Layered spatio-temporal forests for left ventricle segmentation from 4D cardiac MRI data

Ján Margeta[1], Ezequiel Geremia[1], Antonio Criminisi[2], Nicholas Ayache[1]

[1] Asclepios Research Project, INRIA Sophia-Antipolis, France.
[2] Machine Learning and Perception Group, Microsoft Research Cambridge, UK.

Abstract. In this paper we present a new method for fully automatic left ventricle segmentation from 4D cardiac MR datasets. To deal with the diverse dataset, we propose a fully automatic machine learning approach using two layers of spatio-temporal decision forests with almost no assumptions on the data or segmentation problem. We introduce 3D spatio-temporal features to classification with decision forests and propose a method for context aware MR intensity standardization and image alignment. The second layer is then used for the final image segmentation. We present our first results on the STACOM LV Segmentation Challenge 2011 validation datasets.

1 Introduction

The left ventricle plays a fundamental role in circulation of oxygenated blood to the body. To assess its function, several indicators are often calculated in clinical practice. Many of these are based on ventricular volume and mass measurements at reference cardiac phases. To calculate these an accurate delineation of the myocardium and the cavity is necessary. To remove the bias and variance of manual segmentation, and obtain reproducible measurements, an automatic segmentation technique is desirable.

Compared to computed tomography (CT), cardiac magnetic resonance imaging (cMRI) offers superior temporal resolution, soft tissue contrast, no ionizing radiation, and a vast flexibility in image acquisition characteristics. As a disadvantage, cMRI scans often yield significantly lower resolution in the plane orthogonal to the plane of acquisition, the images can suffer from magnetic field inhomogeneities and respiration artifacts can manifest as slice shifts. Moreover, the lack of standard units (compared to the Hounsfield scale in CT) makes it difficult to directly apply most of the intensity based segmentation techniques.

Motivated by the success of Lempitsky et al. [1] in myocardium segmentation from 3D ultrasound sequences in near real time and Geremia et al.[2] for multiple sclerosis lesion segmentation, we propose a fully automated voxel-wise segmentation method based on decision forests (DF) with no assumptions on shape, appearance, motion (except for periodicity and temporal ordering) or knowledge about the cardiac phase of the images in the sequence. The left ventricle segmentation problem is defined as the classification of voxels into myocardium and background.

Instead of robustly registering to an atlas [3], building a model [4] or running a highly specialized segmentation algorithm we leave the learning algorithm to automatically decide the relevant features for solving the segmentation problem using the provided ground-truth only. In principle, any pathology can be learnt once a similar example is represented within the training dataset. The previously used decision forests [1][2] rely on features that work the best when image intensities and orientations are very similar. To tackle the highly variable dataset, we propose a layered learning approach, where the output of each layer serves a different purpose. The first layer is used to prepare the data for a more semantically meaningful and accurate segmentation task in the second layer.

The main contributions of this paper are: a method to use decision forests to solve the MR intensity standardization problem (Section 3.1) and, similarly, perform a context sensitive rigid registration (Section 3.2) to align all images to a reference pose. We also suggest a way to introduce temporal dimension into the currently used 3D random features (Section 2.2). Finally, on the intensity standardized and pose normalized images, we then train a second forest layer (Section 4) using also the spatial information. This helps the trees to automatically build their own latent shape representation.

Dataset. STACOM 2011 LV segmentation challenge data [5] were divided into two sets. Training set (100 3D+t short axis (SA) volumes with manually delineated myocardia at each cardiac phase) and validations sets (5×20 3D+t SA volumes with no delineation provided).

This dataset clearly shows the anatomical variability of heart shape and appearance and some of the main issues of cMRI mentioned above.

2 Layered spatio-temporal decision forests

Decision forests are an ensemble supervised learning method consisting of boosting a set of binary decision trees. The training set contains a set of feature measurements and associated labels (myocardium/background) for each of the voxels in the set.

The trees are built in a top-down fashion, from the root, down to the leaves. At each node, local features and a randomly sampled subset of context-rich features are considered for feature selection. Random sampling of the features leads to increased inter-node and inter-tree variability to improve generalization. Each feature θ can be regarded as a binary decision (in our case $\tau_l < \theta < \tau_h$) that splits the original set into two disjoint subsets. The trees then select the most discriminative features for each split such that the information gain is maximized. The data division then recursively continues until a significant part of the voxels at the node belongs to a single class and the node becomes a leaf. The averaged class distributions of all the leaves in the forest reached by the voxel then represent the posterior probabilities of it belonging to either the myocardium or the background. See Geremia et al. [2] for more details.

2.1 Strategy to learn from spatio-temporal data

The layers are trained one by one, each with the aim to learn to segment. Training with all the 3D+t data was not feasible within the time limits of the challenge, therefore a reduced strategy was designed. This strategy is repeated for each tree:

1. Select a random subset of k 4D volumes from the whole training set
2. Randomly choose a reference 3D frame I^c for each selected 4D volume
3. Select two frames I^{c-o}, I^{c+o} with a fixed offset o on both sides from the reference cardiac image I^c
4. Train the tree using a set of k triplets (I^c, I^{c-o}, I^{c+o})

To reduce the computational time, the size of the subset for each tree was set to $k = 15$, and only one fixed offset $o = 4$ is currently used. The choice of o was made such that the motion between the selected frames is important even when more stable cardiac phases (end systole and diastole) are selected as the reference frame and almost a half of the cardiac cycle could be covered.

2.2 Features

We use several features families to generate the random feature pool operating on the triplets. Their overview can be seen on Figure 2.2).

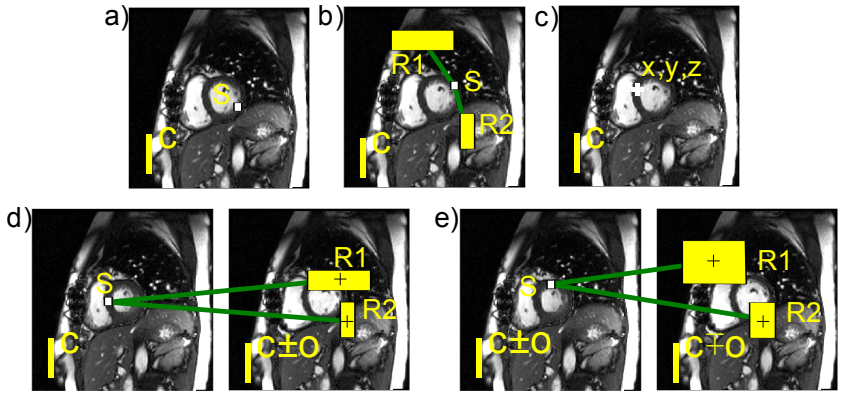

Fig. 1. Illustration of image based features extracted from the images. a) Local features ($3\times3\times3$ box average S around the source voxel in the current frame I^c) [2]. b) Context rich features [2] measuring the difference between source box average S and the sum of remote region averages $R1$ and $R2$. c) Components x,y,z of voxel coordinates as features[1]. d) Spatio-temporal context rich features with the current frame as the source image and offset frame $I^{c\pm o}$ as the remote. e) Spatio-temporal context rich features with one of the offset frames as the source image and the other as remote.

Local features. Proposed in [2] as an average of intensities in the vicinity of the tested voxel to deal with noise in magnetic resonance imaging:

$$\theta^{loc}_{I^c}(x) = \theta^{loc}_{I^c}([x,y,z]) = \sum_{x'=x-1}^{x'\leq x+1} \sum_{y'=y-1}^{y'\leq y+1} \sum_{z'=z-1}^{z'\leq z+1} I^c([x',y',z']) \tag{1}$$

Although these features are not intensity invariant, they can still quite well reject some highly improbable intensities.

Context rich features. Defined also in [2], for multichannel MR acquisitions as a difference between the local source image intensity I^S and box averages of remote regions in image I^R:

$$\theta^{CR}_{I^S,I^R}(x) = I^S(x) - \frac{1}{Vol(R_1)} \sum_{x'\in R_1} I^R(x') - \frac{1}{Vol(R_2)} \sum_{x'\in R_2} I^R(x') \tag{2}$$

The 3D regions R_1 and R_2 are randomly sampled in a large neighborhood around the origin voxel. These capture strong contrast changes and long-range intensity relationships. In our case we define context-rich features as $\theta^{CR}_{I^c,I^c}(x)$.

Spatio-temporal context rich features. The moving heart can be well coarsely extracted by just thresholding the temporal difference magnitude of the image. We propose to exploit this wealth of information and extend the previously context-rich features into the temporal domain by comparing the "current" 3D frame I^c and another frame offset from c by $\pm o$. The temporal context-rich features can be defined as $\theta^{TCR1}_{I^c} = \theta^{CR}_{I^c,I^{c+o}}(x)$ and $\theta^{TCR1}_{I^c} = \theta^{CR}_{I^c,I^{c-o}}(x)$.

Similarly, we measure the differences between the symmetrically offset frames contained in the triplet as $\theta^{TCR2}_{I^c}(x) = \theta^{CR}_{I^{c+o},I^{c-o}}(x)$ and $\theta^{TCR2}_{I^c}(x) = \theta^{CR}_{I^{c-o},I^{c+o}}(x)$. These spatio-temporal features can be seen as an approximation of a temporal differentiation around the center frame. Note that we use both $+o$ and $-o$ to keep some symmetry of the remote region distribution.

Voxel coordinates. Finally, as in [1], we can insert voxel coordinates: $\theta^X_C(x) = x_x$, $\theta^Y_C(x) = x_y$, $\theta^Z_C(x) = x_z$ into the feature pool. However, not until these coordinates have a meaning; which happens later, in the second forest layer when the images are reoriented into the standard pose.

2.3 Data preprocessing

To use fast evaluation of previously defined features based on integral images [6], it is necessary to have consistent spacing. Therefore, all the volumes were resampled to one of the most common spatial spacings in the dataset ($1.56, 1.56, 7.42mm$) and temporal sequence length (20 frames).

Intensity ranges of the images were all linearly rescaled to a fixed range. Similarly to Nyúl et al. [7], we clamp intensities beyond the 99.8 percentile as they usually do not convey much useful information.

3 First layer: Decision forests for image intensity standardization and position normalization

Following the above mentioned training subset selection strategy we can train the first layer of the forests. This is done directly on the images after intensity rescaling i.e. images are brought into the same intensity range but have their original poses. Although short axis scans are often acquired close to a position where the ventricular ring is centered, slice orientation is chosen manually during the acquisition, and precise alignment cannot be guaranteed. Therefore we skip the usage of voxel coordinate features at this step.

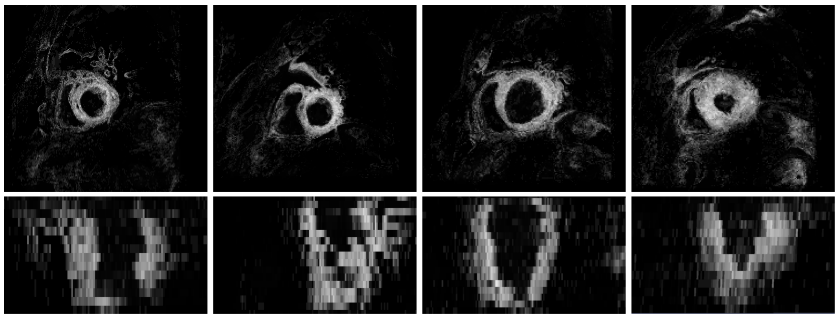

Fig. 2. Short (top) and long (bottom) axis views on the posterior probabilities after the first layer. Brighter value means higher probability.

Several authors (e.g. [3]) have proposed to use Haar like features to detect the heart and crop the heart region. Images can be then registered using the cropped volumes. This removes the influence of background structures and improves the success rate for the registration. However, an extraction of the cropped region will not be necessary to perform a robust registration in our case. We train the first layer of the forests on a rather general scenario, to end up with at least a very rough classification performance (see Figure 2). As we show in the next two sections, using the rough posterior probability map of a tissue belonging to a ventricle this performance can be already good enough for ventricle detection, intensity standardization and alignment onto a reference orientation without any prior knowledge of the data apart from the ground-truth.

3.1 Intensity standardization

MR intensity value differences of the same tissue are significant not only between scanners, acquisition protocols [8] but also for the same followup patients [7]. Therefore good intensity standardization is crucial for any intensity based segmentation algorithm. The variance in median intensities of the myocardia between different cases in the STACOM training set is quite large. There is no

unique mode and the distribution is fairly spread in the whole intensity range $(0, 65535)$. Median myocardial intensities span range $(1954, 36430)$, with standard deviation of 5956 and inter-quantile range 7663). This is a serious problem for any intensity based segmentation method.

Many of the intensity standardization algorithms [9] used today are based on the methods of Nyúl et al. [7][10] and the alignment of histogram based landmarks (e.g. modes, percentiles or statistics of homogeneous connected regions) by rescaling image intensities with a piecewise linear mapping. Many of these methods to work reasonably well for brain images where the white matter is clearly the most dominant tissue. In cMRI, the largest homogeneous regions would most of the time belong to the lungs, liver or cavities, rather than the myocardium.

However, from the rough image first layer classification we already obtain some information about the strength of the belief in the foreground and background object. We propose to remap the source image intensities by a piecewise linear function such that the weighted median (as median is more robust to outliers than the mean) M^c_{source} of the images is transformed to a reference value M_{ref}. The weighted median is defined as follows:

$$M^c_{source} = \arg\min_{\mu} \sum_{x \in I^c} w(x).|I^c(x) - \mu| \tag{3}$$

Where x is the voxel iterator and $w(x)$ are the weights(first layer posterior probabilities). We avoid sorting of all volume intensities by approximating the weighted median with the weighted version of the P^2 algorithm [11][12]. This algorithm dynamically approximates the cumulative probability density function with a piece-wise quadratic polynomial by shifting positions of just five markers as the weighted samples are streamed in. Each of these markers are associated with their position, percentile and an intensity value corresponding to that percentile. The positions are updated such that they correspond to the sum of weights of samples whose intensity value is smaller than the value the markers hold.

3.2 Orientation normalization

In the approach of Lempitsky et al. [1] coordinate features are used directly. This choice cannot be justified without aligning the images onto a reference pose. Moreover, features we use for classification are not rotation invariant. Therefore if all the volumes could be registered to have the same orientation, the classification would certainly benefit from it. The interpatient cardiac registration are generally a difficult problem due to the high variability in the thoracic cage. Shi et al. [3] do first learning based heart detection and then apply a locally affine registration method which they claim to be robust for large differences.

A robust learning based linear inter-patient organ registration was proposed by Konukoglu et al.[13]. Here, each organ is represented with a smooth probability map fit to the bounding boxes obtained as a result from classification. Then, registration of these probability maps is performed. This sigmoid representation

is however rather limiting since it disregards the orientation that we would like to correct for.

Without any assumptions on the shape of the distribution, we propose to use a fast and robust rigid block matching registration technique [14] directly on the myocardium enhanced first layer posterior probability maps instead and obtain the transformation. The reference we used was chosen randomly among the images where the apex was at least partially closed. A better choice of the reference, is currently out of scope of this paper. However, an algorithm similar to Hoogendoorn et al.[15] or a generative technique similar to [16] could be used.

Note, to reduce the computational time, only frames from the middle of the sequence are used to estimate the intensity and pose transformations. The same transformations are then applied to the rest of the frames and also to the ground truths that will be now needed in the second layer.

4 Second layer: Learning to segment with the shape

4.1 Using voxel coordinates

Once the images are registered to a reference volume, the voxel coordinates start to encode spatial relationships with respect to the reference coordinate frame and the coordinate features can be now included in training of the second decision forest layer. Moreover, if the intensity standardization step succeeds, the intensities have more tissue specific meaning (at least for the myocardium).

Thanks to the incorporation of coordinate based features, the tree can completely automatically learn its own latent representation of the possible set of shapes, regularize the classification, and help to remove objects far away from the ventricle. However, this step strongly relies on the success of the previous registration step. Currently, only one reference image is used. Registration to multiple targets should therefore improve robustness and alleviate this problem.

4.2 Transforming the volumes back

After the classification is done in the reference space, the posterior probability maps can be transformed back to the original reference frame and be resampled accordingly. This shows the advantage of a soft classification technique where the final binary mask is obtained by thresholding the transformed non-integer posterior map, thus avoiding some of the interpolation artifacts.

5 Results

Here we show the preliminary results of our method. The forest parameters for the first layer were fixed as follows: 20 trees with depth 20 each. To train each tree, 15 frames (+ their corresponding offset neighbours) from different volumes were randomly selected from the whole training set (91 volumes in total). For the second layer: 27 trees each with depth 20. For each tree 12 frames (+ their

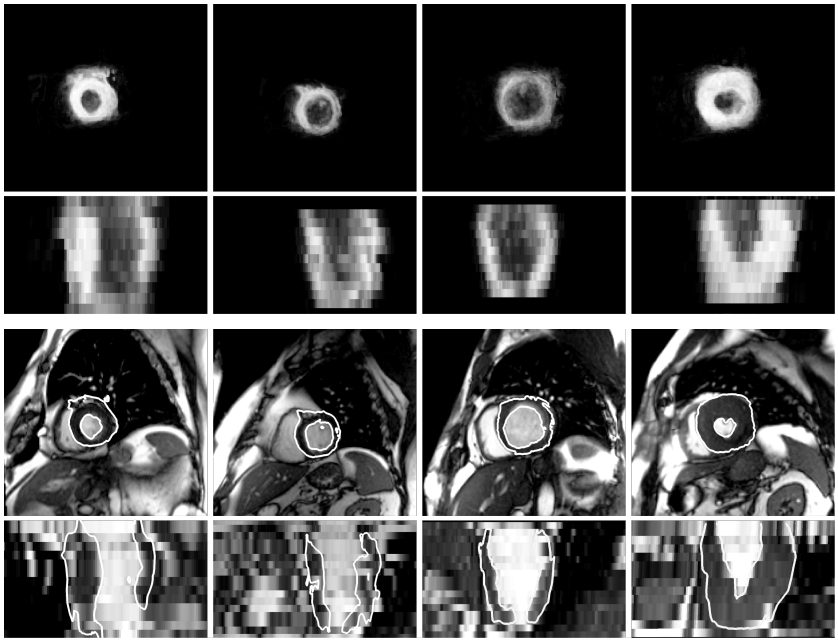

Fig. 3. Short (top) and long (bottom) axis views on the posterior probabilities after the second layer and segmentation results (isocontour of the probability map at 0.5)

corresponding offset neighbours) from different volumes were randomly selected from the training set (91 volumes in total). There is a vast reserve in utilisation of the training set and optimal forest sizes. These parameters were chosen rather empirically to fit into the computational and time limit of the challenge.

After blind evaluation of our classifications on 90 volumes i.e. 25415 slices from the validation dataset by the STACOM LV segmentation workshop organisers following per slice measurements were obtained.

6 Conclusions

We presented our preliminary results of our fully automatic machine learning based algorithm for left ventricle segmentation. The algorithm learnt to automatically select discriminative features for the task using the ground-truth only. The only assumptions we make is that the motion of the object to be segmented is periodic and that the ideal intensity mapping between two different cases can be approximated by a monotonically increasing function. We also introduced a learning based intensity standardization method that allows to do tissue specific remapping of intensities and obtain a more CT like behaviour.

The results were obtained from pure learning, completely automatically, with no interaction and post-processing, and an important reserve on the search of optimal parameters. This should help us to further improve the segmentation (e.g.

	sensitivity $\frac{TP}{TP+FN}$	specificity $\frac{TN}{FP+TN}$	accuracy $\frac{TP+TN}{P+N}$	PPV $\frac{TP}{TP+FP}$	NPV $\frac{TN}{TN+FN}$	dice $\frac{2\|A\cap B\|}{\|A\|+\|B\|}$	jaccard $\frac{\|A\cap B\|}{\|A\cup B\|}$
mean	0.6857	0.9897	0.9861	0.4791	0.9962	0.5045	0.3730
median	0.8099	0.9907	0.9875	0.5234	0.9978	0.5995	0.4281
σ	0.3137	0.0077	0.0077	0.2069	0.0046	0.2571	0.2098

Table 1. Statistics on the per-slice measures of our segmentation results on 90 volumes from the validation dataset. The per-slice measurements strongly penalize voxel misclassifications in the apical and basal areas where the slices contain only very few groundtruth voxels which leads to increased variance in the measures. On the other hand, high specificity is to be expected given the proportion of background on the image compared to the myocardium. Nevertheless, it shows that the algorithm's abilities to most of the time correctly identify the background without much clutter.

number of frames used for training per tree, tree count etc.). The classification is run independently for each voxel with no connectivity nor temporal consistency constraints. Therefore isolated segmentation islets or holes in the resulting binary segmentation can occur as a result of misclassification. However, thanks to the coordinate features, voxels far from the myocardium are normally well discarded. Moreover, in the soft classification, the holes are represented as a drop in the segmentation confidence but rarely fall to zero. This information could be easily considered in the regularization step to further improve the segmentation.

Finally, using a curvature-based iterative hole filling algorithm [17] on the binarized segmentation, we can automatically calculate volumetric measurements and detect the main cardiac phases (see Figure 4).

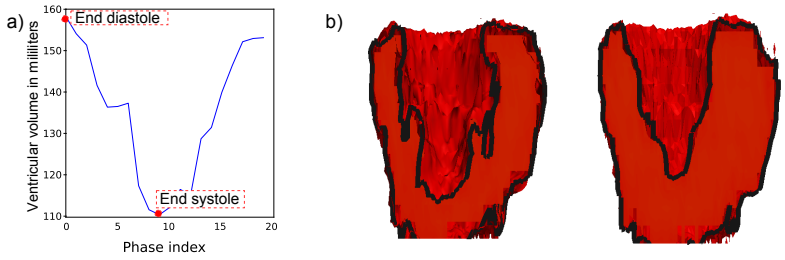

Fig. 4. a) Automatically calculated volume curve from patient DET0026701 during a single cardiac cycle with detected end systole (ES) and end diastole (ED) frames at the volume maximum and minimum respectively. b) Long axis crosssection through the binarized segmentations at ED and ES.

Acknowledgements. This work was partly supported by Microsoft Research through its PhD Scholarship Programme. We used data and infrastructure made available through the Cardiac Atlas Project (www.cardiacatlas.org) [5].

References

1. Lempitsky, V., Verhoek, M., Noble, J., Blake, A.: Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. Functional Imaging and Modeling of the Heart (2009) 447–456
2. Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N.: Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage (2011)
3. Shi, W., Zhuang, X., Wang, H., Duckett, S., Oregan, D., Edwards, P., Ourselin, S., Rueckert, D.: Automatic Segmentation of Different Pathologies from Cardiac Cine MRI Using Registration and Multiple Component EM Estimation. Functional Imaging and Modeling of the Heart (2011) 163–170
4. Lu, X., Wang, Y., Georgescu, B., Littman, A., Comaniciu, D.: Automatic Delineation of Left and Right Ventricles in Cardiac MRI Sequences Using a Joint Ventricular Model. Functional Imaging and Modeling of the Heart (2011) 250–258
5. Fonseca, C., Backhaus, M., Bluemke, D., Britten, R., Chung, J., Cowan, B., Dinov, I., Finn, J., Hunter, P., Kadish, A., Lee, D., Lima, J., Medrano-Gracia, P., Shivkumar, K., Suinesiaputra, A., Tao, W., Young, A.: The Cardiac Atlas Project-an Imaging Database for Computational Modeling and Statistical Atlases of the Heart. Bioinformatics **in press** (2011)
6. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001 (2001) I–511–I–518
7. Nyúl, L.G., Udupa, J.K.: On standardizing the MR image intensity scale. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine **42**(6) (1999) 1072–81
8. Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T.: Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. Medical image analysis **15**(2) (December 2010) 267–282
9. Bergeest, J., Florian Jäger, F.: A Comparison of Five Methods for Signal Intensity Standardization in MRI. Bildverarbeitung für die Medizin 2008 (2008) 36–40
10. Nyúl, L.G., Udupa, J.K., Zhang, X.: New Variants of a Method of MRI Scale Standardization. IEEE transactions on medical imaging **19**(2) (2000) 143–150
11. Jain, R., Chlamtac, I.: The P2 algorithm for dynamic calculation of quantiles and histograms without storing observations. Communications of the ACM **28**(10) (1985)
12. Egloff, D.: Weighted P2 quantile, Boost Accumulators 1.46 (2005)
13. Konukoglu, E., Criminisi, A., Pathak, S.: Robust Linear Registration of CT Images using Random Regression Forests. SPIE Medical (2011)
14. Ourselin, S., Roche, A., Prima, S., Ayache, N.: Block matching: A general framework to improve robustness of rigid registration of medical images. In: Medical Image Computing and Computer-Assisted Intervention, Springer (2000)
15. Hoogendoorn, C., Whitmarsh, T., Duchateau, N., Sukno, F.M., De Craene, M., Frangi, A.F.: A groupwise mutual information metric for cost efficient selection of a suitable reference in cardiac computational atlas construction. In: SPIE. (2010)
16. Iglesias, J., Konukoglu, E., Montillo, A., Tu, Z., Criminisi, A.: Combining Generative and Discriminative Models for Semantic Segmentation of CT Scans via Active Learning. In: Information Processing in Medical Imaging (IPMI). (2011)
17. Krishnan, K., Ibanez, L., Turner, W., Avila, R.: Algorithms, architecture, validation of an open source toolkit for segmenting CT lung lesions. In: Proc. MICCAI Workshop on Pulmonary Image Analysis (Sept. 2009). 365–375