Localization of 3D Anatomical Structures Using Random Forests and Discrete Optimization*

René Donner[1,2], Erich Birngruber[1], Helmut Steiner[1], Horst Bischof[2], and Georg Langs[3]

[1] Computational Image Analysis and Radiology Lab, Department of Radiology, Medical University of Vienna, Austria
rene.donner@meduniwien.ac.at
[2] Institute for Computer Graphics and Vision, Graz University of Technology, Austria
[3] Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, USA

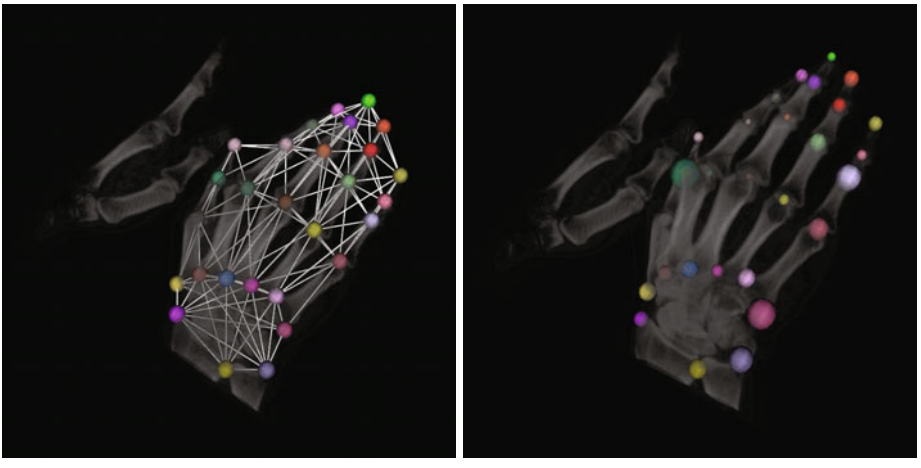
Abstract. In this paper we propose a method for the automatic localization of complex anatomical structures using interest points derived from Random Forests and matching based on discrete optimization. During training landmarks are annotated in a set of example volumes. A sparse elastic model encodes the geometric constraints of the landmarks. A Random Forest classifier learns the local appearance around the landmarks based on Haar-like 3D descriptors. During search we classify all voxels in the query volume. This yields probabilities for each voxel that indicate its correspondence with the landmarks. Mean-shift clustering obtains a subset of 3D interest points at the locations with the highest similarity in a local neighboorhood. We encode these points together with the conformity of the connecting edges to the learnt geometric model in a Markov Random Field. By solving the discrete optimization problem the most probable locations for each model landmark are found in the query volume. On a set of 8 hand CTs we show that this approach is able to consistently localize the model landmarks (finger tips, joints, etc) despite the complex and repetitive structure of the object.

1 Introduction

The reliable, fast segmentation of anatomical structures is a central issue in medical image analysis. It has been tackled by a number of powerful approaches. Among them are Active Shape Models / Active Appearance Models [5], Active Feature Models [12], Graph-Cuts [2], Active Contours [10], or Level-Set approaches [13]. All of these approaches require the correct localization of the

* This work was partly supported by the European Union FP7 Project KHRESMOI (FP7-257528), by the NSF IIS/CRCNS 0904625 grant, the NSF CAREER 0642971 grant, the NIH NCRR NAC P41-RR13218 and the NIH NIBIB NAMIC U54-EB005149 grant. Further supported by the Austrian National Bank grants COBAQUO (12537), BIOBONE (13468) and AORTAMOTION (13497).

B. Menze et al. (Eds.): MICCAI 2010 Workshop MCV, LNCS 6533, pp. 86–95, 2011.



(a) (b)

Fig. 1. (a) Employed data set including the manually annotated landmarks and the connectivities used to build the geometric model. (b) Depicts the localization result with the median residual distance between localization result and ground truth from all leave-one-out runs as radius for each sphere. Note the high accuracy of the localization.

initial model positions or seeds points within or close to the desired object. While most research has focused on the segmentation or analysis of individual regions of interest, the initialization was often performed manually or by application specific and often heuristic approaches. In this paper we propose a generic method that identifies anatomical structures in a global search framework. It learns the local appearance of landmarks, and an elastic shape constraint from annotated training volumes. During search a classifier is used for the generation of candidate points, and the final location is identified by discrete optimization.

The approach proposed in this paper is related to two lines of previous work:

1. Single object localization [18] have demonstrated an efficient voting scheme for localizing anatomical structures - in a sliding window technique each block predicts the location of the object based a priori knowledge learnt from blocks' appearances during a training phase. The result is a very robust estimate of the center of the object and its rotation, but no information is obtained about subparts of the object, although this could be accomplished by using a multi-scale approach narrowing in on the subparts. [6] presented a fast approach to localize individual organs in full-body CTs using 3D feature similar to Haar-wavelets with random offsets. Through these offsets they incorporate long range contextual information which allows to separate similar-looking objects to a certain extent, but the presented results did not comprise any small and repetitive structures. Random forests were used to classify the volumes, which through their parallel nature lend themselves nicely to a GPU-based implementation, resulting in very fast computation.

2. Localizing Complex Structures / Incorporating Subpart Interdependencies [14] parse full body CT data in a hierarchical fashion but are concerned with finding larger scale organs. They first search for one salient slice in each dimension and consequently only localize landmarks within these slices. While greatly speeding up the localization this only works for objects which are rather large in respect to the volume size, as all the objects have to be visible in at least one of the 3 slices. This assumption does not hold true, e. g., in case of the inclined main plain of a hand CT. [17] pose the problem as a dependency graph of individual localization steps, whose order and mutual influence is determined by information theoretic measures. The recently proposed work in [1] is most similar to ours, but randomly selects several candidate interest points within a region of high classification probability, while our mean-shift based approach will yield only one, more stable, interest point. [1] is additionally capable of dealing with missing object parts. [9] uses a GentleBoost based classification approach to find candidate points for heart wall landmarks.

1.1 Sparse MRF Appearance Models

Sparse MRF Appearance Models (SAM) are the most closely related approach to the present work. They match shape and appearance models to query images by solving a Markov Random Field. SAMs are based on interest points and an elastic geometric model of their spatial configuration derived from training images and corresponding landmarks (selected interest points). These selection stems either from manual annotations or from a weakly-supervised learning scheme [8]. The appearance of the anatomical structure is encoded through local descriptors around the interest points and along the connecting edges of a Delaunay triangulation. The model thus encompasses information about the mesh topology, mean and standard deviation of edge lengths, circular statistics of edge directions relative to interest point orientation and the local point and edge descriptors.

To localize the structure in a target image, interest points and descriptors are computed. A Markov Random field is set up with as many nodes as there are model landmarks and with the target image interest point IDs as labels. Node probabilities are derived from point descriptor similarities and edge probabilities from the model's edge features. Solving the MRF yields the most probable match of the model onto the target image.

Despite good results several issues remain. The requirement for a priori chosen interest point detectors presents a delicate design choice. Depending on the structure, different interest point detectors may be needed for the different subparts, as shown in [7]. This greatly increases the number of labels for the MRF inhibiting fast inference and increases memory requirements. One of the prime obstacle for the application of this approach to 3D data is the typically overwhelming number of detected interest points, rendering the straight forward application of the approach unfeasible. Consequently, SAMs have not yet been applied to 3D data sets.

Contribution. The contribution of this paper is a method that learns 3D landmark appearance from training data and computes interest points for each model node of a 3D deformable model. Edge length and orientation probabilities are modeled in a novel, combined representation. The multitude of resulting potential candidate configurations is disambiguated by discrete optimization, yielding a localization of complex and repetitive anatomical structures in a target volume.

The paper is structured as follows: In Sec. 2 we outline how to derive application specific interest points. Sec. 3 details how these target point candidates and the geometric model are combined into a graphical model to perform the matching. In Sec. 4 we present the experimental evaluation of our approach, followed by a conclusion and an outlook in Sec. 5.

2 Domain Specific Interest Points

Instead of using one or several fixed interest point detectors we train a Random Forest classifier on Haar wavelet-like descriptors around the model landmarks. Sampling the resulting classification volumes into point sets and clustering them yields target point candidates specific for each landmark. An overview of the method is depicted in Fig. 2. We derive interest points with support from small, local regions which represent an important feature that will be used as unary node costs during discrete optimization. Additionally, due to the clustering bandwidth it avoids the need for non-maxima suppression and reduces the number of candidate interest points.

2.1 Haar-Like Features

For describing the local appearance around the model landmarks we employ a set of 3D features computed using a basis of filters similar to Haar wavelets. These features [16] can be computed using integral volumes [11] in a highly efficient

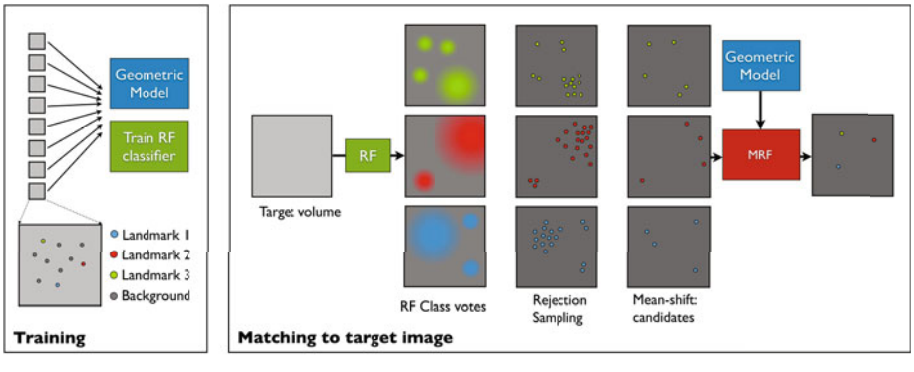


Fig. 2. Outline of the proposed interest point detection and model matching approach

manner. An integral volume $\mathbf{J}$ transforms the information content of the original volume $\mathbf{I}$ such that

$$\mathbf{J}(x,y,z) = \sum_{i=1\ldots x} \sum_{j=1\ldots y} \sum_{k=1\ldots z} \mathbf{I}(i,j,k) \tag{1}$$

This allows to compute the sum s within a cuboid given by the coordinates $(x_1, y_1, z_1, x_2, y_2, z_2)$ by

$$\begin{aligned} s = \mathbf{J}(x_2,y_2,z_2) - \mathbf{J}(x_2,y_1,z_2) - \mathbf{J}(x_1,y_2,z_2) \\ +\mathbf{J}(x_1,y_1,z_1) - \mathbf{J}(x_2,y_2,z_1) + \mathbf{J}(x_2,y_1,z_1) \\ +\mathbf{J}(x_1,y_2,z_1) - \mathbf{J}(x_1,y_1,z_1). \end{aligned} \tag{2}$$

Computing a Haar-like feature then consists of forming 2 such sums with opposing signs. For each of the 3 dimensions, both gradient and ridge features where used, which, together with an average feature derived over the same area as the Haar-features formed the 7 dimensional feature vector for a given wavelet width. 3 different widths wheres used, namely 8, 16 and 32 voxels, yielding a description vector $\mathbf{f}_j$ of dimension 21 for each voxel j in the volume.

2.2 Random Forest Based Appearance Learning and Search

Random Forests [15] are ensemble classifiers. They learn a set of decision trees by randomly sampling from training feature vectors and corresponding labels. During search the decision trees vote for a class label for each query feature vector. Given L landmarks with known positions $\mathbf{x}_l^i$ in all T training volumes ($i = 1, 2, \ldots, T$, $l = 1, 2, \ldots, L$). We assign each training landmark a class label l. In all training volumes we extract local descriptors $\mathbf{f}_j^l$ for all voxels within a 3-voxel radius around landmarks and assign them the corresponding landmark label l. Additionally, we compute descriptors for random background voxels $\mathbf{f}_j^b$. This yields a training set of descriptors, and corresponding labels (L landmark classes, and one background class). A random forest is learnt on the entire set of training examples.

During search on a new target volume, descriptors $\mathbf{f}_j$ are computed for every voxel, and all voxels are classified by the Random Forest. The Random Forests votes are normalized for each class, and yield L volumes $\mathbf{C}^l$ containing the classification probabilities for class l in each voxel. The next step is the generation of landmark candidates from those volumes.

2.3 Mean-Shift Based Interest Point Generation

Mean-shift [3,4] is a method for the density estimation and cluster analysis of a sparse set of points in a, potentially high-dimensional, feature space. Given a d-dimensional dataset $\mathbf{D}$ mean-shift iteratively moves each data point $\mathbf{d}_i$ towards the mean of the data points within a certain distance or bandwidth b (according to a chosen kernel) around $\mathbf{d}_i$. The process is repeated until equilibrium is reached, i. e. when no data point is shifted anymore.

To find the most probable candidates for each model landmark l in the test volume we search for regions within the classification probability volume $\mathbf{C}^l$ with high local support, i. e. regions where their Gaussian weighted integral yields high values. We estimate this by employing mean-shift with a Gaussian kernel with the empirically derived bandwidth $\sigma = 8$ voxels on a rejection sampled version of $\mathbf{C}^l$, i. e. a point set containing the voxels where $\mathbf{C}_l(x, y, z) > r_j$ with r_j being randomly chosen from a uniform distribution between 0 and 1 for each voxel. Each point furthermore carries its probability as weight, leading to means during the mean-shift weighted by $\mathbf{C}_l(x, y, z)$ and the Gaussian kernel.

The result of this process are sets of cluster centers, i. e. interest points, or candidates, $\mathbf{p}_i^l$ for each model landmark l together with estimates of their local support s_i^l consisting of the number of points converged to the cluster center.

3 3D Geometric Model Matching Using Discrete Optimization

In the previous section we have described how to obtain landmark candidates in a search volume. The candidate generation process is based purely on the local appearance learnt from the training cases. In this section, we use the spatial configuration of the landmarks, to constrain the set of landmark candidates, and ultimately find a highly probable landmark assignment in the search volume. The assignment is a tradeoff between local appearance at the landmark position, and the plausibility of the spatial configuration.

We derive an elastic geometric model from the training data, together with confidences about point and edge similarity. The local similarities of this learnt model to the interest points found in the query image are encoded in a Markov Random Field (MRF). The solution of the graphical model yields the match of the model to the query image, i. e. the localization of the anatomical structure in the target volume.

3.1 Modeling Edge Length and Orientation

The landmarks are connected by a set of edges. In contrast to [7] in our experiments we do not employ a Delaunay triangulation but used a manually specified connectivity reflecting the anatomical structure as shown in Fig. 1a, which is almost identical to fully connecting each landmark to its anatomically nearest neighbours. To establish an elastic model from the annotated training landmarks [7] uses mean length and standard deviation of the edges together with circular statistics for relative edge orientations. In [7] both features are used individually to compute confidence values in the range 0 to 1 which are combined to yield similarity confidences between training and target edges.

We propose a more principled approach that combines lengths and orientations and yields proper probabilities through density estimation, as depicted in Fig. 3a **(1-4)**.

For each model edge e from model landmark a to b in the T training volumes two sets P_a, P_b containing the $2T$ endpoints $\mathbf{p}_a^t, \mathbf{p}_b^t$ are known **(1)**. Centering all edges yields the normalized endpoints $\hat{\mathbf{p}}_a^t = \mathbf{p}_a^t - \overline{\langle \mathbf{p}_a^t, \mathbf{p}_b^t \rangle}$ **(2)**.

The mean μ_a and the covariance matrix $\mathbf{\Sigma}_a$ of $\hat{\mathbf{p}}_a^t$, $t = 1, \ldots, T$ now represent the combined length and orientation distribution of edge e: The vector from the origin to μ_a equals the expectation of the orientation $\pm\pi$ and half the length of e **(3)**.

(4) To compute the probability of an edge f between two point candidates $\mathbf{p}_a^f, \mathbf{p}_b^f$ in the target image to be an instance of e we first center $\mathbf{p}_a^f, \mathbf{p}_b^f$: $\hat{\mathbf{p}}_a^f, \hat{\mathbf{p}}_b^f = \mathbf{p}_a^f, \mathbf{p}_b^f - \overline{\langle \mathbf{p}_a^f, \mathbf{p}_b^f \rangle}$. Computing non-normalized Gaussian values d_a, d_b

$$d_a = exp(-\frac{1}{2}(\hat{\mathbf{p}}_a^f - \mu_a)^\top \mathbf{\Sigma_a}^{-1}(\hat{\mathbf{p}}_a^f - \mu_a)) \tag{3}$$

$$d_b = exp(-\frac{1}{2}(\hat{\mathbf{p}}_b^f - \mu_b)^\top \mathbf{\Sigma_b}^{-1}(\hat{\mathbf{p}}_b^f - \mu_b)) \tag{4}$$

and taking the larger of the two values $d_{max}(f, e) = max(d_a, d_b)$ results in the confidence of edge f being from the distribution of edge model e.

3.2 Formulating the MRF

The objective function for matching a model to an example image is

$$Conf(\mathcal{S}) = \sum_{l=1\ldots L} \mathcal{L}(l, \mathcal{S}(l)) + \sum_{e=1\ldots E} \mathcal{E}(e, \mathcal{S}(e)). \tag{5}$$

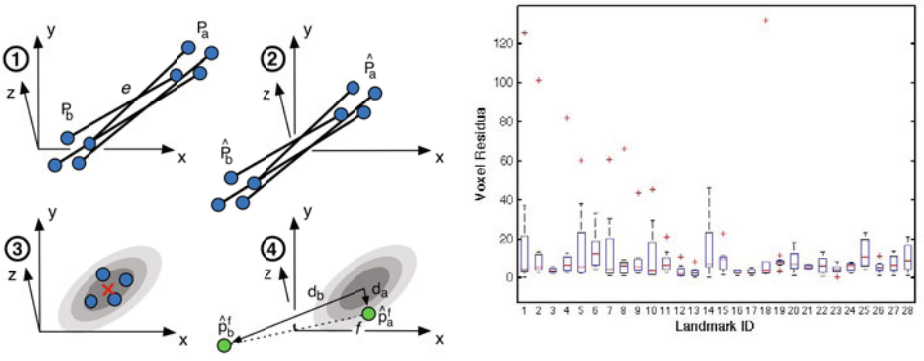
It consists of *unary* terms $\mathcal{L}$ at the nodes of the graphical model describing the L model landmarks to point candidate similarities. *Binary* terms $\mathcal{E}$ capture the similarities of the E model edges to the target edges. Each node has as many labels as that model node has point candidates $\mathbf{p}_i^l$ in the target volume. The confidence for $\mathcal{L}(l, i)$ equals the normalized support $s_i^l / max(s^l)$ and $\mathcal{E}(e, \mathcal{S}(e))$ equals $d_{max}(\mathcal{S}(e), e)$ for the edge between the candidate points $\mathcal{S}(e)$ in relation to model edge e. The MRF's solution, the so called *labeling*

$$\mathcal{S}^* = \underset{\mathcal{S}}{\operatorname{argmax}} Conf(\mathcal{S}) \tag{6}$$

assigns each model node l to one point candidate $\mathbf{p}_i^l$ in the target image, matching the model to the target volume.

4 Experiments

We evaluated the proposed approach on 8 Hand CTs with a resolution of 256 × 384 × 330 voxels. 28 landmarks were manually annotated as shown in Fig. 1a. The dataset is challenging due to repetitive nature of the structures. The varying types of structures to be found impede the use of a single interest point detector.



(a) (b)

Fig. 3. (a) **(1-3)** Combined Gaussian model estimation of edge length and orientation from the training instances of edge e. **(4)** Confidence computation that edge f is an instance of model edge e. (b) Residual distances from all leave-one-out runs for each landmark.

In contrast, the Haar-like features together with the Random Forest classifier are well suited to picking up biological structures at different scales, due their detection of salient ridge-like and edge-like features in the volume, like the finger tips and the small joints. Around 50 candidate interest points were detected on average for each landmark. The experiments were run in a leave-one-out cross validation framework using the manual annotations of 7 training volumes to construct the geometric model. The Random Forrest classifier was trained using 200 trees on the 33250 descriptors extracted around the landmarks and from the backgrounds of the training volumes. The MRF was approximately inferred using a simple random walk approach – a detailed comparison of inference methods for this application is subject of ongoing work. After matching, the residual voxel distances between the selected interest points and the corresponding ground truth landmarks were recorded.

Results. To visualize the result quality Fig. 1b shows one of the hands. At each landmark position the radius of a sphere corresponds to the median residual from all leave-one-out runs. Fig. 3b shows the corresponding boxplot.

The mean / median / std residuals measured 10.13 / 5.59 / 16.99 voxels, and 12.87 / 7.01 / 21.57 mm, respectively. They demonstrate the model matching and localization capability of the proposed approach. While the median shows that the vast majority of points are localized with high accuracy (the average finger width being around 32 voxels) 15 outliers considerably deteriorate the mean and standard deviation values. In one instance the volume is cut-off too close to the carpus, leading the solver to choose an alternative point unrelated to the anatomical structure. In 2 instances fingertips crossed over to the adjacent finger. We suspect the small size of the data set to yield a too restrictive geometric model and the MRF solution not representing the global optimum in certain cases.

The runtime of the proposed approach for a single localization, implemented in Matlab except for the C-based RF, amounted to about 5 minutes.

5 Conclusion and Outlook

We present an approach for localizing complex, potentially repetitive anatomical structures in 3D volumes. Based on Random Forests and Haar-like features, the method detects landmark candidate points in a search volume. The anatomical structure is detected by solving a graphical model defined on the landmark candidates. The method does not rely on predefined interest point detectors but derives the most probable candidate locations for each model landmark automatically. This alleviates the prohibitively high number of candidates encountered when using a fixed descriptor together with a simple distance measure. In localization experiments, the method exhibits very low localization errors, but in a few outlier cases single landmark position estimates were attracted to wrong positions.

Future research will focus on increasing the robustness of the approach, especially with regard to matching missing subparts of the objects. While an initial manual indication of the object of interest is deemed necessary, a learning approach should be able to derive which object parts are representative, eliminating the need for detailed annotation. Evaluation on larger data sets and alternative MRF solving strategies will be performed.

References

1. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A study of parts-based object class detection using complete graphs. Int. J. Comput. Vis. 87(1-2), 93–117 (2010)
2. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proc. ICCV, pp. 105–112 (2001)
3. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE TPAMI 17(8), 790–799 (1995)
4. Comaniciu, D., Meer, P., Member, S.: Mean shift: a robust approach toward feature space analysis. IEEE TPAMI 24, 603–619 (2002)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. PAMI 23(6), 681–685 (2001)
6. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in ct volumes. In: Proc. of MICCAI Workshop on Probabilistic Models for Medical Image Analysis (MICCAI-PMMIA) (2009)
7. Donner, R., Mičušik, B., Langs, G., Bischof, H.: Generalized Sparse MRF Appearance Models (2010)
8. Donner, R., Wildenauer, H., Bischof, H., Langs, G.: Weakly supervised group-wise model learning based on discrete optimization. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 860–868. Springer, Heidelberg (2009)
9. Essafi, S., Langs, G., Paragios, N.: Left ventricle segmentation using diffusion wavelets and boosting. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 919–926. Springer, Heidelberg (2009)

10. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal on Computer Vision 1, 321–331 (1988)
11. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: Proc. ICCV (2005)
12. Langs, G., Peloschek, P., Donner, R., Reiter, M., Bischof, H.: Active Feature Models. In: Proc. ICPR, pp. 417–420 (2006)
13. Paragios, N., Deriche, R.: Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects. IEEE PAMI 22(3) (2000)
14. Seifert, S., Barbu, A., Zhou, S.K., Liu, D., Feulner, J., Huber, M., Suehling, M., Cavallaro, A., Comaniciu, D.: Hierarchical parsing and semantic navigation of full body CT data (2009)
15. Statistics, L.B., Breiman, L.: Random forests. In: Machine Learning, pp. 5–32 (2001)
16. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features, pp. 511–518 (2001)
17. Zhan, Y., Zhou, X.S., Peng, Z., Krishnan, A.: Active scheduling of organ detection and segmentation in whole-body medical images. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 313–321. Springer, Heidelberg (2008)
18. Zheng, Y., Georgescu, B., Ling, H., Zhou, S., Scheuering, M., Comaniciu, D.: Constrained marginal space learning for efficient 3d anatomical structure detection in medical images, pp. 194–201 (2009)