-tier design. First, we present the five SS methods especially designed to deal with IR as a continuous optimization problem.

Diversification generation method. We adopt an approach based on frequency memory [35] to ensure the search space is explored in a uniform manner. The range of each transformation parameter is divided in four sub-ranges of equal size and a frequency counter is associated to each of them. A solution is built in two steps. First, for each transformation parameter, a sub-range is chosen at random with a probability inversely proportional to its frequency count. Then, the actual values of the parameters are selected at random within the selected sub-ranges with uniform probability. Finally, the corresponding frequency counters are increased.

Subset generation method. Our method generates subsets having two elements. First, the solutions from the quality reference set are considered, yielding $b_1(b_1-1)/2$ subsets. Then, the diversity reference set is used, so that other $b_2(b_2-1)/2$ subsets are generated. Finally, we combine solutions from the quality and diversity reference set, adding $b_1 b_2/2$ sets to the result.

Solution combination method. To combine solutions we use the BLX-α crossover operator [36], a common choice in real-coded Evolutionary Algorithms. Two solutions x and y are provided as input. For each position i of their encoding, the operator computes the value $d_i = |x_i - y_i|$ and then it randomly generates a value z in the interval $[\min(x_i, y_i) - \alpha d, \max(x_i, y_i) + \alpha d]$ with uniform probability. The value z is assigned to the ith position of the resulting solution. The parameter α controls the width of the ranges in which the new values are drawn.

Improvement method. Our improvement method is a local search based on a crossover operator. To improve a solution x, we perform crossover with a randomly chosen solution y from either the reference set or P. This results in a number of new solutions $z_1, \ldots, z_m$, depending on the actual crossover operator in use. The best solution among $x, y, z_1, \ldots, z_m$ is then considered as output, and the whole process is repeated for a number of iterations. Note that if all the solutions generated by the crossover are worse than the original solution, the improvement method simply returns its input x.

We use the "parent-centric" version of the BLX-α crossover called PMX-α [37]. In contrast with the BLX-α, the ranges for the new solutions parameters are $[x_i - \alpha d, x_i + \alpha d]$ and $[y_i - \alpha d, y_i + \alpha d]$, which indeed results in solutions that are closer to their parents. This operator has already been used in IR and it yielded the best results in the comparison of memetic approaches to range IR carried out in [38].

Reference set update method. The update policy aims to maintain the highest quality solutions in the quality reference set and the most diverse solutions in the diversity one. The update is performed in three steps. First, a new quality reference set is created using the b_1 having the highest quality values in both the current quality reference set and the pool. Then, the algorithm computes a *diversity* measure over each solution in the pool and in the diversity reference set. Finally, the b_2 solutions having the highest distance values in both the pool and the diversity reference set are placed in the new diversity reference set.

The diversity value of a solution x measures the distance between the solution and the quality reference set S. It is defined as

the minimum of the square difference between x and each solution y in the quality reference set, as in

$$\text{diversity}(x, S) = \min\{\text{distance}(x, y) | y \in S\} \quad (1)$$

$$\text{distance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad (2)$$

where n is the number of components of a solution.

SS$^+$ also includes a number of specific components that exploit the features of the IR problem and are essential to improve the performance of SS.

Duplication control. A common issue in real-coded evolutionary algorithm is that extremely similar solutions can be generated. Despite being different in their representation, such solutions can be so close to each other that for all purposes they encode the same object. This can cause the reference set to contain only almost identical copies of the same solution, which both strongly focus the search around the current best solution and makes pointless the use of multiple solutions. To avoid this behavior, we utilize a duplication control mechanism. We consider two solutions to be different only if their distance (Eq. (2)) is above a given threshold. This condition is checked before the inclusion of a solution in a set (lines 6 and 19 in Algorithm 1).

Multi-resolution strategy. Before the registration, the input images are processed applying both down-sampling and Gaussian smoothing in order to create two image pyramids. A few operations in the optimizer allow it to handle the change from one resolution to the next. The two reference sets are left untouched; as we encoded transformations using physical units rather than voxels, the solutions are not affected by the change of the image's size. However, the quality of a solution changes between one resolution and the next due to the new details that are now visible in the images, therefore we need to compute the fitness of the solutions again.

Restart mechanism. The motivation behind the use of restart is simple. At the end of the first resolution, the algorithm might have found a very low-quality transformation. Refining such transformation is unlikely to produce a good final solution, therefore it is more appropriate to perform again the search for a suitable initial registration by restarting the algorithm. To check whether the best solution obtained at the end of the first resolution is acceptable or not, one might consider to set a threshold on its fitness value (i.e. its similarity metric). However, the fitness value of an appropriate solution depends on the actual content of the input images and is hard to predict. We adopted an alternative approach. The first resolution is performed a fixed number of times, independently of its outcome. At the end of this process, the best solutions found are considered for the second resolution.

The computational cost of performing a restart at the first resolution is quite low, as the images involved are still small. Note that if p is the probability of obtaining a low quality solution during the first resolution, performing n restarts reduces this probability to p^{n+1}, i.e. an exponential decrease. This observation provides a rule of thumb to estimate an appropriate number of restarts. To bound to the probability of obtaining an inappropriate solution to a value r, just set

$$n \geq \frac{\ln(r)}{\ln(p)} - 1$$

Dynamic boundary. This component is used to further take advantage of the use of multiple resolutions. The initial ranges of the transformation parameters are highly dependent on the application, therefore they should be provided by the user. However, as we change from one resolution to the next one, we can assume that the optimal solution for the next resolution lies in the same area of the search space as the best solution found at the current resolution. Therefore, we can improve the intensification of the search by limiting the range of the transformation parameters inside this area. This approach is called *dynamic boundary* [22]. Let x be the best solution found at the current resolution and l, u be, respectively, the vector of the lowest and highest values in the current ranges. Then, the ranges of the transformation parameters for the next resolution are $[l', u']$ where

$$l_{i'} = \max\{x_i - (u_i - l_i)/\gamma, l_i\}$$

$$u_{i'} = \min\{x_i + (u_i - l_i)/\gamma, u_i\}$$

and $\gamma > 1$ is the shrinking factor. Basically, dynamic boundary restricts the parameter ranges around the best solution, i.e. to a γ-times smaller interval centered around x. Note that even for modest values of γ, this can drastically reduce the size of the search space, depending on the number of transformation parameters. For instance, affine transformation has 12 parameters, meaning that for $\gamma = 2$, dynamic boundary reduces the size of the search space by a factor of $2^{12} = 4{,}096$.

The overall structure of SS$^+$, showing also the specific IR components, is given in Algorithm 2. Duplication control is implemented directly inside the SS procedure. The algorithm begins with the first resolution and switches to the second resolution in line 8. During the change, dynamic boundary is applied (line 10). In the first resolution, the restart mechanism is used (lines 2–7), and the best solutions found during the different runs are saved and later used in the second resolution.

Algorithm 2 *(Pseudocode of SS$^+$, highlighting the integration of the general SS optimization procedure with advanced components specific of IR.).*

```
1   Compute the image pyramids of the input images by applying
    downsampling and Gaussian smoothing;
2   Use the first level of the image pyramids;
3   SavedSolutions ←∅;
4   For r ← 1 To NumberofRestarts do
5     Run the Scatter Search optimization procedure;
6     Add the current solution to SavedSolutions;
7   end
8   Use the second level of the image pyramids;
9   Update fitness value of SavedSolutions;
10  Apply dynamic boundary;
11  Run the Scatter Search placing the SavedSolutions into the initial
    reference set;
```

4. Experimental study

The aim of the experimentation is to carry out an objective comparison of our SS proposal and the other state-of-the-art IR methods described in Section 2. To that end, we designed two experiments involving synthetic and real-world medical images. To make the comparison as objective as possible, in addition to performing a visual comparison of the results, the effectiveness of each method is assessed using a *quantitative* validation measure specific to each experiment. Furthermore, as most of the algorithms involved are of non-deterministic nature, we carried out a number of independent runs on each scenario. Our analysis investigates several aspects of the results. First, we measure the performance of the algorithms on each scenario by computing mean and standard deviation of the validation measure and ranking the algorithms accordingly. Next, we assess the overall performance of the algorithms in two ways: by computing the per-scenario mean rank of each algorithm and by counting the number of scenarios in which one outperforms another, called *wins*.

In the last part of the analysis, statistical tests are performed to determine which results are significantly different. We used the tests and the procedures recommended in [39] for comparing

Table 2
The characteristics of the four brain MRI images used in the first experimental study (Fig. 2). The noise value represents the ratio between the standard deviation of the white Gaussian noise and the signal of the brightest tissue. The number of crest line points, used as features, is also reported.

Image	Lesion	Noise	# of features
I_1	No	None	583
I_2	No	1%	393
I_3	Yes	1%	348
I_4	Yes	5%	248

Table 3
Parameters of the similarity transformations we used in the experiments: rotation angle (λ), rotation axis (a_x, a_y, a_z), translation vector (t_x, t_y, t_z) and uniform scaling factor s.

	λ	a_x	a_y	a_z	t_x	t_y	t_z	s
T_1	115	−0.863	0.259	0.431	−26	15.5	−4.6	1
T_2	168	0.676	−0.290	0.676	6	5.5	−4.6	0.8
T_3	235	−0.303	−0.808	0.505	16	−5.5	−4.6	1
T_4	276.9	−0.872	0.436	−0.218	−12	5.5	−24.6	1.2

algorithms over multiple problems. We used non-parametric tests to avoid making (or testing) any assumption about the distribution of the results. The performance of SS$^+$ is compared with that of the remaining algorithms (i.e., a multiple comparison against a control method), a procedure that has more power than a pairwise comparison of all algorithms. The tests we used are Nemenyi's test [40] and sign test. The first is a post hoc procedure of Friedman's rank sum test [41] and is based on the ranks of the algorithms. Sign-test, instead, compares the algorithms using only the number of wins and losses. As multiple comparisons are performed, the p-values of the tests have been adjusted using Holm's method [42] in order to control the family-wise error rate.

For all algorithms, we used the original implementations by the authors. I-ICP, Dyn-GA and SS* are written in C, while SS$^+$, GA$^+$ and ASGD are implemented in C++. Their running times can be fairly compared, as no bias due to the programming language or the execution environment is introduced. All experiments have been performed using an Intel Core i5-2400 CPU with 4 GB of memory.

SS$^+$ was integrated in Elastix [32], a toolbox for intensity-based medical image registration. Elastix is free, open-source and it has been used in over one hundred publications in medical imaging [43]. The software is built on top of the popular "Insight Segmentation and Registration Toolkit" (ITK) [44].

4.1. First experiment: registration of simulated brain magnetic resonance images

The first experiment is similar to the ones carried out in [10,34]; the current proposal extends the study of feature- and intensity-based methods performed in the two previous publications. For this experiment we used four simulated brain magnetic resonance images (MRIs) from a public database. A total of 16 registration scenarios were artificially created by applying to the images a set of four large transformations. On those IR instances, we performed a comparison considering a large, heterogeneous group of IR algorithms.

The images used in this experiment were obtained from the Brainweb database at McGill University [45]. Brainweb provides *simulated* brain MRI along with ground-truth data, therefore it can be easily used to evaluate the performance of various image analysis methods. Indeed, Brainweb has been frequently used by the IR research community [46]. To create scenarios with different difficulties, we added noise and multiple sclerosis lesions to some of the images, as detailed in Table 2. The images are shown in Fig. 2; each image has size $60 \times 181 \times 217$ voxels.

This experiment compares both feature- and intensity-based algorithms, thus some features need to be extracted from the images to provide an input for feature-based algorithms. As in the original comparison, we used crest line points, i.e. points where the surface normal has a sharp variation, detected through the approach described in [47]. SS* also uses the principal curvatures of the crest line points as heuristic information to guide the feature matching. Therefore, this information has been computed and provided to the algorithm.

It is important to remark this difference: while the input of intensity-based methods consists of the whole images data (in case, two images made of $60 \times 181 \times 217 = 2,356,620$ voxels having an 8-bit intensity value), that of feature-based approaches is a set of just a few hundred points (Table 2).

Sixteen IR problem instances were created by choosing pairs of different images among the four available and applying one of the four similarity transformations shown in Table 3. Similarity transformations involve rotation, translation, and uniform scaling. The parameters values of the transformations were chosen to obtain large changes in the object location, orientation and scale. Changes of such magnitude are usually challenging for IR algorithms. The scenarios we considered in the experiments are I_1 versus $T_i(I_2)$, I_1 versus $T_i(I_3)$, I_1 versus $T_i(I_4)$ and I_2 versus $T_i(I_4)$, for $i = 1, 2, 3, 4$.

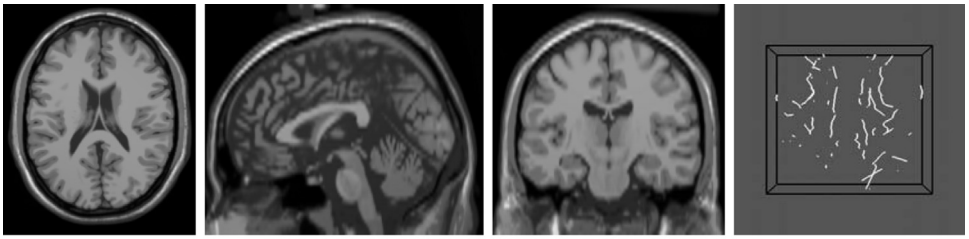
In order to highlight the crucial role played by the specific IR components introduced in Section 3 on the performance of SS$^+$, we first present the preliminary study in Fig. 3. We selected a simple and a complex scenario, in terms of quality of the images and magnitude of the transformations involved, and plotted the results obtained incrementally enabling all the components. In both scenarios, each component brings a noticeable improvement on the results, which can span across a few orders of magnitude. In particular, restart has a major effect on the variability of the results, so is beneficial in terms of robustness, while dynamic boundary affects mostly the accuracy. An analogous pattern was observed throughout the study. Note that, excluding this preliminary study, SS$^+$ is always used with all the components enabled.

In order to provide a uniform comparison of SS$^+$ with respect to our recent results [34], we considered the same algorithms: GA$^+$, ASGD, SS*, I-ICP and Dyn-GA. The parameter settings we used for these methods are the ones corresponding to the best configurations of our previous study. They were manually adjusted for this experiment, starting from the recommended/default values provided by authors. As for SS$^+$, the parameters values were adjusted through a preliminary study, that was performed using an additional registration scenario not included in the experimentation.

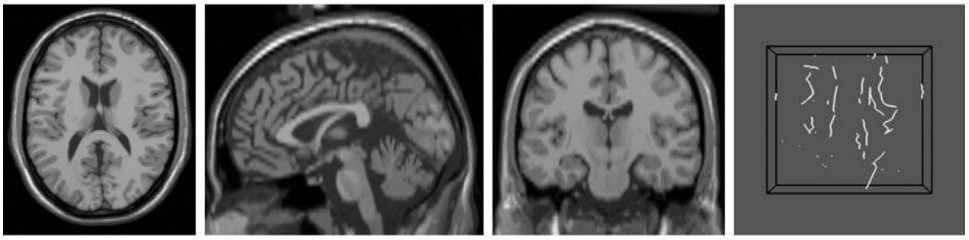
In SS$^+$ the registration is performed in two resolutions; at the first resolution the images are smoothed (Gaussian smoothing, $\sigma = 4$) and downsampled by a factor of 4 in each dimension. The first resolution is repeated two times (i.e. one restart) independently of the results. The algorithm used the same configuration in both resolutions: five elements in the quality and in the diversity reference set, *PSize* set to 32, blend factor $\alpha = 0.3$ for the solution combination method and 50 iterations of PMX with $\alpha = 0.5$ for the improvement method. Mutual information was used as similarity metric.

For all algorithms, the transformation model is similarity transform, and the transformation parameters ranges are $[-30, 30]$ for the translation component and $[0.75, 1.25]$ for the scaling factor. No restriction was applied to the rotation axis or to its magnitude. In SS$^+$, a transform is represented as a real vector with seven elements: three specify the versor of the rotation v, three the translation t and one the scaling s. Valid solutions require $v_x, v_y, v_z \in [-1, 1]$ and $s > 0$.

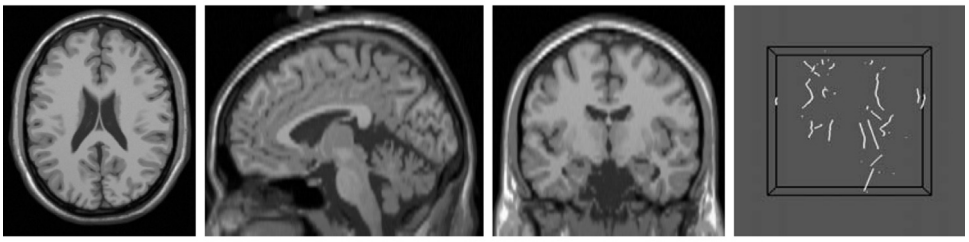
The stopping criterion is running time. It is challenging to design a fair comparison between algorithms having different inputs, in particular inputs with very different size. Indeed, recall that while intensity-based methods use the whole images data (or at least a



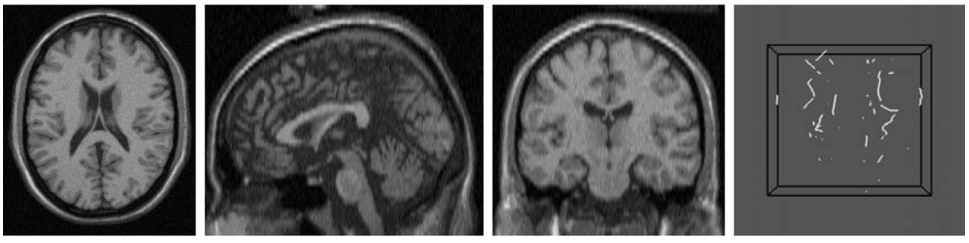
(a) I_1



(b) I_2



(c) I_3



(d) I_4

Fig. 2. The four MRI brain images used in the first experiment. From left to right, axial, sagittal and coronal views, along with the corresponding crest line points used as features.

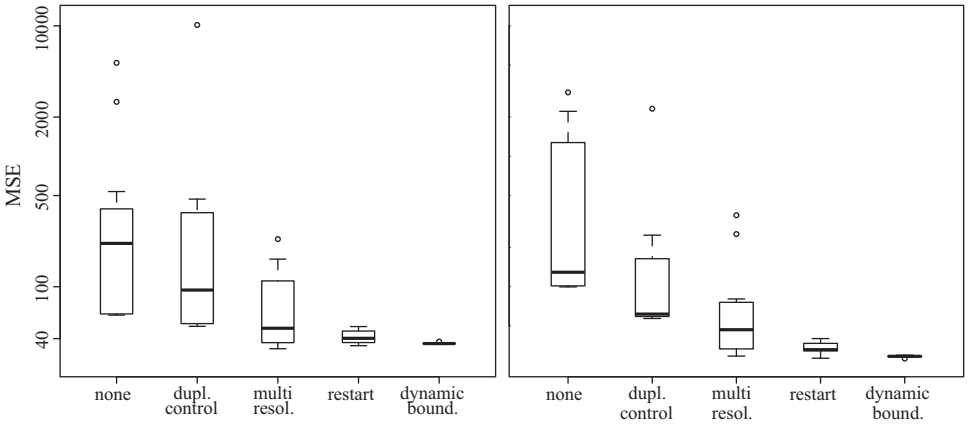


Fig. 3. The effect of the specific IR components of SS$^+$ on its performance. The results refer to the first experimental study, scenarios I_1 vs. $T_1(I_2)$ (left) and I_2 vs. $T_4(I_4)$ (right). From left to right, the components are enable in an incremental fashion, e.g. the boxplot labeled "dynamic boundary" shows the results of SS$^+$ using duplication control, multiple resolutions, restart and dynamic boundary. Logarithmic scale is used to show the improvement brought by each component despite the differences in their order of magnitude.

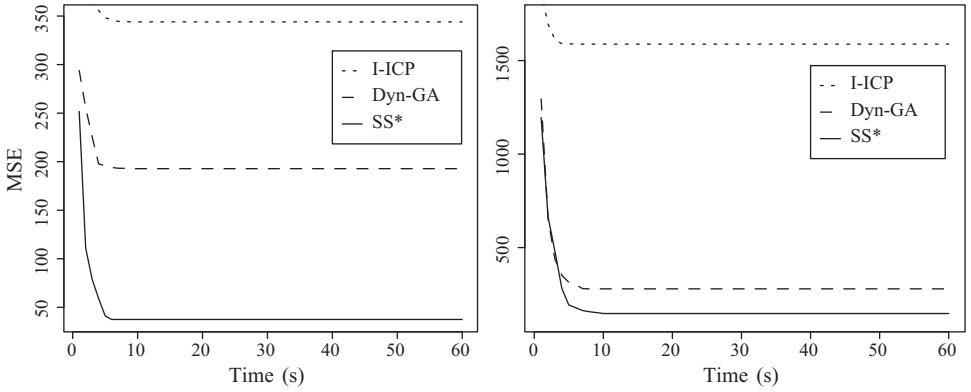


Fig. 4. The MSE scored by I-ICP, Dyn-GA and SS* against running time. Results are averaged over 15 runs and refer to scenarios I_1 vs. $T_1(I_2)$ (left) and I_2 vs. $T_4(I_4)$ (right).

large subset), only a small fraction of the input data is actually used in feature-based techniques. In this case, intensity-based methods use a random subset of 25,000 voxels to compute the similarity metric, while the average number of features in the images is 393. This gives us a proportion of roughly 60, which is used to set the two time limits: 20 s for the feature-based approaches and 20 min for the intensity-based ones. In addition, we studied the behavior of the algorithms over time, to ensure these time limits allow them to reach convergence and deliver representative results. Fig. 4 shows examples of this analysis for the feature-based methods, while the results for intensity-based ones are reported in Fig. 5a.

4.1.1. Results

As in previous works with this dataset, for each registration scenario we performed 15 independent runs of each algorithm. Since we are dealing with algorithms of different natures, and in particular algorithms with different similarity metrics, we cannot simply contrast their values. Instead, we have to agree on a common measure to evaluate all solutions. We used the MSE over the crest line points. For the feature-based algorithms in the comparison, this is simply the similarity metric used by the algorithms. The solutions found by intensity-based algorithms were evaluated in the same way, i.e. by applying the obtained transformation to the scene's features and computing the MSE with respect to the model's features. We expect this choice to introduce a small bias in favor of feature-based algorithms. However, using a similarity metric based on intensities might favor intensity-based methods, therefore as we are proposing an algorithm from the latter class, it seems more appropriate to favor the competitors rather than our approach.

Table 4 reports the results of the first experiment. For each scenario, we reported mean and standard deviation of the MSE values

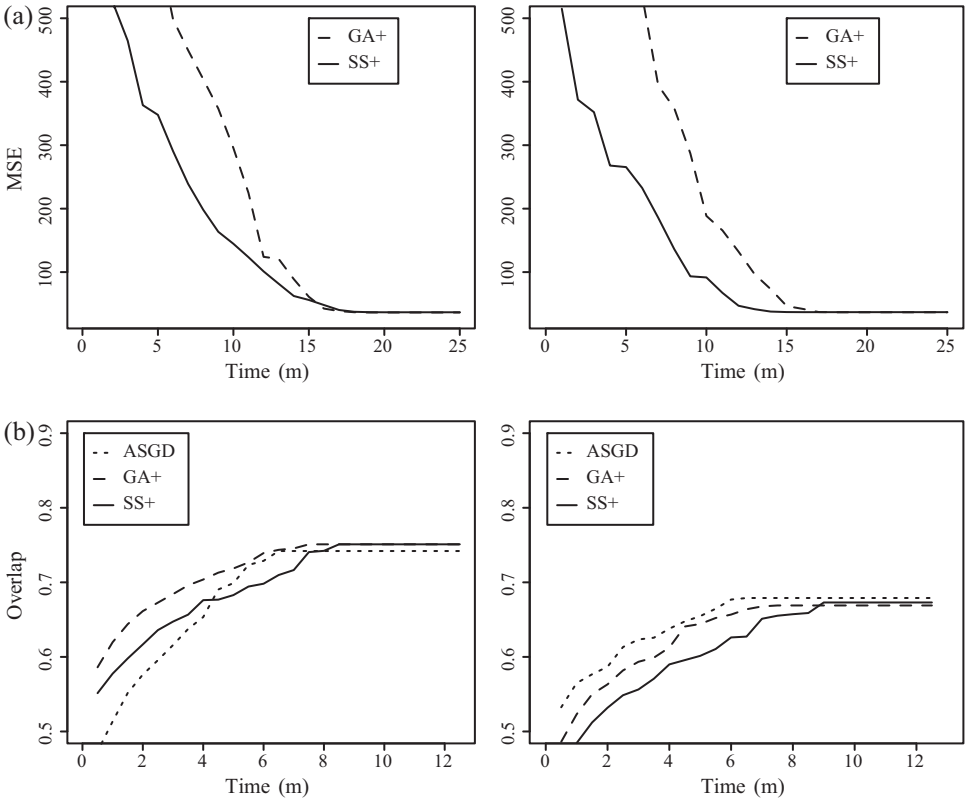


Fig. 5. The results of ASGD, GA⁺ and SS⁺ against running time. Results are averaged over 15 runs and refer to the first two scenarios in the first (top row) and second (bottom row) studies. The score of ASGD in the top row is above 30,000 and it has been omitted.

Table 4
Detailed results of the first experiment. For each scenario, the table reports the average MSE, standard deviation and ranking of the algorithms in the comparison.

Algorithm		MSE Mean	SD	Rank
1	ASGD	61,816.4	795.7	6
	Dyn-GA	194.9	50.5	4
	GA$^+$	36.4	0.3	1
	I-ICP	344.4		5
	SS*	37.0	1.5	3
	SS$^+$	36.7	0.0	2
2	ASGD	34,773.9	238.0	6
	Dyn-GA	107.5	52.1	4
	GA$^+$	36.7	0.4	1
	I-ICP	130.7		5
	SS*	43.4	3.6	3
	SS$^+$	36.9	0.0	2
3	ASGD	111,870.0	154.6	6
	Dyn-GA	211.0	137.3	3
	GA$^+$	1736.5	6216.7	5
	I-ICP	894.3		4
	SS*	63.2	2.9	2
	SS$^+$	41.1	0.0	1
4	ASGD	1233.3	166.8	6
	Dyn-GA	302.0	121.4	4
	GA$^+$	32.7	0.2	2
	I-ICP	631.7		5
	SS*	53.9	2.6	3
	SS$^+$	32.7	0.0	1
5	ASGD	61,063.5	309.1	6
	Dyn-GA	299.3	144.1	4
	GA$^+$	51.4	0.2	2
	I-ICP	517.7		5
	SS*	112.2	12.4	3
	SS$^+$	51.4	0.0	1
6	ASGD	34,796.2	223.8	6
	Dyn-GA	154.0	114.2	4
	GA$^+$	43.8	0.2	1
	I-ICP	330.3		5
	SS*	56.7	4.5	3
	SS$^+$	44.0	0.1	2
7	ASGD	110,131.2	1022.7	6
	Dyn-GA	326.5	174.0	3
	GA$^+$	1091.8	4965.8	5
	I-ICP	437.8		4
	SS*	63.8	46.2	2
	SS$^+$	56.2	0.0	1
8	ASGD	1017.4	252.7	6
	Dyn-GA	354.3	146.9	4
	GA$^+$	44.5	0.3	2
	I-ICP	478.0		5
	SS*	122.7	8.2	3
	SS$^+$	44.3	0.0	1
9	ASGD	58,146.8	661.0	6
	Dyn-GA	255.4	228.2	4
	GA$^+$	52.9	0.3	2
	I-ICP	704.3		5
	SS*	183.6	33.0	3
	SS$^+$	52.9	0.1	1
10	ASGD	35,695.3	2465.3	6
	Dyn-GA	163.1	57.5	3
	GA$^+$	476.4	3648.3	4
	I-ICP	1493.2		5
	SS*	89.2	40.8	2
	SS$^+$	46.7	0.0	1
11	ASGD	111,384.4	574.2	6
	Dyn-GA	224.9	87.3	3
	GA$^+$	2823.9	7863.5	5
	I-ICP	951.3		4
	SS*	82.2	45.1	2
	SS$^+$	57.7	0.0	1
12	ASGD	885.8	356.7	6
	Dyn-GA	414.8	258.2	4
	GA$^+$	47.3	0.4	2
	I-ICP	416.6		5
	SS*	153.9	86.1	3
	SS$^+$	47.2	0.0	1
13	ASGD	56,932.0	568.6	6
	Dyn-GA	179.8	59.5	3
	GA$^+$	35.0	0.2	2
	I-ICP	237.6		5
	SS*	193.1	62.0	4
	SS$^+$	35.0	0.1	1
14	ASGD	31,521.1	6.6	6
	Dyn-GA	105.7	50.8	4
	GA$^+$	30.7	0.3	1
	I-ICP	341.3		5
	SS*	74.9	41.1	3
	SS$^+$	31.0	0.1	2
15	ASGD	112,134.4	1027.4	6
	Dyn-GA	192.2	115.8	3
	GA$^+$	1104.8	5128.7	5
	I-ICP	608.8		4
	SS*	103.8	66.6	2
	SS$^+$	40.4	0.0	1
16	ASGD	512.8	233.2	5
	Dyn-GA	298.1	144.8	4
	GA$^+$	29.5	0.3	2
	I-ICP	1587.8		6
	SS*	150.2	78.3	3
	SS$^+$	29.2	0.0	1

Table 5
First experiment: result of Nemenyi's test comparing SS$^+$ with the remaining algorithms. The table reports the average rankings of the algorithms and the adjusted p-value for each comparison.

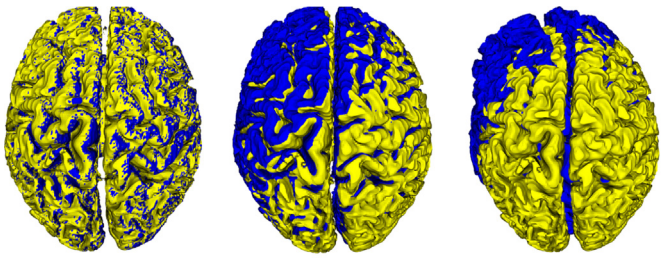
Algorithm	Mean rank	p-Value
SS$^+$	1.25	
GA$^+$	2.38	0.0137
SS*	2.75	0.0020
Dyn-GA	3.62	0.0000
I-ICP	4.81	
ASGD	5.94	

obtained by the algorithms along with their ranks. The average ranks (Table 5) and the count of wins (Table 6) provide a basic view of the results of the comparison. We also include a visual comparison of the average results in the third scenario (Fig. 6).

From the highest to the lowest error ranking is ASGD, I-ICP, Dyn-GA, SS*, GA$^+$ and SS$^+$. ASGD scored the largest MSE values in all but one of the scenarios. Its performance varies greatly depending on the scenario, but in general the mean MSE is at least one order of magnitude away from the best solutions. I-ICP delivered a better, more steady performance, but still with very large MSE values. Comparing the mean values, Dyn-GA scored better than

Table 6
First experiment: the number of scenarios in which the algorithm on the row has a better mean MSE value than that on the column.

	ASGD	Dyn-GA	GA$^+$	I-ICP	SS*	SS$^+$
ASGD	–	0	0	1	0	0
Dyn-GA	16	–	5	16	1	0
GA$^+$	16	11	–	12	11	4
I-ICP	15	0	4	–	0	0
SS*	16	15	5	16	–	0
SS$^+$	16	16	12	16	16	–



(a) SS+ (41.1) (b) SS* (63.2) (c) Dyn-GA (211.0)

Fig. 6. First experiment. A visualization of the overlapping between the scene (yellow) and the model (blue). The figures refer to the third scenario. The solutions used to create the figures are those having the closest MSE to the mean result of the corresponding algorithm. In parenthesis is the MSE value of each solution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 7
First experiment: result of sign test comparing SS+ with the other algorithms. The table lists the algorithms along with their number of scenarios in which they have been outperformed by SS+ (Table 6, bottom row) and the associated adjusted *p*-value.

Algorithm	Losses	*p*-Value
ASGD	16	0.0002
Dyn-GA	16	0.0002
GA+	12	0.0768
I-ICP	16	0.0002
SS*	16	0.0002

I-ICP in all scenarios, with less variability between different scenarios, but the gap with the best results is large nevertheless. Also, the standard deviation is often comparable with the mean, meaning that the quality of the solutions varies a lot even in the same scenario. SS* and GA+ delivered similar performances in terms of ranking: 2.75 and 2.38, respectively. SS* scored constantly quite close to the best results, ranking third or better in all but one scenario. GA+ had a less stable behavior: it either reached very good solutions (11 scenarios) or failed to converge and scored a really high MSE value (the remaining 5 ones). Our SS approach has the lowest rank, 1.25. It found the best solutions in most of the scenarios (12 over 16) and came extremely close in the others, with differences below 1.0. The MSE values vary in a small range, between 29.2 and 57.7.

Table 5 reports the *p*-value of Nemenyi's test comparing SS+ against GA+, SS* and Dyn-GA. We included only the best ranking algorithms to avoid lowering the power of the test. In all three cases the test confirms the performance of SS+ is significantly better than that of the competitors, with the highest *p*-value being that of GA+, 0.0137. The sign-test comparing the number of wins (Table 7) has similar results. All algorithms but GA+ have a *p*-value of 0.0002, while GA+'s is considerably higher, 0.0768. The test confirms the difference between SS+ and the others algorithms is significant, although to a lower degree.

4.2. Second experiment: atlas-based segmentation of real-world MRIs

In the second experiment we used real brain MRI images without applying any transformation. The registration is used to perform atlas-based segmentation of deep brain structures [48]. The quality of the segmentation obtained in this phase is used to assess the effectiveness of the registration methods.

Atlas-based segmentation is a procedure that aims to automatically delineate a region of an image using an atlas (or an "average" image) of a similar subject in which the desired region has been already segmented. The first step is to register the atlas (the scene) to the input image (the model). The transformation resulting from this phase is then used to overlap the segmented region of the atlas to the scene. The region of the scene that overlaps the segmented region of the atlas is the result of the segmentation process. Fig. 7 illustrates the process. Often, atlas-based segmentation is used as preliminary step in a more complex segmentation approach.

Thirteen T_1-weighted brain MRI were retrieved from the NMR database [49]. The deep nuclei structures in each image have been manually delineated by an expert in order to create the ground-truth data used to evaluate the registration. Fig. 7 shows one of the images along with the corresponding deep nuclei. Eighteen registration scenarios were created by selecting pairs of different images at random.

Given the nature of this experiment, we compare only intensity-based algorithms: SS+, GA+ and ASGD. The transformation model is affine transform, which involves rotation, translation, scaling and shearing, and it can be represented using 12 real parameters. Affine transform is a popular choice in registration of medical images [50]. It is flexible enough to represent a wide range of transformations and it does not produce anatomically unrealistic results, as it could happen with deformable models. An affine transform is specified by a linear transformation (a 3×3 matrix $A = \{a_{i,j}\}$) and a translation vector t. In SS+, both elements are combined in a real vector ($a_{1,1}$, $a_{1,2}, \ldots, a_{3,2}, a_{3,3}, t_x, t_y, t_z$) with 12 elements. The valid range of each matrix element $a_{i,j}$ is computed from the limits imposed over rotation, scaling and shearing. As the concrete parameter values of the optimal registration transformations are not known, we estimated parameters intervals considering a big enough range to include all registration solutions for this application. We allowed rotations between -90° and 90°, scaling in the range $[0.9, 1.1]$, shearing in the interval $[-0.1, 0.1]$ and translations between -15 cm and 15 cm.

For GA+ and SS+ we kept the same configuration used in the first experiment. For ASGD we tested several configurations varying the number of resolutions (2, 3 and 4) and iterations (500, 1000 and 2000). In what follows we report only the results obtained with the best configuration, which uses 4 resolutions and 1000 iterations.

The stopping criterion is a time limit of 10 min. Note that no synthetic transformation is applied on the images; the registration aims to compensate for the variability in the pose of the patient during the acquisition of the images. Therefore, with respect to the first experiment, the "magnitude" of the transformations to be found is

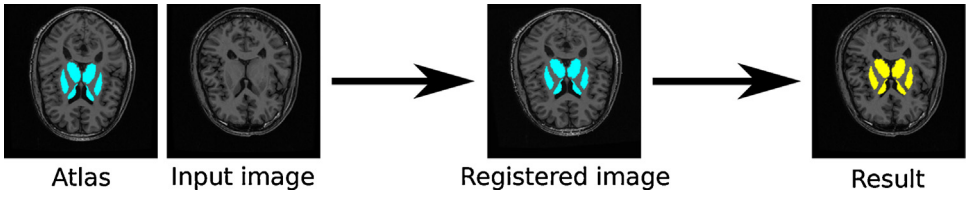


Fig. 7. The atlas-based segmentation of deep brain structures in brain MRI. First, the atlas is registered to the input image and the resulting transformation is applied to the labeled region of the atlas (in blue). The resulting region is overlapped on the input image (in yellow) to determine the output of the process. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

much smaller. This results in a smaller search space, which in turn yields faster convergence for the algorithms, justifying the smaller time limit of this experiment. This phenomenon can be observed throughout the experimental study; two examples are shown in Fig. 5.

Table 8
Detailed results of the second experiment. For each scenario, the table reports the average overlap, standard deviation and ranking of the algorithms in the comparison.

Algorithm		Overlap Mean	SD	Rank
1	ASGD	0.742	0.001	3
	GA$^+$	0.751	0.010	2
	SS$^+$	0.751	0.008	1
2	ASGD	0.679	0.003	1
	GA$^+$	0.669	0.019	3
	SS$^+$	0.673	0.026	2
3	ASGD	0.616	0.005	1
	GA$^+$	0.615	0.033	2
	SS$^+$	0.596	0.026	3
4	ASGD	0.677	0.003	1
	GA$^+$	0.676	0.012	3
	SS$^+$	0.677	0.008	2
5	ASGD	0.682	0.000	1
	GA$^+$	0.671	0.006	2
	SS$^+$	0.670	0.005	3
6	ASGD	0.691	0.001	3
	GA$^+$	0.698	0.011	2
	SS$^+$	0.722	0.018	1
7	ASGD	0.621	0.007	3
	GA$^+$	0.635	0.015	2
	SS$^+$	0.652	0.027	1
8	ASGD	0.756	0.010	2
	GA$^+$	0.755	0.009	3
	SS$^+$	0.773	0.012	1
9	ASGD	0.634	0.006	3
	GA$^+$	0.656	0.014	2
	SS$^+$	0.670	0.022	1
10	ASGD	0.738	0.003	2
	GA$^+$	0.734	0.011	3
	SS$^+$	0.740	0.011	1
11	ASGD	0.717	0.011	3
	GA$^+$	0.736	0.012	2
	SS$^+$	0.750	0.013	1
12	ASGD	0.684	0.005	3
	GA$^+$	0.704	0.024	2
	SS$^+$	0.716	0.022	1
13	ASGD	0.686	0.009	3
	GA$^+$	0.717	0.015	2
	SS$^+$	0.718	0.014	1
14	ASGD	0.680	0.005	3
	GA$^+$	0.713	0.011	1
	SS$^+$	0.693	0.019	2
15	ASGD	0.741	0.001	3
	GA$^+$	0.751	0.020	2
	SS$^+$	0.769	0.023	1
16	ASGD	0.751	0.012	3
	GA$^+$	0.769	0.017	2
	SS$^+$	0.779	0.011	1
17	ASGD	0.754	0.004	2
	GA$^+$	0.745	0.017	3
	SS$^+$	0.756	0.020	1
18	ASGD	0.624	0.004	3
	GA$^+$	0.672	0.044	2
	SS$^+$	0.689	0.030	1

Table 9
Second experiment: result of Nemenyi's post-hoc procedure when comparing SS$^+$ with the remaining algorithms. The table reports the average rankings of the algorithms and the adjusted p-value for each comparison.

Algorithm	Mean rank	p-Value
SS$^+$	1.39	
GA$^+$	2.22	0.0124
ASGD	2.39	0.0054

Table 10
Second experiment: the number of scenarios in which the algorithm on the row has a better mean overlap value than that on the column.

	ASGD	GA$^+$	SS
ASGD	–	7	4
GA$^+$	11	–	3
SS$^+$	14	15	–

4.2.1. Results

The quality of atlas-based segmentation depends closely on the accuracy of the registration step, although the anatomical variability of the target region can limit its effectiveness. In this experiment we validate the results of the registration algorithms by carrying out atlas-based segmentation of deep nuclei. For each scenario we performed 32 independent runs of each algorithm. The model image is used as atlas, while the scene is employed as input image. The segmented region obtained from the registration V_R is then compared with the ground-truth V_{GT}. The overlapping of the two regions is commonly measured using the Dice's coefficient [51], given by

$$\mathrm{Dice}(V_R, V_{GT}) = \frac{2|V_R \cap V_{GT}|}{|V_R| + |V_{GT}|}$$

where $|\cdot|$ is the number of voxels. A value of 1 means perfect overlapping, while 0 means the two regions do not overlap at all.

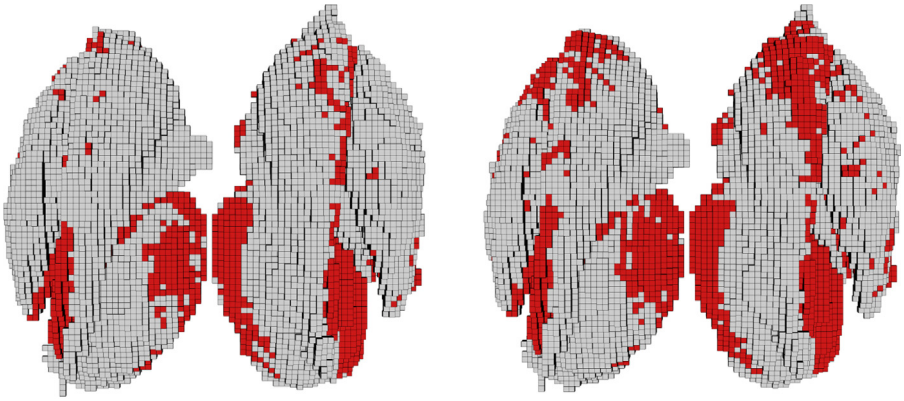
The results of the second experiment are reported in Table 8. We computed the mean and standard deviation of the overlap and ranked the algorithm accordingly for each scenario. Tables 9 and 10 show the mean ranks and the count of wins for the three algorithms in the comparison. Finally, Fig. 8 shows a comparison of the average results in the seventh scenario.

The overlap values can differ considerably across the scenarios, reflecting the fact that the effectiveness of this kind of segmentation can vary depending on the concrete anatomy of the patients. In general, GA$^+$ and ASGD have similar results, whereas SS$^+$ ranked constantly better than the others. In contrast with the previous experiment, even though it ranked last on average, ASGD delivered an acceptable performance. This is due to the lower "magnitude" of the transformations involved, more suitable for a local search method. However, note that the low variability of its overlap values means almost all solutions have a lower quality than the average solution found by the other two algorithms. GA$^+$ had a similar performance: it has almost the same mean rank and it ranked better than ASGD in slightly more than half of the scenarios (11 cases).

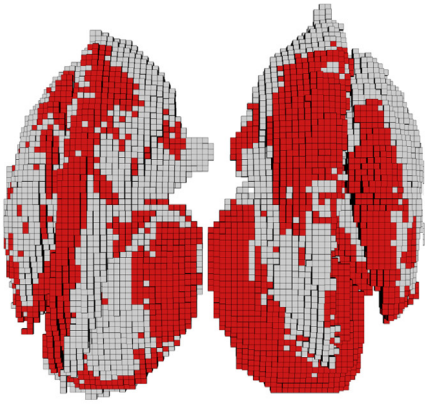
The performance of SS$^+$ is the best one of the comparison. Our proposal outperformed the other algorithms both in terms of mean rank (1.39 against 2.22 and 2.39) and number of wins (14 and 15 against GA$^+$ and ASGD, respectively). The result of Nemenyi's test,

Table 11
Second experiment: result of sign test comparing SS$^+$ with ASGD and GA$^+$. The table lists the algorithms along with their number of scenarios in which they have been outperformed by SS$^+$ (Table 10, bottom row) and the associated adjusted p-value.

Algorithm	Losses	p-Value
ASGD	14	0.0309
GA$^+$	15	0.0151



(a) SS^+ (0.654) (b) GA^+ (0.636)



(c) ASGD (0.619)

Fig. 8. Second experiment. A visualization of the overlapping between the automatically segmented volume (red) and the ground truth (white). The figures refer to the seventh scenario. The solutions used to create the figures are those having the closest overlap value to the mean result of the corresponding algorithm. In parenthesis is the overlap value of each solution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

reported in Table 9, shows the difference with both algorithms is statistically significant to a high degree (0.0124 for GA^+ and 0.0054 for ASGD). The sign-test (Table 11) confirms the thesis, although with slightly higher p-values (0.0151 for GA^+ and 0.0309 for ASGD).

5. Conclusions

We have described the design and implementation of a novel metaheuristic technique to solve medical IR. Our approach is intensity-based and performs the optimization in the space of transformation parameters. The five components of the SS template are specifically designed to tackle IR; in addition, we integrated a multi-resolution strategy with two optimization components. First, the restart mechanism that allows the algorithm to deliver a more robust performance at the cost of a low extra computational effort. Second, the dynamic boundary that focuses the search on the appropriate region of the search space as the optimization progresses.

The merit of our approach is proved experimentally in two separate studies involving synthetic and real-world medical images. Each study included a comparison with other state-of-the-art IR methods using a wide range of approaches to the problem. In both studies, SS^+ delivered the best mean performance, with a statistically significant improvement over the competitors.

The most natural extension to the current work is to tackle deformable registration. This is still an area of on-going research. On one hand, there is an increasing interest in such technology for clinical applications; on the other, automated solutions have not yet reached the same degree of maturity as for rigid or affine registration.

Acknowledgments

Andrea Valsecchi is funded by the European Commission (Marie Curie ITN MIBISOC, FP7 PEOPLE-ITN-2008, GA no. 238819). This work is supported by Ministerio de Economí-a y Competitividad under project SOCOVIFI2 (TIN2012-38525-C02-01 and TIN2012-38525-C02-02) including EDRF funds. The NMR database is the property of CEA/I2BM/NeuroSpin and can be provided on demand to cyril.poupon@cea.fr. Data were acquired with PTK pulse sequences, reconstructed with PTK reconstructor package and postprocessed with Brainvisa Connectomist software, freely available at http://brainvisa.info.

References

[1] Zitová B, Flusser J. Image registration methods: a survey. Image Vis Comput 2003;21:977–1000.
[2] Wang XY, Eberl S, Fulham M, Som S, Feng DD. Data registration and fusion. In: Feng DD, editor. Biomedical information technology. Burlington, USA: Academic Press; 2008. p. 187–210.
[3] Besl PJ, McKay ND. A method for registration of 3D shapes. IEEE Trans Pattern Anal Mach Intell 1992;14:239–56.
[4] Glover F, Kochenberger GA. Handbook of metaheuristics. Boston, USA: Kluwer Academic Publishers; 2003.

[5] Damas S, Cordón O, Santamaría J. Medical image registration using evolutionary computation: an experimental survey. IEEE Comput Intell Mag 2011;6(4):26–42.
[6] Valsecchi A, Damas S, Santamaría J. Evolutionary intensity-based medical image registration: a review. Curr Med Imaging Rev 2013;9(4):283–97.
[7] Glover F. Heuristic for integer programming using surrogate constraints. Decision Sci 1977;8:156–66.
[8] Glover F. A template for scatter search and path relinking. In: Hao J-K, Lutton E, Ronald E, Schoenauer M, Snyers D, editors. Artificial evolution, vol. 1363 of lecture notes in computer science. Berlin/Heidelberg: Springer; 1998. p. 1–51.
[9] Laguna M, Martí R. Scatter search: methodology and implementations in C. Boston, USA: Kluwer Academic Publishers; 2003.
[10] Cordón O, Damas S, Santamaría J, Martí R. Scatter search for the point-matching problem in 3D image registration. INFORMS J Comput 2008;20(1):55–68.
[11] Santamaría J, Cordón O, Damas S, Alemán I, Botella M. A scatter search-based technique for pair-wise 3D range image registration in forensic anthropology. Soft Comput 2007;11:819–28.
[12] Cole-Rhodes A, Johnson K, LeMoigne J, Zavorin I. Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. IEEE Trans Image Process 2003;12:1495–511.
[13] Svedlow M, Mc-Gillem CD, Anuta PE. Experimental examination of similarity measures and preprocessing methods used for image registration. In: Swain P, Morrison D, Parks D, editors. Symposium on machine processing of remotely sensed data, vol. 4(A). 1976. p. 9–17.
[14] Audette MA, Ferrie FP, Peters TM. An algorithmic overview of surface registration techniques for medical imaging. Med Image Anal 2000;4(3):201–17.
[15] Pluim JPW, Maintz JBA, Viergever MA. Mutual-information-based registration of medical images: a survey. IEEE Trans Med Imaging 2003;22(8):986–1004.
[16] Gholipour A, Kehtarnavaz N, Yousefi S, Gopinath K, Briggs R. Symmetric deformable image registration via optimization of information theoretic measures. Image Vis Comput 2010;28(6):965–75.
[17] Maes F, Vandermeulen D, Suetens P. Comparative evaluation of multiresolution optimization strategies for image registration by maximization of mutual information. Med Image Anal 1999;3(4):373–86.
[18] Rouet JM, Jacq JJ, Roux C. Genetic algorithms for a robust 3-D MR-CT registration. IEEE Trans Inform Technol B 2000;4(2):126–36.
[19] Yamany SM, Ahmed MN, Farag AA. A new genetic-based technique for matching 3D curves and surfaces. Pattern Recogn 1999;32:1817–20.
[20] He R, Narayana PA. Global optimization of mutual information: application to three-dimensional retrospective registration of magnetic resonance images. Comput Med Imaging Graph 2002;26:277–92.
[21] Chalermwat P, El-Ghazawi T, LeMoigne J. 2-phase GA-based image registration on parallel clusters. Future Gen Comput Syst 2001;17:467–76.
[22] Chow CK, Tsui HT, Lee T. Surface registration using a dynamic genetic algorithm. Pattern Recogn 2004;37:105–17.
[23] Cordón O, Damas S, Santamaría J, Fast A. Accurate approach for 3D Image Registration using the scatter search evolutionary algorithm. Pattern Recogn Lett 2006;27(11):1191–200.
[24] Cordón O, Damas S, Santamaría J. Feature-based image registration by means of the CHC evolutionary algorithm. Image Vis Comput 2006;22:525–33.
[25] Lomonosov E, Chetverikov D, Ekart A. Pre-registration of arbitrarily oriented 3D surfaces using a genetic algorithm. Pattern Recogn Lett 2006;27(11):1201–8.
[26] Silva L, Bellon ORP, Boyer KL. Robust range image registration using genetic algorithms and the surface interpenetration measure. Singapore: World Scientific; 2005.
[27] Zhang Z. Iterative point matching for registration of free-form curves and surfaces. Int J Comput Vis 1994;13(2):119–52.
[28] Liu Y. Improving ICP with easy implementation for free form surface matching. Pattern Recogn 2004;37(2):211–26.
[29] Rusinkiewicz S, Levoy M. Efficient variants of the ICP algorithm. In: Third International Conference on 3D Digital Imaging and Modeling. Quebec, Canada: IEEE Computer Society; 2001. p. 145–52.
[30] Zambanini S, Sablatnig R, Maier H, Langs Gd. Automatic image-based assessment of lesion development during hemangioma follow-up examinations. Artif Intell Med 2010;50(2):83–94.
[31] Klein S, Pluim J, Staring M, Viergever M. Adaptive stochastic gradient descent optimisation for image registration. Int J Comput Vis 2009;81:227–39.
[32] Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29(1):196–205.
[33] Bäck T, Fogel DB, Michalewicz Z. Handbook of evolutionary computation. Bristol, UK: IOP Publishing Ltd./Oxford University Press; 1997.
[34] Valsecchi A, Damas S, Santamaría J. An image registration approach using genetic algorithms. In: Li X, editor. IEEE congress on evolutionary computation. IEEE; 2012. p. 416–23.
[35] Glover F, Laguna M, Martí R. Scatter search. In: Ghosh A, Tsutsui S, editors. Advances in evolutionary computation: theory and applications. New York: Springer-Verlag; 2003. p. 519–37.
[36] Eshelman LJ. Real-coded genetic algorithms and interval schemata. In: Whitley LD, editor. Foundations of genetic algorithms 2. San Mateo, USA: Morgan Kaufmann; 1993. p. 187–202.
[37] Lozano M, Herrera F, Krasnogor N, Molina D. Real-coded memetic algorithms with crossover hill-climbing. Evol Comput 2004;12(3):273–302.
[38] Santamaría J, Cordón O, Damas S, García-Torres J, Quirin A. Performance evaluation of memetic approaches in 3D reconstruction of forensic objects. Soft Comput 2009;13(8–9):883–904.
[39] Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30.
[40] Nemenyi P. Distribution-free multiple comparisons. Princeton University; 1963 [Ph.D. thesis].
[41] Friedman M. A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 1940;11(1):86–92.
[42] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6(2):65–70.
[43] http://elastix.bigr.nl/wiki/index.php/references,;1; (accessed 01.04.13).
[44] Ibáñez L, Schroeder W, Ng L, Cates J. The ITK software guide. 2nd ed. Kitware, Inc.; 2005. ISBN: 1-930934-15-7.
[45] Collins DL, Zijdenbos AP, Kollkian V, Sled JG, Kabani NJ, Holmes CJ, et al. Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging 1998;17:463–8.
[46] Rogelj P, Kovacic S. Validation of a non-rigid registration algorithm for multimodal data. In: Sonka M, Fitzpatrick JM, editors. SPIE in medical imaging, vol. 4684. 2002. p. 299–307.
[47] Monga O, Benayoun S, Faugeras O. From partial derivatives of 3D density images to ridges lines. In: Computer vision and pattern recognition. Champaign, IL, USA: IEEE; 1992. p. 354–89.
[48] Vemuri BC, Ye J, Chen Y, Leonard CM. Image registration via level-set motion: applications to atlas-based segmentation. Med Image Anal 2003;7(1):1–20.
[49] Poupon C, Poupon F, Allirol L, Mangin J-F. A database dedicated to anatomofunctional study of human brain connectivity. In: 12th annual meeting of the organization for human brain mapping, no. 646. 2006.
[50] Rueckert D, Schnabel JA. Medical image registration. In: Deserno TM, editor. Biomedical image processing, biological and medical physics, biomedical engineering. Berlin/Heidelberg: Springer; 2011. p. 131–54.
[51] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.