A Stochastic Quasi-Newton Method for Non-Rigid Image Registration

Yuchuan Qiao[1], Zhuo Sun[1], Boudewijn P.F. Lelieveldt[1,2], and Marius Staring[1]

[1] Division of Image Processing (LKEB) Department of Radiology,
Leiden University Medical Center, Leiden, The Netherlands
{Y.Qiao,Z.Sun,B.P.F.Lelieveldt,M.Staring}@lumc.nl
[2] Department of Intelligent Systems,
Delft University of Technology, Delft, The Netherlands

Abstract. Image registration is often very slow because of the high dimensionality of the images and complexity of the algorithms. Adaptive stochastic gradient descent (ASGD) outperforms deterministic gradient descent and even quasi-Newton in terms of speed. This method, however, only exploits first-order information of the cost function. In this paper, we explore a stochastic quasi-Newton method (s-LBFGS) for non-rigid image registration. It uses the classical limited memory BFGS method in combination with noisy estimates of the gradient. Curvature information of the cost function is estimated once every L iterations and then used for the next L iterations in combination with a stochastic gradient. The method is validated on follow-up data of 3D chest CT scans (19 patients), using a B-spline transformation model and a mutual information metric. The experiments show that the proposed method is robust, efficient and fast. s-LBFGS obtains a similar accuracy as ASGD and deterministic LBFGS. Compared to ASGD the proposed method uses about 5 times fewer iterations to reach the same metric value, resulting in an overall reduction in run time of a factor of two. Compared to deterministic LBFGS, s-LBFGS is almost 500 times faster.

1 Introduction

Image registration is important in the field of medical image analysis. However, this process is often very slow because of the large number of voxels in the images and the complexity of the registration algorithms [1,2]. A powerful optimization method is needed to shorten the time consumption during the registration process, which would benefit time-critical intra-operative procedures relying on image guidance.

The stochastic gradient descent method is often used to iteratively find the optimum [3]. This method is easy to implement and fast because at each iteration only a subset of voxels from the fixed image is evaluated to obtain gradients. Although it obtains a good accuracy, its convergence rate is poor since only first order derivatives are used. A preconditioning matrix can be used to improve the convergence rate of (stochastic) gradient descent, but this was only proposed in

© Springer International Publishing Switzerland 2015
N. Navab et al. (Eds.): MICCAI 2015, Part II, LNCS 9350, pp. 297–304, 2015.
DOI: 10.1007/978-3-319-24571-3_36

a mono-modal setting [4]. The quasi-Newton method also has a better convergence rate than deterministic gradient descent, but comes at a higher cost in computation time and large memory consumption. Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) takes an advantage in the storage of only a few previous Hessian approximations, however, the computation time is still very long as all voxels are needed for new Hessian approximations [2].

Some approaches to create a stochastic version of the quasi-Newton method are proposed in a mathematical setting, such as online LBFGS [5], careful quasi-Newton stochastic gradient descent [6], regularized stochastic BFGS [7] and stochastic LBFGS [8]. However, there is no application in the image registration field, and applying the stochastic quasi-Newton method to non-rigid image registration is still a challenge. All of the previous methods either used a manually selected constant step size or a fixed decaying step size, which are not flexible when switching problem settings or applications. Moreover, the uncertainty of gradient estimation introduced by the stochastic gradient for Hessian approximation is still a problem. Although Byrd [8] used the exact Hessian to compute curvature updates, which is still difficult to calculate for high dimensional problems. For careful QN-SGD [6], the average scheme may be useless in case of an extremely large or small scaling value for H_0. Mokhtari [7] used a regularized term like Schraudolph [5] did to compensate the gradient difference y from the parameter difference s and introduced a new variable δ, which is not only complex, but also needs to store all previous curvature pairs.

In this paper, we propose a stochastic quasi-Newton method specifically for non-rigid image registration inspired by Byrd *et al.* [8]. Different from Byrd's method, the proposed method employs only gradients and avoids computing second order derivatives of the cost function to capture the curvature. Secondly, we employ an automatic and adaptive scheme for optimization step size estimation instead of a fixed manual scheme. Finally, we propose a restarting mechanism where the optimal step size is recomputed when a new Hessian approximation becomes available, i.e. every L iterations. The proposed method and some variations are validated using 3D lung CT follow-up data using manually annotated corresponding points for evaluation.

2 Methods

Non-rigid image registration aims to align images following a continuous deformation strategy. The optimal transformation parameters are the solution that minimizes the dissimilarity between fixed I_F and moving image I_M:

$$\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}(I_F, I_M \circ \boldsymbol{T}_{\boldsymbol{\mu}}), \tag{1}$$

in which $\boldsymbol{T}_{\boldsymbol{\mu}}(\boldsymbol{x})$ is a coordinate transformation parameterized by $\boldsymbol{\mu}$.

2.1 Deterministic Quasi-Newton

The deterministic quasi-Newton method employs the following iterative form:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma_k \boldsymbol{B}_k^{-1} \boldsymbol{g}_k, \tag{2}$$

where $\boldsymbol{B}_k$ is a symmetric positive definite approximation of the Hessian matrix $\nabla^2 \mathcal{C}(\boldsymbol{\mu}_k)$. Quasi-Newton methods update the inverse matrix $\boldsymbol{H}_k = \boldsymbol{B}_k^{-1}$ directly using only first order derivatives, and have a super-linear rate of convergence. Among many methods to construct the series $\{\boldsymbol{H}_k\}$, Broyden-Fletcher-Goldfarb-Shanno (BFGS) tends to be efficient and robust in many applications. It uses the following update rule for $\boldsymbol{H}_k$:

$$\boldsymbol{H}_{k+1} = \boldsymbol{V}_k^T \boldsymbol{H}_k \boldsymbol{V}_k + \rho_k \boldsymbol{s}_k \boldsymbol{s}_k^T, \tag{3}$$

in which

$$\rho_k = \frac{1}{\boldsymbol{y}_k^T \boldsymbol{s}_k}, \quad \boldsymbol{V}_k = \boldsymbol{I} - \rho_k \boldsymbol{y}_k \boldsymbol{s}_k^T, \quad \boldsymbol{s}_k = \boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k, \quad \boldsymbol{y}_k = \boldsymbol{g}_{k+1} - \boldsymbol{g}_k. \tag{4}$$

Since the cost of storing and manipulating the inverse Hessian approximation $\boldsymbol{H}_k$ is prohibitive when the number of the parameters is large, a frequently used alternative is to only store the latest M curvature pairs $\{\boldsymbol{s}_k, \boldsymbol{y}_k\}$ in memory: limited memory BFGS (LBFGS). The matrix $\boldsymbol{H}_k$ is not calculated explicitly, and the product $\boldsymbol{H}_k \boldsymbol{g}_k$ is obtained based on a 2-rank BFGS update, which uses a two loop recursion [8]. The initial inverse Hessian approximation usually takes the following form, which we also use in this paper:

$$\boldsymbol{H}_k^0 = \theta_k \boldsymbol{I}, \quad \theta_k = \frac{\boldsymbol{s}_{k-1}^T \boldsymbol{y}_{k-1}}{\boldsymbol{y}_{k-1}^T \boldsymbol{y}_{k-1}}. \tag{5}$$

2.2 Stochastic Quasi-Newton

A large part of the computation time of quasi-Newton methods is in the computation of the curvature pairs $\{\boldsymbol{s}_k, \boldsymbol{y}_k\}$. The pairs are computed deterministically using all samples from the fixed image. A straightforward way to obtain a stochastic version of the quasi-Newton method is to construct the curvature pairs using stochastic gradients, using only a small number of samples at each iteration. This however introduces too much noise in the curvature estimation, caused by the fact that stochastic gradients are inherently noisy and for each iteration are also evaluated on different subsets of image voxels, both of which may yield a poor Hessian approximation. This leads to instability in the optimization.

To cope with this problem, Byrd *et al.* [8] proposed a scheme to eliminate the noise by averaging the optimization parameters for a regular interval of L iterations and obtain the curvature through a direct Hessian calculation on a random subset $\mathcal{S}_2$. This is combined with a series of L iterations performing LBFGS using the thus obtained inverse Hessian estimate together with *stochastic* gradients (using a small random subset $\mathcal{S}_1$). Inspired by this scheme, we propose a method suitable for medical image registration and avoiding manual tuning the step size. First, more samples are used for the curvature pair update than for the stochastic gradient evaluation. Second, the curvature information is obtained using a gradient difference instead of second order derivatives evaluated at an identical subset of samples, i.e. $\boldsymbol{y}_t = \boldsymbol{g}(\bar{\boldsymbol{\mu}}_I; \mathcal{S}_2) - \boldsymbol{g}(\bar{\boldsymbol{\mu}}_J; \mathcal{S}_2)$ and the curvature condition $\boldsymbol{y}^T \boldsymbol{s} > 0$ is checked to ensure positive definiteness of the LBFGS update [8]. Finally, the initial step size at the beginning of each L iterations is automatically determined, with or without restarting. Restarting is a recent development [9] showing improved rate of convergence, which in this paper we apply to the step size selection.

Instead of manual constant step size selection as in [8], we employ an automatic method. A commonly used function for the step size which fulfils the convergence conditions [10] is $\gamma_k = \eta a/(t_k + A)^\alpha$, with $A \geq 1$, $a > 0$, $0 \leq \eta \leq 1$ and $0 \leq \alpha \leq 1$. The step size factor a and the noise compensation factor η are automatically determined through the statistical distribution of voxel displacements [11], while $A = 20$ according to [3] and $\alpha = 1$ is theoretically optimal [3]. Different strategies for the artificial time parameter t_k are tested: a constant step size $t_k = 0$, a regularly decaying step size $t_k = k$, and an adaptive step size $t_k = f(\cdot)$. Here, f is a sigmoid function with argument of the inner product of the gradients $\tilde{\boldsymbol{g}}_k^T \cdot \tilde{\boldsymbol{g}}_{k-1}$ for gradient descent. For s-LBFGS, it can be derived that the search direction is needed as argument, i.e. $\boldsymbol{d}_k^T \cdot \boldsymbol{d}_{k-1}$ with $\boldsymbol{d}_k = \boldsymbol{B}_k^{-1} \boldsymbol{g}_k$.

An overview of the proposed s-LBFGS method is given in Algorithm 1.

3 Experiment

The proposed method was integrated in the open source software package `elastix` [12]. The experiments were performed on a workstation with 8 cores running at 2.4 GHz and 24 GB memory, with an Ubuntu Linux OS.

3D lung CT scans of 19 patients acquired during the SPREAD study [13] were used to test the performance. Each patient had a baseline and a follow-up scan with an image size around $450 \times 300 \times 150$ and the voxel size around $0.7 \times 0.7 \times 2.5$ mm. For each image, one hunred anatomical corresponding points were chosen semi-automatically using Murphy's method in consensus by two experts, to obtain a ground truth.

To evaluate the method, each follow-up image was registered to the baseline image using mutual information and a B-spline transformation model. The maximum number of iterations for each resolution was 500. A three-level multi-resolution framework was employed using a Gaussian smoothing filter with standard deviations of 2, 1 and 0.5 mm for each resolution. The grid spacing of the B-spline control points was halved between each resolution resulting in a final grid spacing of 10 mm in each direction. After initial testing, we chose the update frequency $L = 10, 20, 40$ for each resolution, respectively, the memory $M = 5$ from [2,8], the number of samples for stochastic gradient computation $|S_1| = 5000$, and the number of samples for the curvature pair update $|S_2| = 50000$.

To measure the registration accuracy, the anatomical points from each baseline image were transformed using the obtained transformation parameters and

Algorithm 1. Stochastic LBFGS (s-LBFGS) with and without restarting

```
Require: initial parameters $\boldsymbol{\mu}_0$, memory size $M$, update frequency $L$, iteration number $K$
Set $t = 0$, $\bar{\boldsymbol{\mu}}_J = \boldsymbol{\mu}_0, \bar{\boldsymbol{\mu}}_I = \mathbf{0}$
Automatically estimate the initial step size $\lambda_0$    ▷ According to [11]
for $k = 1, 2, 3, \ldots, K$ do
    Compute $\tilde{\boldsymbol{g}}_k(\boldsymbol{\mu}_k; \mathcal{S}_1)$    ▷ stochastic gradient
    $\bar{\boldsymbol{\mu}}_I = \bar{\boldsymbol{\mu}}_I + \boldsymbol{\mu}_k$    ▷ Update the mean parameters
    if $k <= 2L$ then    ▷ ASGD update
        Update the step size $\lambda_k$    ▷ According to [11]
        $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \lambda_k \tilde{\boldsymbol{g}}_k$
    else    ▷ s-LBFGS update
        Compute $\boldsymbol{d}_k = \boldsymbol{H}_t \tilde{\boldsymbol{g}}_k$    ▷ s-LBFGS search direction, see [8] and (2)
        if $\mod(k, L) = 0$ and restarting then
            Automatically estimate the initial step size $\lambda'_0$
            Reset $\lambda_k = \lambda'_0$
        Update the step size $\lambda_k$    ▷ According to [11] but using $\boldsymbol{d}_k^T \cdot \boldsymbol{d}_{k-1}$
        $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \lambda_k \boldsymbol{d}_k$
    if $\mod(k, L) = 0$ then    ▷ Curvature pairs update
        $\bar{\boldsymbol{\mu}}_I = \bar{\boldsymbol{\mu}}_I / L$    ▷ Update the mean parameters
        $\boldsymbol{s}_t = \bar{\boldsymbol{\mu}}_I - \bar{\boldsymbol{\mu}}_J, \quad \boldsymbol{y}_t = \boldsymbol{g}(\bar{\boldsymbol{\mu}}_I; \mathcal{S}_2) - \boldsymbol{g}(\bar{\boldsymbol{\mu}}_J; \mathcal{S}_2)$    ▷ New curvature pair
        $\bar{\boldsymbol{\mu}}_J = \bar{\boldsymbol{\mu}}_I, \bar{\boldsymbol{\mu}}_I = \mathbf{0}, t = t + 1$
return $\boldsymbol{\mu}_K$
```

then compared to the corresponding points of the follow-up image. We used the Euclidean distance between the corresponding points $\boldsymbol{p}_F \in \Omega_F$ and $\boldsymbol{p}_M \in \Omega_M$ to measure the accuracy using the following equation: $ED = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{T}(\boldsymbol{p}_F^i) - \boldsymbol{p}_M^i\|$. For 19 patients, we first obtained the mean distance error of 100 points for each patient then performed Wilcoxon signed rank test to these mean errors. For convergence testing we computed the cost function value after each iteration deterministically, i.e. based on full sampling. The registration time in the first resolution is presented to compare the algorithm speeds.

4 Results

To gain insight in the proposed method, we investigated some aspects that influence registration performance. The restarting scheme (Restart) was compared with a scheme without restarting. We evaluated different step size selection strategies all based on automatic initial step size selection [11]: a constant scheme (Constant, $t_k = 0$), a regularly decaying scheme (Decaying, $t_k = k$), and the proposed adaptive scheme (Adaptive, $t_k = f(\cdot)$). The proposed method is further compared with the ASGD method [3] and with deterministic LBFGS [2].

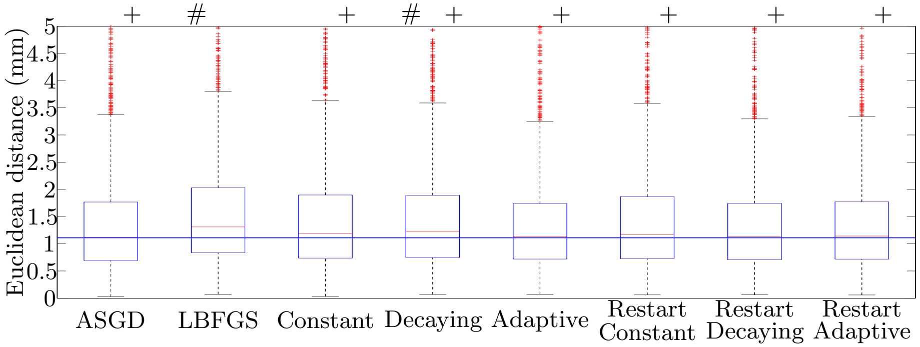
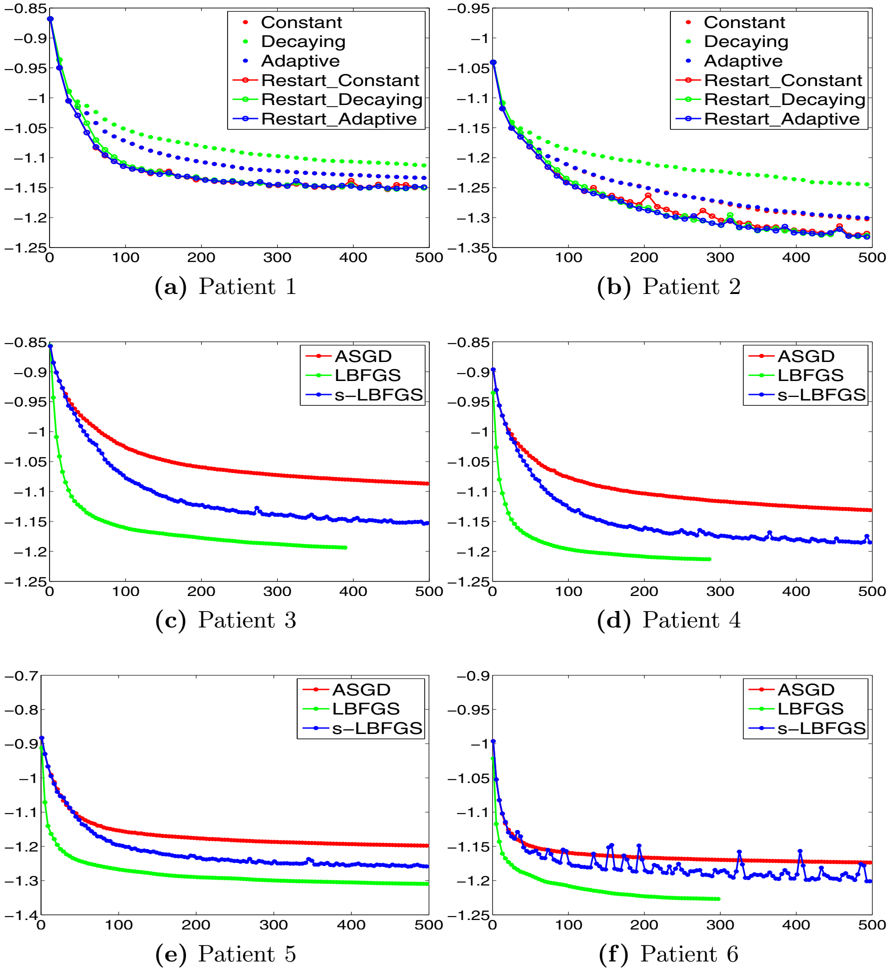


Fig. 1. Euclidean distance error in mm. The symbols # and + indicate a statistically significant difference with ASGD and LBFGS, respectively.

Table 1. Run time in the first resolution. I indicates how many iterations are needed to reach the same metric value as ASGD after 500 iterations. s-LBFGS and s-LBFGS-NR are with and without restarting, both using adaptive step sizes. The speed-up is relative to ASGD.

	Time at $k = 500$	I	Time (s) at $k = I$	Speed-up
ASGD	27.2 ± 0.7	500	-	-
LBFGS	26838 ± 9965	21 ± 1	8081 ± 1580	0.004 ± 0.0005
s-LBFGS-NR	74.3 ± 4.8	190 ± 93	30.6 ± 13.9	1.0 ± 0.4
s-LBFGS	75.8 ± 1.0	107 ± 17	18.1 ± 2.6	1.5 ± 0.2

From Fig. 1 we can see that all methods have very similar final registration error, for LBFGS regularization may improve the results [14]. Fig. 2 shows the convergence plots of the methods for several patients. Comparing the three step size strategies in Fig. 2a and 2b, the regularly decaying method has suboptimal convergence, while the constant and the adaptive scheme behave similarly. The restarting scheme shows a substantial improvement in convergence rate, therefore in Fig. 2c∼2f we only show the result of restarting scheme with adaptive step size (s-LBFGS). Some small spikes are visible in Fig. 2b and Fig. 2f, which we attribute to noise in the curvature pair estimation: an experiment using 1.5 million samples for the curvature estimation yielded smooth results (not shown). In terms of iterations, s-LBFGS always obtains faster convergence than ASGD, but slower than LBFGS. The registration time of ASGD, LBFGS and s-LBFGS is shown in Table 1. The LBFGS method is very costly, as expected. To obtain the same metric value as ASGD at iteration 500, the proposed method always takes fewer iterations resulting in an average speedup of two, while the proposed method without restarting requires more iterations and therefore more time.



(a) Patient 1 **(b)** Patient 2

(c) Patient 3 **(d)** Patient 4

(e) Patient 5 **(f)** Patient 6

Fig. 2. Convergence plots, showing the negated mutual information metric against the iteration number.

5 Conclusion

In this paper, we present for the first time a stochastic quasi-Newton optimization method (s-LBFGS) for non-rigid image registration. It uses the classical limited memory BFGS method in combination with noisy estimates of the gradient. Curvature information of the cost function is estimated robustly once every L iterations and then used for the next L iterations in combination with stochastic gradients. A novel restarting procedure, automatically selecting the optimization step size, is shown to be beneficial for accelerated convergence. The new optimization routine is validated on follow-up data of 3D chest CT scans (19 patients).

Compared to ASGD the proposed method uses about 5 times fewer iterations to reach the same metric value, resulting in an overall reduction in run time of a factor of two. Compared to deterministic LBFGS, s-LBFGS is almost 500 times faster. Future work will focus on developing a stopping condition for stochastic second order procedures, on a more robust estimation of the initial approximation of H_0 more resilient against noise, on alterative quasi-Newton schemes such as the symmetric rank-one update [15], and more extensive validation.

References

1. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Transactions on Medical Imaging 32(7), 1153–1190 (2013)
2. Klein, S., Staring, M., Pluim, J.P.: Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. IEEE Transactions on Image Processing 16(12), 2879–2890 (2007)
3. Klein, S., Pluim, J., Staring, M., Viergever, M.: Adaptive stochastic gradient descent optimisation for image registration. International Journal of Computer Vision 81(3), 227–239 (2009)
4. Klein, S., Staring, M., Andersson, P., Pluim, J.P.W.: Preconditioned stochastic gradient descent optimisation for monomodal image registration. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part II. LNCS, vol. 6892, pp. 549–556. Springer, Heidelberg (2011)
5. Schraudolph, N., Yu, J., Günter, S.: A stochastic quasi-Newton method for online convex optimization (2007)
6. Bordes, A., Bottou, L., Gallinari, P.: SGD-QN: Careful quasi-Newton stochastic gradient descent. The Journal of Machine Learning Research 10, 1737–1754 (2009)
7. Mokhtari, A., Ribeiro, A.: RES: Regularized stochastic BFGS algorithm. IEEE Transactions on Signal Processing 62(23), 6089–6104 (2014)
8. Byrd, R.H., Hansen, S., Nocedal, J., Singer, Y.: A stochastic quasi-Newton method for large-scale optimization. arXiv preprint arXiv:1401.7020 (2014)
9. O'Donoghue, B., Candés, E.: Adaptive restart for accelerated gradient schemes. Foundations of Computational Mathematics, 1–18 (2013)
10. Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23(3), 462–466 (1952)
11. Qiao, Y., Lelieveldt, B., Staring, M.: Fast automatic estimation of the optimization step size for nonrigid image registration. In: SPIE Medical Imaging, International Society for Optics and Photonics, pp. 90341A–90341A (2014)
12. Klein, S., Staring, M., et al.: Elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29(1), 196–205 (2010)
13. Stolk, J., Putter, H., Bakker, E.M., et al.: Progression parameters for emphysema: a clinical investigation. Respiratory Medicine 101(9), 1924–1930 (2007)
14. Sun, W., Niessen, W., van Stralen, M., Klein, S.: Simultaneous multiresolution strategies for nonrigid image registration. IEEE Transactions on Image Processing 22(12), 4905–4917 (2013)
15. Modarres Khiyabani, F., Leong, W.: Limited memory methods with improved symmetric rank-one updates and its applications on nonlinear image restoration. Arabian Journal for Science and Engineering 39(11), 7823–7838 (2014)