Accurate localization of inner ear regions of interests using deep reinforcement learning

Radutoiu, Ana-Teodora; Patou, Francois ; Margeta, Jan ; Paulsen, Rasmus Reinhold; Lopez Diez, Paula

Published in:
Proceedings of the 13th International Workshop on Machine Learning in Medical Imaging

Link to article, DOI:
10.1007/978-3-031-21014-3_43

Publication date:
2022

Document Version
Peer reviewed version

Accurate localization of inner ear regions of interests using deep reinforcement learning

Ana-Teodora Radutoiu[1], François Patou[2], Jan Margeta[3], Rasmus Reinhold Paulsen[1], and Paula López Diez[1]

[1] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
[2] Oticon Medical, Research & Technology group, Smørum, Denmark
[3] KardioMe, Research & Development, Nova Dubnica, Slovakia

Abstract. We propose a novel method for automatic ROI extraction. The method is implemented and tested for isolating the inner ear in full head CT scans. Extracting the ROI with high precision is in this case critical for surgical insertion of cochlear implants. Different parameters, such as CT equipment, image quality, anatomical variation, and the subject's head orientation during scanning make robust ROI extraction challenging. We propose to use state-of-the-art communicative multi-agent reinforcement learning to overcome these difficulties. We specify landmarks specifically designed to robustly extract orientation parameters such that all ROIs have the same orientation and include the relevant anatomy across the dataset. 140 full head CT scans were used to develop and test the ROI extraction pipeline. We report an average overall estimated error for landmark localization of 1.07 mm. Extracted ROI presented an intersection over union of 0.84 and a Dice similarity coefficient of 0.91.

Keywords: Region of interest · Deep reinforcement learning · Computed tomography · Inner ear · Landmarks · Orientation.

1 Introduction

Automatic region of interest (ROI) detection in medical images is a challenging task as medical images generally present high variability between individuals, scanners, and image acquisition and postprocessing protocols. ROI extraction is a necessary step for almost all medical image analysis pipelines. It is also vital for subsequent image processing tasks that rely on input stability. Accurate ROI extraction can not only improve retrieval efficiency but can also help to classify more easily pathological signs within a reduced region especially when the anatomy is particularly challenging for either the clinicians or the processing software to interpret [13].

Cochlear implants (CI) are used to restore the hearing capacities of patients who suffer from severe to profound hearing loss. CIs are common for infants with congenital deafness. Clinicians evaluate each patient using computed tomography (CT) images, the head orientation is important to assess each case and

successfully obtain the relevant measurements for accurate surgical planning. Therefore we had developed an approach to automatically locate this region in clinical full head CT images.

Initially, ROI detection methods were based on bounding boxes from handcrafted features [11,3]. Nowadays, the most common technique is deep learning, where a majority of methods in the literature are designed for 2D medical images [2,13] and other approaches used 2D methods to locate ROIs in 3D [5,16,8]. These methods fail to use the third dimensionality of CT images and need a big amount of annotated data that faithfully represents the anatomical variability and struggle when the anatomy is abnormal. We chose to use a deep reinforcement learning (DRL) based approach that uses landmarks (easier and faster to annotate) to automatically extract the inner ear ROI from CT images. DRL has been successful showing outstanding performance for similar tasks of landmark localization in medical images [1,15]. Other DRL approaches include Navarro et. al. [10], this paper proposes single reinforcement learning agents for organ localization in the torso. They succeeded in effectively finding a ROI around desired organs with a relatively small amount of data. We chose to use a landmark-based approach which we consider will be more robust and could potentially provide more explainability or even help detect abnormalities from an early processing stage [6,13].

2 Data

This study uses 140 full head CT scans from the CQ500 dataset [4] which corresponds to 102 different patients. The dataset has been provided by the Centre for Advanced Research in Imaging, Neurosciences and Genomics (CARING), New Delhi, India [4]. The CT scans are taken from several radiology centers in New Delhi and are collected using various equipment models [4]. All used scans are resampled to have the isotropic voxel spacing 0.5 mm. The image dimensions vary significantly within our dataset. On average the dimensions are $475 \times 475 \times 323$, but the dimensions $x \in [400; 576]$, $y \in [400; 576]$ and $z \in [128; 730]$. All scans are manually labeled with the chosen landmarks using the software 3D Slicer [7]. All the annotations are made public and can be found in `https://github.com/AnaTeodoraR/annotations.git`.

Choosing relevant landmarks is necessary to characterize the inner ear orientation. The landmarks must be uniquely defined within their structure, so they are easily differentiated from other anatomical points nearby. Eleven landmarks are chosen in total, five assigned for each inner ear ROI and one common for both. The five landmarks for each ROI are the same anatomical points but located on their respective side of the CT scan. All landmarks are associated with a number; Numbers 1-5 are in connection to the right ROI, 6-10 for the left, and 11 is common. Two landmarks are within the inner ear; cochlear apex (nr. 1 and 6) and superior semi-circular canal peak (nr. 4 and 9). Additionally, two landmarks in the cochlea nerve; the midpoint below the base of the cochlea (nr. 2 and 7) and of both the CN and FN further down (nr. 3 and 8). To aid in

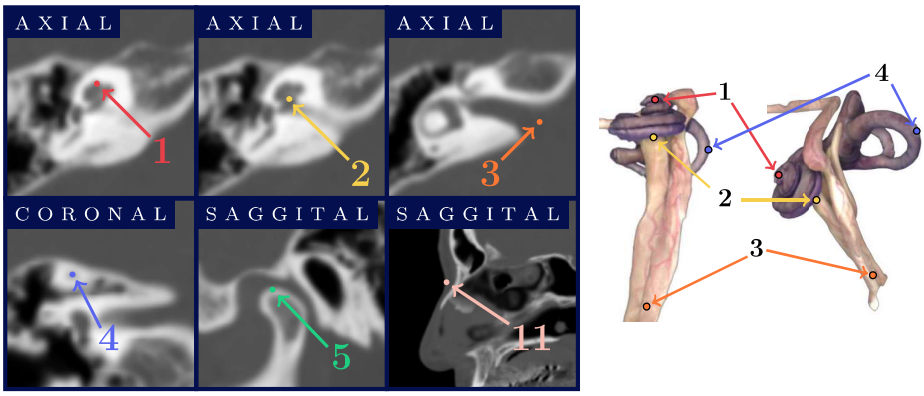


Fig. 1: Landmarks 1-5 and 11 are shown on CT scan nr. 6 and additionally landmarks 1-4 are shown on a 3D representation of an inner ear (edited from [14]). Landmarks 1-5 are for the right inner ear ROI, but the corresponding landmarks 6-10 are placed at the same anatomical points on the left side of the head.

obtaining a similar orientation the remaining landmarks are; At the top of the condyle in the mandible where the temporomandibular joint starts (nr. 5 and 10) and lastly the nasion (nr. 11).

3 Methods

The strategy used for detecting the landmarks is DRL specifically the communicative multi-agent reinforcement learning (C-MARL) model proposed by Leroy et al. [9]. This model uses DRL with multiple agents that communicate and are based on a deep Q-learning network (DQN). The environment is defined as the entire 3D medical scan while an agent is moved voxel by voxel. A DQN is used to predict the optimal Q values given a certain state. As input the network takes states - a 3D patch centered around the agent - and outputs the Q values for each of the possible actions - 3D movement (up, down, left, right, forward, and backward) [9]. The architecture for two agents is shown in figure 2. All the agents share the weights of the convolutional layers (implicit communication) while each agent has its own fully connected (FC) layers only sharing the average output from each FC layer (explicit communication) [9]. The agents use multi-scale which enables them to have a spatially higher view of the image in its state. In our implementation, the agents used four scales including the isotropic voxel spacings 3, 2, 1, and 0.5 mm for the patch to represent the state. Initially, the agents will observe the states with the coarsest spacing, but when they start oscillating the spacing will decrease to the next, finer, resolution. The final model uses 11 agents, one per landmark, thus resulting in 11 FC layers and a discount rate of 0.8. The data is divided into training, validation, and test sets

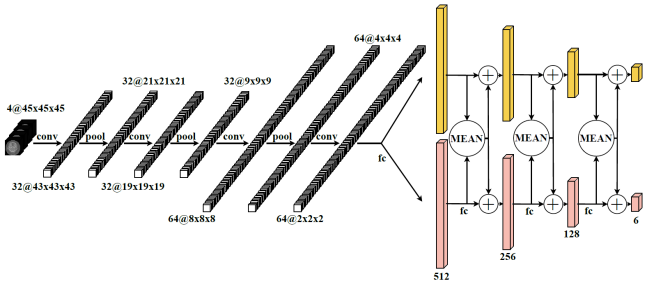


Fig. 2: Visualization of the CNN architecture, when using two agents. The fully connected layers for each agent are colored differently. Edited from [9].

each being 112, 14, and 14 CT scans respectively ($\approx 80\%$, $\approx 10\%$, and $\approx 10\%$). The model was trained on a 12 GB GPU for five days. The training uses an ϵ-greedy strategy ($\epsilon \in [0.1; 1[$). The model which has the best performance on the validation set is further used for testing.

The aim of this study is to extract the two ROIs with similar orientation in a two-step approach. The first step aims to rotate the images to align all heads. The landmarks are used to define the relevant rotation angles that characterize the head orientation. The second step is a regular axis-aligned crop of the inner ear on the newly rotated image. To rotate a 3D medical image, three rotation angles denoted α, β, and γ are used, also known as yaw, roll, and pitch respectively. They describe the head movement in 3D as shown in figure 3.

The rotation angles α and β have the potential to be estimated based on the assumption that the corresponding left and right anatomical structures are symmetrical as is the case for the ear anatomy [12]. Technically, α and β are found by projecting the line between the same two anatomical points onto the axial plane and coronal plane respectively, and finding the deviation from the horizontal line, see figure 3a and 3b. Four independent estimations of these angles are found using the landmark pairs (1,6), (2,7), (3,8), and (5,10) and a median of these estimations is applied as the final rotation angle. Likewise, γ is found assuming landmark 5 or 10 and landmark 11 are at an angle of $\theta = 20°$ when projected on the sagittal plane as seen in figure 3c. Humans are anatomically different but we estimate 20° to be a good estimation of this anatomy. Finally, γ is defined as the angle which corrects the difference between θ and the estimated ones using the landmarks (figure 3c). Once more two different γ values can be estimated, so a median of these will be used as the final γ.

After obtaining the three angles for a single 3D medical scan, the image is rotated accordingly using three individual rotation transformations. The ROI is extracted on the rotated image as a 3D axis-aligned crop from a center point with a customized size. The center point is set to the middle point between two landmarks (landmarks 1 and 2 for the right, 6 and 7 for the left). The radii, r_x, r_y and r_z, of the 3D box, are found as the furthest distance to a set of landmark's respective x, y, and z coordinates (landmarks 1, 2, 3, and 4 for the right ROI,

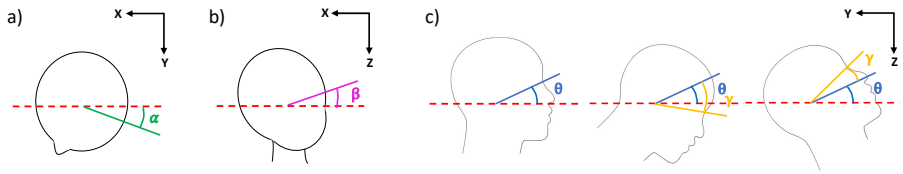


Fig. 3: Sketch of α, β, and γ where the rotation angles have been exaggerated. The angle $\theta = 20°$ describes the desired head position for γ. a) Axial view b) Coronal view c) Sagittal view and γ for the two different orientations.

while 6, 7, 8, and 9 for the left). Moreover, $r_x = r_y = \max(r_x, r_y)$, which results in the ROIs having quadratic axial slices.

A robustness analysis is made by applying a grid of different rotations to an image. The chosen image is the one where the sum of absolute calculated angles (from the annotations) is the smallest, hence the image closest to the reference pose. This image has the estimated angles; $\alpha = 0.19°$, $\beta = 1.08°$, $\gamma = 0.70°$, hence its sum is $1.97°$. For our robustness analysis, we exhaustively sampled all 3D rotations on a uniform grid: $\alpha', \beta' \in \{-15, -10, -5, 0, 5, 10, 15\}$ and $\gamma' \in \{-20, -10, 0, 10, 20, 30\}$, where α', β', and γ' are the applied rotation along the third, second and first axis respectively. This results in $7 \cdot 7 \cdot 6 = 294$ manually rotated images. For all generated images their landmarks will be predicted using the C-MARL model. Furthermore, the angle calculation method is applied using the predicted landmarks and compared to the applied rotation. Consequently, the applied rotation will have the opposite sign of the predicted (if predicted correctly).

4 Results

Different C-MARL models have been trained and it was empirically found that using a discount rate of 0.8 and a single agent per landmark, performed best. Using multiple agents per landmark showed only to have a relevant influence on detecting landmark 11. The results of the final model on the test set can be seen in table 1. Observing the table, the model performs particularly well at detecting landmarks within the inner ear and cochlear nerve (landmarks 1, 2, 3, 4, 6, 7, 8, and 9). For these eight landmarks, the mean distance error is below one millimeter. Depending on the landmark, the estimated errors vary between $0.79 - 2.11$ mm.

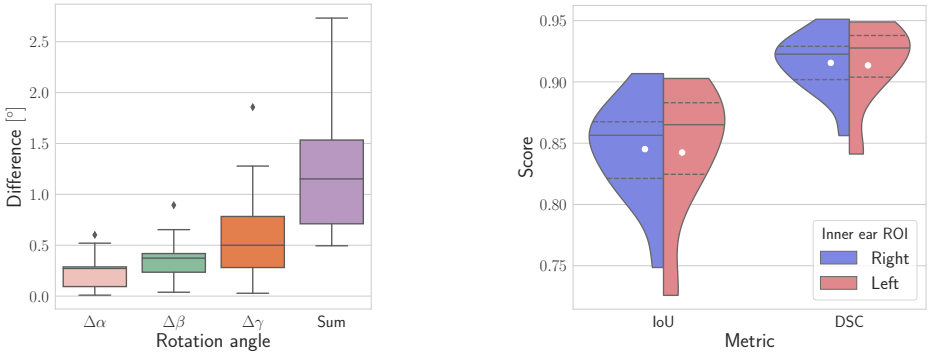
Figure 4a shows box-plots of the rotation angle differences across the test images. Here the rotation angle difference is between the predicted angles and the ones found from the landmark annotations. The maximum deviation of the predicted angle from the estimated (calculated from the annotated landmarks) for α and β is below $1°$ and the difference for γ is capped at $1.8°$, with its upper quartile is below $1°$. Figure 4a additionally illustrates a box-plot of the sum of the three differences in a test image which ranges from $0.36° - 2.68°$.

Now we evaluate the ROI extraction performance. Figure 4b illustrates the intersection over union (IoU) and Dice similarity coefficient (DSC) for both ROIs, additionally, we show the distribution between right and left ROIs. The IoU and DSC are found by comparison with an estimated ROI using the annotated landmarks. Observing Figure 4b, a small difference between the right and left ROI's prediction exists. The scores for the right and left ROI appear unimodal having a single peak.

The results of the robustness analysis are shown in figure 5. Observing the two figures, each 3D point represents a manually rotated image whose three applied rotation angles (α', β', and γ') are its x, y, and z coordinates. The points are color-coded depending on the performance. Looking at figure 5a, many of the rotated images have an estimated error of $1-2$ mm (purple color). A general tendency of more purple colors for smaller and negative γ' is observed and a decline in performance is seen as γ' increases. The lowest accuracy tends to be gathered at highly negative α', positive β', and positive γ' rotations. Similar

Landmark	1	2	3	4	5	6	7	8	9	10	11	Overall
mean(d_{Error}) [mm]	0.84	0.79	0.87	0.87	1.63	0.83	0.85	0.82	0.99	2.11	1.24	1.07
σ_{Error} [mm]	0.38	0.42	0.64	0.44	1.14	0.28	0.39	0.56	0.31	1.04	0.64	0.75
< 1 mm [%]	64.3	78.6	78.6	64.3	35.7	78.6	57.1	85.7	57.1	7.1	50	-

Table 1: The estimated error (mean distance error, d_{Error}) and standard deviation (σ_{Error}) on the test set. The model predicts the landmarks three times on each test image and a median is used as the final prediction. The last row shows the percentage of detections below one millimeter.



(a) Box-plots of angle differences.

(b) Violin-plots of ROI scores.

Fig. 4: **Left:** Box-plot of the angle differences on the test images (both the individual ones and their sum in an image). **Right:** Violin-plot of IoU and DSC on the test images. The line represents the median, the stripes the upper and lower quartile, and the white dot the mean.

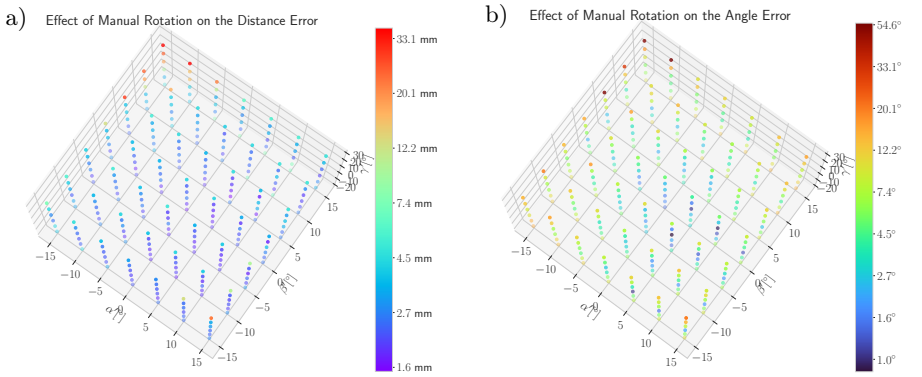


Fig. 5: Each point represents an image that has been manually rotated, the rotation angles used are illustrated as its x, y, and z coordinates. The colors illustrate the estimated error for a landmark (right) and the sum of the angle prediction errors (left). The coloring has a logarithmic scale.

patterns can be observed in the corresponding figure for the angle error (figure 5b). The error here is the sum of differences between the predicted rotation angles and the applied (with opposite sign). Only a few rotated images have an angle error below 2° (blue and black colors), while the general error tends to be around $4 - 10°$ (green, lime, and yellow colors).

5 Discussion

We found the landmark accuracy shown in this work sufficient for the ROI extraction task. Studying the performance of the individual landmarks, a few points can be made. The agents looking for landmarks 1-4 (and 6-9) are searching close by one another, and potentially benefiting more from the communication. Thus, a reason as to why the highest precision for landmark localization is achieved for these eight landmarks. Interestingly, landmark 11 is isolated from the rest, hence why the single agent has more trouble detecting it. Landmarks 5 and 10 emerge as the most difficult to locate. These landmarks mark the peak of the condyle where the jaw joint starts. The peak is a slightly curved surface, moreover, if a patient has a rotating head when taking the CT scan, other parts of the mandible become visible at the same axial slice. These matters make landmarks 5 and 10 difficult to place, both for the annotator and agents.

Multi-scale is of importance [1], especially if agents are looking for a small structure within a big image. It is important that the agents quickly get an idea of which region the landmark is in. The multi-scale enables the agents to view a larger section of the image and quickly narrow down which region their landmark is located. We implemented 4 levels of multi-scale that were sufficient for our task. The preciseness of the landmark localization influences the angle prediction. Moreover, γ is predicted using the landmarks with the

lowest localization precision (landmarks 5, 10, and 11), why this rotation angle has the lowest prediction performance of the three (figure 4a).

Regarding the robustness analysis in figure 5, negative γ' corresponds to rotating the head downwards in the sagittal plane. This is a more common position of the head, why the model performs better at these rotations. Additionally, $\pm 10^\circ$ with α' and β' further worsens the performance, however, it is also less common for a patient to have these rotations when taking a CT scan. Thus, the results of the robustness analysis reveal that rotations that mirror usual head positions perform best. Data augmentation could potentially help the model to be more robust with extreme head rotations.

Considering the ROI extraction, the overall predictions have on average an IoU score of 0.84 and DSC of 0.91. Since the IoU penalizes false positives and false negatives more than the DSC, it explains the lower scores observed for the IoU. Additionally, the regions are not all fixed to be the same size, so it might be the case that a predicted ROI is contained within an estimated (or oppositely). These cases result in a lower IoU and DSC. We compared our ROI extraction method with another state of the art method as seen in table 2. Table 2 compares the highest average IoU (the liver) achieved by Navarro et. al. [10] with the left and right ROI results. Both inner ear ROIs have on average a higher IoU than the liver ROI. Our presented method is specifically designed to extract the inner ears with likewise orientation. On the other hand, Navarro et. al. [10] strives to achieve a general axis-aligned ROI detection framework (using DRL) for organs in the torso.

6 Conclusion

This study successfully used a DRL framework for landmark detection in full head CT scans, and utilize it for ROI extraction of the inner ears. Landmarks were localized with an estimated error between $0.78 - 2.11$ mm (on average 1.07 mm) within this difficult anatomical structure. The defined rotation angles gave the ROIs the desired orientation. Two ROIs were extracted from the detected landmarks with an overall average IoU of 0.84 and DSC of 0.91. The method outperforms other DRL approaches for ROI detection as is the proposed by Navarro et. al. [10]. Through this study, we explored the capability of implementing C-MARL for predicting fine structures in full head CT scans. This paves the way for analysis of inner ear ROIs for surgical use.

	Navarro et. al. [10]	This model right ROI	This model left ROI
Avg IoU	0.80	0.836	0.835

Table 2: Comparing the average IoU for the left and right ROI with the average IoU for the liver (best organ result) from Navarro et. al.

References

1. Alansary, A., Oktay, O., Li, Y., Folgoc, L.L., Hou, B., Vaillant, G., Kamnitsas, K., Vlontzos, A., Glocker, B., Kainz, B., Rueckert, D.: Evaluating reinforcement learning agents for anatomical landmark detection. Medical Image Analysis **53**, 156–164 (2019). https://doi.org/https://doi.org/10.1016/j.media.2019.02.007
2. Bi, L., Kim, J., Kumar, A., Fulham, M., Feng, D.: Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation. The Visual Computer **33**(6), 1061–1071 (2017)
3. Campadelli, P., Casiraghi, E., Esposito, A.: Liver segmentation from computed tomography scans: a survey and a new algorithm. Artificial intelligence in medicine **45**(2-3), 185–196 (2009)
4. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Development and validation of deep learning algorithms for detection of critical findings in head ct scans (2018). https://doi.org/10.48550/ARXIV.1803.05854, dataset: http://headctstudy.qure.ai/dataset
5. De Vos, B.D., Wolterink, J.M., De Jong, P.A., Viergever, M.A., Išgum, I.: 2d image classification for 3d anatomy localization: employing deep convolutional neural networks. In: Medical imaging 2016: Image processing. vol. 9784, pp. 517–523. SPIE (2016)
6. Diez, P.L., Juhl, K.A., Sundgaard, J.V., Diab, H., Margeta, J., Patou, F., Paulsen, R.R.: Deep reinforcement learning for detection of abnormal anatomies. In: Proceedings of the Northern Lights Deep Learning Workshop. vol. 3 (2022)
7. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R.: 3d slicer as an image computing platform for the quantitative imaging network. Magnetic Resonance Imaging **30**(9), 1323–1341 (2012). https://doi.org/10.1016/j.mri.2012.05.001
8. Hiraman, A., Viriri, S., Gwetu, M.: Efficient region of interest detection for liver segmentation using 3d ct scans. In: 2019 Conference on Information Communications Technology and Society (ICTAS). pp. 1–6 (2019). https://doi.org/10.1109/ICTAS.2019.8703625
9. Leroy, G., Rueckert, D., Alansary, A.: Communicative reinforcement learning agents for landmark detection in brain images. In: Kia, S.M., et al. (eds.) Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology. pp. 177–186. Springer International Publishing, Cham (2020), https://arxiv.org/abs/2008.08055
10. Navarro, F., Sekuboyina, A., Waldmannstetter, D., Peeken, J.C., Combs, S.E., Menze, B.H.: Deep reinforcement learning for organ localization in ct. In: Arbel, T., et al. (eds.) Proceedings of the Third Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 121, pp. 544–554. PMLR (06–08 Jul 2020), https://proceedings.mlr.press/v121/navarro20a.html
11. Peng, J., Hu, P., Lu, F., Peng, Z., Kong, D., Zhang, H.: 3d liver segmentation using multiple region appearances and graph cuts. Medical physics **42**(12), 6840–6852 (2015)
12. Reda, F.A., McRackan, T.R., Labadie, R.F., Dawant, B.M., Noble, J.H.: Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients. Medical Image Analysis **18**(3), 605–615 (Apr 2014). https://doi.org/10.1016/j.media.2014.02.001

13. Sudha, S., Jayanthi, K., Rajasekaran, C., Sunder, T.: Segmentation of roi in medical images using cnn-a comparative study. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON). pp. 767–771. IEEE (2019)
14. Trier, P., Noe, K.: The visible ear simulator. `https://ves.alexandra.dk/` (2020)
15. Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B.: Multiple landmark detection using multi-agent reinforcement learning. In: Shen, D., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 262–270. Springer International Publishing, Cham (2019)
16. Ying, W., Cunxi, C., Tong, J., Xinhe, X.: Segmentation of regions of interest in lung ct images based on 2-d otsu optimized by genetic algorithm. In: 2009 Chinese Control and Decision Conference. pp. 5185–5189. IEEE (2009)