This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans

Florin C. Ghesu, Bogdan Georgescu, *Member, IEEE,* Yefeng Zheng, *Senior Member, IEEE,* Sasa Grbic, Andreas Maier, Joachim Hornegger, and Dorin Comaniciu, *Fellow, IEEE*

Abstract—Robust and fast detection of anatomical structures is a prerequisite for both diagnostic and interventional medical image analysis. Current solutions for anatomy detection are typically based on machine learning techniques that exploit large annotated image databases in order to learn the appearance of the captured anatomy. These solutions are subject to several limitations, including the use of suboptimal feature engineering techniques and most importantly the use of computationally suboptimal search-schemes for anatomy detection. To address these issues, we propose a method that follows a new paradigm by reformulating the detection problem as a behavior learning task for an artificial agent. We couple the modeling of the anatomy appearance and the object search in a unified behavioral framework, using the capabilities of deep reinforcement learning and multi-scale image analysis. In other words, an artificial agent is trained not only to distinguish the target anatomical object from the rest of the body but also how to find the object by learning and following an optimal navigation path to the target object in the imaged volumetric space. We evaluate our approach on 1487 3D-CT volumes from 532 patients, totaling over 500,000 image slices and show that we significantly outperform state-of-the-art solutions on detecting several anatomical structures with no failed cases from a clinical acceptance perspective, while also improving the detection accuracy by 20-30%. Most importantly, we improve the detection-speed of the reference methods by 2-3 orders of magnitude, achieving unmatched real-time performance on large 3D-CT scans.

Index Terms—Deep learning, deep reinforcement learning, medical image analysis, multi-scale, scale-space modeling, three-dimensional (3D) object detection, real-time detection, intelligent localization.

1 Introduction

The detection of anatomical landmarks represents a prerequisite for medical image analysis. Many applications for clinical support require the precise, automatic detection of anatomical structures to initialize and constrain mathematical models for volumetric organ segmentation [1], [2], [3], image-to-image registration [4], [5], structure tracking [6], [7], advanced bio-physical modeling and mechanical simulations [8]. As such, enabling accurate and efficient anatomical landmark detection can assist the physician with automated measurements for a more effective and streamlined image reading. We focus on 3D-CT, an imaging technology widely used both interventionally and for diagnosis, e.g. for disease screening, detection of brain hemorrhages, bone fractures, etc. [9].

The current solutions for anatomical landmark detection are typically based on machine learning concepts, which effectively exploit large annotated medical image databases [1], [10], [11], [12]. For this purpose, the detection task is typically decoupled into two independent and sequential stages: the learning of an appearance model and the object search. In the first stage, an appearance model is designed to capture the underlying image information and use it as evidence to identify the anatomical landmark. To enable this, traditional machine learning systems rely on precise feature engineering strategies, using human ingenuity to understand and model the image information [3], [10], [12], [13], [14]. Instead, deep learning solutions propose to learn the image features to better disentangle explanatory attributes of the observed data [15]. This enables the learning models to better capture the complex anatomical variation, ensuring an increased performance and better generalization to unseen data [1], [16], [17], [18]. However, in the second stage of the task, concerning the object search, the aforementioned solutions rely on suboptimal search strategies, e.g. exhaustive scanning [1], [3], [10], [18], one-shot displacement estimation [12], [13], [14] or end-to-end image mapping techniques [16], [17]. These strategies lead in many cases to false-positive detection results and unreasonably high computation times.

In this work, we propose a method which follows a different paradigm, based on the reformulation of the detection task as a behavioral problem for an artificial agent. Using deep reinforcement learning [19] and scale-space theory [20], we learn optimal search strategies for finding anatomical structures, based on the image information at multiple scales. A search strategy generates multi-scale navigation trajectories, which evolve through the voxel-grid of the image at different spatial resolutions, i.e. the scale-space representation of the image [20], and converge to the sought landmark location (see Figure 1).

This formulation exploits in a systematic and natural way the different levels of abstractness encoded in the scale-space representation of an image. The search starts at the

F. Ghesu is with Medical Imaging Technologies, Siemens Healthineers, Princeton, NJ 08540 USA and with the Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany (e-mail: florin.c.ghesu@fau.de)
B. Georgescu, S. Grbic, Y. Zheng and D. Comaniciu are with Medical Imaging Technologies, Siemens Healthineers, Princeton, NJ 08540 USA.
A. Maier and J. Hornegger are with the Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

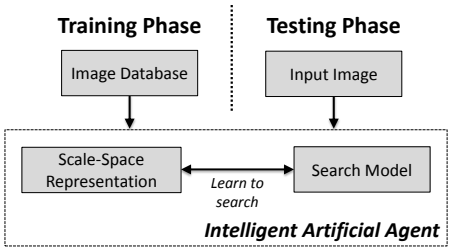

Fig. 1. Schematic overview of the proposed machine learning-based paradigm for anatomical landmark detection. The detection problem is reformulated to learning a navigation strategy, which exploits the scale-space representation of a given image. In other words, an artificial agent learns how to search for an anatomical structure.

coarsest scale level with a global context and continues across scales, capturing increased levels of details when transitioning to finer scales [20]. Such details are used as additional evidence to guide the search. We propose to use independent search models on each level of the scale-space in order to adapt the search to the most discriminative image features, visible on that level. These search models are based on the mechanism of deep reinforcement learning, which we reformulated for the context of anatomical landmark detection in previous work [21]. In this paper, we revisit our formulation [21] and present the following additional contributions:

- We extend the original solution using scale-space theory to exploit multi-scale image representations for a robust and efficient multi-scale object search in 3D-CT scans of arbitrary size in real time.
- We present a simple modification at the core of the DRL system which enables faster and better training (see second paragraph in section 3.2.2).
- We evaluate the performance of our method on detecting 8 landmarks from different types of anatomy, i.e. bone, non-rigid organs, vessel bifurcations, using a set of 1487 3D-CT scans from 532 patients.
- We demonstrate the performance of our method on an additional dataset of 506 3D-MR scout scans acquired from 506 patients. In contrast to the 3D-CT scans, the magnetic resonance images also show a constrained angular variation in the axial plane.
- We provide a detailed comparison with 5 solutions for landmark detection based on the probabilistic boosting tree [3], extremely randomized trees [10] and on deep learning systems [1], [18], [22] – showing that our method outperforms these solutions in terms of number of failures and accuracy.
- We provide a runtime analysis, demonstrating that our method is 20 to 150 times faster compared to these methods, reaching real-time detection speed.
- We include an empirical convergence analysis of our algorithm and discuss the detection of outliers and how to handle cases of missing objects.
- We include a detailed analysis from computer vision perspective, discussing related methods and potential applications of our ideas in this field.

The remaining paper is organized as follows. In section 2 we review previous work on object localization and highlight the limitations of existing technology. In section 3 we present our solution for object detection using multi-scale deep reinforcement learning. Experiments are presented in section 4, and section 5 covers the conclusion of the paper.

2 Background and Motivation

A variety of applications for medical image analysis can benefit from the automatic and accurate detection of anatomical landmarks. In the following, we present several existing solutions for anatomical landmark detection and highlight the main challenges.

2.1 Object Localization in 3D: Challenges

Scanning-based Systems represent the main category of detection solutions, specifically in the context of 3D data. In this case, object detection is formulated as a patch-wise classification problem. A set of discrete hypotheses $\mathbf{H}$, in the form of local volumetric boxes of image intensities, is extracted from any image $\mathbf{I}$ from the training set: $\forall h \in \mathbf{H}, h \sim \mathbf{I}$. This set of hypotheses can be partitioned in a subset of positive hypotheses $\mathbf{H}_+$ which are centered at the landmark location, and negative hypotheses $\mathbf{H}_-$ from the rest of the image: $\mathbf{H} = \mathbf{H}_+ \cup \mathbf{H}_-$. A classifier is trained on this set of hypotheses – essentially learning to distinguish the appearance of the sought object from the rest of the sampled anatomy. One can use traditional machine learning models, e.g. extremely randomized trees [10] and probabilistic boosting trees [23]; or deep learning, e.g. deep convolutional neural networks [24] and sparse adaptive deep neural networks [1]. For the final result, hypotheses aggregation using Hough regression [10] or averaging [1], [11], [24] can be applied. However, in the case of training with a large number of high-resolution whole body 3D-CT scans, the number of hypotheses used for training can surpass the memory capabilities of current GPU-based systems. In addition, at testing time the classifier is used to scan the entire space of hypotheses of a given image – a set which grows exponentially with the dimensionality of the image. In the case of a $200 \times 200 \times 200$ volume, with a hypothesis size of $15 \times 15 \times 15$, the sampling complexity reaches the order of billions. Moreover, this approach can also be sensitive to false-positive classifier responses and is thus often combined with different techniques, such as cascade filtering [1], [25], filter-decomposition [18], simultaneous network propagation [22] or active scheduling [26], [27], to enable effective training and detection in 3D data.

End-to-End Systems, also called image-to-image systems, are inspired by the fully convolutional network (FCN) architectures and learn the mapping between original image and segmentation multi-masks [28] or image codes highlighting the locations of anatomical landmarks [16]. Recently, Dai *et al.* [29] have proposed a region-based FCN approach for efficient object detection. While these solutions enable pixel/voxel-wise classification and support simultaneous detection of multiple landmarks, the training is very complex in terms of both memory management and processing time. For example, a forward-propagation of a single CT scan containing $200 \times 200 \times 200$ voxels through a typical FCN can surpass a memory footprint of 3-4 GB – severely

limiting the size of the sampled batch of images during training. Despite memory issues, these challenges are being addressed within this rapidly advancing line of research.

Regression-based Systems exploit the anatomical context around the landmark in order to learn relative displacement vectors pointing at the landmark location. One can use random regression forests [11], [12], [13], [14], [30], [31], random-ferns [32], deep convolutional neural networks [33] or modern spatial-transformer neural networks [34] to learn the mapping. While such solutions significantly improve the efficiency of scanning-based systems, they are typically difficult to train and not robust to variations in the range of the scan.

Atlas-based Systems Atlas-based registration [35], as well as multi-atlas-based registration methods [2], [36], can also be applied to solve localization tasks. However, the application to large 3D-CT scans poses significant computational challenges with decreased model scalability. Potesil *et al.* [37] address this limitation using graphical models with dense parts-based matching.

2.2 Object Localization in 2D: Related Work

Similar concepts are used to approach object localization also for 2D data (e.g. color images or depth maps). While the lower dimensionality represents a clear advantage in terms of processing complexity, the lack of structure and alignment in comparison to 3D medical data changes the solution perspective. Modeling the intrinsic configuration of object points that need to be detected is thus required. In this context, shape model matching with random forest voting [38] has been proposed. Deformable part models in combination with latent SVMs for partially labeled data [39] have achieved competitive results on the PASCAL challenge. Recently, region-based models using deep convolutional neural networks have achieved state-of-the-art results on the PASCAL VOC dataset [40]. Ren *et al.* [41] further optimized the network architecture by selecting to share the features between the region proposal and the detection network, thereby reducing the computation time. Additional runtime improvements have been achieved by using FCNs [29]. Deep learning solutions have also achieved state-of-the-art results for the task of human pose estimation. Two recent examples are the Convolutional Pose Machines [42] for sequential prediction and the convolutional Stacked Hourglass Networks [43].

3 METHOD

To address the challenges of interpreting 3D data in large CT scans, we propose to reformulate anatomical object detection as a behavioral problem for an intelligent artificial agent that can teach itself how to search [21], [44]. With the keyword *intelligent* we describe the capability of our system to explore and learn the process of finding an object, as opposed to following predefined exhaustive search schemes. To achieve this, we combine concepts of cognitive modeling based on reinforcement learning with scale-space analysis and deep learning.

3.1 Deep Learning: An Overview

Established as a key technological innovation in the field of machine learning, with significant improvements in results for image parsing tasks [1], [28], [45], [46], deep learning systems have replaced the traditional feature handcrafting step with hierarchical, multi-layer models for automatic feature learning [15], [47]. We use the deep convolutional neural network (CNN) as an image feature extractor [48] and universal non-linear function approximator [15]. The network is parametrized by $\theta = [W, b]$, where W denotes the inter-neural connection weights organized as (multi-channel) filter kernels, and b defines the set of neuron bias values. The architecture is inspired by the feed-forward type of information processing observable in the early visual cortex of animals [49]. Convolutional layers exploit local spatial correlations of image voxels to learn translation-invariant convolutional kernels, which capture discriminative image features. Consider a multi-channel signal representation $\mathbf{M}_k$ in layer k, i.e. a channel-wise concatenation of signal representations $\mathbf{M}_{k,c}$ with $c \in \mathbb{N}$. One can generate a signal representation in layer $k+1$ as: $\mathbf{M}_{k+1,l} = \phi\left(\mathbf{M}_k * \mathbf{w}_{k,l} + b_{k,l}\right)$, where $\mathbf{w}_{k,l} \in W$ represents a convolutional kernel with the same number of channels as $\mathbf{M}_k$, the value $b_{k,l} \in b$ represents the bias, l denotes the channel index, and $*$ denotes a convolution operation. The function ϕ represents the non-linear activation function, which is applied point-wise. We use rectified linear unit (ReLU) activations [45]. Depending on the problem, the final network layer is typically fully-connected. In a supervised regression setup, given training data $\mathcal{D} = [(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_N, y_N)]$, i.e. N independent pairs of volumetric image observations with value assignments, one can define the network response function as $\mathcal{R}(\,\cdot\,;\theta)$, and use Maximum Likelihood Estimation to estimate the optimal network parameters ($\mathcal{L}$ denotes the likelihood):

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}\left(\theta; \mathcal{D}\right) = \arg\min_{\theta} \sum_{i=1}^{N} \left(\mathcal{R}(\mathbf{X}_i; \theta) - y_i\right)^2 . \quad (1)$$

This optimization problem is typically solved with stochastic gradient descent (SGD) combined with the backpropagation algorithm to compute the network gradients [50].

3.2 Deep Reinforcement Learning for Intelligent Search

Our approach for intelligent anatomical landmark detection combines the representational power of modern CNN architectures and cognitive modeling through reinforcement learning (RL). Initially, the combination of these two elements was introduced in the literature under the name of *deep reinforcement learning* (DRL), a technique used to train artificial agents to master different ATARI games [19].

3.2.1 Anatomy Detection: A Behavior Learning Problem

We propose to reformulate anatomy detection as a cognitive learning task for an artificial agent. Given a volumetric image $\mathbf{I} : \mathbb{Z}^3 \rightarrow \mathbb{R}$ and the location of an anatomical structure of interest $\vec{p}_{GT} \in \mathbb{R}^3$ within $\mathbf{I}$, the task is to learn a navigation strategy to $\vec{p}_{GT}$ in image space, i.e. the voxel grid of the scan. In other words, we seek voxel-based navigation trajectories from any arbitrary starting point $\vec{p}_0$ to $\vec{p}_k$ within image $\mathbf{I}$, with the property that $\|\vec{p}_k - \vec{p}_{GT}\|$ is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

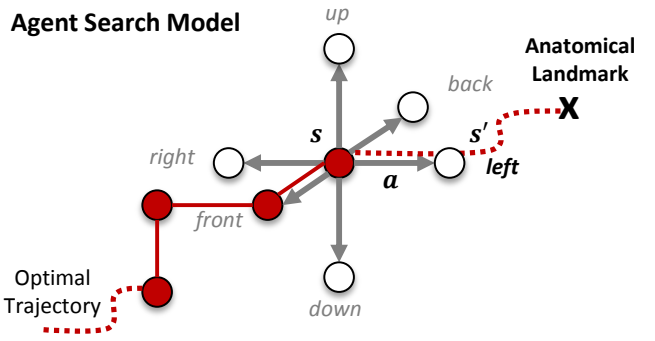

Fig. 2. Schematic visualization of the decision-based search model in state s. Six possible actions allow for voxel-wise movement in the volumetric image space. In this synthetic example the optimal decision with respect to the cumulative future reward is to go *left*, to state s'. The dashed red line represents the optimal search-trajectory to the anatomical landmark, while the circles represent neighboring voxels.

minimal [21]. Reinforcement learning allows us to model this problem using a *Markov Decision Process* (MDP) [51] $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where:

- $\mathcal{S}$ represents a finite set of states, $s_t \in \mathcal{S}$ being the state of the agent at time t. To encode the location of the agent in the imaged volumetric space at time t, we define $s_t = \mathbf{I}(\vec{p}_t)$, which denotes an axis-aligned box of image intensities extracted from $\mathbf{I}$ and centered at the voxel-position $\vec{p}_t$ in image space.
- $\mathcal{A}$ represents a finite set of actions allowing the agent to interact with the environment defined by $\mathbf{I}$, where $a_t \in \mathcal{A}$ is the action the agent performs at time t. We propose a discrete voxel-wise navigation model allowing the agent to move from any voxel position $\vec{p}_t$ to an adjacent voxel position $\vec{p}_{t+1}$ in image space (see Figure 2 for details).
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0; 1]$ is a stochastic transition function, where $\mathcal{T}^{s'}_{s,a}$ describes the probability of arriving in state s', after performing action a in state s.
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is a scalar reward function which drives the behavior of the agent, where $\mathcal{R}^{s'}_{s,a} \in \mathbb{R}$ denotes the expected reward after a state transition. For a state transition $s \rightarrow s'$ at time t from $\vec{p}_t \rightarrow \vec{p}_{t+1}$, we define $\mathcal{R}^{s'}_{s,a} = \|\vec{p}_t - \vec{p}_{GT}\|^2_2 - \|\vec{p}_{t+1} - \vec{p}_{GT}\|^2_2$. Intuitively this represents a distance-based feedback, which is positive if the agent gets closer to the target structure and negative otherwise.
- γ is the discount factor controlling the importance of future versus immediate rewards.

Considering the proposed reward-scheme and an arbitrary trajectory $T = [\vec{p}_0, \vec{p}_1, \ldots, \vec{p}_k]$ in image space, at any time $\hat{t} \in \{0, \ldots, k\}$ the associated cumulative future discounted reward is defined as: $R_{\hat{t}} = \sum_{t=\hat{t}}^{k} \gamma^{t-\hat{t}} r_t$, where the immediate reward at time t is denoted by r_t. In RL theory this is also considered a finite-horizon learning episode of length k [51]. The target is to find optimal trajectories that maximize the associated cumulative future reward (see Figure 2). To achieve this, we define the optimal action-value function $Q^*(\cdot, \cdot)$, which encodes the maximum expected future discounted reward when starting in state s, performing action a, and acting optimally thereafter:

$$Q^*(s, a) = \max_{\pi} \mathbb{E}\left[R_t | s_t = s, a_t = a, \pi\right], \tag{2}$$

where π is an action policy, in other words a probability distribution over actions in any given state. The optimal action-value function gives us the optimal action policy, defining the optimal behavior of the agent in any state:

$$\forall s \in \mathcal{S} : \pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a). \tag{3}$$

One important relation satisfied by the optimal action-value function Q^* is the Bellman optimality equation [52], which represents a recursive formulation of Equation 2:

$$\begin{aligned} Q^*(s, a) &= \sum_{s'} \mathcal{T}^{s'}_{s,a} \left(\mathcal{R}^{s'}_{s,a} + \gamma \max_{a'} Q^*(s', a') \right) \\ &= \mathbb{E}_{s'} \left(r + \gamma \max_{a'} Q^*(s', a') \right), \end{aligned} \tag{4}$$

where s' defines a possible state visited after s, a' the corresponding action and $r = R^{s'}_{s,a}$ represents a compact notation for the current, immediate reward. Viewed as an operator τ, the Bellman equation defines a contraction mapping. In previous work [53] the following property was proven: $\forall Q, \lim_{n \rightarrow \infty} \tau^{(n)}(Q) = Q^*$, which gave rise to the model-based policy iteration algorithm [51].

This standard approach is however not feasible in our case, where the state space is defined by high-dimensional image data. As such, we propose to use a model-free approach based on a non-linear parametrization of Q^* with a deep convolutional neural network. In literature, this is called a deep Q-network (DQN) [19], [53] and is used as a non-linear approximator for the optimal action-value function: $Q(s, a; \theta) \approx Q^*(s, a)$, where $\theta = [W, b]$ are the parameters of the network. Similar to the temporal difference Q-Learning algorithm [53], a deep Q-network can be trained in a RL setup using an iterative approach to minimize the mean squared error based on the Bellman optimality equation (see Equation 4). At any training-iteration i, we can approximate the optimal expected target value for the action-value function using a set of reference parameters $\bar{\theta}^{(i)} := \theta^{(i')}$, based on a previous training iteration $i' < i$:

$$y = r + \gamma \max_{a'} Q(s', a'; \bar{\theta}^{(i)}). \tag{5}$$

As such, we obtain a sequence of well-defined optimization problems, driving the evolution of the network parameters. The error function at each training step i is defined as:

$$\hat{\theta}^{(i)} = \arg\min_{\theta^{(i)}} \mathbb{E}_{s,a,r,s'} \left[\left(y - Q(s, a; \theta^{(i)}) \right)^2 \right]. \tag{6}$$

This is a supervised setup for DL ,which can be approached as described in the beginning of this section. In our framework, we periodically apply stochastic gradient descent steps, approximating the gradient using random sampling:

$$\nabla_{\theta^{(i)}} = \mathbb{E}_{s,a,r,s'} \left[\left(y - Q(s, a; \theta^{(i)}) \right) \nabla_{\theta^{(i)}} Q(s, a; \theta^{(i)}) \right]. \tag{7}$$

Figure 3 shows an example of a learned trajectory defined by the optimal action-value function Q^*. This highlights the difference between our approach and the concept of exhaustive hypotheses scanning.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

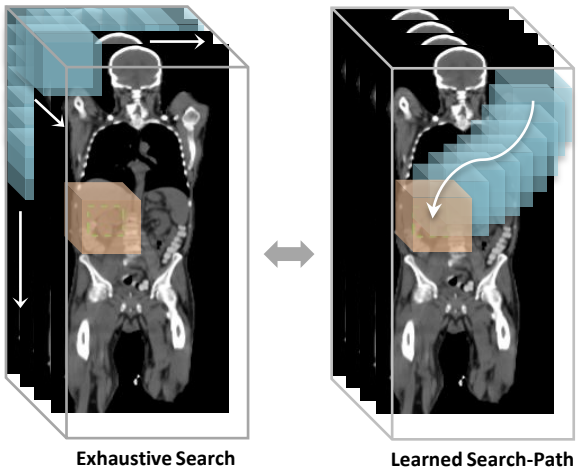

Fig. 3. Visualization of the differences between exhaustive scanning and our proposed method, which learns the search-process. Solutions based on exhaustive scanning, e.g. [1], [10], [18], typically test all hypotheses extracted from the volumetric input and then apply a form of aggregation/clustering of high-probability candidates to obtain a final result. In contrast, our approach learns not only the appearance of the anatomy but also the strategy of how to find a target anatomical landmark. The search starts at a given point $\vec{p}_0$ and defines a 3D trajectory in image space, visualized as a white curve, converging to the sought anatomical landmark location (here the right kidney).

3.2.2 Learning to Search vs. Exhaustive Search

Learning the action-value function Q^* enables the agent to effectively search for objects in the image, as opposed to scanning the volumetric space exhaustively (see Figure 3). This learning process is based on an adequate exploration of the environment, which we ensure through an off-policy ϵ-greedy approach [19]. The variable $\epsilon \in [0, 1]$ controls the randomness in the exploration. This means that during training, actions are selected either uniformly at random with probability ϵ, or deterministically using the current policy with probability $1 - \epsilon$. In our experiments, we linearly anneal ϵ from 1.0 to 0.05. Another important strategy to ensure the training stability is the decorrelation of the training samples using the concept of *experience replay* [54]. During training, the agent maintains an active memory of episodic trajectories $\mathbf{M} = [T_1, T_2, \ldots]$, which is constantly expanded and uniformly sampled to estimate the learning gradient (see Equation 7).

To further accelerate the training, we propose to use an adaptive episode length. Through empirical analysis we observed that by gradually reducing the episode length during training using linear decay, we improve the space exploration by sampling increasing numbers of trajectories that are stored in the active memory. This simple modification not only increased the robustness of the trained policy, but also reduced the training time on average by around 30%. This is due to the fact that as the policy improves during training, sampled trajectories also converge faster to the ground-truth, eliminating the need for long-horizon episodes which can bias the stochastic sampling for the gradient estimation. In our case the initial length of the episode is 1000 and is gradually reduced to 50 steps.

Given this system definition, one can observe a major limitation related to the modeling of the state space $\mathcal{S}$, more specifically to the size of the acquired state representation $s \in \mathcal{S}$, in the form of a box of image-intensities. Acquiring a small-volume box, containing only local information, improves the sampling efficiency, but also increases the complexity of the learning task by disregarding global context. Such context is required to learn an effective navigation policy and avoid local optima. On the contrary, extracting a very large box to represent the state poses significant computational challenges in the 3D space. This trade-off indicates the inability of our preliminary approach to properly exploit the image information at different scales.

3.3 A Scale-space Theoretical Perspective

We propose to address this limitation by using scale-space theory [20]. Given an arbitrary discrete image signal in 3D, defined as: $\mathbf{I} : \mathbb{Z}^3 \rightarrow \mathbb{R}$, the axiomatic formulation of the continuous scale-space of this signal is:

$$L(x; t) = \sum_{\xi \in \mathbb{Z}^3} T(\xi; t)\, \mathbf{I}(x - \xi), \tag{8}$$

where $t \in \mathbb{R}_+$ denotes the continuous scale level, $x \in \mathbb{Z}^3$, $L(x; 0) = \mathbf{I}(x)$ and T defines a one-parameter family of kernels, used to generate the scale-space by convolution. The main property of a scale-space signal representation L, also called the image scale-space, is the non-enhancement of local extrema, which ensures the causality of structure across scales [20, p. 103]. This means that local maximum/minimum points in the image signal at any scale level t_0, do not increase/decrease their value at any higher scale $t > t_0$. Based on this property, as well as the semi-group structure of the family of kernels T, it has been shown that within the class of linear transformations, the scale-space representation L is differentiable, satisfying the differential equation:

$$\partial_t L = \mathcal{A}_{ScSp} L, \tag{9}$$

where $\mathcal{A}_{ScSp}$ is an infinitesimal scale-space generator, based on discrete approximations of the Laplace operator [20]. In this case, one can formulate the change in scale level t as an effect of actions of the agent, e.g. by introducing a new action to increase, decrease or maintain the scale level. In RL theory this can also be formulated as a continuous action that needs to be learned as a step $\Delta t \in \mathbb{R}$, which specifies the change in scale level at each navigation step.

To achieve this, one can redefine the optimal action-value function Q^*, by conditioning the state-representation s and model parameters θ on the scale-space representation L and the current scale level t:

$$Q^*(s, a \mid L, t) = \mathbb{E}_{s'} \left(r + \gamma \max_{a'} Q^*(s', a' \mid L, t') \right), \tag{10}$$

where $t' \in \mathbb{R}_+$ represents the scale level after executing action a. This implies that the object search would occur in continuous image scale-space, allowing the system to exploit structures on different scales. However, since the image dimensionality is preserved across scales, we are still left with the task of effectively addressing the aforementioned trade-off: sampling efficiency versus global context. In addition, recall that the scale-space parameter $t \in \mathbb{R}_+$ is continuous. Since the model parameters θ depend on the scale, one would need to design a learning model that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

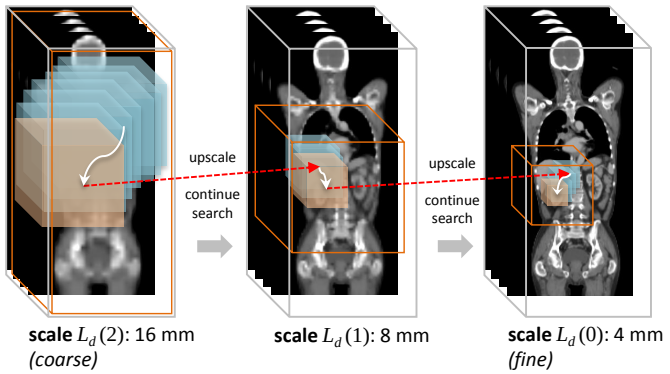

Fig. 4. Visualization of the detection pipeline for the right kidney. The search starts on the coarsest scale level $L_d(2)$. On each scale $L_d(k), k \geq 0$, the agent navigates until convergence. This convergence point (see definition in section 3.4.1) is used as a starting point for the subsequent scale level $L_d(k-1)$ (red dashed arrows indicate the change of scale). The process continues analogously on the following scale levels, with the convergence point on the finest scale marked as the detection result. We visualize with white arrows the optimal 3D search trajectories navigated at each scale. Along the trajectories we visualize the sequence of states, represented as 3D boxes of image context, centered at the location of the agent. The orange frame represents the constrained region sampled and explored during training on each scale. On the coarsest scale, this region always covers the entire 3D volume, with decreasing range on finer scales.

could capture not only the variability in image space but also the variability in scale-space. To avoid this complexity, we propose a discrete approximation L_d of the continuous scale-space L, defined as:

$$L_d(t) = \Psi_\rho(\sigma(t-1) * L_d(t-1)), \tag{11}$$

where $t \in \mathbb{N}$ denotes the discrete scale level, $*$ denotes a convolution, σ represents a scale-dependent Gaussian-like smoothing function and Ψ_ρ denotes a signal operator, reducing the spatial resolution with factor ρ using down-sampling [20]. Similarly to the continuous case, $L_d(0) = \mathbf{I}$. While down-sampling the signal can introduce aliasing effects, they do not affect the learning process, enabling the system state to capture global context on coarse scale and local context on fine scale with similar sampling complexity (see Figure 4).

3.4 Learning Multi-Scale Search Strategies

Given this discrete scale-space definition, we design a navigation model for each scale level: $\mathbf{\Theta} = [\theta_0, \theta_1, \ldots, \theta_{M-1}]$, where M is the number of different scales. While low-level features could arguably be shared across scales to determine a single multi-scale search model, we empirically observed that training a different model on each scale yields optimal results. The motivation for this is that different scale levels are described by different image structures that can be used as robust evidence for the search. Across scales we clone all meta-parameters defining each model: $Q(\cdot,\cdot;\theta_t \mid L_d, t), \forall t < M$, including the range of the state-representation, i.e. the size of the extracted box of image intensities. The search starts at the coarsest scale level $M-1$, where the search-model $Q(\cdot,\cdot;\theta_{M-1} \mid L_d, M-1)$ is trained for convergence from any starting point in the image (we define the convergence criterion in section 3.4.1). On this scale level the field-of-view of the agent is very large, acquiring sufficient global context to ensure an effective navigation. Upon convergence, the scale level is changed to $M-2$ and the search continued from the convergence point at level $M-1$. The process is repeated on the following scales until convergence on the finest scale. Note that on all scales, except the coarsest level $M-1$, the exploration range can be constrained, given the convergent behavior on coarser scales. These search-ranges are robustly determined during training (see Figure 4). Based on the definition of the discrete scale-space L_d and the independence of the search models across scales, we can rewrite Equation 6 and train on each scale level $0 \leq t < M$ according to:

$$\hat{\theta}_t^{(i)} = \arg\min_{\theta_t^{(i)}} \mathbb{E}_{s,a,r,s'} \left[\left(y - Q\left(s, a; \theta_t^{(i)} \mid L_d, t\right) \right)^2 \right], \tag{12}$$

with $i \in \mathbb{N}$ denoting the training iteration. The reference estimate y is determined similarly as in the single-scale solution, using a set of model parameters $\bar{\theta}_t^{(i)} := \theta_t^{(i')}$ from a previous training iteration $i' < i$:

$$y = r + \gamma \max_{a'} Q\left(s', a'; \bar{\theta}_t^{(i)} \mid L_d, t\right). \tag{13}$$

Algorithm 1 describes the training steps for our system.

Algorithm 1 Training Multi-Scale DRL for Detection

Given N training 3D-CT scans: $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N$
Define discrete scale-space: $L_d(t)|_{0 \leq t < M}$
Initialize system memory: $\mathbf{M}(0, \ldots, M-1) = [\,]$
Initialize exploration factor: $\epsilon = 1.0$
Initialize model parameters $\theta_t|_{0 \leq t < M}$ randomly
while $\epsilon > 0.05$ **do**
  for all scale levels $0 \leq t < M$ **do**
    Select random image and starting-point
    Sample ϵ-greedy path T with $Q(\cdot,\cdot;\theta_t \mid L_d, t)$
    $\mathbf{M}(t) \leftarrow \mathbf{M}(t) \cup [T]$
    Train $Q(\cdot,\cdot;\theta_t \mid L_d, t)$ according to Equation 12
  end for
  Decay ϵ - reduce randomness
end while
Output $\mathbf{\Theta} = \left[\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{M-1}\right]$ - estimated models

3.4.1 Empirical Convergence Criterion

Given a test volume $\mathbf{I}$, a discrete scale-space definition L_d and a set of trained multi-scale search models $\mathbf{\Theta}$, two important questions arise: At which location in the image does the search process start? When does the agent know that it has found the object of interest?

The starting point $\vec{p}_0$ is defined based on the expected relative position $\vec{r}$ of the anatomical landmark, which is computed on the training set. Given N training images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N$, we define $\vec{r} \in [0,1]^3$ as $\forall d \in \{1,2,3\}, \vec{r}(d) = \frac{1}{N}\sum_{k=1}^{N} \frac{\text{gtruth}[\mathbf{I_k}]_d}{\text{size}[\mathbf{I_k}]_d}$, where $\text{size}[\mathbf{I_k}]_d \in \mathbb{N}$ denotes the image size, i.e. number of voxels in dimension d, and $\text{gtruth}[\mathbf{I_k}]_d \in \mathbb{R}_+$ denotes the coordinate of the ground-truth landmark annotation in dimension d. Based on $\vec{r}$, we define the starting point as $\vec{p}_0 = \text{size}[\mathbf{I}] \odot \vec{r}$, where the operator $\odot$ denotes an element-wise multiplication. Please note that the choice of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

the starting point is a question of algorithm design, and is not a limitation. As we will explain later in the experiments section, most of the scans are focused on the thorax and abdomen. Using the above definition, the starting point is expected to be close to the landmark location on such cases, considerably reducing the expected detection time. We also demonstrate that similar performance can be achieved by starting in the center of the scan (see Appendix C).

The question of trajectory convergence is implicitly related to the convergence properties of the system. However, the literature shows that there are no theoretical guarantees of global convergence when using a non-linear policy approximator, such as a deep neural network [19], [47]. In practice, several heuristic techniques such as memory replay, delayed updates or random-exploration ensure training stability and convergence. In this context, we formulate our trajectory convergence criterion as follows: given a search-trajectory $T = [\vec{p}_0, \vec{p}_1, \ldots]$, $\exists k, k' \in \mathbb{N}$, with $k' > k \geq 0$ such that $\vec{p}_k = \vec{p}_{k'}$ with $l = k' - k$ minimal. In other words trajectories converge on small, oscillatory-like cycles. Once such a cycle is identified at detection time, we stop the search and yield $\vec{p}_k$ as detection result. We observed that this stopping criterion is robust in practice, that trajectories do not converge on long cycles. We provide an empirical analysis showing that the probability of converging on large cycles is exponentially small (see Appendix A).

3.4.2 Object Not in the Scan Range?

In order to reliably use a machine in a clinical scenario to detect structures and derive automatic measurements that support the radiologist in reading 3D-CT scans, one needs to consider all the different types of such scans, e.g. cardiac, thoracic and abdominal scans, CT scans of the legs and pelvis or head-neck scans. In this general setting, an important question becomes whether the system is capable of recognizing the absence of an anatomical landmark from the captured field-of-view, i.e. the scanned region of the body. In practice, one cannot exclusively rely on meta-information about the scan acquisition to find an answer. Depending on the type of investigation, the medical technical radiology assistant (MTRA) can decide to either increase or decrease the field-of-view when acquiring the scan. To the best of our knowledge, previous solutions for object detection [1], [10], [16], [33] do not consider this scenario. For example, scanning solutions can impose a fixed threshold on the hypothesis probability, and use it as a decision criterion. In our experience, this heuristic is not always accurate.

Our formulation of generic object detection as a search problem represents a principled step towards addressing this challenge. Given an image $\mathbf{I}$, and a structure of interest located outside the field-of-view, i.e. the image space, one can recognize the absence of this structure by following navigation trajectories, which attempt to leave the image space in the direction where the structure was supposed to be located, had the scan captured the whole body. By training the system on differently cropped images, we empirically observed this consistent behavior (see Figure 5). Empirical results are included in section 4 .

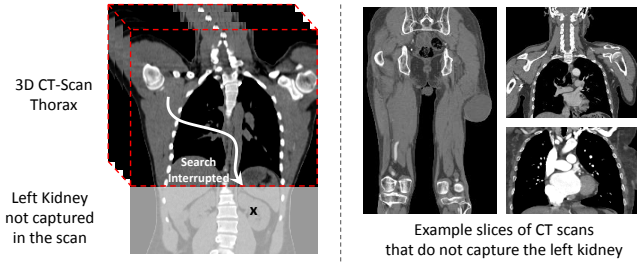

Fig. 5. Visualization of the search-path followed to find the left kidney in a thorax CT scan, which does not capture the left kidney (marked by an **x**). The trajectory leaves the image space, signaling that the organ is missing from the field-of-view. On the right we show several slices of other CT scans from the test set, which do not capture the left kidney. The image on the left is a slice from a CT scan of the legs. On the upper right we show a slice of a thorax CT scan, acquired for lung cancer screening. The lower right slice is from a cardiac CT scan with contrast.

4 Experiments

For the experimental evaluation and comparison, we selected 5 reference methods: scanning with probabilistic boosting trees [3] (PBT), extremely randomized trees with Hough regression [10] (ExtRTrees), the *Overfeat* method adapted from public source-code to 3D data [22], 3D deep learning with filter decomposition [18] (3D-DL) and scanning with cascaded sparse-adaptive deep neural networks [1] (SADNN). All these methods were implemented and evaluated on detecting several anatomical landmarks.

4.1 Dataset

The dataset contains 1487 3D-CT volumes from 532 patients, covering a wide range of scan types with different field-of-views, e.g. cardiac CT scans (with contrast), thoracic scans, abdominal scans, CT scans of the legs and pelvis or CT scans of the head and neck. A large subset of these scans is focused on the thoracic and abdominal region and around 20% cover the whole body. In practice, whole body scans are acquired only rarely, most often to support the fast assessment of injuries in cases of polytrauma patients [9]. In general, different types of CT scans are associated with different diagnostic routines, e.g. head CT scans can be used for diagnosis of brain hemorrhages, brain tumors and aneurysms, lower-limb CT scans for detecting complex bone fractures and tumors in the legs, CT scans of the thorax, abdomen and pelvis for cancer screening, etc. [9]. As such, most of the scans capture challenging anatomical malformations, i.e. large tumors or anomalies. In the preprocessing stage, all volumes were resampled to an isotropic resolution of 2 mm for the finest scale level in the scale-space representation. For our scale-space we used 3 additional levels at isotropic spatial resolutions of 4 mm, 8 mm and 16 mm. We clipped the voxel values to the useful 0-800 HU interval and then normalized this interval to unit-range, i.e. $[0, 1]$.

We selected a set of 8 anatomical landmark points, covering different types of structures, including bone, non-rigid organ, vessel bifurcation and respiratory tract bifurcation. These are the center of the left and right kidneys, the front corner of the left and right hip bones, the bronchial bifurcation, as well as three vessel bifurcations between the aortic arch and the left subclavian artery, the left common carotid

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

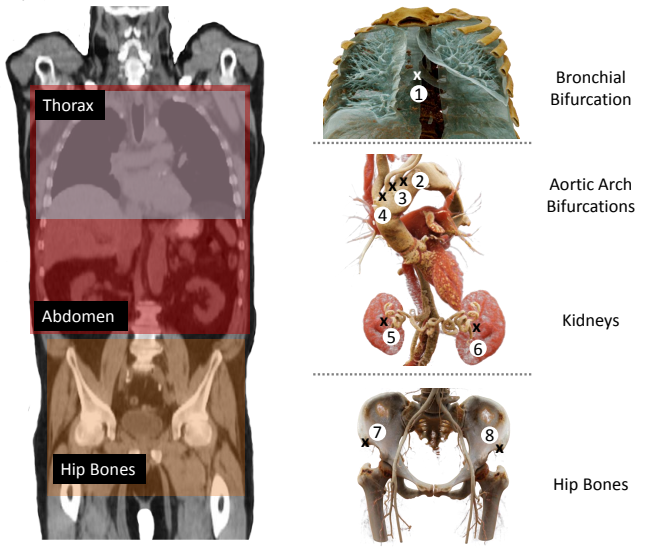

Fig. 6. Visualization of all anatomical landmarks used for evaluation. On the **left** we visualize a section of a whole body 3D-CT scan including neck, thorax, abdomen and pelvis. On the **right** we highlight the different anatomical structures and mark the individual landmark points: bronchial bifurcation (1), bifurcation of left subclavian artery (2), bifurcation of left common carotid artery and left subclavian artery (3), bifurcation of left common carotid artery and brachiocephalic artery (4), center of right kidney (5), center of left kidney (6), front corner of right hip-bone (7), front corner of left hip-bone (8).

artery and the brachiocephalic artery (see Figure 6). This set of landmarks was selected to test the robustness of the solutions to the type/contrast of the anatomical structure, as well as their ability to cope with large variation of non-rigid organs and the confusion of nearby vessel bifurcations with similar appearance. Ground-truth annotations were provided by radiologists. Vessel bifurcation and bone landmarks are anatomically defined very precisely, allowing for very accurate annotations at original resolution. For bone landmarks the average inter-observer variability was reported in the literature at 1 mm, while the intra-observer variability was 0.9 mm [55] (using three expert annotators). This precision level is in agreement with the findings of the study of Chien P.C. *et al.* [56]. In contrast, the variation in the annotation of the kidney center is higher, proportional to the size of the structure. Also for such landmarks, the inter-user variability is in general low in 3D-CT [37].

The validation setup is based on a random split of the annotated volumes in approximately 80% training and 20% unseen testing examples for each landmark. The split was performed at patient level, meaning that all scans of a given patient are either in the training set or the test set. The total number of available ground-truth annotations per landmark are: 1438 (center of left kidney), 1432 (center of right kidney), 552 (front corner of left hip-bone), 1054 (front corner of right hip-bone), 1046 (bronchial bifurcations), 1028 (bifurcation of left subclavian artery), 1048 (bifurcation of left common carotid artery and left subclavian artery), 1048 (bifurcation of left common carotid artery and brachiocephalic artery).

4.2 System Training

We used a scale-space of 3 scale levels at 4 mm (fine) - 8 mm - 16 mm (coarse) isotropic spatial resolutions for the detection of each kidney and include one additional scale level at 2 mm isotropic spatial resolution for the remaining landmarks. We empirically observed that using a finest resolution of 4 mm in this case, yields optimal results both in terms of speed and accuracy. In order to cope with the intensive memory requirements, the SADNN and Overfeat methods were trained on all landmarks using a finest resolution of 4 mm. All other methods, including ours, were trained on a finest isotropic resolution of 4 mm for the kidneys and 2 mm for the remaining landmarks.

We optimized all meta-parameters on the example of one arbitrary landmark – here the front corner of the right hip-bone. Using a patient-based split in training (70%), validation (10%) and testing (20%) sets, we selected the optimal algorithm meta-parameters and the network architecture using a systematic search. Shared on each scale (except the search-range on coarsest scale), the meta-parameters are specified in Table 1. The CNN used to encode the search policy per scale is defined as: conv-layer (32 kernels: 4×4×4, ReLU), pooling (2×2×2), conv-layer (46 kernels: 3×3×3), pooling (2×2×2) and four fully-connected layers (512×256×128×6 units, ReLU). Our implementation is based on the Theano library [57]. The training time per landmark averaged to 4 hours on an Nvidia Titan X GPU. We trained all models in a 16-GPU cluster in around 3 hours.

TABLE 1
Values of all meta-parameters required to train our system.

Description	Value
Training rounds	500
Episode length (linear decay)	1000 → 50
State size	25×25×25 vox.
Max. search range	±10 voxels
Optimal scale-space factor	2
Initial/Final exploration	ϵ = 100% / 5%
Exploration decay	200000
Network update frequency	14
Replay memory size / batch size	100000 / 256
Reference-freeze interval	10000
Min. required memory size	10000
Discount factor	0.9
Optimization method	RMS-prop [19]
Learning rate	0.0005
RMS-decay/epsilon	0.95 / 0.01
Nesterov momentum	0

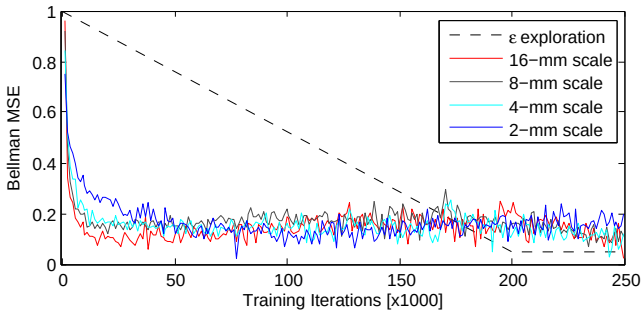

Fig. 7. Visualization of the mean squared error in the Bellman equation on each scale level during training of the right hip-bone landmark. The plot also visualizes the progression of the ϵ variable during training – this value controls the randomness in the exploration and decays from 1 to 0.05. Note that ϵ does not measure an error.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

The training criterion was the Bellman error [53], which measures the quality of the policy on each scale level. Figure 7 shows the evolution of the Bellman error during training for the front corner of the right hip-bone. Overlayed in the same plot is the randomness of the exploration denoted by the variable ϵ. Please note that the target is a low Bellman-error with minimal exploration randomness, i.e. near-deterministic on-policy search. Choosing the randomness decay-rate too high causes the system to fail to train, while a too low decay-rate can lead to overfitting.

4.3 3D Landmark Detection in CT-Scans

Given a trained multi-scale set of search-models for one landmark on M different scales $\mathbf{\Theta} = [\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{M-1}]$, the search process starts on the coarsest scale $M-1$ from the expected location of the landmark, computed on the training set. Using the search model $\hat{\theta}_{M-1}$, the search is executed on the corresponding scale until reaching the convergence point (recall from section 3 the formal definition of convergence point). The process is repeated analogously on all following scales $M-2, \ldots, 0$, using at each scale t as starting point the convergence point from scale $t+1$.

To assess the success rate of each algorithm from an anatomical validity perspective, we imposed hard thresholds which were set to **30 mm** for the center of each kidney and **10 mm** for the remaining landmarks. Any detection above this limit was considered an outlier. Using this failure criterion, we first computed a *failure percentage rate* for all algorithms. Within the set of failed cases, we show the median and maximum error to give a sense of the error magnitude. For the remaining detections, we computed regular statistics (mean, median and standard deviation) to measure the accuracy. Table 2 shows all these metrics for all algorithms and each considered landmark.

Our method achieves 0% failure rate on all considered landmarks, improving the average accuracy of the reference methods by 20-30% (see error distribution in Figure 9). Through the multi-scale search our solution mimics a natural focusing mechanism, starting from global context to low-level image details on the finest scale. This helps to gain robustness to structures with similar anatomical appearance and thereby considerably reduce the number of miss-detections. Please note that in the implementation of the method of Donner *et al.* [10], we did not apply the MRF graph-matching step for hypothesis selection, which exploits the spatial relationship between the landmarks. This was not possible in our case of incomplete data, where several landmarks might be missing from the field-of-view. Instead, we applied an averaging scheme.

For a deeper insight into the performance of additional existing solutions in comparison to our method, please see Table 3. Please note that we did not implement these methods, but simply show their performance, as reported in the literature. The comparison in Table 3 demonstrates that our method not only achieves the highest voxel-accuracy, but also the quickest detection time for the use-case of 3D-CT. More importantly, these results are obtained on the largest dataset reported in literature, consisting of 1487 CT scans from 532 patients. We also measured the performance of our method on an additional dataset of 506 3D-MR scans from 506 patients (see Appendix C).

TABLE 3
Table showing a general comparison against different solutions for anatomical landmark detection in large high-resolution scans. The criteria are the average detection-accuracy and runtime (on CPU), as well as the size of the evaluation set, i.e. number of patients / scans, and data type (CT or MR, i.e. magnetic resonance).

Solution	Dataset Size (Data/Patients)	Accuracy (mm)	Speed (seconds)
Zhan *et al.* [27]	18/18 CT	4.72	4
Fenchel *et al.* [35]	31/31 MR	22.4	20
Criminisi *et al.* [12]	100/– CT	17.60	1
Pauly *et al.* [32]	33/33 MR	14.95	0.8
Cuingnet *et al.* [11]	233/89 CT	10.5	2.8
Donner *et al.* [10]	20/20 CT	5.25	120
Criminisi *et al.* [31]	400/– CT	13.50	4
Chu *et al.* [30]	10/10 CT	1.90 [1]	30
Potesil *et al.* [37]	83/83 CT	4.70	N/A
de Vos *et al.* [24]	100/– CT	4.80	10
Ours	**1487/532** CT	**4.19** [2]	**0.061**

[1] Evaluated only on vertebrae localization with strong priors.
[2] With no failures of clinical significance. All other solutions did not provide any information in this respect.

4.4 Runtime Analysis

In terms of detection speed, the 3D deep learning solution proposed in [18] clocked the highest runtime, averaging 61.56 seconds on CPU to detect one landmark. This high runtime is explained by the fact that scanning was performed on the complete volume on finest resolution. The Overfeat method [22] improved this number by applying an efficient convolutional forward-propagation scheme, running in under 2.5 seconds. A further increase in speed to an average runtime of less than 0.7 seconds was achieved by the SADNN solution [1], by using a cascade of efficient sparse shallow models to pre-filter the large number of negative hypotheses. We emphasize that these results were achieved at a finest isotropic resolution of 4 mm, meaning that the test volumes were 8 times smaller than the same volumes at 2 mm. The PBT-based solution heuristically aggregates multi-scale hypotheses to constrain the search range on fine resolution. This strategy increased the average detection speed by around 50 times, reaching 1 second. In comparison, the method proposed by Donner *et al.* [10] achieved a median detection time of 4.7 seconds.

However, none of these methods could match the detection speed of our solution. Learning the multi-scale search trajectory and evaluating samples only on a single path, as opposed to scanning the image space, led to a median runtime of 33 milliseconds (slowest runtime: 85 ms) – an unmatched real-time performance for landmark detection in 3D-CT. The detection speed was also in similar range on CPU. The improvement against the reference methods was around 2-3 orders of magnitude (see Table 4).

In addition, the detection speed of our method has the property of scaling sublinearly with respect to the scan size. While the runtime of all reference methods increases linearly with the volume size N, i.e. the number of voxels in the volume, in our case the increase is proportional to $\sqrt[3]{N}$ (for more details please see Figure 8 and proof in Appendix B). As such, our method can also easily be applied on higher resolutions, e.g. less than 0.5 mm spatial resolution, where it can achieve similar detection speed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 2
Table showing results on different anatomical landmarks. The first three columns indicate the percentage of failed cases, as well as the median and maximum error within this group. The accuracy is measured on successful detections (excluding failed cases). The error is measured in **mm**.

			Failed Cases			Accuracy (excl. failed cases)	
Structure Type	**Landmark**	**Method**	**% Failed**	**Median**	**Max**	**Mean ± STD**	**Median**
Bone	Hip Bone Right Front Corner	PBT [3]	4.35%	38.89	141.88	3.05 ± 1.60	2.77
		ExtRTrees [10]	8.07%	20.27	460.22	5.01 ± 2.40	4.95
		Overfeat [22]	9.31%	35.64	231.29	4.55 ± 1.98	4.31
		3D-DL [18]	0.62%	10.17	10.17	2.84 ± 1.36	2.53
		SADNN [1]	0%	–	–	3.50 ± 1.63	3.37
		Ours	**0%**	–	–	**2.80** ± 1.46	**2.53**
	Hip Bone Left Front Corner	PBT [3]	3.75%	14.30	28.57	4.03 ± 1.79	3.86
		ExtRTrees [10]	6.25%	13.43	17.44	5.13 ± 2.78	5.08
		Overfeat [22]	3.75%	127.52	242.05	4.19 ± 1.80	4.15
		3D-DL [18]	2.50%	262.81	513.81	3.14 ± 1.49	2.98
		SADNN [1]	1.25%	12.57	12.57	4.44 ± 1.75	4.43
		Ours	**0%**	–	–	**3.07** ± 2.14	**2.85**
Nonrigid Organ	Right Kidney Center	PBT [3]	3.16%	43.54	131.55	11.85 ± 5.97	11.15
		ExtRTrees [10]	4.21%	158.28	166.55	8.06 ± 5.05	6.71
		Overfeat [22]	1.05%	69.62	107.22	7.01 ± 3.94	6.16
		3D-DL [18]	1.58%	47.26	50.42	8.05 ± 4.39	7.59
		SADNN [1]	0%	–	–	6.92 ± 3.97	6.35
		Ours	**0%**	–	–	**6.89** ± 3.65	**5.95**
	Left Kidney Center	PBT [3]	1.11%	89.77	140.03	8.51 ± 4.06	8.02
		ExtRTrees [10]	3.33%	44.43	182.71	8.86 ± 5.72	7.26
		Overfeat [22]	1.67%	40.27	134.05	6.57 ± 3.11	6.40
		3D-DL [18]	0.56%	43.41	43.41	7.78 ± 4.00	7.58
		SADNN [1]	2.22%	50.52	61.97	**6.12** ± 3.07	**5.52**
		Ours	**0%**	–	–	6.72 ± 3.62	6.22
Vessel Bifurcations	Left Com. Carotid Artery Left Subclavian Artery	PBT [3]	7.22%	12.19	34.05	3.96 ± 2.06	**3.36**
		ExtRTrees [10]	15.00%	19.65	35.11	6.17 ± 2.93	5.93
		Overfeat [22]	8.88%	14.36	46.68	5.64 ± 2.33	5.53
		3D-DL [18]	6.11%	11.75	15.93	4.37 ± 2.12	4.02
		SADNN [1]	5.00%	13.21	17.50	4.80 ± 2.23	4.48
		Ours	**0%**	–	–	**3.89** ± 1.95	3.51
	Left Com. Carotid Artery Brachiocephalic Artery	PBT [3]	4.22%	12.88	17.13	3.85 ± 1.85	3.59
		ExtRTrees [10]	18.07%	18.25	48.80	6.57 ± 2.95	6.63
		Overfeat [22]	8.43%	12.04	25.28	5.25 ± 2.52	5.07
		3D-DL [18]	6.02%	12.94	18.65	4.47 ± 2.19	4.09
		SADNN [1]	6.02%	12.76	22.17	4.85 ± 2.29	4.67
		Ours	**0%**	–	–	**3.71** ± 2.01	**3.47**
	Left Subclavian Artery Bifurcation	PBT [3]	3.13%	16.30	19.84	3.67 ± 1.86	3.31
		ExtRTrees [10]	15.63%	17.14	203.38	6.42 ± 2.73	6.24
		Overfeat [22]	7.50%	12.04	34.96	5.34 ± 2.36	5.21
		3D-DL [18]	3.13%	11.85	22.44	3.60 ± 1.76	3.29
		SADNN [1]	6.25%	13.65	44.98	4.59 ± 2.17	4.26
		Ours	**0%**	–	–	**3.09** ± 1.50	**2.86**
Respiratory Tract	Bronchial Bifurcation	PBT [3]	4.79%	13.80	28.20	3.43 ± 1.62	3.25
		ExtRTrees [10]	7.18%	15.66	32.94	5.71 ± 2.65	5.52
		Overfeat [22]	6.58%	11.65	19.05	5.07 ± 2.10	4.92
		3D-DL [18]	2.99%	11.35	15.67	**2.98** ± 1.49	**2.86**
		SADNN [1]	5.39%	11.09	14.96	5.08 ± 2.22	5.00
		Ours	**0%**	–	–	3.35 ± 1.77	3.10

4.5 Object Not in the Scan Range?

We empirically estimated the accuracy of our algorithm in recognizing the absence of the landmark from the field-of-view of the scan. For this, we selected the bronchial bifurcation landmark and 100 random images from the 188 test-images. Each image was randomly cropped along the Z-axis to eliminate the landmark from the field-of-view. The image cut was performed at a distance of **at least 1 cm** from the landmark. During navigation, we investigated whether the search trajectory leaves the image space on any of the scale levels through the correct volume border. We empirically found that, on at least **97%** (an average of **99.2%**) of the images, the search trajectory leaves the image space, regardless of the selected starting point. Similar percentages are achieved also for the hip-bones. i.e. 97.8%, respectively 98.2%. In contrast, for the kidney center the accuracy is lower, i.e. 92.2% for left and 90.5% for the right kidney. The reason for this decrease are border cases. For many thoracic CT scans, the kidney-centers are used as lower-limits for the field-of-view. This results in many challenging test examples in which the kidney center is very close to the border.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

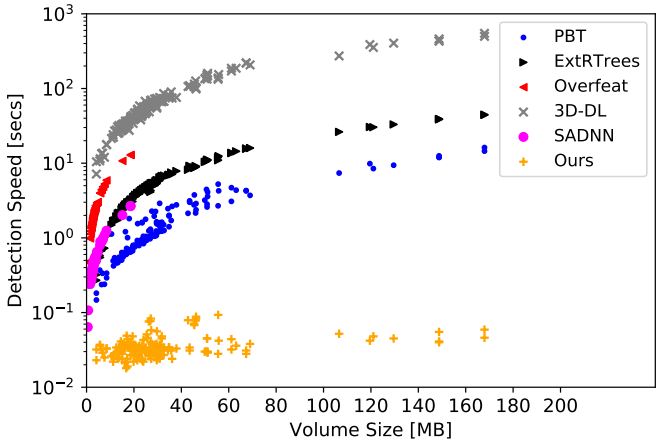

Fig. 8. Scatter plot showing the correlation between scan size and detection speed for all considered methods on the right hip-bone landmark. Note that solutions based on scanning show a linear correlation between volume size and execution time. We remind the reader that for technical reasons the SADNN and Overfeat solutions are evaluated at a finest isotropic resolution of 4 mm while the other methods are evaluated at 2 mm. Our approach not only improves the average speed, but also scales sublinearly with respect to the size of the input volume.

TABLE 4
Table showing the runtime performance on the example of the right hip-bone landmark. For this landmark the scans cover at least 60% of the patient height at an average size of $250 \times 250 \times 500$ voxels at 2 mm isotropic spatial resolution. The speed is measured in **seconds**.

DETECTION TIME per Volume [**seconds**]

Method	Platform	Median	Min	Max
PBT [3]	CPU 8-core	1.051	0.14	16.20
ExtRTrees [10]	CPU 8-core	4.714	0.272	44.535
Overfeat [1] [22]	GPU Titan X	2.175	0.331	14.86
3D-DL [18]	CPU 10-core	57.034	7.161	548.23
SADNN [2] [1]	GPU GTX 1080	0.471	0.064	3.029
Ours	CPU 8-core	0.061	0.035	0.155
Ours	GPU Titan X	**0.033**	**0.018**	**0.085**

[1,2] Evaluated on finest isotropic resolution of 4 mm.

In parallel work, we address this challenge by enforcing the spatial coherence of the visible landmarks using robust statistical shape models. Particularly for border-cases, we investigate the benefits of explicitly training the navigation models on cases where target objects are outside the field-of-view. These ideas fall out of the scope of this paper.

4.6 A Computer Vision Perspective

Although motivated in the context of anatomical landmark detection in large 3D-CT scans, our method can also be applied to different 2D problems from both the medical domain [21] and computer vision – where data is unstructured and objects are often occluded. Several very recent publications demonstrate competitive results on a variety of computer vision tasks:

Object Localization Using convolutional neural networks as pre-trained feature extractors, deep Q-learning can be applied to learn a policy for hierarchical object localization [58], [59]. Using tree-structured parsing schemes, Jie *et al.* [60] propose to use DRL for the sequential search of multiple objects. Reinforcement learning is also used to learn stochastic policies for how and where to apply detectors for object localization [61]. Reported results on the PASCAL VOC 2012 dataset show significant speed-improvements over conventional solutions.

Tracking Trajectory learning can also be formulated for image sequences over time, e.g. video frames, to support active object tracking [62]. By using deep recurrent models, the learned tracking policy captures motion patterns over time. Reinforcement learning is also used to model the lifetime and visibility of objects in online tracking [63].

Visual Navigation Deep actor-critic models are effectively applied to visual navigation tasks [64], such as robot navigation or autonomous driving. In this formulation the actor model learns the action selection policy using the feedback from the critic, which estimates the long-term reward.

The principles of *multi-scale deep reinforcement learning*, introduced in this work, might address the limitations of the aforementioned methods and increase their performance. First, using an explicit scale-space model to represent the state space allows for an improved system scalability for different object scales. This property is important for localization, tracking and navigation tasks based on photographs, for which there is typically no prior information about scale [20]. In addition, as demonstrated in our applications, modeling the object search across scales significantly increases the effectiveness of the exploration, leading to a more robust and globally consistent navigation policy, that is invariant to ambiguous local image information. This is particularly important e.g. for visual navigation for autonomous driving or tracking in high-dynamic scenes [62], [64]. Second, coupling the concept of navigation with a scale-space representation increases the detection speed, which scales sub-linearly with respect to the scan size. This enables the effective exploitation of raw high-resolution data to increase the detection accuracy while maintaining real-time performance – an important requirement for tracking and online navigation systems. Finally, training the system end-to-end leads to a high policy performance. This was emphasized in [19] in the context of game playing, and also empirically observed in our experiments on medical image data. In contrast, Caicedo *et al.* [58] and Bellver *et al.* [59] apply pre-trained feature extractors to get an embedding of the state, which is used as input for the policy network.

5 Conclusion

In this work, we presented a novel method for accurate real-time 3D anatomical landmark detection in CT scans. Based on the reformulation of the problem as a generic behavioral learning task, we combine the concept of deep reinforcement learning with multi-scale image analysis to enable an artificial agent to systematically learn optimal strategies for finding anatomical structures. Experiments show that our method is robust against outliers and achieves an average accuracy improvement of 20-30% over the selected reference solutions. At the same time, the detection speed of our algorithm is 2-3 orders of magnitude faster, reaching real-time performance on high-resolution 3D-CT volumes. Our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2782687, IEEE Transactions on Pattern Analysis and Machine Intelligence

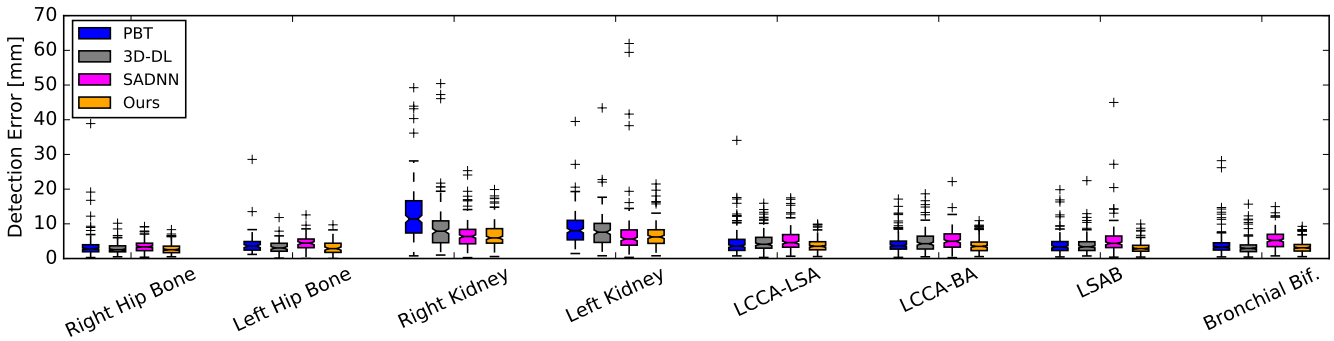

Fig. 9. Comparison of the four best performing solutions (regarding accuracy and failures). For each of the considered anatomical landmarks our method reduces the number of failed detections to zero and improves the average and median error by around 20-30%. Note that the plot displays the distribution of detection errors that are smaller than 70 mm and does not show very large outliers above this value (noted in Table 2).

solution can also elegantly handle the case of absent objects and can be extended to support the simultaneous detection of multiple objects. Through high robustness and real-time performance, the proposed method might represent an important component of next-generation clinical technologies that contribute to better, faster and more reproducible patient diagnosis, therapy and disease management.

Disclaimer This feature is based on research, and is not commercially available. Due to regulatory reasons its future availability cannot be guaranteed.

Acknowledgment We thank David Liu for his contributions to this work during his time at Medical Imaging Technologies, Siemens Healthineers, Princeton NJ.

REFERENCES

[1] F. C. Ghesu, E. Krubasik, B. Georgescu, V. Singh, Y. Zheng, J. Hornegger, and D. Comaniciu, "Marginal space deep learning: Efficient architecture for volumetric image parsing," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1217–1228, 2016.

[2] A. M. Pouch, H. Wang, M. Takabe, B. M. Jackson, J. H. G. III, R. C. Gorman, P. A. Yushkevich, and C. M. Sehgal, "Fully automatic segmentation of the mitral leaflets in 3D transesophageal echocardiographic images using multi-atlas joint label fusion and deformable medial modeling," *Medical Image Analysis*, vol. 18, no. 1, pp. 118–129, 2014.

[3] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features," *IEEE Transactions on Medical Imaging*, vol. 27, no. 11, pp. 1668–1681, 2008.

[4] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 809–817.

[5] B. Glocker, D. Zikic, and D. R. Haynor, "Robust registration of longitudinal spine CT," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 8673. Springer International Publishing, 2014, pp. 251–258.

[6] C. R. Hatt, M. A. Speidel, and A. N. Raval, "Robust 5DOF transesophageal echo probe tracking at fluoroscopic frame rates," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9349. Springer International Publishing, 2015, pp. 290–297.

[7] L. Yang, B. Georgescu, Y. Zheng, Y. Wang, P. Meer, and D. Comaniciu, "Prediction based collaborative trackers (PCT): A robust and accurate approach toward 3D medical object tracking," *IEEE Transactions on Medical Imaging*, vol. 30, no. 11, pp. 1921–1932, 2011.

[8] J. R. Cebral and R. Lohner, "Efficient simulation of blood flow past complex endovascular devices using an adaptive embedding technique," *IEEE Transactions on Medical Imaging*, vol. 24, no. 4, pp. 468–476, 2005.

[9] E. P. Hess, L. R. Haas, N. D. Shah, R. J. Stroebel, C. R. Denham, and S. J. Swensen, "Trends in computed tomography utilization rates: A longitudinal practice-based study," *Journal of Patient Safety*, vol. 10, no. 1, pp. 52–58, 2014.

[10] R. Donner, B. H. Menze, H. Bischof, and G. Langs, "Global localization of 3D anatomical structures by pre-filtered Hough forests and discrete optimization," *Medical Image Analysis*, vol. 17, no. 8, pp. 1304–1314, 2013.

[11] R. Cuingnet, R. Prevost, D. Lesage, L. D. Cohen, B. Mory, and R. Ardon, "Automatic detection and segmentation of kidneys in 3D CT images using random forests," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 7512. Springer Berlin Heidelberg, 2012, pp. 66–74.

[12] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in CT studies," in *Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging*, vol. 6533. Springer Berlin Heidelberg, 2010, pp. 106–117.

[13] D. Štern, T. Ebner, and M. Urschler, "From local to global random regression forests: Exploring anatomical landmark localization," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9901. Springer International Publishing, 2016, pp. 221–229.

[14] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization with cascaded global-to-local regression and shape prior," *Medical Image Analysis*, vol. 23, no. 1, pp. 70–83, 2015.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[16] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using CNNs," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9901. Springer International Publishing, 2016, pp. 230–238.

[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9901. Springer International Publishing, 2016, pp. 424–432.

[18] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, "3D Deep learning for efficient and robust landmark detection in volumetric data," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9349. Springer International Publishing, 2015, pp. 565–572.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[20] T. Lindeberg, *Scale-Space Theory in Computer Vision*. Springer US, 1994, ch. Scale-space for N-D discrete signals, pp. 101–122.

[21] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu, "An artificial agent for anatomical landmark detection in medical images," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9902. Springer International Publishing, 2016, pp. 229–237.

[22] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Le-

Cun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," *Computing Research Repository*, vol. abs/1312.6229, 2013.

[23] Z. Tu, "Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 1589–1596.

[24] B. D. de Vos, J. M. Wolterink, P. A. de Jong, M. A. Viergever, and I. Išgum, "2D image classification for 3D anatomy localization: employing deep convolutional neural networks," in *Medical Imaging: Image Processing*, vol. 9784, 2016, pp. 9784–9791.

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[26] D. Liu, K. S. Zhou, D. Bernhardt, and D. Comaniciu, "Search strategies for multiple landmark detection by submodular maximization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2831–2838.

[27] Y. Zhan, X. S. Zhou, Z. Peng, and A. Krishnan, "Active scheduling of organ detection and segmentation in whole-body medical images," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 5241. Springer Berlin Heidelberg, 2008, pp. 313–321.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[29] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 379–387.

[30] C. Chu, D. L. Belavý, G. Armbrecht, M. Bansmann, D. Felsenberg, and G. Zheng, "Fully automatic localization and segmentation of 3D vertebral bodies from CT/MR images via a learning-based method," *PLOS ONE*, vol. 10, no. 11, pp. 1–22, 2015.

[31] A. Criminisi, D. Robertson, O. Pauly, B. Glocker, E. Konukoglu, J. Shotton, D. Mateus, A. Martinez Möller, S. G. Nekolla, and N. Navab, *Decision Forests for Computer Vision and Medical Image Analysis*. Springer London, 2013, ch. Anatomy Detection and Localization in 3D Medical Images, pp. 193–209.

[32] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A. M. Möller, S. Nekolla, and N. Navab, "Fast multiple organ detection and localization in whole-body MR Dixon sequences," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 6893. Springer Berlin Heidelberg, 2011, pp. 239–247.

[33] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2155–2162.

[34] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, pp. 2017–2025.

[35] M. Fenchel, S. Thesen, and A. Schilling, "Automatic labeling of anatomical structures in MR fastView images using a statistical atlas," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 5241. Springer Berlin Heidelberg, 2008, pp. 576–584.

[36] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. van Ginneken, "Multi-atlas-based segmentation with local decision fusion: Application to cardiac and aortic segmentation in CT scans," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1000–1010, 2009.

[37] V. Potesil, T. Kadir, G. Platsch, and M. Brady, "Personalized graphical models for anatomical landmark localization in whole-body medical images," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 29–49, 2015.

[38] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1862–1874, 2015.

[39] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[42] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[43] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, vol. 9912. Springer International Publishing, 2016, pp. 483–499.

[44] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Pearson Education, 2003.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[46] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 2843–2851.

[47] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[48] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[49] D. H. Hubel and T. N. Wiesel, "Shape and arrangement of columns in cat's striate cortex," *The Journal of Physiology*, vol. 165, no. 3, pp. 559–568, 1963.

[50] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Explorations in the microstructure of cognition: Foundations," in *Parallel Distributed Processing*. MIT Press, 1986, vol. 1.

[51] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. MIT Press, 1998.

[52] R. Bellman, *Dynamic Programming*, 1st ed. Princeton University Press, 1957.

[53] C. J. C. H. Watkins and P. Dayan, "Q-Learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.

[54] L.-J. Lin, "Reinforcement learning for robots using neural networks," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, USA, 1992.

[55] J. Victor, D. Van Doninck, L. Labey, B. Innocenti, P. Parizel, and J. Bellemans, "How precise can bony landmarks be determined on a CT scan of the knee?" *The Knee*, vol. 16, no. 5, pp. 358–365, 2009.

[56] P. Chien, E. Parks, F. Eraso, J. Hartsfield, W. Roberts, and S. Ofner, "Comparison of reliability in anatomical landmark identification using two-dimensional digital cephalometrics and three-dimensional cone beam computed tomography in vivo," *Dentomaxillofacial Radiology*, vol. 38, no. 5, pp. 262–273, 2009.

[57] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, 2016.

[58] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2488–2496.

[59] M. Bellver, X. Giró i Nieto, F. Marqués, and J. Torres, "Hierarchical object detection with deep reinforcement learning," *Computing Research Repository*, vol. abs/1611.03718, 2016.

[60] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan, "Tree-structured reinforcement learning for sequential object localization," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 127–135.

[61] S. Mathe, A. Pirinen, and C. Sminchisescu, "Reinforcement learning for visual object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2894–2902.

[62] D. Zhang, H. Maei, X. Wang, and Y. Wang, "Deep reinforcement learning for visual object tracking in videos," *Computing Research Repository*, vol. abs/1701.08936, 2017.

[63] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4705–4713.

[64] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," *Computing Research Repository*, vol. abs/1609.05143, 2016.

Florin-Cristian Ghesu is a PhD student at the Pattern Recognition Lab of the Friedrich-Alexander University Erlangen-Nuremberg, Germany and is supervised by Prof. Dr.-Ing. Joachim Hornegger. His research is conducted in close collaboration with the Medical Imaging Technologies team at Siemens Healthineers, Princeton, New Jersey, where his work is overseen by Dr. Dorin Comaniciu and Dr. Bogdan Georgescu. Holder of an elite scholarship from the German Academic Exchange Service, he has won several international awards including gold and silver medals in algorithmic competitions of the Association for Computing Machinery. He is a winner of the MICCAI 2017 Young Scientist Award and the BVM 2015 Award, the latter recognizing the best graduation thesis in the field of medical image analysis in Germany. His work focuses on developing technologies for medical image understanding.

Bogdan Georgescu is a Principal Expert Scientist with Siemens Medical Solutions USA, Inc. in Princeton, New Jersey. He received his Dipl. Engn. Degree in 1996, the M.Sc. degree in 1997 in electrical engineering from the Bucharest Polytechnic Institute, Bucharest, Romania, the M.Sc. degree in 2001 and Ph.D. degree in 2004 in computer science from Rutgers University, Piscataway, New Jersey. His current focus is providing robust solutions for advanced medical image analytics with applications to 2D/3D/4D echocardiography, 4D CT and 4D MR. His research interests are in robust computer vision, machine learning, image understanding, object detection, tracking and information fusion.

Yefeng Zheng is Principal Key Expert at Siemens Healthineers Medical Imaging Technologies in Princeton, New Jersey. He received a Ph.D. degree from University of Maryland, College Park, USA, in 2005 with a dissertation on handwritten document image analysis. Before that, he received B.E. and M.E. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1998 and 2001, respectively. After graduation, he joined Siemens Corporate Research in Princeton, New Jersey, USA. He is now a Principal Key Expert working on 2D/3D object detection and segmentation problems in medical imaging. His research interests include medical image analysis, document image analysis, pattern recognition, and computer vision. He has published over 100 papers on various top journals and conferences in the above fields. His work has been cited more than 3000 times, resulting in an h-index of 28. Furthermore, he has invented more than 50 patents (granted and pending applications), including a patent on Marginal Space Learning based cardiac chamber segmentation, which won the Thomas A. Edison Patent Award in 2011. He is a major contributor for several prestigious awards, including the Techno-College Innovation Award of the European Association for Cardio-Thoracic Surgery (EACTS) in 2010. As a co-developer of an Asian Optical Character Recognition (OCR) system, he won the National Scientific and Technological Progress Award (2nd class) of China in 2003. He is a senior member of the IEEE and an Associate Editor of IEEE Journal of Biomedical and Health Informatics.

Sasa Grbic is Staff Scientist at Siemens Healthcare in Princeton, NJ. He received a M.Sc. in Computer Science from the Graz University of Technology in 2008, and a Ph.D. in Computer Science from the Technical University in Munich in 2014. He has been conducting research towards the development of machine learning methods for detection and segmentation of cardiac structures from CT, MRI and Ultrasound. He won the young scientist award at the Medical Image Computing and Computer Assisted Interventions Conference in 2010 and was nominated as runner-up in 2011.

Andreas Maier is Head of the Pattern Recognition Lab at Friedrich-Alexander University Erlangen-Nuremberg, Germany. Born on the 26th of November 1980 in Erlangen, he studied computer science, graduated in 2005, and received his PhD in 2009. His major research subject was medical signal processing in speech data. In this period, he developed the first online speech intelligibility assessment tool - PEAKS - that has been used to analyze over 4.000 patient and control subjects so far. From 2009 to 2010, he started working on flat-panel C-arm CT as post-doctoral fellow at the Radiological Sciences Laboratory in the Department of Radiology at the Stanford University. From 2011 to 2012 he joined Siemens Healthcare as innovation project manager and was responsible for reconstruction topics in the Angiography and X-ray business unit. In 2012, he returned to the University of Erlangen-Nuremberg as head of the Medical Reconstruction Group at the Pattern Recognition Lab. In 2015 he became professor and head of the lab.

Joachim Hornegger is President of Friedrich-Alexander University Erlangen-Nuremberg. He was born in 1967, studied computer science with mathematics at Friedrich-Alexander University and completed his doctoral degree in 1996 with his thesis on statistical object modeling and recognition. From 1997 to 1998, Professor Hornegger was a guest researcher at the Massachusetts Institute of Technology (MIT) and the Computer Science Department at Stanford University. After this period abroad, he accepted a position in industry as a development engineer at Siemens Medical Solutions. In 2001, he became a manager for medical imaging in the Angiography and Radiography Systems division and in 2003 he took on overall responsibility for imaging system development. During his time at Siemens, Professor Hornegger was a guest lecturer at the universities of Erlangen-Nuremberg (1998-1999), Eichstätt-Ingolstadt (2000) and Mannheim (2000-2003). As part of his management training at Siemens, he obtained an additional degree in advanced management from Duke University. Professor Hornegger became head of the Chair of Pattern Recognition at the Faculty of Engineering and a secondary member of the Faculty of Medicine at Friedrich-Alexander University in 2005. He was Vice Dean of Computer Science from 2009 to 2011. Professor Hornegger was a Vice President of Friedrich-Alexander University and part of the Executive Board from 2011 to 2015. In this capacity he was responsible for research and young researchers. He has been President of Friedrich-Alexander University since April 1st, 2015.

Dorin Comaniciu serves as Vice President at Siemens Healthineers, having global responsibility for Medical Imaging Technologies. His team specializes in using large collections of data to build artificial intelligence applications for healthcare. His scientific contributions to medical imaging and machine intelligence translated into multiple clinical products focused on improving the quality of care, specifically in the fields of diagnostic imaging, image-guided therapy, and personalized medicine. A Top Innovator of Siemens, Dr. Comaniciu is a Fellow of the IEEE, the Medical Image Computing and Computer Assisted Intervention Society, and the American Institute for Medical and Biological Engineering. He is recipient of multiple honors, including the 2010 IEEE Longuet-Higgins Prize for fundamental contributions to computer vision. Comaniciu holds 250 patents in the areas of machine intelligence, medical imaging and computer vision. He has co-authored 300 peer-reviewed publications, and co-wrote the book Marginal Space Learning for Medical Image Analysis. His publications have 35,000 citations according to Google Scholar. A graduate of the Advanced Management Program of University of Pennsylvania's Wharton School, Comaniciu received a doctorate in electrical and computer engineering from Rutgers University and a doctorate in electronics and telecommunications from Polytechnic University of Bucharest.