Genetic Algorithms for Voxel-based Medical Image Registration

Andrea Valsecchi and Sergio Damas
European Centre for Soft Computing
Mieres, Spain
{andrea.valsecchi, sergio.damas}@softcomputing.es

José Santamaría
University of Jaén
Jaén, Spain
jslopez@ujaen.es

Linda Marrakchi-Kacem
Neurospin, CEA, Gif-Sur-Yvette, France
CRICM, UPMC Université Paris 6, France
linda.marrakchi@gmail.com

***Abstract*—Image registration (IR) – the task of aligning different images having a common content – is a fundamental problem in computer vision. In particular, IR is one of the key steps in medical imaging, with applications ranging from computer assisted diagnosis to computer aided therapy and surgery. As IR can be formulated as an optimization problem, a large family of metaheuristics methods can be used to improve the results obtained by classic gradient-based, continuous optimization techniques. In this work, we extend our previous intensity-based image registration (IR) technique based on a real-coded genetic algorithm with a more appropriate design. The performance evaluation of an heterogeneous group of state-of-the-art IR techniques is also extended to two experimental studies on both synthetic and real-word medical IR problems. The results prove the accuracy and applicability of our new method.**

I. Introduction

A large number of applications in image processing require the integration of information from multiple images of the same or similar subjects obtained under different conditions (time, viewpoint, sensor or any combination of the latter). Hence, the images need to be properly aligned in order to put in correspondence the common content. This task is called *image registration* (IR) [1]: given two images, image registration aims to find the geometric transformation leading to the best possible overlapping.

IR approaches usually fall into two categories: *intensity-based* (or *voxel-based*) and *feature-based* methods. The former make use of the entire images while the latter employ only salient, distinctive objects such as lines, corners and contours, detected in a preprocessing step. Feature-based techniques are faster, as they use only a fraction of the imaging data, but usually suffer from inevitable errors in the feature detection process. Independently of their nature, IR techniques involve an iterative *optimization procedure* that explores the space of possible transformations. Registration approaches based on *evolutionary algorithms* (EAs) have proven to be a promising solution to overcome the drawbacks of traditional gradient-based algorithms [2]–[6]. In fact, they are considered global optimization approaches able to perform a robust search in complex search spaces like those arising in IR. In particular, evolutionary methods have successfully tackled feature-based medical IR [7].

In [8] we introduced an intensity-based IR method based on Genetic Algorithms. Although the proposal was competitive with other approaches, the algorithm exhibited convergence issues that occasionally led to low quality solutions. In this work, we address those issues with a new design of the optimizer component. Also, we extend the experimental comparison of a number of IR methods by including a second medical IR task and a statistical analysis of the results.

The paper is structured as follows. Section II introduces the IR problem in detail. Section III describes the proposed approach, while Section IV introduces the experimental comparison, the test problems and the analysis of their results. Finally, Section V provides conclusions and directions for future work.

II. Image registration

In a typical problem instance we are provided with two images: a reference image, the *model*, and the image that will be transformed to reach the model geometry, called *scene* [1]. We will denote these two images by I_M and I_S respectively. The result of the registration process is a transformation f such that the model I_M and the transformed scene $f(I_S)$ are as similar as possible.

IR methods can be characterized by their three main components: the *transformation model*, the *similarity metric* and the *optimization process*. The transformation model determines what kind of transformation is used to align the images. For instance, a rigid transformation is a combination of translation and rotation operations, while similarity transformations also allow scaling. Their degrees of freedom for 3-D images are 6 and 7, respectively. B-splines and thin-plate splines are instead examples of *elastic* (or *non-rigid*) transformations models, able to represent local deformations (warping). In applications, the appropriate transformation model depends on both the nature of the images and the particular application involved.

A similarity metric is a function F that measures the quality of a solution of an IR problem. The final performance of any IR method depends on the accurate estimation of the alignment of the images, therefore the similarity metric is considered a crucial component [9]. To evaluate a solution f, the scene image I_S is transformed according to f and then the degree of resemblance between the transformed scene image $f(I_S)$ and the model image I_M, denoted by Ψ, is computed, so $F(I_M, I_S, f) = \Psi(I_M, f(I_S))$. Several choices for Ψ can be found in the literature, depending on the nature

978-1-4673-5919-1/13/$31.00

of the considered images. In feature-based approaches, metrics are usually based on the distance between corresponding geometric primitives [10], such as mean square error (MSE), which is the average square distance between corresponding feature (in this case points) in the scene and the model images. To compute the MSE, each point of the model is assigned to the closest point in the transformed scene, regardless of whether the latter had been already assigned to another model point. That is, MSE $= \frac{1}{r}\sum_{i=1}^{r} \|x_i - c_i\|^2$ where c_i is the point of $f(I_S)$ that is closest to x_i.

In intensity-based approaches, common choices are sum of squared differences, normalized correlation (NC) and mutual information (MI) [11]. In particular, MI is specially suited for multi-modal registration and other scenarios in which the images have different intensity distributions. It is defined as

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

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

Finally, the optimization procedure is the component responsible for finding an appropriate transformation to carry out the registration. Figure 1 shows the flow chart of the whole registration process. The search strategy adopted depends on the nature of the algorithm. In matching-based algorithms, once the images features have been detected, the optimizer looks for a matching between them and the transformation parameters are derived from the match. The process is iterated until reaching convergence within a tolerance threshold of the concerned similarity metric.

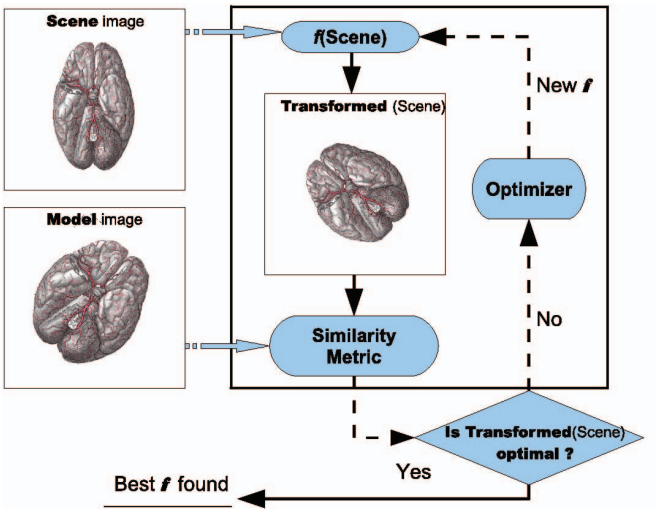


Figure 1. The interactions among the components of a registration technique.

Instead, in parameters-based methods the search is performed directly in the transformation parameters space. Classic numerical optimization algorithm like gradient descent, Newton's method, Powell's method and discrete optimization [12] are among the most common choices, together with approaches based on EC and other meta-heuristics [2]–[6], [13]–[15]. It is common to start the registration process using a "simpler" version of the images obtained through smoothing and downsampling. The registration is divided in multiple stages, called *resolutions*, in which increasingly larger and more detailed versions of the input images are used.

III. Genetic Algorithms for Image Registration

This section describes the methodology proposed in this work, named r-GA. Following the intensity-based approach, our method tackles IR using the images in their whole, rather than considering only some of their features. The registration is carried out through a search in the space of transformation parameters, therefore an individual encodes a transformation directly.

The design of the algorithm can be trivially adapted to different kind of transformations. Let us consider a 3D similarity transformation as an example. Such transformation has a rotation, a translation and a scaling component, so it can be represented by seven real numbers: three to specify the versor of the rotation v, three for the translation t and one for the scaling factor s. An individual of the GA is a real vector with seven elements. Valid solutions require $v_x, v_y, v_z \in [-1, 1]$ and $s > 0$. Note that the translation component is specified in spatial units (e.g, millimeters), rather than in number of voxels.

The operators used in our method are common choices for real-coded genetic algorithms: blend crossover (BLX-α) [16] and random mutation [17]. The random mutation operator randomly picks one the of individual genes and replace it with a random value in the gene's range. Both random choices are made using uniform probability. Blend crossover is more complex. Given to two individuals x and y, called "parents", for each position i of the parents' coding, the algorithms computes the value $d = |x_i - y_i|$ and then randomly generates two values a, b in the interval

$$[\ \min(x_i, y_i) - \alpha d,\ \max(x_i, y_i) + \alpha d\ ]$$

with uniform probability. The values a and b are assigned to the i-th positions of the two offspring. The value α is a positive value controlling the width of the ranges in which the new genes' values are drawn. The fitness value of a solution f is simply the similarity between the two input images when registered using f, i.e. $f \rightarrow \Psi(I_M, f(I_S))$, where Ψ is a similarity metric (e.g. MI) and I_M, I_S are the scene and the model images.

In addition to the GA-based optimizer, r-GA makes use of multiple resolutions and a *restart* mechanism. The algorithm performs the search in two resolutions. In the first one, the algorithm uses a small, low-detail version of the input images that allow the algorithm to quickly obtain a coarse approximation of the desired transformation. In the second one, this approximation is refined using the high-detail input images. 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 is more appropriate to perform again the search for a suitable initial registration by restarting the algorithm. The 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 value). This is the approach we used with the original algorithm presented in [8]. However, the fitness value of a good solution depends on the actual content of the input images and it is hard to predict.

In this work, we propose 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. As the first resolution deals with a low-resolution version of the input images, this stage of the registration is cheap in terms of the total computational effort.

A second improvement is obtained altering the search space in the second resolution. As this phase is meant to be a refinement phase, we focus the search by restricting the search space around the parameters value of the best solution found in the first resolution. For each transformation parameter , the original range $[\ l,\ u\ ]$ is replaced by $[\ b-(b-l)/h,\ b+(u-b)/h\ ]$, where b is the value of the parameter in the best solution and h is the shrinking factor. This ensures the search is performed in an area of high quality solutions.

IV. Experimental Study

The aim of the experimentation is to carry out an objective comparison of the r-GA proposal and other state-of-the-art IR methods. As competitors, we considered an heterogeneous group of algorithms to represent a wide range of approaches to the IR problem, including our recent method GA$^+$ introduced in [8]. They are listed in Table I. Note that the algorithms differ in nature (feature- or intensity-based) as well as in the search strategy (based on matching or transform parameters). Also, different kind of optimization process are used: classic gradient-based techniques, evolutionary computation, metaheuristics.

Table I
The IR algorithms included in the experimental study.

	Nature	Strategy	Optimizer	Ref.
I-ICP	feature	matching	Gradient Descent	[18]
Dyn-GA	feature	parameters	Genetic Algorithms	[2]
SS*	feature	matching	Scatter Search	[19]
ASGD	intensity	parameters	Gradient Descent	[20]
GA$^+$	intensity	parameters	Genetic Algorithms	[8]

We designed two experiments involving synthetic and real-world medical images. To make the comparison as objective as possible, 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 investigate 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 [21] for comparing 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 r-GA 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 test we used is Nemenyi's test [22], which is a post hoc procedure of Friedman's rank sum test [23] and is based on the ranks of the algorithms. As multiple comparison are performed, the p-values of the tests have been adjusted using Holm's method [24] in order to control the family-wise error rate.

For all algorithms, we used the original implementation by the authors. r-GA has been written in C++ and integrated in Elastix [25], 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 [26]. The software is built on top of the popular Insight Segmentation and Registration Toolkit (ITK) [27].

A. First experiment: registration of simulated brain MRIs

The first experiment is similar to the ones carried out in [8], [19]; 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 sixteen 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.

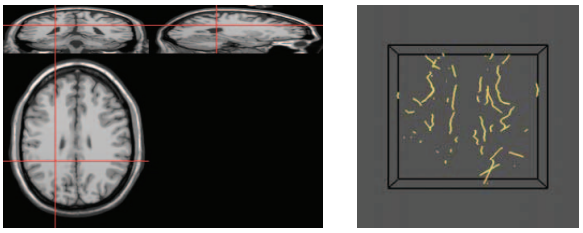
1) Setup:

a) Images: The images used in this experiments were obtained from the BrainWeb database at McGill University [28]. 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 [29]. To create scenarios with different difficulties, we added noise and multiple sclerosis lesions to some of the images, as detailed in Table II. The images are shown in Figure 2; each image has size $60 \times 181 \times 217$ voxels.

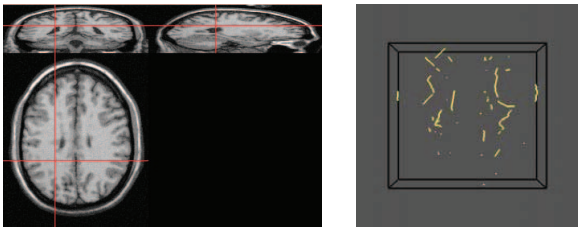
This experiments compares both feature- and intensity-based algorithms, thus some features need to be extracted from the images to provide an input for feature-based algorithms. In the original comparison, the authors computed the isosurfaces and extracted the crest line points with relevant curvature information [30]. It is important to remark this difference: while the input of intensity-based methods consists of the whole images

Table II
THE NOISE LEVEL AND THE PRESENCE OF LESION IN THE FOUR BRAIN MRI IMAGES USED IN THE EXPERIMENTAL STUDY. 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



(a) I_1



(b) I_4

Figure 2. Two MRI brain images (left) used in the first experiment, along with the corresponding crest line points with relevant curvature information (right).

data (in case, two images made of $60 \times 181 \times 217 = 2356620$ voxels having an 8-bit intensity value), that of feature-based approaches is a set of just a few hundred of points (Table II).

b) Registration scenarios: 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 III. 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 $i = 1, 2, 3, 4$.

Table III
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

c) Algorithms and parameters settings: In order to provide a uniform comparison of r-GA with respect to our recent results [8], 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. As for r-GA, 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 five times (i.e. four restart) independently of the results. The population was evolved for 50 generations in the first resolution and 25 in the second. The rest of the configuration was the same configuration in both resolutions: population size of 500 individuals, mutation probability of 0.1, crossover probability of 0.5, blend factor (α) 0.3 and tournament size equal to 3.

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. The stopping criteria needs some discussion. It is challenging to design a fair comparison between algorithms having different inputs, in particular inputs with very different size. In [8] the algorithms were allowed to run for a fixed amount of time: 20 seconds for feature-based algorithms and 20 minutes for intensity-based one. The two amounts of time match the proportion between the size of the inputs for the two kinds of algorithms: the number of voxels in the images is roughly 60 times the number of features.

d) Validation procedure: 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 crestline 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.

2) Analysis of results: Table IV reports the results of the first experiment. For each scenario, we reported mean and standard deviation of the MSE values obtained by the algorithms along with their ranks. The average ranks (Table V) and the count of wins (Table VI) provides another view of the results of the comparison.

From the highest to the lowest average rank is ASGD, I-ICP, Dyn-GA, SS*, GA$^+$ and r-GA. ASGD scored the largest MSE values in all but one of the scenarios. Its performance varies

Table V
FIRST EXPERIMENT: RESULT OF NEMENYI'S TEST COMPARING R-GA 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
r-GA	1.31	
GA$^+$	2.56	0.0285
SS*	2.75	0.0033
Dyn-GA	3.62	0.0000
I-ICP	4.81	
ASGD	5.94	

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

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. Dyn-GA scored better than I-ICP in all scenarios, with less variability between different scenarios, but the gap with the best results is large nevertheless. SS* scored constantly quite close to the best results, ranking third or second in 15 over 16 scenarios. GA$^+$ exhibits an inconsistent behavior. In 11 scenarios, it either scored best or extremely close to the best, while in the remaining 5 scenarios the mean MSE value is really large ($>$1000). As shown in Figure 3, the high average MSE is due to a few solutions having extremely high MSE. This points to the convergence problems we addressed with the new restart mechanisms in r-GA. Indeed, r-GA consistently got either the best mean MSE (11 scenarios) or came really close (i.e. less than 1.0 from the best one). Also, the standard deviation values is always less than 3.0, confirming r-GA is robust and no run of the algorithm has produced a low quality solution.

Table V reports the p-value of Nemenyi's test comparing r-GA 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 r-GA is significantly better than those of the competitors, with the highest p-value being that of GA$^+$, 0.0285.

B. 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 [31]. The quality of the segmentation obtained in this phase is used to assess the effectiveness of the registration methods.

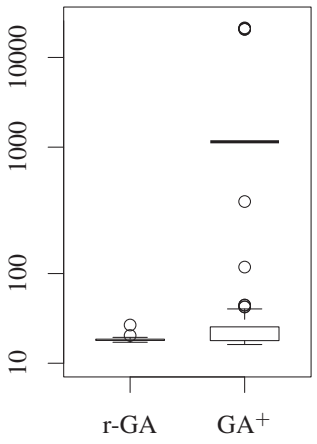


Figure 3. First experiment: boxplots of the results in the third scenario. While the majority of GA$^+$'s results are in an acceptable range, some solutions have very high MSE, explaining the high average MSE (the thick line) scored by the algorithm.

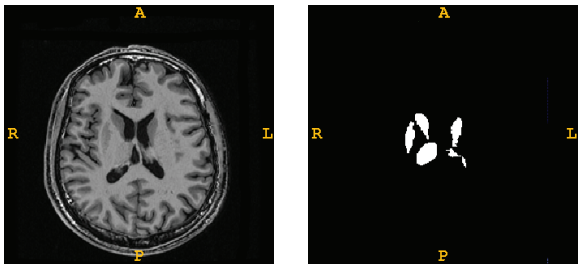


Figure 4. A slice of a 3D MRI brain image used in the second experiment (left) and the corresponding deep brain structure (right).

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 input image (the scene) to the atlas (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 with the segmented region of the atlas is the result of the segmentation process. Often, atlas-based segmentation is used as preliminary step in a more complex segmentation approach.

1) Setup:

a) Images: Thirteen T_1-weighted brain MRI were retrieved from the NMR database [32]. 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. Figure 4 shows one of the images along with the corresponding deep nuclei.

b) Registration scenarios: Nine registration scenarios were created by selecting a pair of different images at random.

Table IV
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	MSE sd	Rank
1	ASGD	61816.4	795.7	6
	Dyn-GA	194.9	50.5	4
	r-GA	36.3	0.5	1
	GA$^+$	36.4	0.3	2
	I-ICP	344.4		5
	SS*	37.0	1.5	3
2	ASGD	34773.9	238.0	6
	Dyn-GA	107.5	52.1	4
	r-GA	36.5	0.5	1
	GA$^+$	36.7	0.4	2
	I-ICP	130.7		5
	SS*	43.4	3.6	3
3	ASGD	111870.0	154.6	6
	Dyn-GA	211.0	137.3	3
	r-GA	41.7	1.6	1
	GA$^+$	1736.5	6216.7	5
	I-ICP	894.3		4
	SS*	63.2	2.9	2
4	ASGD	1233.3	166.8	6
	Dyn-GA	302.0	121.4	4
	r-GA	32.7	0.1	1
	GA$^+$	32.7	0.2	2
	I-ICP	631.7		5
	SS*	53.9	2.6	3
5	ASGD	61063.5	309.1	6
	Dyn-GA	299.3	144.1	4
	r-GA	51.1	0.5	1
	GA$^+$	51.4	0.2	2
	I-ICP	517.7		5
	SS*	112.2	12.4	3
6	ASGD	34796.2	223.8	6
	Dyn-GA	154.0	114.2	4
	r-GA	43.7	0.3	1
	GA$^+$	43.8	0.2	2
	I-ICP	330.3		5
	SS*	56.7	4.5	3
7	ASGD	110131.2	1022.7	6
	Dyn-GA	326.5	174.0	3
	r-GA	56.6	0.7	1
	GA$^+$	1091.8	4965.8	5
	I-ICP	437.8		4
	SS*	63.8	46.2	2
8	ASGD	1017.4	252.7	6
	Dyn-GA	354.3	146.9	4
	r-GA	45.2	0.1	2
	GA$^+$	44.5	0.3	1
	I-ICP	478.0		5
	SS*	122.7	8.2	3
9	ASGD	58146.8	661.0	6
	Dyn-GA	255.4	228.2	4
	r-GA	53.1	0.2	2
	GA$^+$	52.9	0.3	1
	I-ICP	704.3		5
	SS*	183.6	33.0	3
10	ASGD	35695.3	2465.3	6
	Dyn-GA	163.1	57.5	3
	r-GA	46.5	0.2	1
	GA$^+$	476.4	3648.3	4
	I-ICP	1493.2		5
	SS*	89.2	40.8	2
11	ASGD	111384.4	574.2	6
	Dyn-GA	224.9	87.3	3
	r-GA	58.6	2.4	1
	GA$^+$	2823.9	7863.5	5
	I-ICP	951.3		4
	SS*	82.2	45.1	2
12	ASGD	885.8	356.7	6
	Dyn-GA	414.8	258.2	4
	r-GA	47.7	0.2	2
	GA$^+$	47.3	0.4	1
	I-ICP	416.6		5
	SS*	153.9	86.1	3
13	ASGD	56932.0	568.6	6
	Dyn-GA	179.8	59.5	3
	r-GA	35.5	0.3	2
	GA$^+$	35.0	0.2	1
	I-ICP	237.6		5
	SS*	193.1	62.0	4
14	ASGD	31521.1	6.6	6
	Dyn-GA	105.7	50.8	4
	r-GA	30.5	0.2	1
	GA$^+$	30.7	0.3	2
	I-ICP	341.3		5
	SS*	74.9	41.1	3
15	ASGD	112134.4	1027.4	6
	Dyn-GA	192.2	115.8	3
	r-GA	40.7	1.2	1
	GA$^+$	1104.8	5128.7	5
	I-ICP	608.8		4
	SS*	103.8	66.6	2
16	ASGD	512.8	233.2	5
	Dyn-GA	298.1	144.8	4
	r-GA	29.7	0.1	2
	GA$^+$	29.5	0.3	1
	I-ICP	1587.8		6
	SS*	150.2	78.3	3

No transformation was applied on the images; however, the location of the brain in each image is different due to the variability in the pose of the patient during the acquisition of the images.

c) Algorithms and parameters settings: Given the nature of this experiment, we compare only intensity-based algorithms, i.e. r-GA, 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 [33]. It is flexible enough to present a wide range of transformations and it does not produce anatomically unrealistic results, as it could happen with deformable models.

In this experiment we have ground-truth data to evaluate the registration, but we do not know the concrete parameters values of the optimal registration transformations. Therefore, we estimated parameters values intervals considering a big enough range to include all registration solutions for this application. We allowed rotations between -90 and 90 degrees, scaling in the range $[0.9, 1.1]$, shearing in the interval $[-0.1, 0.1]$ and

translations between -15 and 15 centimeters.

For r-GA, we kept the same configuration used in the first experiment. For ASGD we tested several configuration 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 criteria is again time; as the magnitude of the transformations involved is smaller than in the previous experiment, the time limit was set to 10 minutes.

p

d) Validation procedure: 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 [34], given by $\text{Dice}(V_R, V_{\text{GT}}) = 2|V_R \cap V_{\text{GT}}|/(|V_R| + |V_{\text{GT}}|)$ 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.

2) Analysis of results: The results of the second experiment are reported in Table VII. We computed the mean and standard deviation of the overlap for each scenario. Table IX shows the count of wins for the algorithms in the comparison.

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. ASGD and GA$^+$ had a similar performance. They have almost identical mean rank values (2.33 and 2.44) and a similar number of wins against each other (5 and 4). Again, GA$^+$ occasionally has quite large standard deviation values compared to the others, e.g. scenario number 2. r-GA ranked first in 8 out of 9 scenarios and came second in the last one, delivering the best performance both in terms of ranking and number of wins. The results Nemenyi's test (Table VIII) show the advantage of r-GA over the other algorithms is statistically significant. The adjusted p-values of the tests are both 0.019.

V. Conclusions

In this work, we extend our previous intensity-based IR technique based on genetic algorithms. Our method uses a modern, real-coded design for solutions and genetic operators, as well as a multi-resolution strategy, allowing the registration to be performed in multiple stages with increasing complexity. By improving the restart mechanism and focusing the search to the right area of the search space, we overcame the convergence problems experienced by the original method while improving the precision of the algorithm.

The merit of this new approach is proved experimentally in two separate studies involving synthetic and real-world medical images. Each study included a comparison with other

Table VII
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	Overlap sd	Rank
1	ASGD	.742	.001	3
	r-GA	.755	.012	1
	GA$^+$	.751	.010	2
2	ASGD	.616	.005	2
	r-GA	.618	.007	1
	GA$^+$	.615	.033	3
3	ASGD	.677	.003	2
	r-GA	.679	.008	1
	GA$^+$	.676	.012	3
4	ASGD	.691	.001	3
	r-GA	.706	.016	1
	GA$^+$	.698	.011	2
5	ASGD	.756	.010	2
	r-GA	.760	.009	1
	GA$^+$	.755	.009	3
6	ASGD	.738	.003	2
	r-GA	.739	.005	1
	GA$^+$	.734	.011	3
7	ASGD	.686	.009	3
	r-GA	.729	.004	1
	GA$^+$	.717	.015	2
8	ASGD	.741	.001	3
	r-GA	.750	.007	2
	GA$^+$	.751	.020	1
9	ASGD	.754	.004	1
	r-GA	.749	.014	2
	GA$^+$	.745	.017	3

Table VIII
SECOND EXPERIMENT: RESULT OF NEMENYI'S POST-HOC PROCEDURE WHEN COMPARING R-GA 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
r-GA	1.22	
ASGD	2.33	0.0190
GA$^+$	2.44	0.0190

Table IX
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	r-GA	GA$^+$
ASGD	-	1	5
r-GA	8	-	8
GA$^+$	4	1	-

state-of-the-art IR methods using a wide range of approaches to the problem. In both studies, our approach delivered an excellent performance and it was able to outperform all other algorithms in almost all the scenarios.

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 registrations.

ACKNOWLEDGMENT

This work is supported by the European Commission with the contract No. 238819 (MIBISOC Marie Curie ITN). 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 post-processed with Brainvisa/Connectomist software, freely available at http://brainvisa.info.

REFERENCES

[1] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image Vision Comput.*, vol. 21, pp. 977–1000, 2003.

[2] C. K. Chow, H. T. Tsui, and T. Lee, "Surface registration using a dynamic genetic algorithm," *Pattern Recogn.*, vol. 37, pp. 105–117, 2004.

[3] O. Cordón, S. Damas, and J. Santamaría, "A Fast and Accurate Approach for 3D Image Registration using the Scatter Search Evolutionary Algorithm," *Pattern Recogn. Lett.*, vol. 27, no. 11, pp. 1191–1200, 2006.

[4] O. Cordón, S. Damas, and J. Santamaría, "Feature-based image registration by means of the CHC evolutionary algorithm," *Image Vision Comput.*, vol. 22, pp. 525–533, 2006.

[5] E. Lomonosov, D. Chetverikov, and A. Ekart, "Pre-registration of arbitrarily oriented 3D surfaces using a genetic algorithm," *Pattern Recogn. Lett.*, vol. 27, no. 11, pp. 1201–1208, 2006.

[6] L. Silva, O. R. P. Bellon, and K. L. Boyer, *Robust range image registration using genetic algorithms and the surface interpetenetration measure*. World Scientific, 2005.

[7] S. Damas, O. Cordón, and J. Santamaría, "Medical image registration using evolutionary computation: An experimental survey," *IEEE Computational Intelligence Magazine*, vol. 6, no. 4, pp. 26 –42, nov. 2011.

[8] A. Valsecchi, S. Damas, and J. Santamaría, "An image registration approach using genetic algorithms," in *Proceedings of the IEEE World Congress On Computational Intelligence - Congress on Evolutionary Computation 2012*, 2012, pp. 1–8.

[9] M. Svedlow, C. D. Mc-Gillem, and P. E. Anuta, "Experimental examination of similarity measures and preprocessing methods used for image registration," in *Symposium on Machine Processing of Remotely Sensed Data*, vol. 4(A), Indiana, USA, 1976, pp. 9–17.

[10] M. A. Audette, F. P. Ferrie, and T. M. Peters, "An algorithmic overview of surface registration techniques for medical imaging," *Med. Image Anal.*, vol. 4, no. 3, pp. 201–217, 2000.

[11] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE T. Med. Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[12] F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for image registration by maximization of mutual information," *Med. Image Anal.*, vol. 3, no. 4, pp. 373–386, 1999.

[13] J. M. Rouet, J. J. Jacq, and C. Roux, "Genetic algorithms for a robust 3-D MR-CT registration," *IEEE T. Inf. Technol. B.*, vol. 4, no. 2, pp. 126–136, 2000.

[14] R. He and P. A. Narayana, "Global optimization of mutual information: application to three-dimensional retrospective registration of magnetic resonance images," *Comput. Med. Imag. Grap.*, vol. 26, pp. 277–292, 2002.

[15] P. Chalermwat, T. El-Ghazawi, and J. LeMoigne, "2-phase GA-based image registration on parallel clusters," *Future Gener. Comp. Sy.*, vol. 17, pp. 467–476, 2001.

[16] L. J. Eshelman, "Real-coded genetic algorithms and interval schemata," in *Foundations of Genetic Algorithms 2*, L. D. Whitley, Ed. San Mateo, USA: Morgan Kaufmann, 1993, pp. 187–202.

[17] T. Bäck, D. B. Fogel, and Z. Michalewicz, *Handbook of Evolutionary Computation*. IOP Publishing Ltd and Oxford University Press, 1997.

[18] Y. Liu, "Improving ICP with easy implementation for free form surface matching," *Pattern Recogn.*, vol. 37, no. 2, pp. 211–226, 2004.

[19] O. Cordón, S. Damas, J. Santamaría, and R. Martí, "Scatter search for the point-matching problem in 3D image registration," *INFORMS Journal on Computing*, vol. 20, no. 1, pp. 55–68, 2008.

[20] S. Klein, J. Pluim, M. Staring, and M. Viergever, "Adaptive stochastic gradient descent optimisation for image registration," *International Journal of Computer Vision*, vol. 81, pp. 227–239, 2009.

[21] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[22] P. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton University, 1963.

[23] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940. [Online]. Available: http://dx.doi.org/10.2307/2235971

[24] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: http://dx.doi.org/10.2307/4615733

[25] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[26] "Elastix webpage," http://elastix.bigr.nl, 2012.

[27] L. Ibanez, W. Schroeder, L. Ng, and J. Cates, *The ITK Software Guide*, 2nd ed., Kitware, Inc. ISBN 1-930934-15-7, 2005.

[28] D. L. Collins, A. P. Zijdenbos, V. Kollkian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, "Design and construction of a realistic digital brain phantom," *IEEE T. Med. Imaging*, vol. 17, pp. 463–468, 1998.

[29] P. Rogelj and S. Kovacic, "Validation of a Non-Rigid Registration Algorithm for Multimodal Data," in *SPIE in Medical Imaging*, M. Sonka and J. M. Fitzpatrick, Eds., 2002, pp. 299–307.

[30] O. Monga, S. Benayoun, and O. Faugeras, "From partial derivatives of 3D density images to ridges lines," in *Computer Vision and Pattern Recognition*. Champaign, Illinois, USA: IEEE, 1992, pp. 354–389.

[31] B. C. Vemuri, J. Ye, Y. Chen, and C. M. Leonard, "Image registration via level-set motion: Applications to atlas-based segmentation," *Medical Image Analysis*, vol. 7, no. 1, pp. 1–20, 2003.

[32] C. Poupon, F. Poupon, L. Allirol, and J.-F. Mangin, "A database dedicated to anatomo-functional study of human brain connectivity," in *Proceedings of the 12th Annual Meeting of the Organization for Human Brain Mapping*, no. 646, Florence, Italy, 2006.

[33] D. Rueckert and J. A. Schnabel, "Medical image registration," in *Biomedical Image Processing*, ser. Biological and Medical Physics, Biomedical Engineering, T. M. Deserno, Ed. Springer Berlin Heidelberg, 2011, pp. 131–154.

[34] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.