Validation of DRAMMS among 12 Popular Methods in Cross-Subject Cardiac MRI Registration

Yangming Ou, Dong Hye Ye, Kilian M. Pohl, and Christos Davatzikos

Section of Biomedical Image Analysis (SBIA),
Department of Radiology, University of Pennsylvania

Abstract. Cross-subject image registration is the building block for many cardiac studies. In the literature, it is often handled by voxel-wise registration methods. However, studies are lacking to show which methods are more accurate and stable in this context. Aiming at answering this question, this paper evaluates 12 popular registration methods and validates a recently developed method DRAMMS [16] in the context of cross-subject cardiac registration. Our dataset consists of short-axis end-diastole cardiac MR images from 24 subjects, in which non-cardiac structures are removed. Each registration method was applied to all 552 image pairs. Registration accuracy is approximated by Jaccard overlap between deformed expert annotation of source image and the corresponding expert annotation of target image. This accuracy surrogate is further correlated with deformation aggressiveness, which is reflected by minimum, maximum and range of Jacobian determinants. Our study shows that DRAMMS [16] scores high in accuracy and well balances accuracy and aggressiveness in this dataset, followed by ANTs [13], MI-FFD [14], Demons [15], and ART [12]. Our findings in cross-subject cardiac registrations echo those findings in brain image registrations [7].

Keywords: Image Registration, Validation, Evaluation, Cardiac MRI.

1 Introduction

Cross-subject image registration rests in the core of many cardiac studies. Examples include atlas construction [3], atlas-based segmentation [4], and morphologic study to understand disease patterns [5].

In literature, cross-subject cardiac image registration is often handled by voxel-wise registration methods [6]. Voxel-wise registration methods rely on image information only, and do not require anatomic information or human intervention. Therefore, they can be applied to various organs including the heart [6]. Some basic question remains, however: 1) which voxel-wise registration methods are more accurate and more stable in cross-subject cardiac registration context; 2) whether those more accurate methods in cardiac registration coincide with those in brain image registrations (e.g., as found in [7]). The answers to these questions are not immediately clear, largely because the heart is usually imaged

B.M. Dawant et al. (Eds.): WBIR 2012, LNCS 7359, pp. 209–219, 2012.

with lower resolution, lower signal-to-noise ratio (SNR), more severe moving artifacts, and has a very different shape than the brain.

Towards answering these questions, this paper evaluates 12 commonly-used and publically-available registration methods and validates a recently developed method DRAMMS [16] in the context of cross-subject cardiac registrations. We have collected short-axis end-diastole magnetic resonance (MR) images of 24 subjects. By permuting source and target images, this dataset results in 552 possible pair-wise registrations for each of those 12 registration methods. The large number of experiments (perhaps largest to date in cardiac context) is **the first feature** of this study. **The second feature** of this study is the comprehensive evaluation criteria. Unlike other evaluation studies (e.g. [7]) that only measure accuracy, we measure both accuracy and aggressiveness of deformations, and visualize their relationship in a joint plot. A deformation is considered more "aggressive" if it leads to self-foldings at more locations, and if it takes greater expansions/shrinkages to capture cross-individual variations. Aggressiveness and accuracy are usually a pair of trade-off. Higher accuracy often comes from increased aggressiveness in deformation. On the other hand, too aggressive deformation will undesirably break topology. An ideal method should achieve high accuracy while accurately preserving topology. Measuring both accuracy and aggressiveness will help reveal which methods better balance the two. **The third feature** of this study is that, instead of using only one set of parameters, we have examined two parameter settings for the four more accurate methods – one more aggressive and one smoother version. This is important, because different cardiac studies will have different requirements on aggressiveness levels of deformation. It also helps reveal which methods achieve consistently high accuracy when aggressiveness levels change.

In the rest of the paper, we present evaluation protocol in Section 2 and evaluation results in Section 3. We discuss and conclude the paper in Section 4.

2 Evaluation Protocol

This section describes our evaluation protocol. It contains three parts: description of dataset (Section 2.1), brief review of registration methods included in this study (Section 2.2), and description of evaluation criteria (Section 2.3).

2.1 Dataset for Evaluation

We now describe the dataset and pre-processings. Three-dimensional short-axis cardiac MR images of 24 subjects are collected at end-diastole phase. The image dimension is $120 \times 120 \times 12$ and voxel size is $1.25 \times 1.25 \times 8.0mm^3$. Common preprocessing steps include respiratory motion correction [19] and N3-based bias field correction [20]. Non-cardiac structures are removed by a semi-automatic process. In this process, the heart is first automatically outlined by a public software "Segment" [18]. Then, a cardiovascular expert refined the separation of cardiac and non-cardiac structures. Removal of non-cardiac structures is similar to skull-stripping in brain image registrations. The purpose is to remove

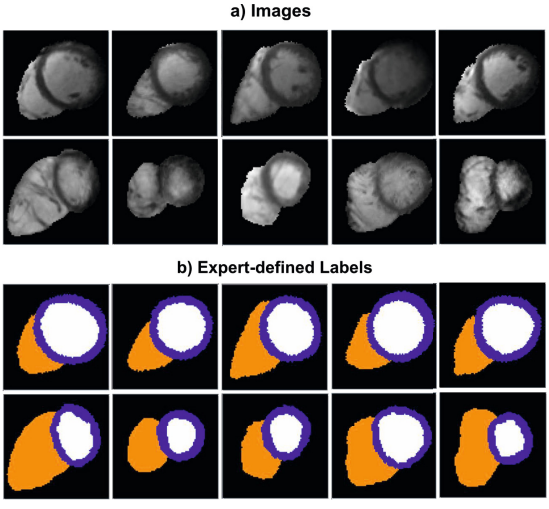

Fig. 1. Images (a) and expert-annotation of structures (b) for some 10 typical subjects from the dataset used in this study. Subjects in the first row in (a) are healthy controls and in the second row are with tetralogy-of-fallot (TOF) defect. In the expert annotation, white, orange and blue regions are LV, RV and myocardium, respectively.

unnecessary challenges, especially when different images may contain different non-cardiac structures due to different fields of view. Each cardiac image is further annotated by the same cardiologist into three structures – left ventricle (LV), right ventricle (RV) and myocardium. Some typical intensity images and expert-annotation images are shown in Fig. 1. We note that, except for removing non-cardiac structures, those expert annotations of LV/RV/myocardium are in no means used as any part of the registration process. They are only used to evaluate registration accuracy.

This dataset represents the common challenges in cardiac registrations – lower resolution, lower SNR, more severe moving artifacts and quite different shape from the brain. Besides, 11 out of 24 subjects have tetralogy-of-fallot (TOF) defect, hence having irregular ventricle shapes largely different from the remaining 13 normal subjects (Fig. 1).

2.2 Registration Methods to Be Evaluated

A total of 12 widely-used and publically-available methods are included in this study (Table 1). We note that they are only a small fraction of the vast number of registration methods developed in the community. The pool can be always expanded in the future to include other widely-acknowledged methods. In general, we chose those 12 methods because of the wide variety they represent. That is, they have different similarity measures, different deformation models and different optimization strategies, which are the most important components for registration algorithms (see Table 1). Out of those 12 registration methods, 9 methods were included in a recent brain registration evaluation study [7].

In addition, we have included three registration methods that were not included in that brain study [7]. Those three methods are: Demons [15] (a widely-used, ITK-based, public and fast software), DRAMMS [16] (our method that matches images by voxel-wise texture attributes instead of intensities), and DROP [17] (a novel discrete optimization strategy that is fast and accurate).

To encourage objectivity in evaluation, we need to take special care of parameters for different methods. In some previous evaluation studies [7,8], parameters are provided by authors of each method. However, this is not without problem. One issue is the lack of comparability in their aggressiveness levels, and hence possible unfairness to those methods that generate smoother deformations. Actually, almost all methods can score higher accuracy at more aggressive deformations. Ideally, we should require similar aggressiveness level for all methods, and then compare their accuracies. A second issue is the lack of information about sensitivity of accuracy with regard to parameter changes. With only one set of best parameters, it is hard to tell sensitivity.

To cope with those two issues and to promote objectivity, we set parameters by the following two rules. To settle the first issue, we tune parameters not just for best accuracy, but for best accuracy at similar aggressiveness level. Specifically, we start from parameters in a method's user manual or past papers. In each iteration, we keep other methods' parameters fixed, and slightly adjust one method's parameters until its deformations are at similar level with most other methods (few or no self-foldings, similar min, max and range of Jacobian determinants). We iterate on every method until they all converge to similar aggressiveness level. This provides common ground for more objectively evaluating their accuracies. To settle the second issue, we provide two sets of parameters, instead of only one most accurate set, for the four most accurate methods. One aggressive set for generally higher accuracy but increased risk of self-folding; and one smooth set for generally smoother deformation but lower

Table 1. Registration methods to be evaluated in this paper (diff.–diffeomorphism; MI – mutual information; NMI – normalized MI; SSD – sum of squared difference; SAD – sum of absolute difference; MSD – mean squared difference; CC – correlation coefficient; NCC – normalized CC)

Method	Deformation Model	Similarity	Regularization
flirt [9]	affine	SSD/(N)MI/CC	–
fnirt [10]	cubic B-spline	SSD	bending energy
AIR [11]	5^{th} polynomial	MSD	by polynomial
ANTs [13]	symmetric diff.	CC	Gaussian smoothing
ART [12]	homeomorphism	NCC	Gaussian smoothing
CC-FFD [14]	cubic B-spline	CC	bending energy
MI-FFD [14]	cubic B-spline	MI	bending energy
SSD-FFD [14]	cubic B-spline	SSD	bending energy
DROP [17]	cubic B-spline	SAD	bending energy
Demons [15]	optical flow	SSD	Gaussian smoothing
Diff. Demons [15]	diff. optical flow	SSD	Gaussian smoothing
DRAMMS [16]	cubic B-spline	SSD of attributes	bending energy

accuracy. This reveals consistency of accuracy as parameters change. All parameters used in this paper can be found at `http://www.seas.upenn.edu/ ouya/documents/research/Ou12_WBIR_Supplementary.pdf`.

To avoid bias in template selection, we have considered all possible images as source and target in registration. This results in a total of 552 ($= 24 \times 23$) possible pair-wise registrations for each registration method.

2.3 Evaluation Criteria

This sub-section presents the criteria for evaluating both deformation accuracy and aggressiveness. Specifically, accuracy is implied by Jaccard Overlap between deformed expert-annotation of source image and the expert-annotation of target image. We measure overlaps in 3 regions: LV, RV, and myocardium. Larger overlap often indicates greater spatial alignment between subjects [7,21].

A deformation is considered more "aggressive" if it has self-foldings at more locations, and if it takes greater expansions/shrinkages to capture cross-individual variability. In measuring deformation aggressiveness, we have used Jacobian determinants. Jacobian determinant measures voxel-wise volumetric change ratio. It is > 1 for expansion, between 0 and 1 for shrinkage and < 0 if self-folding occurs. In particular, we measure 4 Jacobian-based metrics: 1) the number of deformations having negative Jacobian determinants; 2) the percentage of voxels having negative Jacobian determinants; 3) minimum and 4) maximum Jacobian determinants in a deformation. Finally, we use one metric, the range of Jacobian determinants (=maxJac-minJac), to quantify deformation aggressiveness.

For fairness, we used a standard ITK calculator to compute Jacobians of deformation. This requires converting deformation files from different software into a standard ITK-compatible MetaImage format. We carefully checked to assure the conversion reproduces the same exact warped images.

3 Results and Observations

We now present evaluation results (accuracy, aggressiveness, and their correlation) in this section. Observations follow each set of results. Average computational time of each method is listed in Appendix of this paper.

3.1 Deformation Accuracy Indicated by Jaccard Overlap is shown in Fig. 2 for myocardium, LV and RV. Several observations can be made:

a) in general, voxel-wise registration methods evaluated in this paper have obtained 0.6-0.9 Jaccard (roughly 0.75-0.95 Dice) overlap in left and right ventricles, and 0.4-0.7 Jaccard (roughly 0.55-0.85 Dice) overlap in myocardium.

b) DRAMMS scores highest Jaccard overlap in all three structures in this dataset – average 0.85 Jaccard (0.9 Dice) in LV and RV, 0.7 Jaccard (0.8 Dice) in myocardium. The margin is bigger in myocardium regions. A plausible explanation is that DRAMMS uses texture attributes other than solely intensity information to define similarity at each voxel.

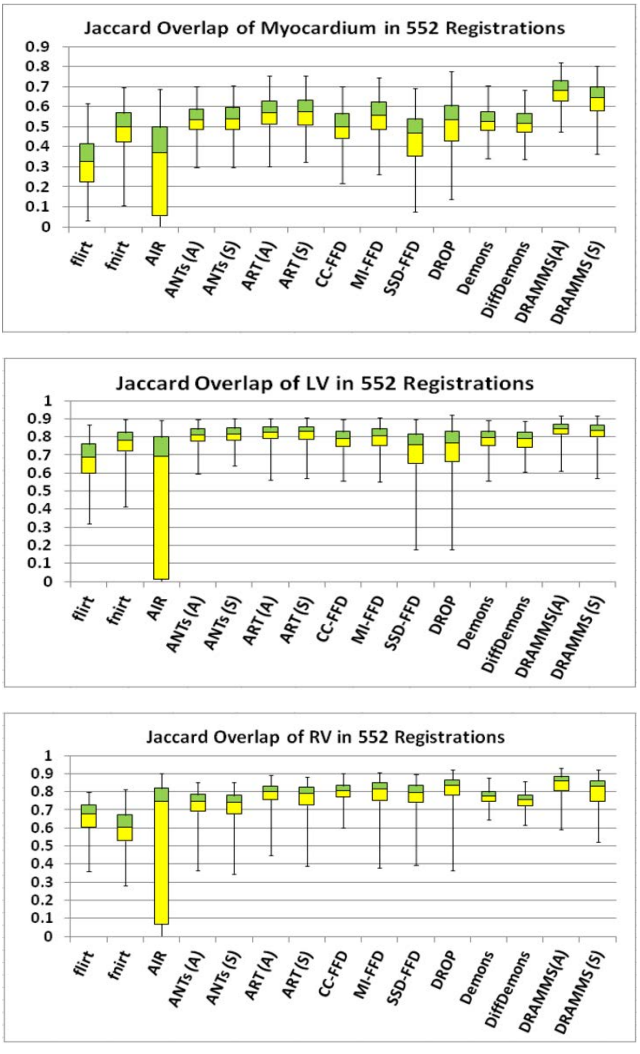

Fig. 2. Box-and-Whisker plots: accuracy indicated by Jaccard overlap in 3 expert-annotated structures. From top to bottom, results for myocardium, LV and RV regions. Letter "A" stands for aggressive version and "S" for smooth version of a method.

c) ANTs, MI-FFD, Demons, and ART also obtained high overlaps in this cardiac dataset. This echoes findings in brain registration evaluation study [7].

d) Methods using intensity differences (SSD) as similarity metric have reasonable Jaccard overlap on average. However, they have larger variations, and suffer in difficult cases. This shows that SSD metric is less likely to consistently capture large anatomical variations. One solution is to combine intensity difference with deformation mechanism of more degrees of freedom (like in ART and Demons).

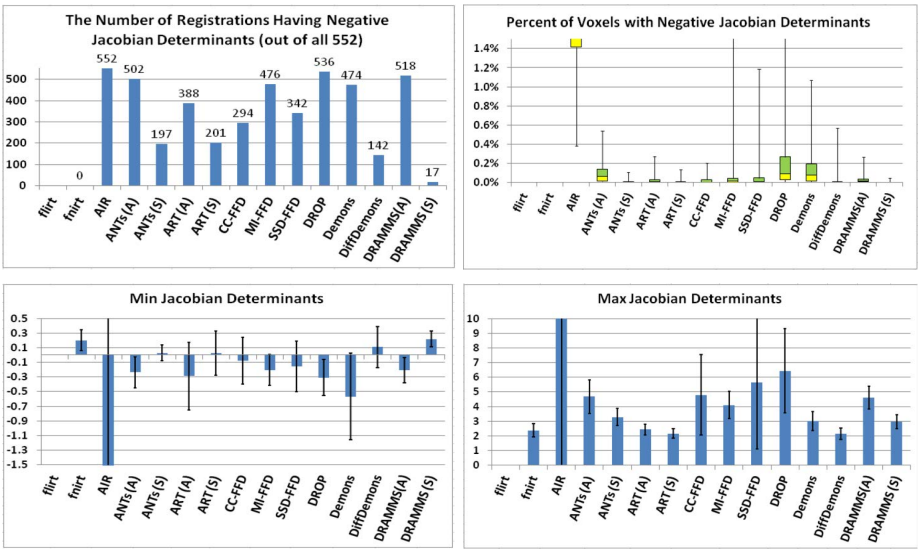

Fig. 3. Jacobian-based metrics to indicate deformation aggressiveness. Upper left: number of deformations (out of all 552) that have negative Jacobian determinants; Upper right: box-and-whisker plot of percentage of voxels having negative Jacobian determinants in a deformation; Lower row: min (left) and max (right) Jacobian determinants.

A perhaps better solution is to replace it with more robust similarity metric, such as correlation (like in ANTs), mutual information (like in MI-FFD), or attribute-based similarity (like in DRAMMS).

3.2 Deformation Aggressiveness is indicated by the four sets of results shown in Fig. 3. From left to right, top to bottom, they are: number of deformations with negative Jacobian determinants; percent of voxels having negative Jacobian determinants; minimum and maximum Jacobian determinants in deformations. We observe the following from those results:

a) From the top row in Fig. 3, fnirt is the only non-rigid registration method that guarantees diffeomorphism in this dataset. Diffeomorphism means no existence of negative Jacobian determinants (i.e. no self-folding) in deformations. It is a nice property that preserves topology and one-to-one forward and backward correspondences. fnirt guarantees diffeomorphism by directly checking and removing negativity in Jacobian map. However, this is at the cost of overlap-indicated registration accuracy, as reflected in Fig. 2. Actually, whether cross-subject deformation is a diffeomorphism is an unknown matter, especially when there are large anatomic variations.

b) DRAMMS(A), ANTs(A), MI-FFD, Demons and ART(A) scored higher overlap in Fig. 2. Interestingly, results in lower row of Fig. 3 show they have quite different deformation styles. In particular, DRAMMS(A), ANTs(A) and MI-FFD

have greater maximum Jacobian determinants, trying to capture individual variability with larger expansions. Demons and ART(A) have more negative minimum Jacobian determinants, trying to capture individual variability with more self-foldings in deformations.

3.3 Correlation between Accuracy and Aggressiveness Surrogates is depicted in Fig. 4. Here y-axis is the mean Jaccard overlap over all 3 structures and all 552 registrations, indicating overall accuracy of a registration method. X-axis is the mean range of Jacobian determinants (=mean(maxJacobianDet-minJacobianDet)) over all 552 registrations, indicating aggressiveness of a method. Three observations can be made from this figure:

a) Methods score higher overlap at more aggressive deformations.

b) An ideal registration method should obtain highest possible overlap while preserving diffeomorphism. Combining Fig. 4 with upper left part of Fig. 3, DRAMMS(S), the smooth version of DRAMMS, obtained second highest overlap and preserved diffeomorphism in almost all but 3% (17/552) deformations.

c) In Fig. 4, we used dashed lines to connect the smooth and aggressive versions of four top-ranking methods. As a result, we observe that DRAMMS is general high in accuracy. More importantly, it has greater increase when going from smooth to aggressive version. It therefore offers wider range of choices for varying needs. That is, the aggressive version, DRAMMS(A), seems a good choice for single-/multi-atlas-based segmentation, where overlap is the focus. The smoother version, DRAMMS(S), is perhaps a better choice for finding common disease pattern in a population, where the key is to maximum possibly remove global difference and meanwhile preserve disease-induced individual variability.

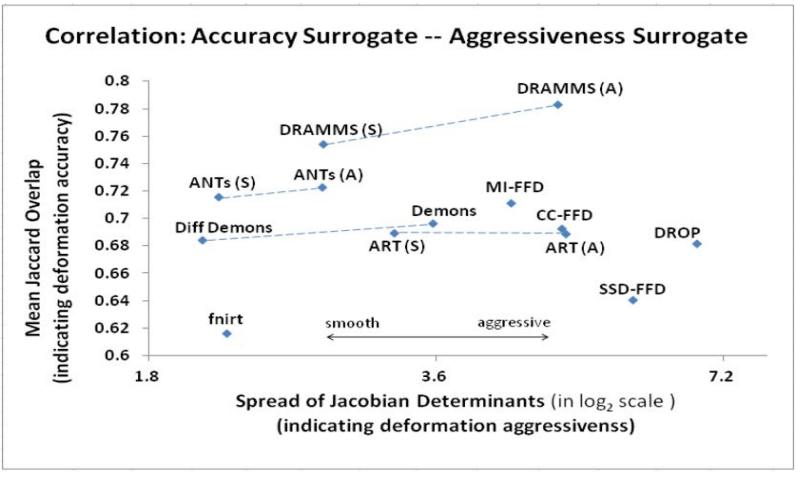

Fig. 4. Correlation between accuracy and aggressiveness surrogates. Letter "A" stands for aggressive and "S" for smooth versions for some methods.

4 Discussion

This paper evaluates 12 voxel-wise registration methods within the context of cross-subject cardiac registrations in a dataset of 24 subjects. Results show that those top-ranking registration methods – DRAMMS, ANTs, MI-FFD, Demons, ART – obtained average Jaccard overlap of 0.7-0.9 (i.e. Dice of 0.82-0.95) in left and right ventricles, and 0.5-0.7 (i.e., Dice of 0.66-0.82) in myocardium. In the following, we will discuss those important aspects of the paper.

Objectivity is a critical issue. In our study, it is encouraged by looking at accuracies when most methods are at similar aggressiveness levels. Deformation accuracy and aggressiveness are often a pair of trade-off. Reporting both and correlating them are a more comprehensive set of criteria than purely accuracy criterion. Their results (Figs. 2,3) and their correlation (Fig. 4) show that the smooth version of DRAMMS achieves best balance – high overlap and maximum preservation of diffeomorphism. ANTs, MI-FFD, Demons and ART also perform well in this cardiac dataset. This echoes findings in brain registration study [7].

On the note of similarity metrics, intensity difference is less stable than correlation (like in ANTs), mutual-information (like in MI-FFD) or attribute-based similarity (like in DRAMMS). On transformation models, different behaviors are observed. Transformation models behind DRAMMS, ANTs and MI-FFD tend to capture individual variability by larger deformation expansions and less severe self-foldings. Models behind Demons and ART tend to behave reversely.

One surprising observation is regarding diffeomorphism. fnirt is the only one that guarantees diffeomorphism in this dataset, as it directly checks and removes negative Jacobian determinants. Non-diffeomorphism occurs for many methods, although some were theoretically designed diffeomorphic. Numerical issues might be one reason. Or, perhaps the process of deforming subjects with large anatomical variability itself is not completely diffeomorphic in nature.

Future work includes additional validations that consist of additional registration methods, cardiac datasets, and accuracy surrogates like surface distance.

Acknowledgement. The project described was supported in part by Grant UL1RR024134 from the National Center for Research Resources, and in part by the Institute for Translational Medicine and Therapeutics (ITMAT) Transdisciplinary Awards Program at the University of Pennsylvania. We thank Dr. Litt Harold, from Cardiovascular Imaging Section of Hospital of the University of Pennsylvania, for annotating cardiac structures that serve as ground truth for our evaluation.

References

1. Chandrashekara, R., Rao, A., Sanchez-Ortiz, G.I., Mohiaddin, R.H., Rueckert, D.: Construction of a Statistical Model for Cardiac Motion Analysis Using Nonrigid Image Registration. In: Taylor, C.J., Noble, J.A. (eds.) IPMI 2003. LNCS, vol. 2732, pp. 599–610. Springer, Heidelberg (2003)

2. Isola, A., Grass, M., Niessen, W.J.: Fully automatic nonrigid registration-based local motion estimation for motion-corrected iterative cardiac CT reconstruction. Med. Phy., 1093–1109 (2010)
3. Perperidis, D., Mohiaddin, R., Rueckert, D.: Spatio-temporal free-form registration of cardiac MR image sequences. MedIA 9, 441–456 (2005)
4. Zhuang, X., Rhode, K.S., Razavi, R.S., Hawkes, D.J., Ourselin, S.: A Registration-Based Propagation Framework for Automatic Whole Heart Segmentation of Cardiac MRI. TMI, 1612–1625 (2010)
5. Ye, D.H., Litt, H., Davatzikos, C., Pohl, K.M.: Morphological Classification: Application to Cardiac MRI of Tetralogy of Fallot. In: Metaxas, D.N., Axel, L. (eds.) FIMH 2011. LNCS, vol. 6666, pp. 180–187. Springer, Heidelberg (2011)
6. Makela, T., Clarysse, P., Sipila, O., Pauna, N., Pham, Q., Katila, T., Magnin, I.E., Axis, L.L.: A review of cardiac image registration methods. TMI 21, 1011–1021 (2002)
7. Klein, A., et al.: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786–802 (2009)
8. Murphy, K., van Ginneken, B., Reinhardt, J.M., et al.: Evaluation of Registration Methods on Thoracic CT: The EMPIRE10 Challenge. TMI 30, 1901–1920 (2011)
9. Jenkinson, M., Smith, S.: A global optimisation method for robust affine registration of brain images. MedIA 5(2), 143–156 (2001)
10. Andersson, J., Smith, S., Jenkinson, M.: FNIRT–FMRIB's non-linear image registration tool. Human Brain Mapping (2008)
11. Woods, R., Grafton, S., Holmes, C., Cherry, S., Mazziotta, J.: Automated image registration: I. general methods and intrasubject intramodality validation. JCAT, 139–152 (1998)
12. Ardekani, B., Guckemus, S., Bachman, A., Hoptman, M.J., Wojtaszek, M., Nierenberg, J.: Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans. J. Neu. Methods. 142, 67–76 (2005)
13. Avants, B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. MedIA 12, 26–41 (2008)
14. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast MR images. TMI 18, 712–721 (1999)
15. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient nonparametric image registration. NeuroImage 45(1), 61–72 (2009)
16. Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C.: DRAMMS: Deformable Registration via Attribute Matching and Mutual-Saliency Weighting. MedIA, 622–639 (2011)
17. Glocker, B., Komodakis, N., Tziritas, G., Navab, N., Paragios, N.: Dense image registration through MRFs and efficient linear programming. MedIA, 731–741 (2008)
18. Heiberg, E., Sjogren, J., Ugander, M., Carlsson, M., Engblom, H., Arheden, H.: Design and Validation of Segment - a Freely Available Software for Cardiovascular Image Analysis. BMC Medical Imaging 10, 1 (2010)
19. Zhang, H., Wahle, A., Johnson, R., Scholz, T., Sonka, M.: 4D Cardiac MR Image Analysis: Left and Right Ventricular Morphology and Function. TMI, 350–364 (2010)
20. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. TMI 17(1), 87–97 (1998)

21. Christensen, G.E., Geng, X., Kuhl, J.G., Bruss, J., Grabowski, T.J., Pirwani, I.A., Vannier, M.W., Allen, J.S., Damasio, H.: Introduction to the Non-rigid Image Registration Evaluation Project (NIREP). In: Pluim, J.P.W., Likar, B., Gerritsen, F.A. (eds.) WBIR 2006. LNCS, vol. 4057, pp. 128–135. Springer, Heidelberg (2006)

Appendix: Computational Time

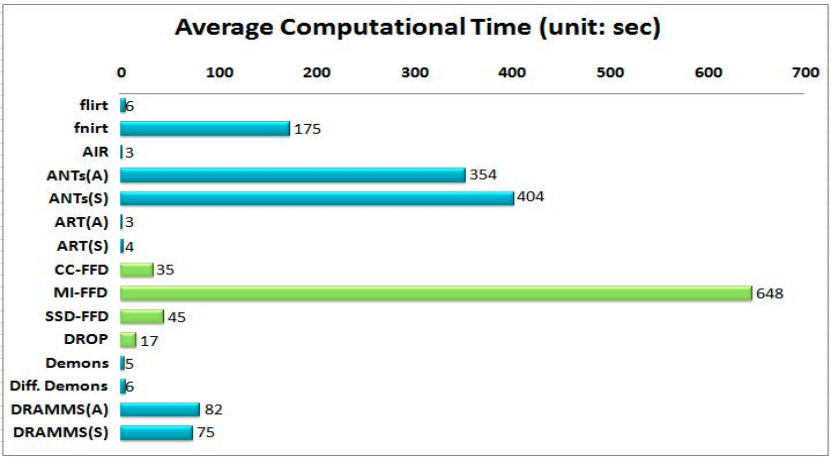

Fig. 5. Average computation time to register a pair of cardiac images in our dataset ($120 \times 120 \times 12 voxels$, $1.25 \times 1.25 \times 8.0 mm^3/voxel$). Blue bars are times in Linux centOS-5 Operating System, Xeon 2.80GHz CPU, 48GB memory. Green bars are times in Windows 7 Operating System, Intel i7 2.93GHz CPU, 4GB memory.