GAS: a Genetic Atlas Selection Strategy in Multi-Atlas Segmentation Framework

Michela Antonelli[a,*], M. Jorge Cardoso[d,b], Edward W. Johnston[c], Mrishta Brizmohun Appayya[c], Benoit Presles[a], Marc Modat[d,b], Shonit Punwani[c], Sebastien Ourselin[d,b]

[a] *Centre for Medical Image Computing, University College London, U.K.*
[b] *Dep. of Medical Physics and Biomedical Engineering, University College London, U.K.*
[c] *Centre for Medical Imaging, University College London, U.K.*
[d] *School of Biomedical Engineering and Imaging Science, King's College London, U.K.*

Abstract

Multi-Atlas based Segmentation (MAS) algorithms have been successfully applied to many medical image segmentation tasks, but their success relies on a large number of atlases and good image registration performance. Choosing well-registered atlases for label fusion is vital for an accurate segmentation. This choice becomes even more crucial when the segmentation involves organs characterized by a high anatomical and pathological variability. In this paper, we propose a new genetic atlas selection strategy (GAS) that automatically chooses the best subset of atlases to be used for segmenting the target image, on the basis of both image similarity and segmentation overlap. More precisely, the key idea of GAS is that if two images are similar, the performances of an atlas for segmenting each image are similar. Since the ground truth of each atlas is known, GAS first selects a predefined number of similar images to the target, then, for each one of them, finds a near-optimal subset of atlases by means of a genetic algorithm. All these near-optimal subsets are then combined and used to segment the target image. GAS was tested on single-label and multi-label segmentation problems. In the first case, we considered the segmentation of both the whole prostate and of the left ventricle of the heart from magnetic

*Corresponding author
Email address: m.antonelli@ucl.ac.uk (Michela Antonelli)

resonance images. Regarding multi-label problems, the zonal segmentation of the prostate into peripheral and transition zone was considered. The results showed that the performance of MAS algorithms statistically improved when GAS is used.

Keywords: Atlas selection, Genetic Algorithm, Multi-atlas based segmentation, Multi-parametric MRI, Prostate segmentation

1. Introduction

Multi-Atlas based Segmentation (MAS) algorithms have been successfully applied to a wide range of medical image segmentation tasks (Isgum et al., 2009; Cardoso et al., 2013; Aljabar et al., 2009). Their success relies on the introduction of *a priori* knowledge using a set of pre-segmented images. An atlas consists of a medical image and a corresponding segmented label image. According to the subdivision proposed in Iglesias and Sabuncu (2014), a MAS algorithm can be implemented following four sequential steps: first each atlas image is registered to the target image (registration step), second either all or a subset of atlases are chosen (atlas selection step) and their label images are propagated into the target space (label propagation step). Finally the propagated labels are fused to generate the segmentation of the target (label fusion step). Several types of label fusion algorithm can be found in Sabuncu et al. (2010), while in Iglesias and Sabuncu (2014) the authors present an extensive review of MAS algorithms.

The performance of a MAS algorithm depends greatly on the registration step: if an atlas is badly registered, considering it in the fusion step could misguide the segmentation of the target image and decrease its accuracy.

The use of more complex label fusion techniques can mitigate the above problem. In Artaechevarria et al. (2009), the authors investigate how different types of atlas combinations affect the segmentation accuracy. The authors studied the use of both global and local voting strategies for the label fusion step and their results showed that, in general, there is no strategy that can recover when

the number of poorly-registered atlases is high. For this reason, it is important to apply an intermediate step between label propagation and label fusion that selects from all the atlases a near-optimal subset of them to propagate.

In the literature, several approaches have been proposed to select the best subset of atlases to propagate and use in the label fusion step. Aljabar et al. (2009), investigated several selection strategies based on the rank of the atlases. The selection of the atlases from a database is performed according to their suitability for a given target image. First, the atlases are ranked according to their image similarity with the target, expressed by means of metrics such as cross-correlation (CC) and normalized mutual information (NMI). Then only the first k atlases from the top ranked list are selected and used in the fusion step. The results showed that MAS algorithms that apply atlas similarity selection obtained better segmentations compared of those that use either the same number of randomly-selected atlases or all the atlases of the database.

Klein et al. (2008) proposed a MAS algorithm which first calculates the NMI value between the target image and each atlas, then selects and fuses only the atlases with NMI greater than a fixed threshold. Y. Ou and Davatzikos (2012) applied the same selection strategy but used the mutual information in place of NMI. Langerak et al. (2010) proposed a label fusion method (SIMPLE) that can be used as an atlas selection strategy: first the NMI between each atlas and the target image is calculated and only the atlases with a value of NMI greater than a fixed threshold are selected and fused to generate an initial estimation of the target segmentation. Then, each atlas is re-evaluated by considering a binary overlap measure between the estimated segmentation and the registered image label. If this value is lower than a fixed threshold the atlas is not considered for the next re-estimated segmentation. This process is repeated until the estimated segmentation converges.

All of the above mentioned approaches apply atlas selection after the registration step. Alternatively, in van Rikxoort et al. (2010), instead of registering all the atlases to the target, and then applying the selection step, the authors first performed a fast registration (computationally cheap) between the target

and all the atlases, then they select the atlases depending on the difference between the target and the registered atlas images. On the contrary, in Langerak et al. (2013) the authors estimate the performance of an atlas before registering it to the target image: first groups of atlases are formed on the basis of the results of pairwise registration among the atlases themselves. Then, the groups that are too dissimilar to the target image are discarded before the registration step. In this way, in addition to an improvement of the performance, they also save computational time.

However, most of the existing MAS algorithms that perform atlas selection rank the atlases using a measure of image similarity between each atlas and the target image, and then they apply a threshold either on the size of the subset of the selected atlases or on the value of the image similarity measure. The similarity between the registered atlas images and the target may not always be a good estimate of the performance of the atlas. Also, optimizing the values of the thresholds may be not trivial.

Atlas selection is a crucial step when the segmentation concerns organs such as the prostate that are characterized by a high variability in both shape and surrounding structures. In this case, it is not always possible to obtain accurate registrations between the atlases and the target. Since each registered atlas image can be considered as a voxel-wise classifier, using a poorly-registered atlas can be regarded as using a classifier with a very low performance.

Prostate segmentation is a vital step of almost every computer-aided diagnosis systems (CAD) for the diagnosis and treatment of prostate cancer. Although multi-parametric magnetic resonance imaging (mpMRI) has been shown to be very effective in detecting prostate cancer and in its treatment (Graham et al., 2014), interpreting prostate MRI images requires a high level of expertise and is time consuming. For this reason, there has been an increasing interest in the development of these CAD systems (S. Wang and Summers, 2014).

Whether the CAD system based on mpMRI is used for diagnosis or treatment, the first step is the extraction of the prostate region from T2-weighted (T2w) images as the target region of interest for further feature extraction.

Moreover, if the CAD system is used for cancer detection and diagnosis, it is crucial not only to identify the prostate region, but also to divide this region into the two main zones: the peripheral zone (PZ) and the transition zone (TZ). Cancers belonging to these two zones show different behaviour, so for a CAD system to be effective, it is important to apply different algorithms to the two zones. Therefore, the segmentation step, in addition to the whole prostate (WH) identification, has to identify PZ and TZ.

This step is extremely important as its accuracy determines the success of the following CAD stages: a wrong segmentation may cut out prostate regions containing lesions, or generate errors in volumetric calculations. Manual segmentation may be an option, but it is very time-consuming and prone to intra- and inter-observer variability.

Several approaches have been proposed to automatically segment the whole prostate in T2w MRI images. Makni et al. (2014), used an adaptation of the Evidential C-Means clustering where the optimization process takes into account voxels' spatial neighborhood information. In Guo et al. (2016) the authors first used deep learning to extract the latent features from prostate MR images. Then, based on the learned features, a prostate likelihood map is inferred by means of a sparse patch matching method. Mahapatra and Buhmann (2014), applied a supervoxel segmentation to identify a volume of interest followed by a random forest classification to generate probability maps for voxels giving their likelihood of being prostate or non-prostate. In Yan et al. (2015); Martin et al. (2008); Y. Ou and Davatzikos (2012); Klein et al. (2008) multi-atlas based approaches were applied to the whole prostate segmentation. Although automatic whole prostate segmentation has already been addressed in the literature, very few works have been proposed for zonal segmentation (Chilali et al., 2016; Makni et al., 2014; Qiu et al., 2014; Toth et al., 2013) and none of them use multi-atlas based methods.

In this paper we propose a new genetic atlas selection strategy (GAS), in the framework of MAS algorithm, that automatically chooses the best subset of atlases on the basis of both image similarity and Dice coefficient. GAS works

under the hypothesis that if image A is similar to image B, the performance of an atlas that segments A is similar to the performance of the same atlas for segmenting B. Since the manual segmentation of each atlas is known, we first select the n most similar atlas images to the target, then we apply a genetic algorithm to find the best subset of atlases that segment each of these n atlas image. Finally, these n best subsets are combined and used to segment the real target image.

We tested the proposed GAS strategy on both single-label and multi-label segmentation problems. For single-label segmentation, we evaluated GAS on the segmentation of both the whole prostate and the left ventricle of the heart. We compared the results of four MAS algorithms without GAS, with GAS and with the atlas selection strategy presented in Langerak et al. (2010) (SIMPLE).

The four MAS algorithms consisted of two standard MAS algorithms characterized by two different label fusion methods, and two state-of-the-art algorithms, namely STEPS (Cardoso et al., 2013), which is an extension of the well-known STAPLE, and the MAS algorithm introduced in Wang et al. (2013), which is based on a new joint label fusion method.

For multi-label segmentation, we validated GAS on the segmentation of the prostate into peripheral and transition zones, comparing the results obtained by the four MAS algorithms with and without GAS.

The results showed that GAS statistically significantly improves the performance for all the four MAS algorithms and for both the single-label and multi-label segmentations. Furthermore, it generates segmentations that are statistically more accurate than the ones obtained by the four MAS algorithms which use SIMPLE as atlas selection strategy.

This paper is organized as follows: Section II describes the MAS framework used and the new genetic atlas selection strategy. Section III shows the experimental results and Section IV draws some final conclusions.

2. Material and methods

In this section we introduce the standard multi-atlas based segmentation (MAS) framework and the associated notation, then we present the new genetic atlas selection strategy.

2.1. Multi-atlas based segmentation framework

Let I_T be the target image to be segmented and $A = \{A_1, \ldots, A_N\}$ the set of N atlases available in the database. Each atlas A_i is defined by the pair (I_i, L_i) where I_i is the raw intensity image and and L_i the corresponding image segmentation or "image label" .

Generally, a MAS algorithm is divided into three main sequential steps, namely the registration of the N image atlases to I_T, the propagation of the corresponding N labels into the target space and the combination of the propagated labels into the final segmentation of I_T.

The registration step computes the spatial correspondences between each image I_i and I_T. Here, for each atlas A_i, with $i = \{1, \ldots, N\}$, we calculate the transformation function by applying first an affine registration(Ourselin et al., 2000), followed by a cubic B-spline non-rigid registration(Modat et al., 2010). Both registration types are intensity-based, hence they search for the transformation that maximizes a measure of the similarity in intensity between corresponding pixels. The convolution-based fast local normalized correlation coefficient (LNCC) (Cachier et al., 2003) is used as similarity measure for the non-rigid registration to enable robustness to Bias Field inhomogeneity.

For the $i - th$ atlas, the output of the registration step is a transformation function, T_i, such that $T_i(I_i) \approx I_T$, which will be used to propagate the atlas labels L_i to the target image coordinate space.

Once T_i with $i = 1, \ldots, N$ are calculated, the classic MAS algorithm propagates each image label L_i into the image target space, applying the corresponding transformation T_i. In this way, each atlas produces a segmentation for I_T. Finally, the last step of a MAS algorithm is the fusion of all these candidate

segmentations into a single segmentation. One possible label fusion strategy is majority voting: for each voxel v of I_T it counts the votes considering each registered atlas label and chooses the label with the highest score.

Since not all the atlases produce equally accurate segmentations, to improve the fusion step and, consequently, the segmentation of I_T, a local weighted voting strategy can be used. This type of label fusion assigns a weight to each voxel of each atlas that reflects the similarity between the registered atlas image and the target around that voxel. In this way, more accurate registrations, and subsequent segmentations, have more influence on the label choice.

In our experiments the weights are based on LNCC: for each voxel v of the $i-th$ atlas, $w_i(v)$ is computed by the following equation

$$w_i(v) = exp(1 - \frac{LNCC(T_i(I_i), I_T)_v^2}{\sigma}) \quad (1)$$

where σ determines how much the value of LNCC influences the weight.

2.2. Genetic Atlas Selection strategy

The success of a MAS algorithm heavily relies on the success of the registration step: if the registration between the atlas and the target image fails, considering that atlas in the fusion step only introduces noise. The weighted voting strategy described in sub-section 2.1 may reduce the effect of considering that poorly-registered atlas, but the LNCC between the atlas image and target image does not always indicate the accuracy of the registered atlas image and the corresponding segmentation.

For this reason, to select the best subset of atlas S_{I_T} to be used for segmenting I_T, we proposed a selection strategy based on both image similarity and Dice coefficient. First, image similarity is used to select the n most similar images to the target. Then, each image in turn is considered as pseudo-target image and the best subset of atlases is calculated using the Dice coefficient as measure of goodness of the subset. Finally, these n best subsets are combined and used to segment the real target I_T.

To select the n most similar images to I_T, we first registered each atlas image to the target using the affine registration followed by the non-rigid registration. Then, using equation 1, we calculated LNCC for each pair (atlas image, target) and selected the n atlases with highest LNCC values.

Let $I^* = (I_1^*, \ldots, I_n^*)$ be the set of n most similar images to I_T and $L^* = (L_1^*, \ldots, L_n^*)$ the corresponding labels. The aim is, for each I_i^*, to calculate the subset $S_{I_i^*}$ of atlases that produces an automatic segmentation $\widehat{L_i^*}$ such that to maximize the Dice coefficient $D_i = D(\widehat{L_i^*}, L_i^*)$. The segmentation of the pseudo-target L_i^* is generated by using majority voting as label fusion strategy.

If the size N of the atlas database is very small, this subset could be found by exhaustively generating the automatic segmentation using all the possible combinations of the $N-1$ atlases of the database and choosing the combination that generates the automatic segmentation with highest Dice coefficient. However, as the number of all possible subsets grows combinatorially with N, for larger values of N this exhaustive search is not feasible. Thus a heuristic has to be introduced to avoid generating all the possible 2^{N-1} segmentations.

We propose a genetic algorithm (GA) as a heuristic to find $S_{I_i^*}$. GAs are optimization algorithms inspired by the concepts of natural selection and evolution (Beasley et al., 1993). They usually start from a random population of individuals, called chromosomes, each representing a candidate solution to the given optimization problem. A fitness score is then assigned to each individual depending on how good the solution is for the problem. Individuals with high fitness values are given higher chances to reproduce and create a new population of offspring than the ones with lower fitness values.

In this way, highly fit individuals are more likely to be selected for reproduction and to pass their good characteristics to the offspring, while the least fit die out. The population at the next generation is obtained by selecting the best individuals from the current population and new population of offspring. This process, repeated over generations, allows the exploration of the more promising area of the search space until the GA converges to an optimal solution of the problem.

In order to use a GA for generating $S_{I_i^*}$ two main aspects have to be defined, namely the chromosome representation and the fitness function.

With regards to the chromosome representation, a binary string with N-1 genes is used where N is the total number of atlases of the database. Each gene is associated with an atlas: if the $k-th$ gene is equal to 1, then the $k-th$ atlas is used for generating the segmentation of I_i^*, if it is equal to 0 then that atlas is not used. Thus, each chromosome constitutes a subset of atlases to be used by the MAS for segmenting I_i^*.

The fitness function of a chromosome represents how well the subset of atlases coded in the chromosome segments I_i^*. As we know the manual segmentation L_i^* of I_i^*, the fitness function of each individual is the value of the Dice coefficient $D(\widehat{L_i^*}, L_i^*)$, where $\widehat{L_i^*}$ is the automatic segmentation obtained applying a MAS that uses only the atlases with the corresponding genes in the chromosome set to 1.

Figure 1 shows the pseudo-code of the GA. The algorithm starts with a randomly generated population P_0. At each iteration t, the fitness function of each individual in the population is evaluated, and, using the roulette wheel selection mechanism, individuals are selected for reproduction . This type of selection associates a selection probability to each chromosome proportional to its fitness value. Thus, the evolutionary process consists first on the selection of two parents, p_1 and p_2, then, two offspring are generated applying one-point crossover and mutation operators to p_1 and p_2, with probability P_{cr} and P_{mut}, respectively. Crossover takes copies of p_1 and p_2, randomly selects a point in the binary string, and swaps sub-strings of equal length between their chromosomes. Mutation randomly selects a gene in the chromosome and flips its value .

The two new offspring are added to the temporary population P_{temp}. This is repeated until P_{temp} contains a number of individuals equal to the population size pop_size. The population P_{t+1} at generation $t+1$ is generated by choosing the best $pop\_size/2$ individuals from both P_{temp} and P_t.

The chromosome with the highest fitness value contained on the population after a fixed number of generations ($maxNumGeneration$) determines the

```
//GAS routine for atlas I_x
P_0 = generate_random_population(pop_size)
loop j = 0 to pop_size
   f_j = fitness_evaluation(p_j)
endloop
loop t = 0 to maxNumGeneration
   P_temp = create_empty_population(pop_size)
   loop i = 1 to pop_size /2
      [p_1] = wheel_parent_selection(P)
      [p_2] = wheel_parent_selection(P)
      if (rand() < P_cr)
         [o_1,o_2] = crossover(p_1,p_2)
      else
         P_mut = 1
      endif
      if (rand() < P_mut)
         [o_1]=mutation(p_1); [o_2]=mutation(p_2)
      endif
      [f_1] = fitness_evaluation(o_1)
      [f_2] = fitness_evaluation(o_1)
      add_to_population(P_temp,o_1,o_2)
   endloop
   P_t+1 = combine_best_elements(P_t,P_temp)
endloop
return (best_element(P_t+1))


//mutation(p_i)

o_i = p_i
gene = rand(1,size(p_i))
o_i[gene] = not(o_i[gene])
return(o_i)
```

```
//crossover(p_i, p_j)
n = size(p_1) = size(p_j)
o_1 = p_i
o_2 = p_j
cut = rand(1,n)
o_1[cut:n] = p_2[cut:n]
o_2[1:cut-1] = p_1[1:cut-1]
return(o_1,o_2)


//fitness_evaluation(p_i)
s_i = create_segmentation(p_1)
f_i = Dice(s_i,I_x)
return(f_1)


//wheel_parent_selection(P)
F = sum_all_fitness(P)
sumPr = 0
loop i = 1 to pop_size
   prob_i = sumPr + fitness_evaluation(p_i)/F
   sumPr += prob_i
endloop
r = rand()
loop i = 1 to pop_size - 1
   if (r > prob_i and r <= prob_i+1)
      return p_i
   endif
endloop
```

Figure 1: Pseudo-code of the selection strategy explained in section 2.2.

atlases in $S_{I_i^*}$.

Once $S_{I_i^*}$ is generated for each I_i^*, the subset of atlases S_{I_T} used to segment the actual target I_T is calculated by combining all the n subsets $S_{I_i^*}$. In particular, let $S_{I^*} = \cup_{i=1}^{n} S_{I_i^*}$ be the union with repetition of the n subsets of atlases found by the genetic algorithm. S_{I_T} will contain along with the n most similar atlases, the atlases that are contained at least $\lceil n/2 \rceil$ times in S_{I^*}. If no atlas has been selected $\lceil n/2 \rceil$ times, S_{I_T} will contain only the n most similar atlases ($n = 9$ in the experiment).

As an example, let $A = \{A_1, ..., A_{20}\}$ be the set of atlases, A_2, A_7, and A_{11} the n most similar atlases ($n = 3$), and $S_{I_2^*} = \{A_5, A_6, A_9, A_{16}\}$, $S_{I_7^*} = \{A_2, A_4, A_9\}$, and $S_{I_{11}^*} = \{A_1, A_3, A_9, A_{16}\}$ the subset of atlases found by the genetic algorithm. The final subset of atlases to segment the target will be $S_{I_T} = \{A_2, A_7, A_9, A_{11}, A_{16}\}$.

3. Results and Discussion

In this section we describe the datasets used to evaluate GAS, the experimental setup and the results obtained.

3.1. Dataset

Three datasets have been used in our analysis. For each dataset, the manual segmentation of an experienced radiologist was provided. All annotations were performed on a slice-by-slice basis using a contouring tool. The datasets used are:

- PROMISE12 (Litjens et al., 2014), the MICCAI 2012 prostate segmentation challenge dataset (http://promise12.grand-challenge.org/). Scans were collected from four centres under different clinical settings and acquisition protocols. Each centre provided 25 transverse T2w MR images for a total of 100 MRI scans, split into 50 training cases, 30 test cases and 20 live challenge cases. We used only the 50 training cases as they include the reference segmentations of the whole prostate. The reference segmentations were checked and corrected by a second radiologist who has read more than 1000 prostate MRIs.

- PICTURE (Simmons et al., 2017), a single centre, diagnostic cohort study undertaken at the University College London Hospital NHS Foundation Trust (UCLH). The study recruited from a population of men who are at risk of prostate cancer and who have already undergone a standard diagnostic transrectal ultrasound guided (TRUS) biopsy and where diagnostic uncertainty remains. We selected 90 T2w MRI scans from 90 patients, where each subject had a manual segmentation of both the whole prostate and the PZ and TZ regions.

- Sunnybrook Cardiac Data (SCD) (Radau et al., 2009), the 2009 Cardiac MR Left Ventricle Segmentation Challenge. The data consists of 45 cine MR images of healthy subjects (9), and patients with different pathologies,

namely hypertrophy (12) , heart failure with infarction (12) and heart failure without infarction (12). For each scan an experienced radiologist manually contoured the endocardium (EN) and epicardium (EP) for every diastole phase. We performed a single-phase segmentation considering only the images related to the end-systolic phase. The phase selection was manually performed.

3.2. Experimental setup

To assess the effectiveness of the proposed genetic atlas selection (GAS) strategy on MAS segmentation accuracy, two types of segmentation problems have been addressed using four state-of-the-art MAS algorithms.

First, we evaluated GAS on single label segmentation problems. We considered the segmentation of the whole prostate from T2w MRI images and the segmentation of the left ventricle og the heart from cine MRI. For both cases, we first compare the results obtained by four state-of-the-art MAS algorithms both with and without GAS, then we compare the results of the four MAS with GAS and with a different atlas selection strategy.

We further validated the proposed method on a multi-label segmentation problem. Here, we evaluated the results obtained by the four MAS algorithms when GAS is applied to the segmentation of the prostate into peripheral and transition zones.

The four state-of-the-art MAS algorithms used for both the analyses are:

- Majority Voting (MV): a standard MAS algorithm which uses a majority voting strategy as fusion label method;
- Local Weighted Voting (WV): a standard MAS algorithm which uses a local weighted voting strategy as fusion label method;
- STEPS (Cardoso et al., 2013): an extension of the classical STAPLE algorithm (Warfield et al., 2004), where, at each voxel location, the atlases are ranked on the basis of the LNCC between the atlas image and the target image and only a fixed number NC of atlases are used by STAPLE;

- JLF (Wang et al., 2013): a MAS characterized by a joint label fusion strategy where the weighted voting is formulated as an optimization problem over unknown voting weights by minimizing the total expectation of the labelling error.

In the following subsection we will denote the four algorithms with MV, WV, STEPS, and JLF if all the atlases are used during the label fusion steps, and with GAS-MV, GAS-WV, GAS-STEPS, and GAS-JLF if only the atlases selected by GAS are exploited during the label fusion step.

For all the analyses a leave-one-out strategy (LOO) strategy is applied for validation: one scan is used as the target image to be segmented and the other scans are used as atlases.

For STEPS and JLF, the default parameters were used as described in the corresponding papers (Cardoso et al. (2013), Wang et al. (2013)). To optimize the GAS parameters, 10 patients were randomly selected from the 90 patients in the PICTURE dataset. The Dice score between the manual segmentation and the segmentation obtained by using GAS-WV with different values of the parameters was calculated. The parameters corresponding to the maximum Dice coefficient value were fixed and used for the experiments on both the PROMISE12 and PICTURE datasets.

For the genetic algorithm, on the basis of our previous experience, we chose the crossover and mutation probability values that allow a good trade off between exploration and exploitation of the search space. Regarding the number of generations, since this value is more dependent on the size of the search space and on the type of optimization problem tackled by the GA, we tried different values and chose the smallest value which ensured convergence of the algorithm.

Since the registration step is the same for all the MAS algorithms, its parameters were not optimized. Instead they were set on the basis of the same 10 patients registered one to each other and evaluated by visual inspection. Table 1 shows the parameters used.

The segmentation accuracy of each algorithm is assessed by using one over-

Table 1: Values of the optimized parameters used in the experiments.

Image registration	
Size of Gaussian Kernel used for LNCC ($\boldsymbol{\sigma_G}$)	5 mm
Control point spacing	2.5mm
Bending Energy	0.1
MAS	
Size of Gaussian Kernel used for LNCC ($\boldsymbol{\sigma_G}$)	5 mm
Temperature used for weighted label fusion ($\boldsymbol{\sigma}$)	0.3
GAS	
Size of the Gaussian Kernel used for LNCC ($\boldsymbol{\sigma_G}$)	5 mm
Number of most similar image ($\boldsymbol{n}$)	9
Size of the population ($\boldsymbol{pop\_size}$)	16
Number of generation ($\boldsymbol{maxNumGeneration}$)	40
Crossover probability ($\boldsymbol{P_{cr}}$)	0.6
Mutation probability ($\boldsymbol{P_{mut}}$)	0.2

lapping metric and two distance based metrics, namely, Dice Similarity Coefficient the (DSC)(Dice, 1945), that measures the spatial overlap between two masks, the 95% Hausdorff Surface Distance (HSD), that measures the largest difference between two contours (the 5% percentile outliers are discarded), and the Symmetric Mean Absolute Surface Distance (MSD) (Gerig et al., 2001), that is the mean of the sum of the shortest Euclidean distance (for each voxel) between segmentation contours.

To assess if there are statistical differences between the performance obtained by each algorithm with and without GAS, and between the performance obtained by GAS and the other selection strategy, a statistical analysis of the three metrics is performed. The Wilcoxon signed-rank test for pairwise comparison (Sheskin, 2007) is applied to the distributions of each metrics calculated using LOO.

3.3. Single-label segmentation

In this sub-section, GAS is tested on single-label segmentation problems, namely, on the segmentation of the whole prostate from T2w MR images and on the segmentation of the left ventricle from cine MRI. We compared the results obtained by the four MAS algorithms without atlas selection, with GAS, and

with the atlas selection strategy proposed in Langerak et al. (2010) (SIMPLE). SIMPLE uses an iterative strategy where at each iteration an estimation of the ground truth segmentation is performed and badly performing atlases are discarded. The iterative method ends when no atlases are discarded for two consecutive iterations.

3.3.1. Whole prostate segmentation

For the segmentation of the whole prostate we performed two different types of validation. First an intra-dataset validation was considered applying the LOO analysis to the PROMISE12 dataset. Then, to test the robustness of GAS when target and atlases belong to different datasets, we applied a inter-dataset validation using images from PROMISE12 and PICTURE as target and atlases, respectively, and vice versa.

Intra-dataset validation

This subsection shows the results obtained applying a LOO segmentation to the PROMISE12 dataset: 49 scans were used as atlas images and one scan as a target image. Table 2 reports the mean and standard deviation of DSC, MSD, and HSD obtained by the four MAS algorithms when different selection strategies are applied. The mean values of the number of atlases selected by GAS and SIMPLE were 21.4 and 29.1, respectively. Table 2 suggests that the use of the GAS always improves the performance of the four MAS algorithms. Indeed, the algorithms with GAS obtain better mean values for all three metrics. This improvement increases as the label fusion technique becomes simpler: as an example, for GA-MV, DSC improves by 6% with respect to MV and by 2% with respect to SIMPLE-MV. When more complex and robust label fusion techniques are considered, the performance improvement reduces, GAS-JLF improves by 2% and 1% with respect to JLF and SIMPLE-JLF, respectively.

The box plot in Fig. 2 confirms the results of Table 2. MAS with GAS are in general more robust: the values of the three indexes are characterized by a smaller number of outliers and a smaller interquartile range of variations (IQR).

Table 2: Mean ± standard deviation of DSC, MSD, and HSD obtained by the four algorithms with the different selection strategies when applied to the whole prostate segmentation on the PROMISE12 dataset. For each metric the best results are in bold, the superscript † represents significantly worse (**p-value**< 0.05) results when compared to GAS.

	DSC	**MSD**	**HSD**
MV	$0.77 \pm 0.13^{\dagger}$	$1.30 \pm 1.45^{\dagger}$	$4.33 \pm 5.54^{\dagger}$
SIMPLE-MV	$0.81 \pm 0.11^{\dagger}$	$1.19 \pm 1.32^{\dagger}$	$3.24 \pm 4.09^{\dagger}$
GAS-MV	$\mathbf{0.83 \pm 0.10}$	$\mathbf{1.13 \pm 1.22}$	$\mathbf{2.45 \pm 2.7}$
WV	$0.79 \pm 0.12^{\dagger}$	$1.26 \pm 1.42^{\dagger}$	$3.14 \pm 3.26^{\dagger}$
SIMPLE-WV	$0.82 \pm 0.11^{\dagger}$	$1.17 \pm 1.28^{\dagger}$	$2.83 \pm 3.62^{\dagger}$
GAS-WV	$\mathbf{0.83 \pm 0.10}$	$\mathbf{1.09 \pm 1.19}$	$\mathbf{2.07 \pm 2.17}$
STEPS	$0.81 \pm 0.10^{\dagger}$	$1.14 \pm 1.19^{\dagger}$	$2.85 \pm 2.96^{\dagger}$
SIMPLE-STEPS	0.83 ± 0.10	1.16 ± 1.18	2.22 ± 2.75
GAS-STEPS	$\mathbf{0.84 \pm 0.08}$	$\mathbf{1.08 \pm 1.16}$	$\mathbf{2.00 \pm 1.66}$
JLF	$0.82 \pm 0.10^{\dagger}$	$1.12 \pm 1.25^{\dagger}$	$2.20 \pm 1.69^{\dagger}$
SIMPLE-JLF	$0.83 \pm 0.12^{\dagger}$	$1.07 \pm 1.18^{\dagger}$	$2.3 \pm 2.47^{\dagger}$
GAS-JLF	$\mathbf{0.84 \pm 0.09}$	$\mathbf{1.04 \pm 1.12}$	$\mathbf{1.83 \pm 1.36}$

To test if there is a statistically significant difference among the distribution of the three metrics, the Wilcoxon signed-rank test for pairwise comparison is applied between each MAS with GAS and both MAS with SIMPLE and MAS without any atlas selection.

For the comparison MAS vs. GAS-MAS, the values of DSC, MSD and HSD were found to be statistically better for all the four MAS algorithms when the GAS is applied. As expected, the highest improvement is obtained for MAS algorithms with the simplest label fusion strategy, i.e., MV and WV (*p-value* lower than 0.0001 for all the three indexes). Although the improvements on the mean values are slightly lower for STEPS and JLF compared to those of MV and WV, the null hypothesis is still rejected. This confirms that when a state-of-the-art local label fusion technique is applied, the global selection of atlases in MAS still significantly improves the segmentation results.

Regarding the comparison of GAS-MAS vs. SIMPLE-MAS, the null hypothesis is always rejected in favour of GAS, except for STEPS.

Finally, to illustrate the segmentations obtained using the four MAS algorithms with and without GAS, Figure 3 shows the segmentation results of the

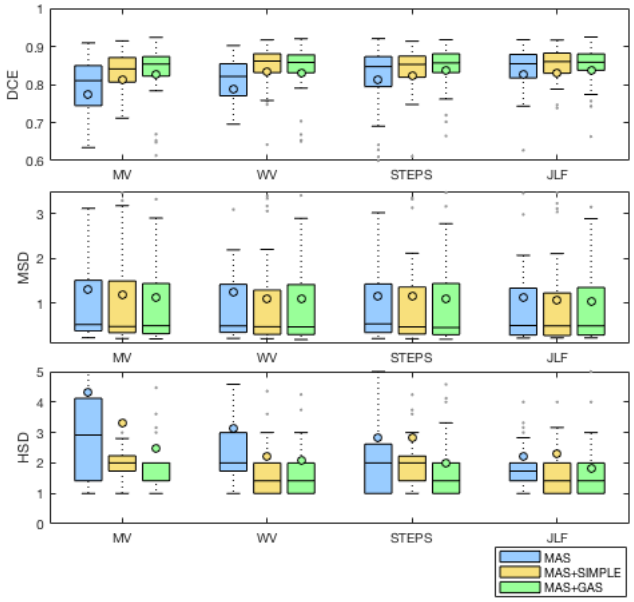

Figure 2: Box plot of the DSC, MSD, and HSD values obtained by the four MAS without atlas selection, with SIMPLE and with GAS, for the PROMISE12 dataset.

central slice obtained by applying the MAS algorithms with and without GAS to four different patients. For comparison, the first column shows the manual segmentation of the same slice.

Among the 50 patients we showed the following cases:

- row (a) the patient with maximum DSC mean value calculated among the four MAS algorithms with GAS ($Max\_\overline{DCE}_{GAS}$).
- row (b) the patient with minimum DSC mean value calculated among the four MAS algorithms with GAS ($Min\_\overline{DCE}_{GAS}$).
- row (c) the patient with maximum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS ($Max\_\overline{DCE}_{diff}$).
- row (d) the patient with minimum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS

$(Min\_\overline{DCE}_{diff})$.

The two extreme examples shown on rows (c) and (d) represent the case where applying GAS produces the best improvement (c) and the biggest deterioration (d) of the DSC value. From Figure 3(c) and (d), we can confirm the numerical results shown in the previous tables: GAS is able to improve the segmentation accuracy and, in the worse case the deterioration of the MAS performance is not significant, i.e., row (d) shows that the segmentations produced with and without GAS are very similar.

From Figure 3 we observe that when MV is applied, removing poorly-registered atlases is crucial as they will affect the final segmentation with the same weight as the well-registered atlases. If the label fusion is based on local weights (WV and JLF), using GAS still substantially improves the results: firstly, although poorly-registered atlases will influence the segmentation with a low weight, if there are a considerable number of these atlases, the noise they introduce affects the final segmentation in a significant way; secondly, poorly-registered atlases can be locally very similar in appearance (local minima in the registration), meaning that image similarity based atlas-selection methods (local and global WV) can give high weight to poorly-registered atlases; thirdly, in applications where image registration is complex, a large portion of the atlases can be consistently badly registered, resulting in a biased consensus and affecting methods such as STAPLE and JLF. For algorithms such as STEPS, which includes both consensus estimation and an image-appearance-based atlas selection, the use of GAS improves the results to a lesser degree.

Figure 3 shows also how the use of GAS affects differently the segmentation results depending on the label fusion.

Inter-dataset validation

In this subsection we compare the performance of the four MAS algorithms with the different selection strategies when the target image and the atlas database belong to different datasets. More specifically, we used the 50 scans from PROMISE12 as target images and the 80 scans from PICTURE as atlas

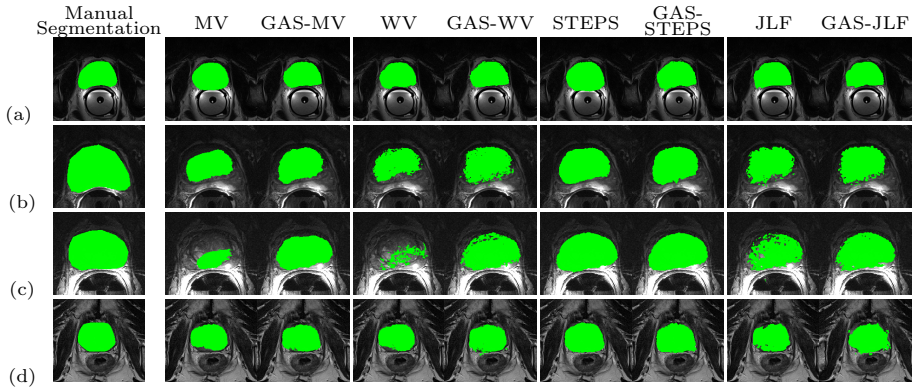

Figure 3: Examples of segmentation of the whole prostate obtained by the four MAS algorithms with and without GAS on four different patients of the PROMISE12 dataset; (a) patient with maximum DSC mean value calculated among the four MAS algorithms with GAS, (b) patient with minimum DSC mean value calculated among the four MAS algorithms with GAS, (c) patient with maximum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS, (d) patient with minimum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS.

database and vice versa. Table 3 shows the values of the three indexes for both cases and Figure 4 shows the corresponding boxplots. The mean values of the number of atlases selected by GAS and SIMPLE were, respectively,14.8 and 18.5 when PROMISE12 was used as atlas database and 28.0 and 35.1 when the atlases are from PICTURE.

When the target images are from PROMISE12, although the results are worse than the ones obtained for the intra-dataset validation, using GAS still improves the performance of each of the MAS algorithms. Indeed, MAS with GAS obtained statistically better values of all the three indexes when compared to both MAS with SIMPLE and MAS without atlas selection. The performance improvement is even more substantial than the one obtained for the inter-dataset validation. In this case using more complex decision fusion techniques does not generate better segmentation results as the number of inaccurate registrations is too high. Furthermore, using GAS increases the robustness of the MAS independently of the applied fusion technique used.

The same conclusions are obtained when PROMISE12 is used an atlas database and the target images are from PICTURE. Also in this case MAS with GAS obtained statistically better results than MAS without atlas selection. Regarding SIMPLE-MAS, the segmentations generated by our method are always statistically more accurate, except for STEPS. GAS-STEPS and SIMPLE-STEPS are not statistically different for any of the three indexes. This could be due to the additional global selection included in STEPS.

Table 3: Mean ± standard deviation of DSC, MSD, and HSD obtained by the MAS algorithms when the atlases are from PICTURE and the targets are from PROMISE12 (on the left), and vice versa (on the right). For each metric the best results are in bold, the superscript † represents significantly worse (p-$value$< 0.05) results when compared to the GAS.

	Atlases from PICTURE Targets from PROMISE12			Atlases from PROMISE12 Targets from PICTURE		
	DSC	**MSD**	**HSD**	**DSC**	**MSD**	**HSD**
MV	$0.68 \pm 0.15^{\dagger}$	$1.55 \pm 1.67^{\dagger}$	$6.25 \pm 6.62^{\dagger}$	$0.74 \pm 0.09^{\dagger}$	$1.27 \pm 0.30^{\dagger}$	$5.49 \pm 5.07^{\dagger}$
SIMPLE-MV	$0.69 \pm 0.17^{\dagger}$	$1.51 \pm 1.64^{\dagger}$	$6.27 \pm 6.83^{\dagger}$	$0.76 \pm 0.09^{\dagger}$	$1.23 \pm 0.29^{\dagger}$	$4.95 \pm 4.86^{\dagger}$
GAS-MV	$\mathbf{0.74 \pm 0.13}$	$\mathbf{1.40 \pm 1.50}$	$\mathbf{4.38 \pm 4.35}$	$\mathbf{0.78 \pm 0.06}$	$\mathbf{1.21 \pm 0.26}$	$\mathbf{3.90 \pm 3.48}$
WV	$0.70 \pm 0.15^{\dagger}$	$1.50 \pm 1.63^{\dagger}$	$5.80 \pm 6.03^{\dagger}$	$0.76 \pm 0.08^{\dagger}$	$1.20 \pm 0.27^{\dagger}$	$4.27 \pm 2.70^{\dagger}$
SIMPLE-WV	$0.70 \pm 0.17^{\dagger}$	$1.48 \pm 1.61^{\dagger}$	$5.90 \pm 6.27^{\dagger}$	$0.77 \pm 0.10^{\dagger}$	$1.17 \pm 0.28^{\dagger}$	$\mathbf{4.11 \pm 3.90}$
GAS-WV	$\mathbf{0.75 \pm 0.12}$	$\mathbf{1.33 \pm 1.43}$	$\mathbf{3.81 \pm 3.54}$	$\mathbf{0.79 \pm 0.06}$	$\mathbf{1.14 \pm 0.25}$	4.22 ± 4.59
STEPS	$0.71 \pm 0.14^{\dagger}$	$1.48 \pm 1.60^{\dagger}$	$5.18 \pm 4.72^{\dagger}$	$0.72 \pm 0.09^{\dagger}$	$1.44 \pm 0.28^{\dagger}$	$4.08 \pm 2.70^{\dagger}$
SIMPLE-STEPS	$0.70 \pm 0.17^{\dagger}$	$1.49 \pm 1.61^{\dagger}$	$5.98 \pm 6.51^{\dagger}$	$\mathbf{0.75 \pm 0.09}$	$\mathbf{1.31 \pm 0.32}$	4.05 ± 3.925
GAS-STEPS	$\mathbf{0.75 \pm 0.12}$	$\mathbf{1.34 \pm 1.42}$	$\mathbf{3.81 \pm 3.73}$	$\mathbf{0.75 \pm 0.08}$	1.36 ± 0.29	$\mathbf{3.38 \pm 2.02}$
JLF	$0.71 \pm 0.14^{\dagger}$	$1.47 \pm 1.64^{\dagger}$	$5.53 \pm 5.31^{\dagger}$	$0.82 \pm 0.06^{\dagger}$	$1.01 \pm 0.23^{\dagger}$	$2.69 \pm 2.31^{\dagger}$
SIMPLE-JLF	$0.70 \pm 0.12^{\dagger}$	$1.45 \pm 1.60^{\dagger}$	$5.72 \pm 5.97^{\dagger}$	$0.81 \pm 0.07^{\dagger}$	$1.31 \pm 0.33^{\dagger}$	$2.92 \pm 2.72^{\dagger}$
GAS-JLF	$\mathbf{0.75 \pm 0.12}$	$\mathbf{1.32 \pm 1.44}^{\dagger}$	$\mathbf{4.14 \pm 4.13}$	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.98 \pm 0.20}^{\dagger}$	$\mathbf{2.23 \pm 1.60}$

3.3.2. Left ventricle segmentation

In this subsection we compare the results obtained by the four MAS algorithms with the different atlas selection strategies. LOO was applied to the 45 patients of SCD. Table 4 shows the mean and standard deviation of the three metrics for both the endocardical and epicardial segmentation and the corresponding boxplots. Although we did not optimize the registration parameters and the parameters of GAS are not optimized for this type of segmentation, we obtained good results, comparable with those obtained during the challenge(Radau et al., 2009). On average, GAS and SIMPLE selected 14.4 and

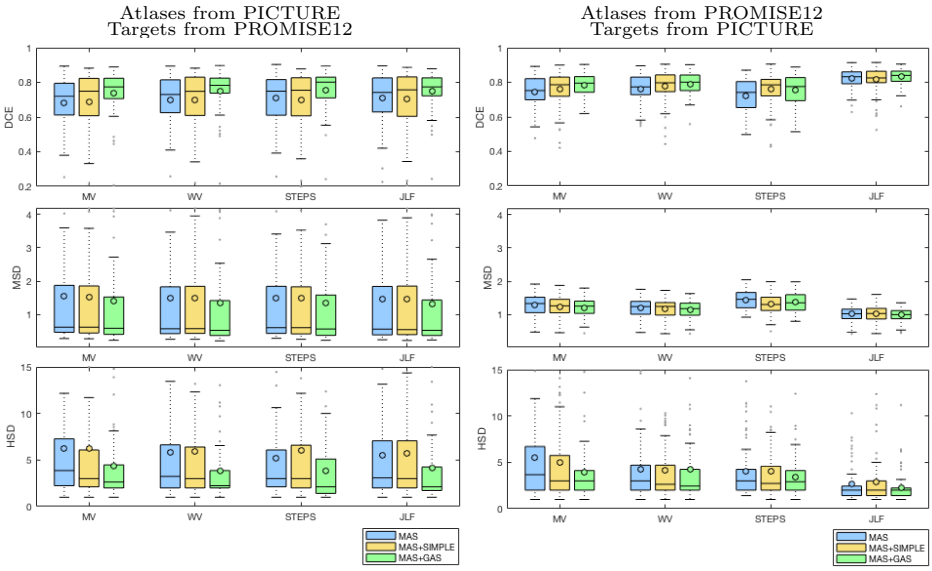

Figure 4: Box plot of the DSC, MSD, and HSD values obtained by the four MAS with the different atlas selection strategies, when the target images are from PROMISE12 and the atlases from PICTURE (left), and vice versa (right).

11.2 atlases, respectively.

Table 4: Mean ± standard deviation of DSC, MSD, and HSD obtained by the four algorithms with the different selection strategies when applied to SCD. For each metric the best results are in bold, the superscript † represents significantly worse (p-$value < 0.05$) results when compared to GAS.

	Endocardial segmentation			Epicardial segmentation		
	DSC	**MSD**	**HSD**	**DSC**	**MSD**	**HSD**
MV	$0.87 \pm 0.06^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$3.58 \pm 2.17^{\dagger}$	$0.87 \pm 0.05^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$3.33 \pm 2.20^{\dagger}$
SIMPLE-MV	$0.88 \pm 0.06^{\dagger}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{2.63 \pm 1.69}$	$0.88 \pm 0.06^{\dagger}$	$\mathbf{0.02 \pm 0.01}$	$3.44 \pm 2.87^{\dagger}$
GAS-MV	$\mathbf{0.89 \pm 0.05}$	$\mathbf{0.02 \pm 0.01}$	2.65 ± 1.63	$\mathbf{0.89 \pm 0.04}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{3.18 \pm 1.88}$
WS	$0.87 \pm 0.06^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$3.75 \pm 3.09^{\dagger}$	$0.87 \pm 0.06^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$3.41 \pm 3.56^{\dagger}$
SIMPLE-WS	$0.87 \pm 0.07^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$3.17 \pm 2.33^{\dagger}$	$0.87 \pm 0.07^{\dagger}$	$\mathbf{0.02 \pm 0.01}$	2.65 ± 1.57
GAS-WS	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{2.99 \pm 2.72}$	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{2.64 \pm 1.96}$
STEPS	$0.86 \pm 0.07^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$2.86 \pm 1.75^{\dagger}$	$0.85 \pm 0.05^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$2.92 \pm 2.88^{\dagger}$
SIMPLE-STEPS	$\mathbf{0.88 \pm 0.06}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{2.58 \pm 1.60}$	$\mathbf{0.86 \pm 0.06}$	$\mathbf{0.02 \pm 0.01}$	$2.83 \pm 2.38^{\dagger}$
GAS-STEPS	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.02 \pm 0.01}$	2.62 ± 1.57	$\mathbf{0.86 \pm 0.06}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{2.64 \pm 3.81}$
JLF	$0.90 \pm 0.05^{\dagger}$	$0.02 \pm 0.01^{\dagger}$	$3.22 \pm 3.44^{\dagger}$	$0.89 \pm 0.05^{\dagger}$	$0.03 \pm 0.01^{\dagger}$	$2.82 \pm 2.88^{\dagger}$
SIMPLE-JLF	$0.90 \pm 0.06^{\dagger}$	$0.02 \pm 0.01^{\dagger}$	$\mathbf{2.58 \pm 1.89}$	$0.89 \pm 0.07^{\dagger}$	0.02 ± 0.01	$2.77 \pm 2.55^{\dagger}$
GAS-JLF	$\mathbf{0.91 \pm 0.05}$	$\mathbf{0.01 \pm 0.01}$	2.56 ± 3.58	$\mathbf{0.90 \pm 0.05}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{2.60 \pm 2.40}$

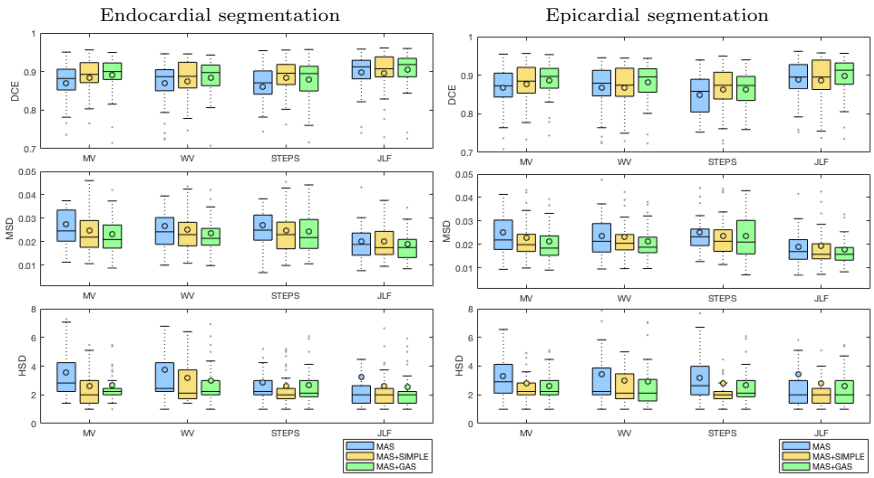

Figure 5: Box plot of the DSC, MSD, and HSD values obtained by the four MAS without atlas selection, with SIMPLE and with GAS, for the SCD dataset.

All four MAS with GAS generated statistically significantly better results for all the three indexes, than the ones obtained without any atlas selection strategy for both the endocardial and epicardial segmentation. However, the improvement on the algorithm's performance is less substantial than the one achieved for prostate segmentation. This is probably due to the higher anatomical variability in prostate images compared to the heart.

Regarding the comparison with SIMPLE, GAS still obtained statistically better performance in terms of Dice coefficient for all the MAS algorithms except STEPS. Indeed, GAS-STEPS and SIMPLE-STEPS are not statistically different on any the three indexes. Since the selection has a lower impact on SDC, also the difference between SIMPLE and GAS is less critical. However, as the boxplot of Figure 5 shows, MAS with atlas selection, and with GAS in particular, generates more robust results than those generated by MAS without atlas selection.

Finally, Figure 6 shows the segmentations obtained by MAS and GAS-MAS on the four patients selected following the same criteria listed in section 3.3.1

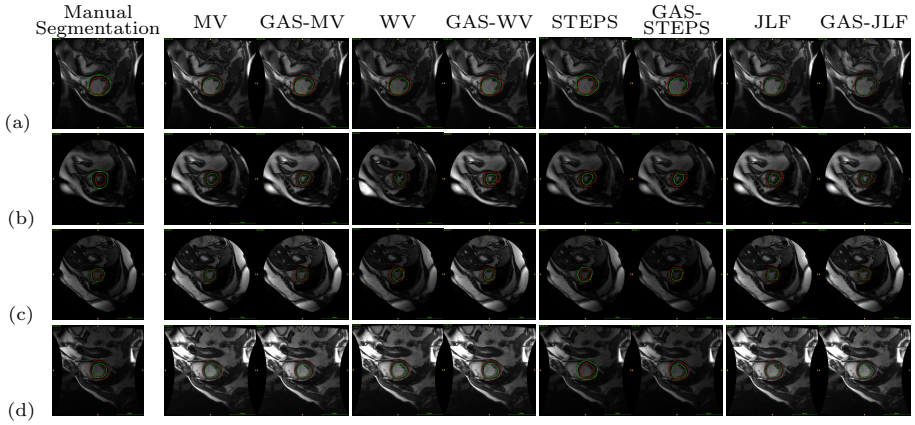

Figure 6: Examples of both endocardial (red) and epicardial (green) segmentations of the left ventricle obtained by the four MAS algorithms with different selection strategies, on four different patients of the SCD dataset;(a) patient with maximum DSC mean value calculated among the four MAS algorithms with GAS, (b) patient with minimum DSC mean value calculated among the four MAS algorithms with GAS, (c) patient with maximum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS, (d) patient with minimum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS.

3.4. Multi-label segmentation

To further validate the proposed selection strategy, we applied GAS to a multi-label scenario: the segmentation of the prostate into peripheral zone (PZ) and transition zone (TZ). We compared the results obtained by the four MAS algorithms with and without GAS when segmenting the prostate into PZ and TZ. LOO analysis was applied to the 80 patients of the PICTURE dataset that had not been used for parameter optimization: 79 patients were used as the atlas database and one patient as the target image to be segmented.

Table 5 and Figure 7, show the mean and standard deviation of the three indexes and the box-plot, respectively, obtained by the four MAS algorithms with and without GAS for the PZ and TZ segmentation. GAS selected, on average 38.2 atalses to segment PZ and TZ

Table 5: Mean ± standard deviation of DSC, MSD, and HSD obtained by the four algorithms with and without GAS when applied to the PZ and TZ segmentations on the PICTURE dataset. For each metric the best results are in bold, the superscript † represents significantly worse ($p\text{-}value < 0.05$) results when compared to GAS.

	PZ segmentation			TZ segmentation		
	DSC	**MSD**	**HSD**	**DSC**	**MSD**	**HSD**
MV	$0.67 \pm 0.08^{\dagger}$	$0.85 \pm 0.21^{\dagger}$	5.39 ± 4.06	$0.81 \pm 0.06^{\dagger}$	$0.91 \pm 0.22^{\dagger}$	$3.12 \pm 2.78^{\dagger}$
GAS-MV	$\mathbf{0.69 \pm 0.07}$	$\mathbf{0.81 \pm 0.20}$	$\mathbf{4.11 \pm 4.41}$	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.88 \pm 0.21}$	$\mathbf{2.76 \pm 2.40}$
WV	$0.68 \pm 0.08^{\dagger}$	$0.82 \pm 0.20^{\dagger}$	$4.52 \pm 3.17^{\dagger}$	$0.82 \pm 0.05^{\dagger}$	$0.88 \pm 0.21^{\dagger}$	$2.97 \pm 2.50^{\dagger}$
GAS-WV	$\mathbf{0.70 \pm 0.07}$	$\mathbf{0.78 \pm 0.20}$	$\mathbf{3.45 \pm 1.51}$	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.85 \pm 0.20}$	$\mathbf{2.74 \pm 2.32}$
STEPS	$0.71 \pm 0.07^{\dagger}$	$0.76 \pm 0.20^{\dagger}$	$2.81 \pm 1.91^{\dagger}$	$0.82 \pm 0.05^{\dagger}$	$0.87 \pm 0.23^{\dagger}$	$2.60 \pm 1.98^{\dagger}$
GAS-STEPS	$\mathbf{0.72 \pm 0.08}$	$\mathbf{0.74 \pm 0.19}$	$\mathbf{2.68 \pm 1.51}$	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.84 \pm 0.20}$	$\mathbf{2.28 \pm 1.65}$
JLF	$0.70 \pm 0.07^{\dagger}$	$0.77 \pm 0.19^{\dagger}$	$3.25 \pm 1.63^{\dagger}$	$0.83 \pm 0.05^{\dagger}$	$0.84 \pm 0.19^{\dagger}$	2.85 ± 3.09
GAS-JLF	$\mathbf{0.71 \pm 0.07}$	$\mathbf{0.76 \pm 0.18}$	$\mathbf{3.11 \pm 1.63}$	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.83 \pm 0.18}$	$\mathbf{2.55 \pm 2.06}$

The values of Table 5 confirm the results obtained for single-label segmentations: when GAS is applied, the four MAS algorithms generate more accurate segmentations in terms on DSC, MSD and HSD. However, for the PICTURE dataset, this performance improvement is less evident than for the PROMISE12 dataset. Since the scans come from the same centre and are acquired with the same scanner, they are very similar to each other, therefore it is less likely that the registration will fail. For this reason the atlas selection step is still necessary

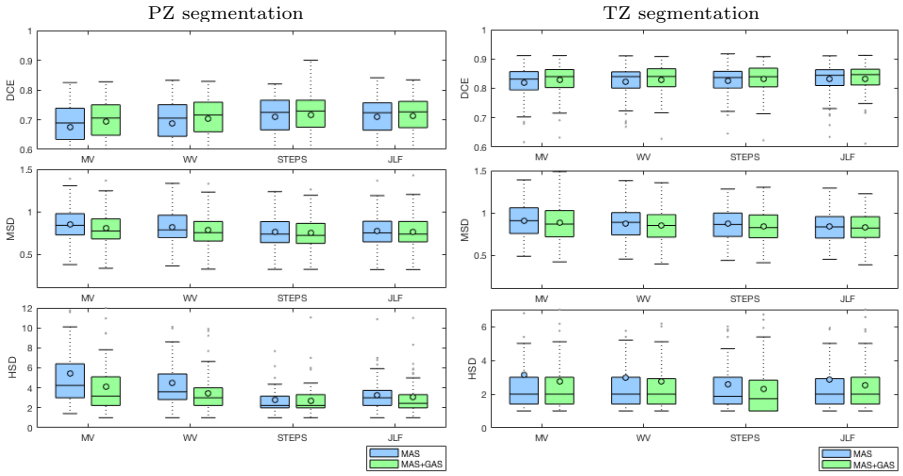

Figure 7: Box plot of the DSC, MSD, and HSD values obtained by the four MAS with and without GAS on the PICTURE dataset for PZ and TZ segmentation.

as this always improves the results, but it affects the segmentations performance in a less significant way.

From Figure 7 we observe that the distributions of the three indexes were found to have better mean values, and smaller values of the IQR when GAS is applied for both PZ and TZ segmentations.

The Wilcoxon signed-rank test applied on the mean distributions confirms the performance improvement on the three indexes. The test results for both the PZ and TZ segmentation showed that the use of MAS algorithms with GAS generates segmentations characterized by statistically significant better values for all the three metrics than the ones generated by the same MAS without any atlas selection.

We can conclude that GAS improves the results also when applied to a multi-label segmentation problems. The magnitude of the improvement depends on the similarity among the scans of the used dataset: if the dataset contains heterogeneous scans, the registration step is more likely to fail and the atlas selection will impact more strongly on the results.

Finally, Figure 8 shows the segmentations obtained by MAS and GAS-MAS

on the four patients selected following the same criteria listed in section 3.3.1.

To further investigate GAS in the multi-label scenario, we compared the results of MAS with and without GAS when applied to the parcellation of the brain in T1 MRI. The dataset used consists of 130 T1 MRI images (80 for the atlas database, 50 for the target images). All the 130 scans have been automatically segmented using GIF (MJ et al., 2015) and manually corrected and quality controlled by either an experienced radiologist or a trained neuroanatomist. As expected, for brain images, the difference between MAS without atlas selection and MAS with GAS is less critical. The reason for this is that inter-subject registration for brain does not fail very often, as the image variability is lower when compared to prostate or heart images.

Although the results of MAS and GAS-MAS are very similar, GAS-MAS still generated segmentations that are statistically significantly more accurate than those generated by MAS, for all the algorithms except JLF. In appendix A the numerical results obtained when applying GAS to MAS with the simplest (MV) and the complex (JLF) selection strategy, are shown.

4. Conclusion

In this paper we have presented a new genetic selection strategy to automatically choose the best subset of atlases within a MAS framework.

We applied our method to single-label and multi-label segmentation problems. For single label problem, we considered the segmentation of the whole prostate from T2w MRI and of the left ventricle of the heart from cine MRI. For multi-label problem, we evaluated GAS on the zonal segmentation of prostate into PZ and TZ. To assess the effectiveness of the GAS strategy we compare the results of four state-of-the-art MAS algorithms without any selection strategy, with GAS, and with SIMPLE.

The statistical analysis of the results showed that, MAS algorithms which use GAS obtained statistically better segmentations than the ones obtained by MAS without GAS for both single-label and multi-label segmentations. The

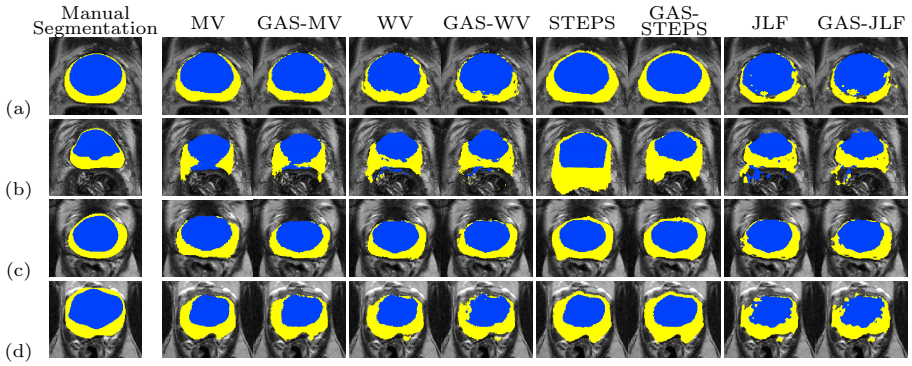

Figure 8: Examples of PZ and TZ segmentations obtained by the four MAS algorithms with and without GAS on four different patients of the PICTURE dataset. (a) patient with maximum DSC mean value calculated among the four MAS algorithms with GAS, (b) patient with minimum DSC mean value calculated among the four MAS algorithms with GAS, (c) patient with maximum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS, (d) patient with minimum difference value between the DSC mean value calculated among the four MAS algorithms with and without GAS.

inter-dataset validation applied to PROMISE12 and PICTURE datasets showed that the improvement of the segmentation accuracy is more substantial when the target images and the atlases are from different datasets. For this reason the performance improvement obtained applying GAS to the multi-center dataset PROMISE12 is more substantial of the one obtained on PICTURE. Thus, with GAS the MAS algorithm becomes more robust to noise, and this is more evident when the complexity and robustness of the label fusion step of the MAS algorithm decreases. We also showed that the use of GAS reduces the impact of the fusion step on the final segmentation. Indeed, when using GAS the performance of the four MAS algorithms become more similar to each other. By selecting the right set of atlases, MAS with a simple local weighed fusion strategy achieves the same performance as MAS algorithms which apply complex label fusion methods such as JLF.

Finally, the statistical comparisons between GAS and SIMPLE demonstrated that GAS, by taking into account not only the performance of a single atlas but

also the interaction of the atlases among themselves, can generate statistically significantly better segmentations for both whole prostate and left ventricle segmentation problems.

5. Acknowledgements

Special thanks go to Catherine Scott for her overall assistance. This work is supported by EPSRC, the National Institute for Health Research University College London Hospitals Biomedical Research Centre (BRC), Cancer Research UK (CRUK), and by the Comprehensive Cancer Imaging Centre (CCIC).

6. Appendix A. Brain parcellation in T1 MRI

For the brain parcellation, we used a dataset of 130 T1 MRI images (80 for the atlas database, 50 for the target images) automatically segmented by GIF (MJ et al., 2015) and manually corrected and quality controlled by either an experienced radiologist or a trained neuroanatomist.

Table 6 shows the mean DSC value obtained by MAS and GAS-MAS for two label fusion strategies: the simplest (MV) and the most complex (JLF). The results related to nine main clinically relevant brain structures are reported; if a region has a left and a right side then the mean value of both sides is shown. The best results are in bold, and the * represents significantly better results.

The difference between MAS and MAS-GAS is less noticeable for brain segmentation, and JLF and GAS-JLF have almost the same performance. Although GAS-MV generates statistically better results than MV, the difference between the two algorithms in terms of Dice values is small. This result confirms that the improvement achieved by applying GAS is related to the quality of the registration step.

Table 6: Mean ± standard deviation of DSC obtained by the MAS with the simplest (MV) and the most complex label fusion strategy with and without GAS when applied to the brain parcelation. The best results are in bold, the superscript † represents significantly worse ($p\text{-}value < 0.05$) results when compared to GAS.

	MV	**GAS-MV**	*p-value*	**JLF**	**GAS-JLF**	*p-value*
Hippocampus	$0.833 \pm 0.008^{\dagger}$	$\mathbf{0.844 \pm 0.006}$	< 0.0001	$\mathbf{0.918 \pm 0.005}$	$\mathbf{0.918 \pm 0.006}$	0.689
Amygdala	$0.864 \pm 0.007^{\dagger}$	$\mathbf{0.870 \pm 0.003}$	< 0.0001	$\mathbf{0.899 \pm 0.003}$	$\mathbf{0.899 \pm 0.004}$	0.662
Caudate	$0.827 \pm 0.004^{\dagger}$	$\mathbf{0.848 \pm 0.004}$	< 0.0001	$\mathbf{0.920 \pm 0.002}$	$\mathbf{0.920 \pm 0.002}$	0.942
Accumbens Area	$0.866 \pm 0.006^{\dagger}$	$\mathbf{0.871 \pm 0.005}$	0.0200	$\mathbf{0.901 \pm 0.004}$	0.900 ± 0.004	0.932
Putamen	$0.919 \pm 0.007^{\dagger}$	$\mathbf{0.924 \pm 0.005}$	0.0003	$\mathbf{0.936 \pm 0.004}$	0.935 ± 0.005	0.424
Thalamus	$0.892 \pm 0.002^{\dagger}$	$\mathbf{0.902 \pm 0.002}$	< 0.0001	0.935 ± 0.001	$\mathbf{0.936 \pm 0.001}$	0.698
Globus pallidus	$0.902 \pm 0.007^{\dagger}$	$\mathbf{0.907 \pm 0.006}$	0.0003	0.926 ± 0.004	$\mathbf{0.927 \pm 0.003}$	0.936
Brain Stem	$0.916 \pm 0.004^{\dagger}$	$\mathbf{0.918 \pm 0.003}$	< 0.0001	$\mathbf{0.947 \pm 0.001}$	$\mathbf{0.947 \pm 0.001}$	0.496
Cerebellum	$0.786 \pm 0.005^{\dagger}$	$\mathbf{0.795 \pm 0.005}$	< 0.0001	$\mathbf{0.882 \pm 0.003}$	$\mathbf{0.882 \pm 0.003}$	0.324
Lateral Ventricle	0.853 ± 0.007	$\mathbf{0.869 \pm 0.005}$	0.09	$\mathbf{0.957 \pm 0.001}$	0.956 ± 0.002	0.494
Corpus Callosum	$0.883 \pm 0.002^{\dagger}$	$\mathbf{0.902 \pm 0.002}$	< 0.0001	0.939 ± 0.001	$\mathbf{0.940 \pm 0.001}$	0.043

References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage 46, 726 – 738.

Artaechevarria, X., Muoz-Barruti, A., Ortiz-de-Solrzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imag. 28, 1266–1277.

Beasley, D., Bull, D.R., Martin, R.R., 1993. An overview of genetic algorithms: Part 1, fundamentals. University Computing 15, 58–69.

Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the PASHA algorithm. Computer Vision and Image Understanding 89, 272 – 298.

Cardoso, M.J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S., 2013. Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Medical Image Analysis 17, 671 – 684.

Chilali, O., Puech, P., Lakroum, S., Diaf, M., Mordon, S., Betrouni, N., 2016. Gland and zonal segmentation of prostate on T2W MR images. Journal of Digital Imaging 29, 730–736.

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26, 297–302.

Gerig, G., Jomier, M., Chakos, M., 2001. Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation. Springer Berlin Heidelberg, Berlin, Heidelberg.

Graham, J., Kirkbride, P., Cann, K., Hasler, E., Prettyjohns, M., 2014. Prostate cancer: summary of updated NICE guidance. BMJ : British Medical Journal 348.

Guo, Y., Gao, Y., Shen, D., 2016. Deformable MR prostate segmentation via deep feature learning and sparse patch matching. IEEE Trans. Med. Imaging 35, 1077–1089.

Iglesias, J.E., Sabuncu, M.R., 2014. Multi-atlas segmentation of biomedical images: A survey. Med. Image Anal. 24, 205–219.

Isgum, L., Staring, M., A. Rutten, M.P., Viergever, M., Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion. application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging 28, 1000 – 1010.

Klein, S., van der Heide, U.A., Lips, I.M., van Vulpen, M., Staring, M., Pluim, J.P.W., 2008. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Medical Physics 35, 1407–1417.

Langerak, R., van der Heide, U.A., Kotte, A.N.T.J., Viergever, M.A., van Vulpen, M., Pluim, J.P.W., 2010. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans. Med. Imaging 29, 2000–2008.

Langerak, T.R., Berendsen, F.F., Van der Heide, U.A., Kotte, A.N.T.J., Pluim, J.P.W., 2013. Multiatlas-based segmentation with preregistration atlas selection. Medical Physics 40.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., Zhang, J., Strand, R., Malmberg, F., Ou, Y., Davatzikos, C., Kirschner, M., Jung, F., Yuan, J., Qiu, W., Gao, Q., Edwards, P., Maan, B., van der Heijden, F., Ghose, S., Mitra, J., Dowling, J., Barratt, D., Huisman, H., Madabhusi, A., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge 18, 359–373.

Mahapatra, D., Buhmann, J.M., 2014. Prostate MRI segmentation using learned semantic knowledge and graph cuts. IEEE transactions on biomedical engineering 61, 756 – 764.

Makni, N., Betrouni, N., Colot, O., 2014. Introducing spatial neighbourhood in evidential c-means for segmentation of multi-source images: Application to prostate multi-parametric MRI. Information Fusion 19, 61 – 72.

Martin, S., Daanen, V., Troccaz, J., 2008. Atlas-based prostate segmentation using an hybrid registration. International Journal of Computer Assisted Radiology and Surgery 3, 485–492.

MJ, C., M, M., R, W., A, M., D, C., D, R., S, O., 2015. Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion. IEEE Trans Med Imaging 34, 1976–88.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine 98, 278–284.

Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N., 2000. Recon-

structing a 3D structure from serial histological sections. Image and Vision Computing 19, 25–31.

Qiu, W., Yuan, J., Ukwatta, E., Sun, Y., Rajchl, M., Fenster, A., 2014. Dual optimization based prostate zonal segmentation in 3D MR images. Medical Image Analysis 18, 660–673.

Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G., 2009. Evaluation framework for algorithms segmenting short axis cardiac MRI. `http://hdl.handle.net/10380/3070` .

van Rikxoort, E.M., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M.A., Pluim, J.P., van Ginneken, B., 2010. Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. Medical Image Analysis 14, 39 – 49.

S. Wang, K. Burtt, B.T.P.C., Summers, R.M., 2014. Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research. BioMed Research International .

Sabuncu, M.R., Yeo, B.T.T., Leemput, K.V., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging 29, 1714–1729.

Sheskin, D.J., 2007. Handbook of Parametric and Nonparametric Statistical Procedures. 4 ed., Chapman & Hall/CRC.

Simmons, L., Kanthabalan, A., Arya, M., Briggs, T., Barratt, D., Charman, S., Freeman, A., Gelister, J., Hawkes, D., Hu, Y., Jameson, C., McCartan, N., Moore, C., Punwani, S., Ramachandran, N., van der Meulen, J., Emberton, M., Ahmed, H.., 2017. The PICTURE study: diagnostic accuracy of multiparametric MRI in men requiring a repeat prostate biopsy. Br J Cancer 116, 1159–1165.

Toth, R., Ribault, J., Gentile, J., Sperling, D., Madabhushi, A., 2013. Simultaneous segmentation of prostatic zones using active appearance models with

multiple coupled levelsets. Computer Vision and Image Understanding 117, 1051 – 1060.

Wang, H., Suh, J.W., Das, S.R., Pluta, J., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35, 611–623.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23, 903–921.

Y. Ou, J. Doshi, G.E., Davatzikos, C., 2012. Multi-atlas segmentation of the prostate: A zooming process with robust registration and atlas selection. Promise Miccai 2012 Grand Challenge on Prostate MR Image Segmentation , 60–65.

Yan, P., Cao, Y., Yuan, Y., Turkbey, B., Choyke, P.L., 2015. Label image constrained multiatlas selection. IEEE Trans. Cybernetics 45, 1158–1168.