SEGMENTING THE KIDNEY ON CT SCANS VIA CROWDSOURCING

Paras Mehta[1], Veit Sandfort[1], Daan Gheysens[2], Gert-Jan Braeckevelt[2], Jonathan Berte[2], Ronald M. Summers[1]

[1]Imaging Biomarkers and Computer-aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD USA 20892-1182
[2]Robovision AI, Gent, Belgium

ABSTRACT

Organ segmentation, or annotation, is an essential step for a variety of radiologic purposes such as automated organ detection, automated lesion detection, and radiotherapy. Convolutional Neural Networks (CNN) are a class of neural network that requires large amounts of training data for sensitive and specific image analysis. Medical image annotation of reference standard training data is costly and time consuming for relevant clinically experienced professionals. Here, we evaluate the feasibility of using crowdsourcing from untrained workers as a viable modality for large-scale data annotation. This pilot study evaluates the accuracy and usage viability of crowdsourced kidney segmentations. 42 CT scans were labeled by 72 users on the Robovision AI platform and their submissions averaged. Primary validation was conducted by comparing the crowd's submissions to reference segmentations. Crowdsourced segmentations and expert-labeled segmentations were then used individually and together as training data for separate CNN models. We found that the performance of the model trained on crowdsourcing data (Dice score = 0.904 ± 0.026) was not significantly different (P = 0.50) compared to the performance of the expert-labeled model (Dice Score = 0.885 ± 0.112). When trained on a combined set, the CNN performance achieved a comparable result (Dice Score = 0.932 ± 0.040). These data suggest that untrained workers can be used as cost-effective alternatives to expert segmentation in radiologic kidney segmentation. This presents a new modality for scalable, medical imaging data generation.

Index Terms—crowdsourcing, organ segmentation, CNN, kidney

1. INTRODUCTION

Computer-aided diagnosis (CAD) is a valuable tool for the acceleration of medical imaging analysis. A method of current great interest for medical image CAD is a Convolutional Neural Network (CNN). CNNs are powerful tools for image recognition but necessitate large quantities of annotated images as training data to increase performance to clinically relevant standards. Fully autonomous diagnostic CNNs for single conditions like diabetic retinopathy train on over 1 million training images [1]. To achieve dermatologist-level classification on skin cancer lesions, a CNN was trained on over 100,000 skin lesions [2].

Artificial neural networks have an established use in the medical field, but radiologic segmentation has preconditions that differentiate it from other types of annotated images. In segmentation, the organ of interest must be carefully delineated along its border in all three dimensions. This often requires hundreds of data points per organ to accurately segment its location, magnifying our data requirement by several orders of magnitude

Due to a variety of factors, some of which have been described, public availability of large-scale expert-annotated CT datasets is rare, presenting the major hurdle in developing clinically significant advances in CAD. This hurdle is currently the subject of investigation via techniques such as Generative Adversarial Networks (GAN). However, annotations of existing patient CT scans would be most useful for training purposes, as they exhibit the variations in human anatomy.

Segmentation is not a cost-effective use of time for a clinically trained professional, and individual expert annotations are not a scalable approach to annotate thousands of CT scans. Crowdsourcing has long been considered a reliable method for generating large quantities of data for machine learning purposes [3]. Crowdsourcing has additionally been proven to be useful in medical applications [4, 5]. Human supervision of the training data confers high levels of differentiation to otherwise untrained models. The possibility of crowdsourcing thousands of CT scan annotations with segmentations for each organ is an especially promising technique that is poised to change drastically the quantity of training data available for CAD.

We seek to measure untrained, human worker performance as a viable alternative for generating large amounts of kidney segmentations in a cost-effective manner while reducing the need for expert annotation. This study will elucidate whether outsourcing the task of annotation to the crowd is a feasible, accurate, and scalable modality to generate large amounts of properly annotated data.

2. METHODS

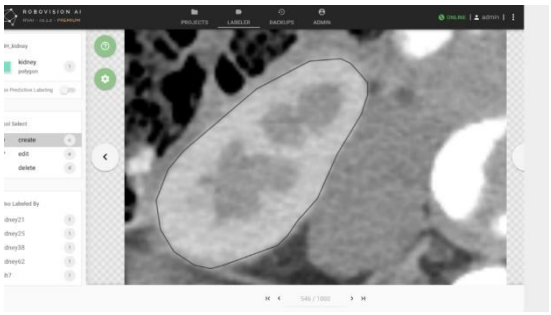

Figure 1: The RVAI Platform where the user can segment the kidney by means of a polygon

42 CT scans were taken from "Pancreas-CT" from *The Cancer Imaging Archive* [6, 7, 8]. The dataset contains CT scans of NIH Clinical Center patients with a mean age of 46.8 ± 16.7. Selected CT scans were screened for major abdominal pathologies by an experienced radiologist. The organ designated for crowd labeling during this pilot segmentation study was the kidney, selected for its ease of identification. 2D axial images were extracted from the CT scans and set to a viewing level of 40 Hounsfield Units (HU) with a window width of 400 HU. To avoid redundancy of similar images and to minimize cost, every fifth slice of 42 CT scans comprised the 1000 images sent to the contracted partner Robovision AI (RVAI) for use of their crowdsourcing platform.

One part of the RVAI platform is a portal which enables crowdsourcing tasks for machine learning purposes. RVAI additionally integrates model training into their platform, but this feature was not utilized during this study. The images were loaded onto this platform for annotation by polygon construction. **(Figure 1)** RVAI is populated by a user base of approximately 60,000 untrained labelers. After a short training module made by the RVAI team for this project outlining the position of the kidney in relation to the spine, the workers completed 15 to 20 test segmentations to become qualified. 5 to 10 qualified users drew polygons outlining all kidneys in each 2D image. To account for random and user error, a pixel-by-pixel voting system was used to determine whether a certain pixel was considered kidney or not. The threshold for a positive kidney value was >70% of respondents labeling said pixel as the kidney. This generated 1000 averaged crowd segmentations.

The GrabCut algorithm by Rother et al. was tested in post-processing to remedy the potential coarseness of the annotation due to the used polygon tool [9]. For the input of the algorithm, the center of the mask was considered to be certain foreground, the region around the annotation edge as uncertain and the remaining image as certain background. This produced 1000 crowd (processed) segmentations. **(Figure 2)**

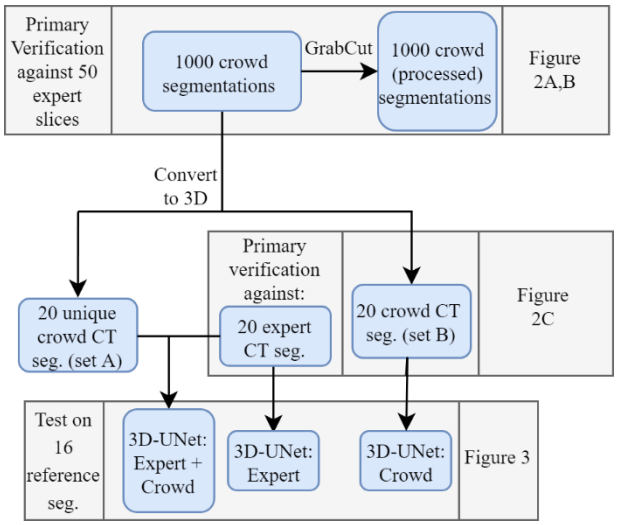

Figure 2: Flowchart after user submission averages. 50 slices from the crowd dataset and 50 slices from the crowd (processed) dataset were verified against 50 expert slices in Fig 2A, B. Set B (20 of the crowd 3D CT scans) was verified against expert segmentations of the corresponding 20 CT scans in Fig 2C. Set B, the corresponding 20 expert segmentations, and the 20 expert segmentations + Set A were trained on separate CNNs and their performance measured in Figure 3.

A representative sample of 50 kidney segmentations was created by a medical student to serve as the standard with which to verify crowd submissions. The corresponding 50 segmentations from each of the two datasets, both crowd and crowd (processed), were then verified against this representative sample by measuring Dice scores.

Since only every fifth slice per CT scan was crowd annotated, the annotated slices were then interpolated by the method outlined in Schenk et al. to fill in the missing 4 slices [10]. After interpolation, 42 segmentations were converted into 3D segmentations of the original CT scans. 20 were selected for direct comparison on a CNN (Set B) and 20 were set aside for combined testing (Set A). 3 datasets were compared: the crowd CT dataset (n=20: Set B), the expert CT dataset (n=20), and a combined dataset (20 expert CTs and 20 crowd CTs from Set A: n=40). Each dataset was then used as training data for a segmentation CNN.

For segmentation, a PyTorch implementation of a modified 3D U-Net with 128x128x128 voxels and 16 base filters was used (architecture described in Isensee et al. and concepts described in Kayalibay et al.) [11, 12]. In addition, group normalization was used instead of instance normalization. For image augmentation, we used 3D rotation and nonrigid deformation. Training was performed for 200 epochs (expert, crowd datasets) or 100 epochs (combined dataset) to result in the same total number of iterations. Performance of all three models was measured by Dice score.

3. RESULTS

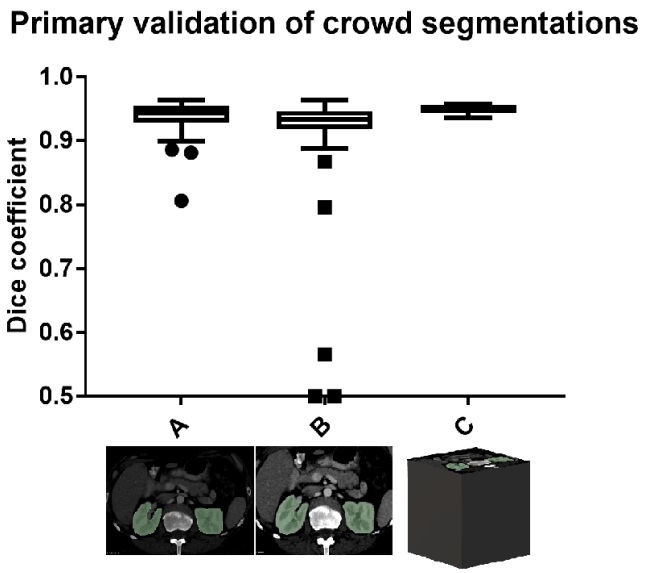

Figure 3: Primary validation of crowd annotations. A: Dice scores of 56 representative 2D crowd segmentations when compared with corresponding 2D reference standards. B: Dice scores of 56 crowd (processed) images when compared with corresponding 2D reference standard data. C: Dice scores of 20 fully crowdsourced 3D scans compared with corresponding reference standard data.

72 users contributed to 7947 annotations of 1000 slices from 42 patients, with an average of 110 slice annotations per user. An average of 1.7 users annotated each patient. Due to differences in patient height, each patient had varying numbers of slices to label. After compilation of the majority-voted averages of all user submissions, 1000 crowd segmentations corresponding to the original dataset were isolated. These segmentations were then tested on the GrabCut post-processing algorithm. Representative samples of 50 segmentations of each were compared to expert-labeled reference standard data. **(Figure 3)** The 50 crowd segmentations had an average Dice score of 0.938 ± 0.027 while the 50 images from the crowd (processed) set had an average Dice score of 0.895 ± 0.163, showing the GrabCut was detrimental to accuracy.

The 3D CT scans constructed from the crowd-labeled images subsequently achieved a Dice score of 0.949 ± 0.006 when compared with the corresponding expert-labeled 3D CT scans. This indicated a high level of similarity between the crowd's segmentations and the reference standard expert segmentations. We compared the crowdsourced (Set B) CNN vs the expert trained CNN. **(Figure 4)** After testing both CNNs on 16 kidney segmentations, we found that the crowdsourced CNN achieved an average Dice score of 0.904 ± 0.026 on the testing data while the expert labeled CNN achieved an average Dice score of 0.885 ± 0.112.

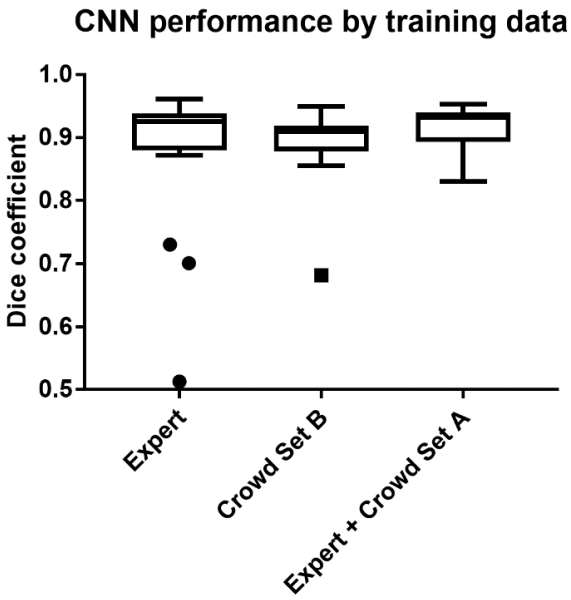

Figure 4: The Dice scores of CNN-generated segmentations separated by different types of training data. All CNNs were tested on the same set of 16 reference standard kidney segmentations.

Using a two-tailed t-test, no significant difference between results obtained by the CNN trained on crowdsourced segmentations and the CNN trained on expert segmentations (P = 0.50).

Training a single CNN on an n=40 dataset of combined expert and crowdsourced (Set A) data for 100 epochs (resulting in the same number of iterations compared with 200 epochs for the single modality CNNs where n=20) resulted in an average Dice score of 0.932 ± 0.040.

Lastly, the time required for manual expert annotation of both kidneys was an average of 37 minutes per CT scan. The RVAI platform collected 13998 annotations of kidneys spanning 7681 axial images in 48 hours.

4. DISCUSSION

The primary validation data proved that by averaging user submissions and employing a majority voting approach, crowdsourced segmentations have a high level of similarity to the reference standard. These crowdsourced submissions were then used to train a separate CNN and the performance compared to a CNN trained only on expert annotations. A two-tailed t-test revealed no significant difference between the means of the CNN trained with expert data and the CNN trained with crowd data. This establishes that crowdsourced data is viable training data for medical imaging neural networks. A third CNN was equivalently trained with both 20 crowd segmentations and 20 expert segmentations, and achieved a smaller standard deviation, showing that a

combination of expert data and crowd data is a highly effective combination.

The primary validation data also proved that GrabCut processing and edge adhering is not beneficial to segmentation accuracy. With a wider error margin, the crowd (processed) set had many comparisons with lower Dice scores. This may be attributed to the similarity of renal vessels to the kidney. The inconsistency of the GrabCut segmentations in Figure 3 strengthens our study's goals as it demonstrates humans are able to differentiate features, such as blood vessels, that the GrabCut algorithm cannot. This speaks to the valuable ability of humans to perform tasks at an accuracy higher than that of non-machine learning and some machine learning algorithms.

Human-made labels are expensive and time-consuming. If the label collection were simply classification or binary (disease present vs disease absent) then labeling would be a simpler task. However, each segmentation requires hundreds of data points itself, therefore a large dataset of annotated organs would contain millions of data points. This pilot study used a small sample size of healthy kidneys and served as a proof-of-concept for the use of crowdsourcing in medical image labeling. The results collected from the RVAI platform demonstrate the crowd can perform segmentations of the kidney as accurately as expert annotators with no difference in neural network outcomes.

A pilot study seeking to study crowdsourced liver segmentations was recently published by Heim et al. with similar methods [13]. Heim found, as we did, that after majority voting between user submissions, the segmentations generated from the crowd matched expert segmentations in quality. This serves to further establish that crowdsourcing is a viable tool for large scale organ segmentation projects.

The speed and effort required to generate these high accuracy segmentations via crowdsourcing are virtually negligible due to the automation of the process compared to the numerous hours of concerted effort required for manual expert annotation.

5. FURTHER WORK

Future pilot studies would require more difficult to identify organs, such as the pancreas, to see whether untrained individuals can perform at a similar level.

6. ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center.

7. REFERENCES

[1] Abràmoff, Michael D., et al. "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices." npj Digital Medicine vol. 1, no. 1, pp. 39, August 2018.

[2] Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." Nature vol 542, no. 7639 February 20

[3] Gebru, Timnit, et al. "Scalable annotation of fine-grained categories without experts." Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, March 2017.

[4] Park, Ko, and Swerlick. "Crowdsourcing Dermatology: DataDerm, Big Data Analytics, and Machine Learning Technology." Journal of the American Academy of Dermatology vol. 78, no. 3, pp. 643-44, March 2018

[5] Irshad, Humayun, Eun-Yeong Oh, Daniel Schmolze, Liza M. Quintana, Laura Collins, Rulla M. Tamimi, and Andrew H. Beck. "Crowdsourcing scoring of immunohistochemistry images: Evaluating Performance of the Crowd and an Automated Computational Method." Scientific Reports vol.7, pp. 43286, February 2017.

[6] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". Journal of Digital Imaging, Vol.26, No.6, pp 1045-1057, December, 2013.

[7] Roth HR, Lu L, Farag A, Shin H-C, Liu J, Turkbey EB, Summers RM. DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 556–564, June 2015

[8] Holger R. Roth, Amal Farag, Evrim B. Turkbey, Le Lu, Jiamin Liu, and Ronald M. Summers. "Data From Pancreas-CT". The Cancer Imaging Archive. 2016

[9] Rother, Carsten, Vladimir Kolmogorov, and Andrew Blake. "Grabcut: Interactive foreground extraction using iterated graph cuts." In ACM transactions on graphics (TOG), vol. 23, no. 3, pp. 309-314. ACM, August 2004.

[10] Schenk, Andrea, Guido Prause, and Heinz-Otto Peitgen. "Efficient semiautomatic segmentation of 3D objects in medical images." In International Conference on Medical Image Computing and Computer-Assisted Intervention, vol.1935, pp. 186-195, October 2000.

[11] Isensee, Fabian, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. "Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge." In International MICCAI Brainlesion Workshop, pp. 287-297, February 2017.

[12] Kayalibay, Baris, Grady Jensen, and Patrick van der Smagt. "CNN-based segmentation of medical imaging data." arXiv preprint arXiv:1701.03056, January 2017.

[13] Heim, Eric, Tobias Roß, Alexander Seitel, Keno März, Bram Stieltjes, Matthias Eisenmann, Johannes Lebert et al. "Large-scale medical image annotation with crowd-powered algorithms." Journal of Medical Imaging vol. 5, no. 3, September 2018