Robust Multi-Scale Anatomical Landmark Detection in Incomplete 3D-CT Data

Florin C. Ghesu[1,2], Bogdan Georgescu[1], Sasa Grbic[1],
Andreas Maier[2], Joachim Hornegger[2], and Dorin Comaniciu[1]

[1] Medical Imaging Technologies, Siemens Healthineers, Princeton NJ, USA
[2] Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen, Germany
florin.c.ghesu@fau.de

Abstract. Robust and fast detection of anatomical structures is an essential prerequisite for the next-generation automated medical support tools. While machine learning techniques are most often applied to address this problem, the traditional object search scheme is typically driven by suboptimal and exhaustive strategies. Most importantly, these techniques do not effectively address cases of incomplete data, i.e., scans taken with a partial field-of-view. To address these limitations, we present a solution that unifies the anatomy appearance model and the search strategy by formulating a behavior-learning task. This is solved using the capabilities of deep reinforcement learning with multi-scale image analysis and robust statistical shape modeling. Using these mechanisms artificial agents are taught optimal navigation paths in the image scale-space that can account for missing structures to ensure the robust and spatially-coherent detection of the observed anatomical landmarks. The identified landmarks are then used as robust guidance in estimating the extent of the body-region. Experiments show that our solution outperforms a state-of-the-art deep learning method in detecting different anatomical structures, without any failure, on a dataset of over 2300 3D-CT volumes. In particular, we achieve 0% false-positive and 0% false-negative rates at detecting the landmarks or recognizing their absence from the field-of-view of the scan. In terms of runtime, we reduce the detection-time of the reference method by $15-20$ times to under 40 ms, an unmatched performance in the literature for high-resolution 3D-CT.

1 Introduction

Accurate and fast detection of anatomical structures is a fundamental step for comprehensive medical image analysis [1, 2, 5]. In particular for automatic support of clinical image reading, where the field-of-view of the acquired CT scan is typically unknown, ensuring the accurate detection of the visible landmarks and recognizing the absence of missing structures pose significant challenges. Addressing these limitations is essential to enable artificial intelligence to increase and support the efficiency of the clinical workflow from admission through diagnosis, clinical care and patient follow-up. In this context, state-of-the-art deep learning solutions based on hypothesis scanning [1] or end-to-end learning [5]

typically propose to threshold the detection confidence to handle cases of incomplete data — a suboptimal heuristic in terms of accuracy.

In this work we present a solution for robust anatomical landmark detection and recognition of missing structures using the capabilities of deep reinforcement learning (DRL) [4]. Inspired by the method introduced in [2], we choose to learn the process of finding an anatomical structure and use it as a natural mechanism to recognize its absence by signaling the divergence of search trajectories outside the image space. To increase the system robustness and avoid suboptimal local convergence, we propose to use scale-space theory [3] to enable the system to hierarchically exploit the image information. In addition, we ensure the spatial coherence of the detected anatomical structures using a robust statistical shape-model fitted with M-estimator sample consensus [7]. Based on the robust detections, we infer the vertical range of the body captured in the 3D-CT scan.

2   Background and Motivation

2.1   Challenges of 3D Landmark Detection in Incomplete Data

Deep scanning-based systems represent the main category of recent solutions [1]. Here the problem is reformulated to a patch-wise classification between positive $h \in \mathbf{H}_+$ and negative hypotheses $h \in \mathbf{H}_-$, sampled as volumetric boxes of image intensities. Alternatively, end-to-end deep learning systems based on fully convolutional architectures approach the problem by learning a direct mapping $f(\mathbf{I}) = \mathbf{M}$ between the original image $\mathbf{I}$ and a coded map $\mathbf{M}$ highlighting the locations of anatomical landmarks [5]. However, in the case of over thousands of large range 3D-CT scans at high spatial resolution (e.g. 2 mm or less), the training of such systems becomes infeasible due to the excessive memory requirements and the high computational complexity. In particular for incomplete data, all these systems share a common limitation, i.e., they rely on suboptimal and inaccurate heuristics such as probability thresholding to recognize whether an anatomical landmark is visible in the field-of-view of the 3D scan.

2.2   Learning to Search using Deep Reinforcement Learning

A different perspective on the general problem of landmark detection in 3D data is presented in [2]. The task is reformulated as an intrinsic behavior learning problem which asks the question of *how to find a structure?* given image evidence $\mathbf{I} : \mathbb{Z}^3 \to \mathbb{R}$. To model the system dynamics and enable the navigation in image space, a Markov Decision Process (MDP) [6] $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$ is defined, where: $\mathcal{S}$ represents a finite set of states over time with $s_t \in \mathcal{S}$ being the state of the agent at time t – a constrained axis-aligned box of image intensities centered at position $\boldsymbol{p}_t$ in image space; $\mathcal{A}$ represents a finite set of actions allowing the agent to navigate voxel-wise within the environment (± 1 voxels in each direction); $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a stochastic transition function, where $\mathcal{T}_{s,a}^{s'}$ describes the probability of arriving in state s' after performing action a in

state s; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a scalar reward function to drive the behavior of the agent, where $\mathcal{R}_{s,a}^{s'} = \|\boldsymbol{p}_t - \boldsymbol{p}_{GT}\|_2^2 - \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_{GT}\|_2^2$ denotes the expected distance-based reward for transitioning from state s to state s', i.e., $\boldsymbol{p}_t$ to $\boldsymbol{p}_{t+1}$ while seeking the ground-truth position $\boldsymbol{p}_{GT}$ of the landmark; and $\gamma \in (0, 1)$ is the discount-factor controlling future versus immediate rewards [2].

In this context the optimal action-value function $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined, which quantifies the maximum expected future reward of an optimal navigation policy π^* starting in s with action a: $Q^*(s, a) = \max_\pi \mathbb{E}[R_t | s_t = s, a_t = a, \pi]$. A recursive formulation of this function based on the dynamic state-graph defines the so called Bellman criterion [6]: $Q^*(s, a) = \mathbb{E}_{s'}(r + \gamma \max_{a'} Q^*(s', a'))$. Using a deep neural network with parameters θ to approximate this complex non-linear function $Q^*(s, a) \approx Q(s, a; \theta)$, one can learn optimal trajectories in image-space that converge to the sought anatomical structures with maximum reward [2,4,8].

Learning the navigation policy replaces the need for exhaustive and sub-optimal search strategies [1,5]. More importantly, this formalism can elegantly address the question of missing structures with trajectories that leave the image space — a natural ability of the system in contrast to explicit post-processing heuristics. However, in the context of large incomplete volumetric 3D scans this approach suffers from several inherent limitations. The first is related to the spatial coverage of the acquired state descriptor $s \in \mathcal{S}$. Acquiring limited local information improves the sampling efficiency at the cost of local optima. On the contrary, extracting a very large context to represent the state, poses significant computational challenges in the 3D space. This indicates the inability to properly exploit the image information at different scales. Secondly, the system fails to exploit the spatial distribution of the landmarks to further increase robustness.

3   Proposed Method

To address these limitations, we propose to use scale-space theory [3] and robust statistical shape modeling for multi-scale spatially-coherent landmark detection.

3.1   A Discrete Scale-Space Model

In general, the continuous scale-space of a 3D image signal $\mathbf{I} : \mathbb{Z}^3 \to \mathbb{R}$ is defined as: $L(x; t) = \sum_{\xi \in \mathbb{Z}^3} T(\xi; t) \mathbf{I}(x - \xi)$, 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 kernel-family. The main properties of such a scale-space representation are the non-enhancement of local extrema and implicitly the causality of structure across scales [3]. These properties are essential for the robustness of a search process, starting from coarse to fine scale. We propose to use a discrete approximation of the continuous space L while best preserving these properties. We define this discrete space as:

$$L_d(t) = \Psi_\rho(\sigma(t-1) * L_d(t-1)), \qquad (1)$$

where $L_d(0) = \mathbf{I}$, $t \in \mathbb{N}_0$ denotes the discrete scale-level, σ represents a scale-dependent smoothing function and Ψ_ρ denotes a signal operator reducing the spatial resolution with factor ρ using down-sampling [3].
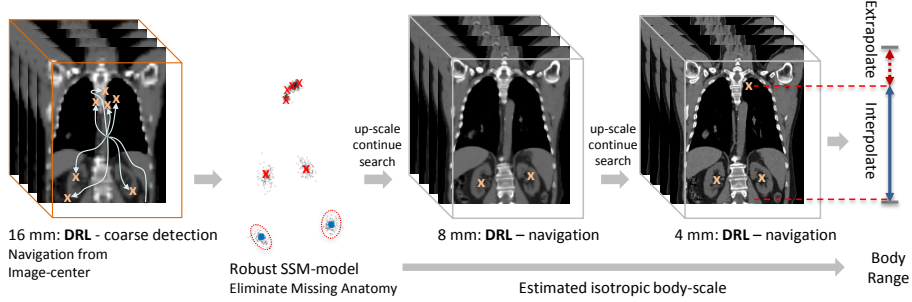
Fig. 1. Visualization of the complete system pipeline.

3.2   Learning Multi-Scale Search Strategies

Assuming w.l.o.g. a discrete scale-space of M scale-levels with $\rho = 2$, we propose a navigation model across scales — starting from the coarsest to the finest scale-level. For this we redefine the optimal action-value function Q^* by conditioning the state-representation s and model parameters θ on the scale-space L_d and the current scale $t \in [0, \cdots, M-1]$: $Q^*(s, a \mid L_d, t) \approx Q(s, a; \theta_t \mid L_d, t)$. This results in M independent navigation sub-models $\boldsymbol{\Theta} = [\theta_0, \theta_1, \cdots, \theta_{M-1}]$, one for each scale-level. Each model is trained on each individual scale-level as proposed in [2], i.e., by optimizing the Bellman criterion on each level $t < M$:

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

with $i \in \mathbb{N}_0$ denoting the training iteration. The reference estimate y is determined using the update-delay [4] technique: $y = r + \gamma \max_{a'} Q\left(s', a'; \bar{\theta}_t^{(i)} \mid L_d, t\right)$, where $\bar{\theta}_t^{(i)} := \theta_t^{(j)}$ represents a copy of the model parameters from a previous training step $j < i$. This significantly increases the training stability [2].

The detection workflow is defined as follows: the search starts in the image center at the coarsest scale-level $M-1$. Upon convergence the scale-level is changed to $M-2$ and the search continued from the convergence-point at $M-1$. The same process is repeated on the following scales until convergence on the finest scale. We empirically observed that optimal trajectories converge on minimal (oscillatory) cycles. As such, we define the convergence-point as the center of gravity of this cycle. The search-model $Q(\cdot, \cdot; \theta_{M-1} \mid L_d, M-1)$ is trained for global convergence while the models on any of the following scales $t < M-1$ are trained in a constrained range around the ground-truth. This range is robustly estimated from the accuracy upper-bound on the previous scale $t+1$. Note that the spatial coverage of a fixed-size state $s \in \mathcal{S}$ is increasing exponentially with the scale. This multi-scale navigation model allows the system to effectively exploit the image information and increase the robustness of the search (see Figure 1).

Missing Landmarks: We propose to explicitly train the global search model

θ_{M-1} for missing landmarks to further improve the accuracy for such cases. Assuming the CT-scans are cut only horizontally, the system is trained to constantly reward the trajectories to leave the image space through the correct volume border. For this we require for each missing landmark a ground-truth annotation on whether it is outside above the field of view, or below.

3.3   Robust Spatially-Coherent Landmark Detection

To ensure the robust recognition of missing anatomical structures and outliers we propose to model the spatial distribution of the considered anatomical landmarks using robust statistical shape modeling. This step constrains the output of the global search model θ_{M-1} (see the complete pipeline visualized in Figure 1). Assuming a complete set of N anatomical landmarks, we normalize the distribution of these points on all complete training images to zero mean and unit variance. In this space, we model the distribution of each individual landmark $i \in [0, \cdots, N-1]$ as a multi-variate normal distribution $\boldsymbol{p}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are estimated using maximum likelihood. This defines a mean shape-model for the landmark set, defined as $\boldsymbol{\mu} = [\boldsymbol{\mu}_0, \cdots, \boldsymbol{\mu}_{N-1}]^\top$. Given an unseen configuration of detected points at scale $M-1$ as $\tilde{\boldsymbol{P}} = [\tilde{\boldsymbol{p}}_0, \cdots, \tilde{\boldsymbol{p}}_N]^\top$, one can approximate $\tilde{\boldsymbol{P}}$ with a translated and isotropic-scaled version of the mean model using least linear squares as: $\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}=[t,s]} \|\tilde{\boldsymbol{P}} - \boldsymbol{t} - s\boldsymbol{\mu}\|_2^2$. However, for the case of incomplete data the cardinality of $|\tilde{\boldsymbol{P}}| \leq N$. In addition, outliers can corrupt the data. To enable the robust fitting of the shape-model, we propose to use M-estimator sample consensus [7]. Based on random 3-point samples from the set of all triples $I_3(\tilde{\boldsymbol{P}})$ one can obtain the mean-model fit $\hat{\boldsymbol{\omega}} = [\boldsymbol{t}, s]$. The target is to optimize the following cost function based on the redescending M-estimator [7] and implicitly maximize the cardinality of the consensus set $\hat{\mathcal{S}}$:

$$\hat{\mathcal{S}} \leftarrow \arg\min_{S \in I_3(\tilde{\boldsymbol{P}})} \sum_{i=0}^{|\tilde{\boldsymbol{P}}|} \min\left[\frac{1}{Z_i} \left(\phi(\tilde{\boldsymbol{p}}_i) - \boldsymbol{\mu}_i\right)^\top \boldsymbol{\Sigma}_i^{-1} \left(\phi(\tilde{\boldsymbol{p}}_i) - \boldsymbol{\mu}_i\right), 1 \right], \qquad (3)$$

where $\phi(\boldsymbol{x}) = \frac{\boldsymbol{x} - \boldsymbol{t}}{s}$ is a projector to normalized shape-space with the estimated fit $\hat{\boldsymbol{w}} = [\boldsymbol{t}, s]$ on set S. The normalization coefficient $Z_i \in \mathbb{R}_+$ defines an oriented ellipsoid – determining the outlier-rejection criterion. We use the χ_3^2-distribution to select Z_i such that less than 0.5% of the inlier-points are incorrectly rejected.

Detect Outliers and Reset: Enforcing spatial coherency not only corrects for diverging trajectories by re-initializing the search, but also significantly reduces the false-negative rate by correcting for border cases. These are landmarks very close to the border $(< 2\,\mathrm{cm})$, falsely labeled as missing at scale $M-1$.

Scan-range estimation: The robust fitting of the shape-model also enables the estimation of the body-region captured in the scan. We propose to model this as a continuous range within normalized z-axis, to ensure consistency among different patients. For a set of defined landmarks $\boldsymbol{P}$ in normalized shape-space, the point $\boldsymbol{p}_{min} = \min_{\boldsymbol{p}_i \in \boldsymbol{P}}[\boldsymbol{p}_i^z]$ determines the 0%-point, while the point $\boldsymbol{p}_{max} = \max_{\boldsymbol{p}_i \in \boldsymbol{P}}[\boldsymbol{p}_i^z]$ the 100%-point. Assume for a given set of landmarks $\tilde{\boldsymbol{P}}$ that the
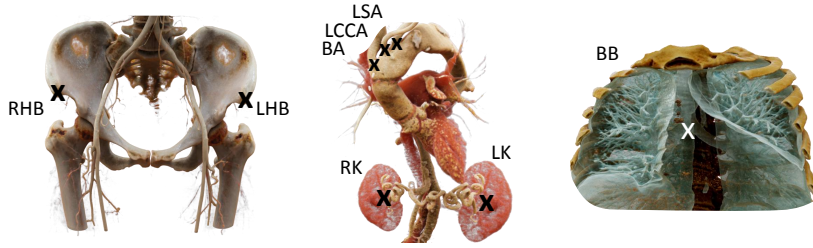
Fig. 2. The anatomical landmarks used for evaluation. These are the front corner of the left (LHB) and right hip bones (RHB), the center of the left (LK) and right kidneys (RK), the bifurcation the left common carotic artery (LCCA), brachiocephalic artery (BA) and left subclavian artery (LSA) and the bronchial bifurcation (BB).

fitted robust subset is represented by $\hat{P} \subseteq \tilde{P}$. Using our definition of range, the span of the point-set $\hat{P}$ can be determined between $0\% - 100\%$ in normalized shape-space. This also allows the linear extrapolation of the body-range outside the z-span of the point-set $\hat{P}$ (more details follow in Section 4).

4   Experiments

Dataset: For evaluation we used a dataset of 2305 3D-CT volumes from over 850 patients. We determined a random split in 1887 training volumes and 418 test volumes, ensuring that all scans from each patient are either in the training or the test-set. We selected a set of 8 anatomical landmarks with annotations from medical experts (see Figure 2). Each volume was annotated with the location of all visible landmarks. To allow the fitting of the shape-model, we selected scans with at least 4 annotations (this is not a limitation since our target for future work is to cover more than 100 landmarks). This resulted in a $70\% - 30\%$ split of the annotations for each landmark into training and testing. Over the entire dataset the distribution of visible−missing landmarks was approximately as follows: $80\% - 20\%$ for kidneys, $60\% - 40\%$ for hip-bones, and $50\% - 50\%$ for the rest. We refer to false-positive (FP) and false-negative (FN) rates to measure the accuracy in detecting landmarks or recognizing their absence from the scan.

System Training: A scale-space of 4 scales was defined at isotropic resolutions of $2 - 4 - 8$ and $16\,\text{mm}$. For the kidney center the fine resolution was set to $4\,\text{mm}$, given the higher variability of the annotations. For each scale and landmark the network structure was the same: conv-layer (40 kernels: $5 \times 5 \times 5$, ReLU), pooling ($2 \times 2 \times 2$), conv-layer (58 kernels: $3 \times 3 \times 3$), pooling ($2 \times 2 \times 2$) and four fully-connected layers ($512 \times 256 \times 6$ units, ReLU). The compact model-size under 8 MB per scale-level enables efficient loading and evaluation. Also the meta-parameters for training were shared across scales and landmarks: training-iterations (750), episode-length (1500), replay-memory size (10^5), learning rate (0.25×10^{-2}), batch-size (128) and discount-factor $\gamma = 0.9$. The dimensionality of the state was also fixed across scales to $25 \times 25 \times 25$ voxels. Recall that

Table 1. Comparison with state-of-the-art deep learning [1]. Accuracy is in **mm**.

		Anatomical Structures							
		LK	RK	LHB	RHB	LCCA	BA	LSA	BB
FP-rate	MSDL [1]	1.5%	1.5%	1.1%	1.2%	1.0%	1.0%	1.1%	1.0%
	Ours	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**
FN-rate	MSDL [1]	13.9%	9.4%	1.2%	0.4%	10.8%	11.3%	7.2%	4.9%
	Ours	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**
Mean	MSDL [1]	6.17	6.36	4.92	3.66	4.78	5.05	5.25	5.10
	Ours	6.83	6.98	3.61	2.63	4.02	4.26	4.23	4.07
Median	MSDL [1]	5.64	5.80	4.70	3.44	4.17	4.54	4.62	4.53
	Ours	6.32	6.63	2.83	2.49	2.86	3.46	3.21	3.77
STD	MSDL [1]	3.32	3.06	2.09	1.83	3.30	3.02	3.51	2.82
	Ours	3.52	3.83	2.08	1.53	3.33	2.97	3.37	2.16

on all scales except $M-1$ the training is performed in a constrained image range around the ground-truth $\boldsymbol{p}_{GT} \pm \boldsymbol{r}$. Depending on scale and landmark: $\boldsymbol{r} \in [-12, 12]^3$ voxels. The training time for one landmark averages to 4 hours on an Nvidia Titan X GPU. We train all models in a 16-GPU cluster in 2.5 hours.

Robust Multi-Scale Navigation: Given trained multi-scale models for each landmark: $\boldsymbol{\Theta}_0, \cdots, \boldsymbol{\Theta}_8$, the search starts on the lowest scale in the center of the scan. Let $\tilde{\boldsymbol{P}}$ be the output of the navigation sub-models on coarsest scale. Robust shape-model fitting was performed on $\tilde{\boldsymbol{P}}$ to eliminate outliers and correct for misaligned landmarks to a robust set $\hat{\boldsymbol{P}}$. This reduced the FP and FN rates from around 2% to under 0.5%. Applying the training range $\boldsymbol{r}$ to bound the navigation on the following scales $[M-2, \cdots, 0]$, we empirically observed that the shape-constraint was preserved while the FP-/FN-rate were reduced to zero.

Result Comparison: In contrast to our method, the reference solution proposed in [1] uses a cascade of sparse deep neural networks to scan the complete image space. Missing structures are detected using a fixed cross-validated threshold on the hypothesis-probability. The operating point was selected to maintain a FP-rate of less than 1.5%. Table 1 shows the obtained results. Our method significantly outperforms [1] in recognizing the presence/absence of structures (see FP and FN rates). In terms of accuracy, the improvement is statistically significant (paired t-Test p-value $< 10^{-4}$), averaging $20-30\%$ on all landmarks except the kidneys. The increased apparent performance on the kidney center of the method [1] might be explained by the high FN-rate as well as the robust candidate aggregation [1], which accounts for the high variability of the annotations. *Please note*: A comparison with the method [2] is not possible on this large volumetric data. Training the detector only on the finest scale, as proposed in [2], is only possible within a limited range around the ground-truth (e.g. ± 15 cm). This highlights the importance of using a scale-space model and robust shape M-estimation, which enable training in large-range incomplete 3D data.

Runtime: Learning the multi-scale search trajectory leads to real-time detection. With an average speed of $35 - 40\,$ms per landmark, our method is $15 - 20$ times faster than MSDL [1] which achieved an average speed of around $0.8\,$sec.

Body-region Estimation: We defined a continuous range-model based on the landmark set with the LHB corner at 0% and the LCCA bifurcation at 100%. The levels of the remaining landmarks were determined in normalized shape-space using linear interpolation. Using the robust detections $\hat{\boldsymbol{P}}$ as reference range, we extrapolated the body-range above the LCCA or below the hip bones. Qualitative evaluation shows that the scan in Figure 1 extends from 21.3% to 109.0%.

5   Conclusion

In conclusion, this paper presents an effective approach for multi-scale spatially coherent landmark detection for incomplete 3D-CT data. Learning multi-scale search trajectories and enforcing spatial constraints ensure high robustness and reduce the false-positive and false-negative rates to zero, significantly outperforming a state-of-the-art deep learning approach. Finally we demonstrate that the detected landmarks can be used to robustly estimate the body-range.

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

References

1. Ghesu, F.C., Krubasik, E., Georgescu, B., Singh, V., Zheng, Y., Hornegger, J., Comaniciu, D.: Marginal Space Deep Learning: Efficient architecture for volumetric image parsing. IEEE TMI 35(5), 1217–1228 (2016)
2. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An artificial agent for anatomical landmark detection in medical images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016, Part III. LNCS. vol. 9902, pp. 229–237. Springer International Publishing (2016)
3. Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, Norwell, MA, USA (1994)
4. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature 518(7540), 529–533 (2015)
5. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using cnns. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016, Part II. LNCS. vol. 9901, pp. 230–238. Springer International Publishing (2016)
6. Sutton, R.S., Barto, A.G.: Introduction to Reinforcement Learning. MIT Press, 1st edn. (1998)
7. Torr, P.H.S., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Elsevier CVIU 78, 138–156 (2000)
8. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning 8(3), 279–292 (1992)