Large-scale medical image annotation with crowd-powered algorithms

Eric Heim
Tobias Roß
Alexander Seitel
Keno März
Bram Stieltjes
Matthias Eisenmann
Johannes Lebert
Jasmin Metzger
Gregor Sommer
Alexander W. Sauter
Fides Regina Schwartz
Andreas Termer
Felix Wagner
Hannes Götz Kenngott
Lena Maier-Hein

Large-scale medical image annotation with crowd-powered algorithms

Eric Heim,^a Tobias Roß,^a Alexander Seitel,^a Keno März,^a Bram Stieltjes,^b Matthias Eisenmann,^a Johannes Lebert,^c Jasmin Metzger,^d Gregor Sommer,^b Alexander W. Sauter,^b Fides Regina Schwartz,^b Andreas Termer,^c Felix Wagner,^c Hannes Götz Kenngott,^c and Lena Maier-Hein^{a,*}

^aGerman Cancer Research Center (DKFZ), Computer Assisted Medical Interventions, Heidelberg, Germany

^bUniversity of Basel, University Hospital Basel, Radiology and Nuclear Medicine Clinic, Basel, Switzerland

^cUniversity of Heidelberg, Department of General, Visceral and Transplant Surgery, Heidelberg, Germany

^dGerman Cancer Research Center (DKFZ), Medical Image Computing, Heidelberg, Germany

Abstract. Accurate segmentations in medical images are the foundations for various clinical applications. Advances in machine learning-based techniques show great potential for automatic image segmentation, but these techniques usually require a huge amount of accurately annotated reference segmentations for training. The guiding hypothesis of this paper was that crowd-algorithm collaboration could evolve as a key technique in large-scale medical data annotation. As an initial step toward this goal, we evaluated the performance of untrained individuals to detect and correct errors made by three-dimensional (3-D) medical segmentation algorithms. To this end, we developed a multistage segmentation pipeline incorporating a hybrid crowd-algorithm 3-D segmentation algorithm integrated into a medical imaging platform. In a pilot study of liver segmentation using a publicly available dataset of computed tomography scans, we show that the crowd is able to detect and refine inaccurate organ contours with a quality similar to that of experts (engineers with domain knowledge, medical students, and radiologists). Although the crowds need significantly more time for the annotation of a slice, the annotation rate is extremely high. This could render crowdsourcing a key tool for cost-effective large-scale medical image annotation. © 2018 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.5.3.034002]

Keywords: crowdsourcing; segmentation; statistical shape models.

Paper 18078R received Apr. 13, 2018; accepted for publication Jul. 26, 2018; published online Sep. 8, 2018.

1 Introduction

The accurate segmentation of organs in medical images is highly relevant for different clinical applications, e.g., radiation therapy, the planning of surgical interventions and follow-up of tumor diseases. In the clinical routine, a vast number of segmentations are still performed manually with interactive segmentation methods, which can be very time-consuming in the case of three-dimensional (3-D) medical image modalities, such as computed tomography (CT) scans. Recent advances of different automatic segmentation methods have shown great potential in this context.^{1–6} The major bottleneck of most of those techniques, especially with the rise of deep learning algorithms, is the accurate and cost-effective annotation of the often large amount of required training data.⁷ Crowdsourcing has become popular in this context as it is based on outsourcing cognitive tasks to many anonymous, untrained individuals, so-called workers, from an online community.⁸ It has proven itself a valuable tool for cost-effective large-scale image annotation,^{9–11} in particular when the data cannot be processed by computers and is too large to be annotated by individuals. Today, crowdsourcing has been successfully applied to various problems from the field of medical research.¹² Especially in view of the lack of publicly available reference data in the field of medical image analysis, crowd-sourced annotation of radiological images could have a huge impact on the whole research field.² In contrast to the everyday images “recognizable by a 4-year-old”¹¹ that

are typically annotated by crowdsourcing techniques, the accurate interpretation of radiological images may require trained medical experts with years of expertise.^{13–15} The lack of medical expertise is normally compensated by pointing the workers to the structures of interest,¹⁶ training the crowd workers,¹⁷ and abstraction of the real data by different rendering techniques^{18,19} or a large number of redundant annotations.¹⁸

This paper presents a hybrid crowd-algorithm-based annotation framework to address the problem of large-scale cost-effective organ segmentation in 3-D medical image volumes. The hypothesis of this work is that it is possible to create expert-level organ segmentations in 3-D radiological image volumes with nonexpert online workers acquired through crowdsourcing at a fraction of the time required by medical experts. A validation study performed on publicly available dataset of abdominal CT scans addresses the following research questions: (1) Can anonymous nonexpert crowd workers without any previous medical education detect inaccurate organ segmentations in CT images? (2) Can crowdsourcing be used to correct inaccurate organ segmentations in CT images? (3) How well do the crowd workers perform in comparison with trained medical experts with a large amount of domain-specific knowledge?

Our hybrid crowd-algorithm approach, an initial version of which was presented (in German) at a recent conference,²⁰ was integrated into a medical imaging platform combining the best of both worlds: the reliability and processing speed of algorithms from the domain of medical image analysis, with the cognitive skills of humans acquired through crowdsourcing.

*Address all correspondence to: Lena Maier-Hein, E-mail: l.maier-hein@dkfz.de

The paper is outlined as follows: previously published related work is presented in Sec. 2. Section 3 presents the proposed hybrid crowd-algorithm annotation framework for medical images. The design of the experiments for the performed validation study is presented in Sec. 4 followed by the results (Sec. 5) and the discussion (Sec. 6).

2 Related Work

The related work particularly relevant for the present paper was divided into the following three categories: approaches to annotation of two-dimensional (2-D) medical image data, approaches to annotation of 3-D medical image data, and the comparison of crowd-based annotations to those of medical experts.

2.1 Two-Dimensional Data Annotation

One of the first approaches to crowd-based medical image annotation was presented by Maier-Hein et al.²¹ The approach to segmentation of surgical instruments in endoscopic video data involved a microtask-based concept, where multiple segmentations per image are acquired through the crowdsourcing platform Amazon Mechanical Turk (MTurk). Task instructions and examples of accurate and inaccurate segmentations are included in the task. Finally, to assure the segmentation quality, the segmentations of surgical instruments are created by merging multiple segmentations from different workers with a pixel-wise majority voting approach. The authors²² later on extended their work with a hybrid crowd-algorithm system to create segmentations of surgical instruments. It uses the concept of atlas forest²³ classifiers to create initial instrument segmentations. The segmented regions with high uncertainties are further refined with crowdsourcing and later on used to retrain and improve the classifier. A similar hybrid crowd-algorithm approach was applied by Albarqouni et al.²⁴ for breast cancer detection in pathological images. The approach involves training a convolutional neural network (CNN) that is steadily refined with crowd-sourced annotations in a multistage annotation pipeline. If a new image enters the pipeline it gets classified by the CNN. In the next step, the initial result created by the CNN is further refined by the crowd and then again used to train and further improve the CNN. Bittel et al.²⁵ used a hybrid crowd-algorithm approach to create today's largest publicly available annotated endoscopic dataset. The approach uses a CNN to create initial instrument segmentations that are then further refined with crowdsourcing. Gurari et al.¹⁶ applied a similar approach for the segmentation of structures in a phase contrast microscopy, fluorescence microscopy, and single slices from small animal magnetic resonance imaging (MRI) images. They point the workers to the structure of interest by clamping the images using bounding boxes computed from the reference segmentations. Another commonly used technique to compensate the lack of medical expertise of crowd workers applied across various medical imaging modalities is to provide predefined selections the workers can choose from.^{26–29}

2.2 Three-Dimensional Data Annotation

One of the earliest works in the context of crowdsourced 3-D medical image segmentation was proposed by Chávez-Aragón et al.³⁰ who introduced a crowdsourcing-based method for femur segmentation in MRI images. The crowdsourcing was performed with a small group of students from an engineering faculty and only selected 2-D slices were segmented, not full

volumes. No experiments with untrained online workers were performed. Cheplygina et al.³¹ proposed a method to create segmentations of airways in lung CT scans. They point the worker to the structure of interest by extracting 2-D images with a size of 50×50 pixels that include the airway to be segmented out of the volume. Approximately 30% of the crowd-sourced segmentation was unusable, but not created by spammers. They were traced back to annotations of incorrectly identified structures, such as blood vessels. Also here multiple segmentations per image were acquired in conjunction with majority voting to assure segmentation quality. Rajchl et al.³² presented a method for liver segmentation in CT scans using different types of weak annotations. The basic idea behind the approach is to create segmentations of the liver in some slices of the CT volume that are used later on by an algorithm to compute the segmentation for the remaining slices of volume. However, the validation study was performed with simulated crowd annotations. No further experiments showed how the method would perform using online workers acquired through common crowdsourcing platforms. In addition to segmentation, crowdsourcing was applied for different classification tasks in 3-D radiological images, including colorectal polyp detection in CT scans¹⁹ and the classification of different tissue types in CT scans of the lung.^{29,33}

2.3 Comparison to Medical Experts

Generally, crowd-sourced annotations are evaluated against annotations created by medical experts. No further comparison with the medical experts is performed. Mavandadi et al.³⁴ pointed out that the crowd was able to identify false positive (FP) classifications in the reference data created by medical experts. Gurari et al.¹⁶ compared the segmentations created by the crowd against a reference dataset and segmentations of single medical experts. However, they state that their validation study was biased toward the used annotation software. The medical experts were able to choose between three different professional segmentation tools, while a web-based annotation tool was provided to the crowd. Depending on the type of tool, the experts created significant differences in the segmentation quality. Furthermore, the medical experts were also the creators of the reference segmentations. Maier-Hein et al.³⁵ presented a crowdsourcing approach to find correspondences in endoscopic images. The crowd was evaluated against medical expert annotators using the same web-based annotation software. Further, a new concept to automatically detect false correspondence was proposed. However, the method is tailored to the task of correspondence search and cannot be readily applied for the evaluation of crowdsourced segmentations.

In conclusion, the authors are not aware of any prior work (1) on crowd-algorithm collaboration in 3-D medical image segmentation and (2) on systematically comparing 3-D image annotations from crowd workers to those created by medical domain experts.

3 Annotation Approach

This section gives a conceptual overview of the employed multistage segmentation pipeline (Sec. 3.1) and introduces the software architecture to acquire the crowd-sourced organ segmentations (Sec. 3.2). An automatic preprocessing step to create an initial segmentation and convert the medical images to formats suitable for online distribution is presented in Sec. 3.3. The crowdsourcing approach to detect and correct

Fig. 1 Segmentation pipeline for crowd-sourced organ segmentation. Initially, the input volume is segmented with an automatic segmentation method. In the next step, the segmentation is distributed between the crowd workers over the Internet. The workers detect inaccurate segmentations and refine them if required. In the last step, the final segmentation is created by merging the annotations from different crowd workers.

inaccurate segmentations is presented in Secs. 3.4 and 3.5, respectively. Finally, Sec. 3.6 presents the method used to merge annotations from different workers to further improve segmentation quality.

3.1 Annotation Concept Overview

The concept for the accurate segmentation of organs in 3-D medical image volumes using nonexpert annotators recruited through crowdsourcing incorporates a hybrid crowd-algorithm approach integrated into a multistage annotation pipeline. Figure 1 shows a schematic overview of the different steps applied in the proposed multistage annotation pipeline.

To gain access to a large-scale crowd of untrained nonexpert workers, the concept was implemented using a microtask-based annotation approach that can be used in conjunction with common crowdsourcing platforms. In a microtask-based crowdsourcing platform,³⁶ crowd workers can freely choose their task in an online market place and get a small monetary reward upon completing a task that typically takes several minutes.³⁷ The tasks are distributed with a web application over the market place that can be accessed by the crowd workers. Due to the 3-D nature of medical image volumes consisting of several successive image slices, the data require more hardware resources in terms of space, computation time, and network bandwidth than 2-D images do. Such volumes are, therefore, not well suited for online distribution compared with commonly used 2-D graphic formats specially designed for usage in the world wide web (WWW) like the portable network graphics (PNG)³⁸ format. Furthermore, 3-D medical image volumes require complex software platforms for visualization.^{39–41} With the recent advances in web technologies, several highly specialized toolkits for the visualization of 3-D medical image volumes in web applications have emerged.^{42–45} Unfortunately, these toolkits have higher hardware requirements to render the data on the client side than common websites have. It should also be added that a large amount of workers in microtask-based crowdsourcing platforms reside in developing countries⁴⁶ and therefore might not have access to the same network infrastructure or computer hardware as workers from more developed countries have. To fully leverage the potential of the crowdsourcing platform, it is mandatory to keep the hardware requirements and network traffic low to reach as many crowd workers as possible. To achieve this goal, a 2-D slice-wise annotation concept is applied in this study. It consists of a four-stage hybrid crowd-algorithm annotation pipeline incorporating the following steps: (1) In the first step, the volume is preprocessed with an automatic segmentation algorithm to create an initial

segmentation of the target organ, split into multiple slices, converted and prepared for online distribution (Sec. 3.3). (2) In the next step, the crowd workers detect inaccurate slice-wise segmentation outlines in the initial automatic segmentation (Sec. 3.4). (3) The slice-wise segmentation outlines rated as inaccurate are further refined by the crowd workers (Sec. 3.5). (4) Finally, the results from different workers are merged to further improve segmentation quality and create the final organ segmentation (Sec. 3.6).

The web-based annotation software to detect [Fig. 2(a)] and refine [Fig. 2(b)] inaccurate segmentations is implemented as a website that can be accessed with any common web browser. The prototype implementation of the concept is realized as module within the Medical Imaging and Interaction Toolkit (MITK),⁴⁷ enabling access to a variety of image processing tools from the field of medical image analysis. Implementation details as well as the employed software architecture are presented in Sec. 3.2.

3.2 Architecture for Crowd-Sourced Image Annotation

This section presents a prototype implementation of the annotation approach previously introduced in Sec. 3.1. Figure 3 shows a swim lane flowchart incorporating the different components used to implement the previously introduced annotation pipeline shown in Fig. 1. It consists of a medical imaging platform, a web server including a database, and the crowdsourcing platform. The medical imaging platform communicates directly with the web server and the crowdsourcing platform through a web service. It is responsible for creating the crowdsourcing tasks, approving the results, paying the workers upon task completion, uploading the task data to the web server, and aggregating the results created by the crowd. The web server hosts an annotation software that is embedded into the crowdsourcing platform as an external web application. When a worker starts a new task, the crowdsourcing platform requests a new task in the annotation software hosted on the external web server. As soon as the worker completes the task, the results are transferred back to the web server where they can be accessed by the medical imaging platform. Through the integration into a medical imaging platform, it is possible to combine the crowdsourcing functionality with a variety of different algorithms from the domain of medical imaging. The presented architecture is designed to be open to the way the different components can be implemented. In this work, the following components are used to implement the proposed annotation pipeline.

Select contours that do not accurately enclose the organ!

Cancer treatment typically involves destroying a tumor while sparing healthy tissue. For this purpose, different anatomical structures have to be precisely located in medical image data. In this task you are presented a set of medical images that show the contour of an organ in green. Your task is to decide whether this contour accurately encloses the organ. Positive and negative examples are provided below. Please mark those images that do not provide an accurate contour by clicking on the image. You can click again to reset the picture to its original valid status.

Once you are finished, please verify your selections and submit the form using the button below.

Submit

Examples

Positive examples

In these cases, the green line is located on the organ boundary.

Negative examples

In these cases, the green lines do not perfectly enclose the organ.

(a)

Delineate human organs!

Cancer treatment typically involves destroying a tumor while sparing healthy tissue. For this purpose, different anatomical structures must be precisely located in medical image data. In this task you are presented a set of medical images that show the contour of an organ in green. Your task is to modify the contour such that it accurately encloses the organ. Positive and negative examples are provided below.

Object

Instructions

- Double click to remove points on the polygon.
- Click on gray square to add new points between two existing points.
- Click on a point and drag to move individual points.
- Click the image and drag to pan the image.
- Use the mouse wheel to zoom in and out.

Once you are finished, please verify your segmentation and submit the form using the button below.

Submit

Examples

Input (initial segmentation you receive)

Output (improved segmentation you have to make)

(b)

Fig. 2 The crowdsourcing tasks are distributed to the crowd workers throughout web-based applications that can be embedded into MTurk. Instructions as well as examples of accurate and inaccurate segmentations are included in the tasks. (a) To detect inaccurate segmentations the workers get displayed successive slices of a CT volume in a website. They can be marked as valid or invalid by clicking on the image. (b) The segmentation application enables the workers to modify or delete existing segmentation outlines and add new segmentation outlines.

3.2.1 Medical imaging platform

The medical imaging platform component is implemented with MITK.⁴⁷ The web service interface used to communicate with the crowdsourcing platform and the web server is based on the open-source python implementation of the Amazon web services (AWS) library⁴⁸ and integrated as a module within MITK using the available python wrapping interface. The module itself provides a cpp microservice⁴⁹ that is dynamically instantiated at runtime.

3.2.2 Web server

Figure 2 shows the user interface of the annotation software embedded into the crowdsourcing platform. The annotation software is implemented as a web application with a classical three-tier architecture⁵⁰ using JavaScript and hypertext markup language (HTML) for the client side application, hypertext pre-processor (PHP)⁵¹ for the application layer, and a MySQL⁵² database as a persistence layer.

3.2.3 Crowdsourcing platform

The prototype is implemented using the crowdsourcing platform MTurk.⁵³ MTurk enables the task provider to distribute micro-tasks—so-called human intelligence tasks (HIT)—to a crowd of

untrained workers via an online market place. Furthermore, MTurk provides programmatic access to the crowdsourcing platform over a representational state transfer⁵⁴ application programming interface to customize and control the HITs. In addition to the programmatic access, MTurk provides a seamless integration of web-based applications for custom tasks hosted on external web servers.

3.3 Automatic Contour Initialization

Before the image volume is distributed to the crowd, several pre-processing steps are applied with MITK to assure that the images are displayed within their correct proportions in the web-based annotation software shown in Fig. 2. The pre-processing steps consist of an initial automatic segmentation and the conversion of the medical data into formats suitable for online distribution over the WWW.

In the first step, an initial binary segmentation V of the target organ located in the volume A is created by applying the SSM approach presented by Heimann et al.⁵⁵ After the volume is segmented, a polygon model P is created from the binary segmentation V by applying the marching cubes algorithm.⁵⁶ Only those slices of the volume that contain parts of the segmentation are considered for further processing. Therefore, the subvolume $\hat{A} \in A$ is created by reslicing the volume A in the longitudinal viewing direction (along the y -axis), extracting every slice

Fig. 3 Swim lane flowchart depicting the interaction between the different components used to implement the annotation pipeline presented in Fig. 1. The medical imaging platform manages the crowdsourcing platform and the data on the web server. The annotation software is implemented as a web application on an external web server that is remotely accessed by the workers acquired through the crowdsourcing platform.

$a_i \in A$ that intersects with the created binary segmentation $a \cap V$. Afterward, a clipping plane is created for each slice to extract 2-D segmentation outlines from the polygon model. To speed up the data transfer to the web server and simplify the modification of the outlines for the crowd workers, the number of vertices in the extracted segmentation outlines is reduced with the Douglas–Peucker algorithm.⁵⁷ The algorithm successively removes vertices according to an error tolerance to preserve the original topology of the outline. To display the extracted 2-D slices matching their true proportions in the web application, the slices are resampled in terms of world coordinates and not index coordinates of the image grid to a unified spacing of $\hat{s} = (1, 1, 1)$ millimeter per voxel and translated to the origin of the 3-D coordinate system $(x, y, z) = (0, 0, 0)$ by

applying the operations described in Ref. 58. After altering the slices, the vertices of the segmentation outlines are projected onto the resampled and translated image planes. Finally, the extracted images slices and their corresponding segmentation outlines are converted and exported to formats suitable for on-line distribution over the WWW. The segmentation outlines are exported as a set of successive points into JavaScript object notation files. To enhance the contrast of the target organ in the extracted CT slices, the gray values are modified by applying a level window according to the Hounsfield units of the tissue from the target organ, for example, the liver.^{59–61} To convert the CT slices into the PNG format, the values of each pixel in the extracted slices have to be rescaled matching the value range of an unsigned byte (0 to 255).⁶²

Fig. 4 Example of (a) an accurate and (b) inaccurate segmentation outline included in the detection task presented in Fig. 2(a). As shown in (b), the state of a slice can be toggled to valid or invalid by clicking on the image.

3.4 Detection of Inaccurate Segmentation Outlines

To detect inaccurate segmentations, the workers are asked to label accurate and inaccurate segmentations using the interface from the website shown in Fig. 2(a). The task contains several successive slices of a CT volume incorporating the organ of interest with the corresponding outlines from the initial automatic segmentation. By clicking on an image, the worker can mark a segmentation as valid or invalid (Fig. 4). Instructions with examples of accurate and inaccurate segmentation outlines are included in the website.

3.5 Refinement of Inaccurate Segmentation Outlines

For the refinement of inaccurate segmentations, one slice-wise initial segmentation of the CT volume is included in the web-based annotation tool [Fig. 2(a)]. The annotation tool to refine the segmentation outlines is based on the Leaflet⁶³ JavaScript library. As shown in Fig. 5, the workers can refine the segmentation outlines by selecting and dragging vertices to the desired position. For further refinement, it is possible to remove vertices by a double click or add an additional vertex between two existing vertices by clicking on a grayed out suggested vertex. An integrated zoom function enables the workers to perform more detailed corrections. Additionally, the workers can keep track of the region of interest where the refinement is performed, by panning the image. Segmentation outlines that do not belong to the organ of interest can be removed by selecting them in the

Fig. 5 (a) Overview of the segmentation tool for crowd-based organ segmentation. (b) Segmentation outlines can be refined by moving, deleting, or adding new vertices. (c) In the delete mode existing outlines can be removed. (d) Workers can add new outlines by switching into the polygon mode. (e) If the contour intersects with itself an error message is displayed.

Fig. 6 Schematic visualization how majority voting is applied to merge the slice-wise organ segmentations of multiple crowd workers. In the first step, the crowd segmentations of an organ are merged into a frequency map. Afterward, majority voting is applied to the frequency map. Every pixel that is contained in the segmentations from the majority of crowd workers is segmented as “organ” and added to the final segmentation.

delete mode. Furthermore, it is possible to create custom free-hand segmentations by switching into the polygon mode. When the polygon mode is enabled, clicks on the image are successively connected by lines, resulting in the final segmentation contour. If the created contour intersects with itself, a warning is displayed and the color changes. In this case the worker has to correct the outline, otherwise the contour cannot be created. For enhanced usability an undo function is included, enabling the workers to easily correct their mistakes.

3.6 Merging Multiple Crowd-Sourced Annotations

The refined slice-wise segmentation outlines from multiple crowd workers are merged to the final segmentation by applying the pixel-wise majority voting approach introduced in Maier-Hein et al.²¹ Figure 6 shows a schematic overview of how majority voting is applied to the slice-wise segmentations refined by the crowd. To apply the pixel-wise majority voting, the refined segmentation outlines are converted back to binary image masks using a scan-line algorithm-based approach.⁶⁴ In the next step, the images are summed up into a frequency map. Every pixel in the frequency map that is included in the majority of the segmentations created by the crowd workers is considered as “organ” and added to the final binary segmentation mask.

4 Experiments

The proposed method was evaluated on the case study of liver segmentation in abdominal CT scans using a publicly available dataset. All experiments were performed on the training data of the MICCAI Liver Segmentation Competition 2007 (SLIVER07),⁶⁵ where the available reference segmentations of the training data served as a baseline to measure the performance of the acquired crowd-sourced segmentations. In the experiments, the results from nonexpert online workers acquired through the crowdsourcing platform MTurk were compared with three groups of medical experts familiar with medical CT scans and a high amount of knowledge about the morphology of the liver. The medical experts are divided into the following groups: a group of four radiologists. Four engineers that worked on algorithms for automatic liver segmentation in CT volumes. Finally, a group of five medical students completing their practical year in the department of general, visceral, and graft surgery of a surgical clinic.

4.1 Crowd-Sourced Annotations

4.1.1 Detection of inaccurate segmentation outlines

The detection of inaccurate segmentations was performed with the web-based annotation software presented in Sec. 3.4. Initially, the liver was segmented in all CT volumes included in the SLIVER07 training dataset using the SSM segmentation approach presented by Heimann et al.⁵⁵ In the next step, the slice-wise segmentations were extracted from the segmented volumes, converted into data formats suitable for online distribution, uploaded to the web server and distributed as microtasks to the crowdsourcing marketplace (Sec. 3). Each HIT for the detection of inaccurate segmentation outlines contained 10 successive slice-wise segmentations from the same CT volume to give the worker an overview of the progression throughout the volume. The following configuration was chosen for the detection HITs in MTurk: reward: \$0.02 US, maximum runtime per HIT: 10 min, and 10 assignments per HIT. To avoid spammers, the HITs were restricted to workers having 95% positive rating on their accomplished HITs. Based on the rating of different workers, a slice-wise segmentation was classified as inaccurate if the majority of $m \geq 6$ workers with $m \in \{1, \dots, 10\}$ rated the segmentation as inaccurate.

4.1.2 Refinement of inaccurate segmentation outlines

The web-based annotation tool presented in Sec. 3.5 was used for the refinement of segmentation outlines. All slice-wise segmentations that were rated as inaccurate by $m \geq 6$ crowd workers were refined in the correction step of the segmentation pipeline presented in Fig. 1. Every HIT contained one slice-wise segmentation for refinement. The following configuration was used for every correction HIT on MTurk: reward: 0.10 \$ US, maximum runtime per HIT: 10 min, 10 assignments per HIT, and restricted to workers that accomplished 95% of their HITs with a positive rating. The final segmentation of a slice was created by applying the pixel-wise majority voting approach presented in Sec. 3.6 to the results of all assignments from one slice-wise segmentation. A pixel was segmented as “liver” when it was contained in the contours of at least $k \geq 6$ workers. For further comparisons with the different groups of medical expert annotators, the simultaneous truth and performance-level estimation (STAPLE) algorithm,⁶⁶ that was designed to merge multiple segmentations of experts, was applied to merge the refined slice-wise segmentations.

4.2 Annotations from Medical Experts

The annotations from the different groups of medical experts were acquired with the same web-based annotation software used by the nonexpert crowd workers in Sec. 4.1. Instead of rolling out the HITs to a crowdsourcing market place, the expert annotators were granted direct access to the annotation software hosted on an external web server.

4.2.1 Detection of inaccurate segmentation outlines

The detection of inaccurate initial segmentations was carried out the same way as for the nonexpert crowd workers in Sec. 4.1. Each HIT contained 10 successive slices of the initial SSM segmentation. Every expert annotator had to rank all slices distributed to the crowd. A slice-wise segmentation was classified as

inaccurate if the majority of $m \geq 3$ annotators within an expert group ranked the segmentation as inaccurate.

To keep the workload feasible for the limited availability of the medical expert annotators, only a subset of the slice-wise segmentations refined by the crowd was distributed for refinement to the different groups of medical experts. The subset contained slice-wise segmentations with a different degree of difficulty, ranging from segmentations the crowd was not able to improve up to segmentations that were improved by the nonexpert crowd workers. To select the subset, all slice-wise segmentations refined by the crowd were sorted by the absolute improvement achieved on the initial SSM segmentation after majority voting was applied. As a measure for segmentation quality, the DICE similarity coefficient (DSC)⁶⁷ was used, which is defined by comparing a segmentation V against its reference segmentation U [Eq. (1)]

$$\text{DSC} = \frac{2|V \cap U|}{|V| + |U|}. \quad (1)$$

Afterward, the sorted set of slice-wise segmentations was divided into 10 buckets and the segmentation at the median of each bucket was chosen for further refinement. This procedure resulted in the subset $\{S_1, \dots, S_{10}\} \in \hat{A}$ containing the 10 slice-wise liver segmentations shown in Fig. 10. In addition to the different degrees of difficulties, the subset included segmentations with the following properties: five slices contained an equal amount of contours in the initial SSM and the reference segmentation. Three slices contained less contours in the initial SSM than the reference segmentation. Two slices contained more contours in the initial SSM than the reference segmentation, where one slice had no contour in the reference segmentation. Each slice-wise segmentation included in the subset was refined once by each expert annotator. In each expert group, pixel-wise majority voting and the (STAPLE) algorithm were applied to merge the refined slice-wise segmentations and further improve segmentation quality.

4.3 Evaluation

The following aspects were investigated in the conducted experiments: (1) Can nonexpert online workers acquired through a microtask crowdsourcing platform detect and refine inaccurate segmentations in radiological images? (2) How well do the nonexpert online workers perform compared with trained medical experts with a vast amount of domain-specific knowledge?

4.3.1 Detection of inaccurate segmentations

The DSC of each slice-wise segmentation was used to assess the performance for the detection of inaccurate segmentation outlines in the four different annotator groups: radiologists, students, engineers, and crowd. Based on the DSC, a slice-wise segmentation was classified as inaccurate [true positive (TP) vote] if the DSC was <0.9 and as accurate [true negative (TN) vote] with a DSC ≥ 0.95 . Slice-wise segmentations with a DSC <0.9 , which were rated as accurate segmentations after majority voting was applied, were considered as false negative (FN) votes and segmentations with a DSC ≥ 0.95 rated as inaccurate, considered as FP votes.

4.3.2 Refinement of segmentation outlines

The performance of the crowd was assessed by comparing the absolute improvement in the refined slice-wise segmentations created with pixel-wise majority voting against the initial SSM segmentation in terms of the DSC. Each expert group was compared against the initial SSM segmentation and the crowd segmentations created with pixel-wise majority voting. Furthermore, the raw segmentations without any further processing were used to assess the performance of the individual expert annotators. Finally, the performance of all four annotator groups consisting of radiologists, engineers, students, and the crowd was compared against each other using the slice-wise segmentations created with pixel-wise majority voting and the STAPLE algorithm.

5 Results

The results for the detection of inaccurate segmentation outlines are presented in Sec. 5.1 followed by the results for the refinement of inaccurate segmentation outlines in Sec. 5.2.

5.1 Detection of Inaccurate Segmentation Outlines

A total of 364 different slice-wise segmentations were used in the experiments for the detection of inaccurate segmentation outlines. With 10 assignments per HIT, this resulted in 3640 rated slice-wise segmentations by the crowd, 1820 rated slices by the group of medical students, and 1456 rated slices by the group of radiologists and engineers, respectively. All four annotator groups achieved similar results in the detection of inaccurate segmentation outlines (Fig. 7). The annotators from the expert groups achieved a similar accuracy in detecting inaccurate segmentations, while there is a high variation in the accuracy of the different crowd workers [Fig. 7(b)]. Engineers, students and the crowd achieved approximately the same results, while the radiologists had a slightly higher amount of FP classifications. These FP votes are slice-wise segmentations with

Fig. 8 Example of an accurate slice-wise segmentation that the radiologists rated as inaccurate (FP). The FP segmentations were traced back to a CT scan where the outline of the reference segmentation crossed the vena cava inferior (1). Furthermore the lobus caudatus (2) was not accurately segmented and the outline contained minor inaccuracies (3).

a DSC >0.95 that the radiologists rated as inaccurate. These FP classifications were traced back to a CT scan, where the radiologists criticized small inaccuracies of the segmentation outline at the lobus caudatus (Fig. 8). Furthermore, this case had a questionable reference segmentation where the segmentation outline intersects with the vena cava inferior. In contrast to the other groups, the radiologists had the lowest rate of inaccurate segmentations classified as accurate segmentation outlines (FN votes).

The detection of inaccurate segmentation outlines was performed by 49 different nonexpert workers acquired through crowdsourcing. They required 47 s on average to complete a HIT with a total elapsed time of 15 h, measured from distribution of the HITs to completion of the last HIT (Table 1). Compared with the crowd, the radiologists required approximately the same amount of time to complete one HIT. In contrast to the radiologists, the students required only 47% and the engineers 62% of the time compared with the crowd, while achieving similar results. Despite the fast completion of a single HIT in the group of students and engineers, compared with the crowd, the elapsed time from distribution to completion of all

Fig. 7 (a) Results for the detection of inaccurate segmentation outlines in each annotator group after majority voting was applied and (b) the detection accuracy of each worker. The distribution of correctly identified inaccurate segmentation outlines (TP), correctly identified accurate segmentations (TN), inaccurate segmentation outlines rated as accurate (FN), and accurate segmentation outlines rated as inaccurate (FP) are similar for all annotator groups.

Table 1 Mean $\bar{T}$, median (interquartile range) $\tilde{T}$ (IQR), minimum $T_{\min}$ and maximum $T_{\max}$ time to process one detection HIT containing ten slice-wise segmentations. The total elapsed time ΔT is measured from the distribution to completion of all HITs. Compared with the crowd, the different groups of medical expert achieved low annotation rates due to the low availability of the individual annotators.

Detection						
Group	$\bar{T}$ (s)	$\tilde{T}$ (IQR) (s)	$T_{\min}$ (s)	$T_{\max}$ (s)	ΔT	Annotations per hour
Radiologist	46	27 (16,44)	7	1571	19 days 3 h	3
Students	22	16 (10,27)	4	125	14 days	6
Engineers	29	21 (14,132)	5	236	16 days 4 h	4
Crowd	47	35 (20,62)	8	267	15 h	243

HITs was 26 to 30 times higher for the different groups of medical experts. Especially when considering the annotation rates in terms of rated slice-wise segmentations per hour (each HIT contains ten slice-wise segmentations), the crowd was able to achieve ~ 41 to 81 times higher annotation rates compared with the different groups of medical experts.

5.2 Refinement of Inaccurate Segmentation Outlines

The mean, median [interquartile range (IQR)] DSC of the pool of initial slice-wise SSM segmentations refined by the crowd was improved from 0.84, 0.9 (IQR: 0.82, 0.96) to 0.88, 0.92

Fig. 9 DSC of the refined crowd segmentation outlines merged with majority voting compared to the initial SSM segmentation. The dotted lines inside of the violins represent the median and IQR.

(IQR: 0.89, 0.95) (Fig. 9). A paired t -test showed that the improvement in the refined segmentation outlines was statistically significant at a p -value of 0.004. A total of 193 different nonexpert online workers acquired through crowdsourcing refined the segmentation outlines. They required 159 s on average to refine one slice-wise segmentation (Table 2). It required 39 h from distribution to completion of all HITs. This resulted in an annotation rate of 35 annotations per hour. Compared with the crowd, the radiologists and the medical students required only two-thirds of the time for the refinement of one slice-wise segmentation. Again, compared with the crowd the total time to process all HITs was way higher for the different groups of medical experts. Here, also, the crowd significantly outperformed the medical expert annotators with 233 to 350 times higher annotation rates.

As illustrated in the frequency maps displayed in Fig. 10, there is a high variability in the segmentation outlines refined by the crowd. In addition to the initial properties of the chosen validation subset (Sec. 4.2), the following different segmentation types were identified: multiple contours, low gradients, questionable reference segmentation, and single contours. Compared with the other annotator groups, the crowd is more error prone toward segmentations containing multiple contours, e.g., removing additional contours or creating new ones. Figure 11 shows the number of annotators that were not able to improve the initial SSM segmentation of the validation subset. Overall, the crowd includes higher amount of workers that are not able to improve the initial segmentations, whereas the medical experts only had problems to improve the segmentations of S_1 , S_2 , S_3 , and S_4 . It is worth noting that no annotator was able to correctly annotate the slice S_4 . Due to partial volume effects

Table 2 Mean $\bar{T}$, median (interquartile range) $\tilde{T}$ (IQR), minimum $T_{\min}$, and maximum $T_{\max}$ time to refine one slice-wise segmentation. The total elapsed time $T_{\max}$ is measured from distribution to completion of all HITs in each annotator group. Compared with the crowd, the different groups of medical experts achieved low annotation rates due to the low availability of the individual annotators.

Refinement						
Group	$\bar{T}$ (s)	$\tilde{T}$ (IQR) (s)	$T_{\min}$ (s)	$T_{\max}$ (s)	ΔT	Annotations per hour
Radiologist	117	83 (47,137)	7	546	19 days 2 h	0.1
Students	102	49 (31,75)	4	1388	14 days	0.15
Engineers	135	100 (64,139)	28	678	16 days 3 h	0.1
Crowd	158	107 (64,193)	8	849	1 days 15 h	35

Fig. 10 Validation subset of slice-wise segmentations with their corresponding reference segmentation (green outline). The colored boxes highlight slices with the following properties: More contours in the SSM than the reference segmentation (red), equal amount of contours (blue), and more contours in the reference than the SSM segmentation (green). Each slice is further associated with a segmentation type that is displayed with its initial SSM segmentation, a frequency map of all crowd segmentations and the final result created with pixel-wise majority voting. As S_4 did not contain any reference segmentation, it cannot be associated to one of the error classes.

not even the radiologists correctly identified the segmentation outline that was not part of the liver. Slice S_3 is related to the same type of error caused by questionable reference segmentations that resulted in a higher amount of FP classifications in

Fig. 11 Number of annotators for each slice-wise segmentation were not able to improve the initial SSM segmentation. Due to partial volume effects, none of the annotators were able to identify and remove the wrong contour that was added by the SSM segmentation to S_4 .

the group of radiologists (Sec. 5.1). All medical experts that did not improve the initial contour split the initial outline into two different outlines around the portal vein, whereas only one outline is included in the reference data (see Fig. 10). In contrast to the medical experts, the crowd workers from Fig. 11 did not split the outline of S_3 but extended the part of the segmentation outline that did not accurately enclose the liver and included a part of the colon into their segmentation.

Specifically, the cases requiring multiple outlines (S_1 , S_5 , S_6 , and S_8) turned out to be more error prone for crowd workers, while all medical experts were able to improve this type of segmentation. Although it contained only a single outline, S_2 turned out to be a difficult case among all annotator groups. The slice is located at the border of the segmentations and due to low gradients the border to nearby structures is barely visible. The crowd performed especially well in cases with a single contour in regions with high gradients where the outline of the liver can clearly be distinguished from other organs (S_7 , S_9 , and S_{10}). With a few exceptions, all crowd workers were able to improve the DSC of those cases.

Figure 12 shows selected examples demonstrating the lack of medical expertise from crowd workers that did not understand the task. The figure incorporates the slices S_1 , S_6 , and S_9 with the crowd segmentation, the corresponding initial SSM

Fig. 12 Selected examples for the slices S_1 , S_6 , and S_9 depicting the lack of medical expertise from nonexpert crowd workers. Red: initial SSM segmentation, green: reference segmentation, and cyan: segmentation from crowd workers with a low level of medical expertise. S_1 : The crowd worker did not remove the second (wrong) contour. S_6 and S_9 : instead of refining the suggested outline of the liver, the crowd worker created a segmentation of the abdomen.

segmentation, and the reference segmentation. None of the crowd workers correctly identified and removed the wrong additional contour in the initial segmentation of S_1 . In contrast to the crowd, three out of four radiologists, three out of four engineers, and three out of five students correctly identified and removed the contour. The initial SSM segmentation of S_6 contained only one segmentation outline. Two of the radiologists, three of the engineers, none of the students, and none of the crowd workers split the initial outline of slice S_6 into two separate outlines.

Fig. 13 Statistics of the slice-wise segmentations created by the individual annotators from the different expert groups compared with the initial SSM segmentation and the crowd segmentations created with pixel-wise majority voting. Except for one engineer, all expert annotators created segmentations with a similar mean, median (IQR) DSC for the subset introduced in Sec. 4.2.

Fig. 14 DSC of the slice-wise segmentations from the validation subset introduced in Sec. 10 merged with pixel-wise majority voting and the STAPLE algorithm.

For further statistical validation purposes, the slice S_4 was excluded from the pool of segmentations as no reference segmentation was available. As shown in Fig. 13, all annotators except for one in the group of engineers were able to improve the mean, median (IQR) DSC from the subset of initial slice-wise SSM segmentations introduced in Sec. 4.2. Some of the expert annotators create slightly more accurate segmentations than the crowd with pixel-wise majority voting. Further comparisons with multiple Wilcoxon signed-rank tests⁶⁸ and p -value correction with Bonferroni–Holm α adjustment⁶⁹ showed that compared with the crowd, none of the individual expert annotators were found to create statistically significant differences in the segmentation quality at a significance level of 0.05.

When merging multiple annotations with pixel-wise majority voting and the STAPLE algorithm, all groups of medical experts were able to slightly outperform the crowd in terms of the mean and median DSC achieved on the subset of initial SSM segmentations (Fig. 14). In contrast to the individual expert annotators, comparisons with multiple Wilcoxon signed-rank tests and p -value correction after Bonferroni–Holm α adjustment yielded in p -values for each expert group that were statistically significant at a significance level of 0.05.

6 Discussion

The key findings of this paper can be summarized as follows:

- Crowdsourcing can, in principle, be used for both, detecting and refining inaccurate organ segmentations. Problems occur mainly (1) at organ boundaries due to lack of 3-D context and/or the convexity of organs and (2) due to lack of training/unclear conventions (e.g., should the segmentation include the Vena Cava?). These challenges should be considered when transferring our approach to different anatomical targets.
- The quality of the annotations was not statistically significantly different in the four groups investigated (crowd after merging, engineers with domain knowledge, medical students, and radiologists). Yet, medical students were the fastest annotators, with the untrained crowd requiring much more time per slice (detection: 114%, refinement: 55% more on average).

6.1 Discussion of Experiments

With pixel-wise majority voting, it was possible to create crowd-sourced organ segmentations that match the quality of those created by individual medical experts. However, each group of medical experts was able to slightly outperform the crowd when segmentations from multiple expert annotators were merged. Especially in the case of segmentations incorporating multiple contours (e.g., Fig. 12, S_1), none of the crowd workers were able to correctly identify which contour belonged to the liver. Due to prior anatomical knowledge, the majority of expert annotators correctly identified the additional wrong contour in S_1 and removed it. The missing medical expertise of the crowd is also displayed in the incorrect examples shown in Fig. 12. Two different workers that did not understand the task segmented the whole abdomen instead of the liver. These workers cannot be considered as spammers that tried to cheat the system to maximize their monetary income as it most likely required more effort to create a segmentation of the abdomen than to refine the initial SSM outlines. Compared with the segmentations created by these workers, the initial contour in S_9 would only have required minor adjustments of two vertices to create an accurate segmentation. None of the annotator groups were able to correctly identify the additional contour in S_4 that was not part of the liver. However, the case might not have been identified due to missing 3-D context information. In radiological image viewers, the expert annotators would have probably been able to identify the additional outline by slicing through the 3-D volume and adjusting the gray values to their personal preferences. In our study, the same web-based annotation tool was provided to all annotator groups to exclude any bias resulting from different annotation tools providing different functionalities. Compared with the crowd, the radiologists required less time to complete contour refinement tasks. This might be related to their prior medical expertise and training, as these annotators already know what they have to pay attention to. Despite the fast accurate refinement of single slice-wise segmentations, all groups of medical experts required more than two weeks to complete the refinement of the selected subset that only consisted of 10 slice-wise segmentations. The crowd in contrast achieved high annotation rates and required <2 days to refine all of the available slice-wise segmentations, not only the subset. However, the low annotation rates might be related to the low availability of the individual expert annotators, as they were performing the annotations in addition to their daily job.

6.2 Future Work

Future research questions to reduce the annotation costs and further exploit the potential of crowdsourcing in the context of medical image segmentation include the following.

6.2.1 Crowd-algorithm collaboration

As the focus of this paper was on the performance of the crowd in detecting and refining inaccurate contours, we did not optimize the individual components of our hybrid crowd-algorithm framework (e.g., by using a deep learning approach to initialize the contour^{70,71}). Future work, however, should be directed to exploiting the potential of the crowd along with the most recent algorithm developments to generate new tools that can be used for large-scale high-quality medical data annotation.

6.2.2 Annotation software

How can 3-D context information efficiently be integrated into a crowdsourcing task? At the cost of higher hardware requirements, radiological image volumes can be visualized in web applications,^{42–44} providing the same functionalities as for radiological viewers. The functionalities of such radiological image viewing frameworks would clearly benefit the groups of medical experts. On the other hand, the complexity of such platforms and the hardware requirements might have the opposite effect on untrained non-expert online workers.

6.2.3 Training of crowd workers

Clear task instructions and training of crowd workers can increase the quality of crowdsourcing tasks.^{17,72} How can crowd workers be trained for the task of organ segmentation in medical image volumes? A possibility would be to include step wise tutorials, where the workers have to successfully solve one case order to proceed to the paid tasks.

In conclusion, the proposed hybrid crowd-algorithm framework demonstrates that crowdsourcing can be used to create accurate organ segmentations with a similar quality to those created by single medical expert annotators. The high annotation rate, scalability,^{73,74} and the ability to create segmentations that match the quality of those created by single medical experts makes crowdsourcing a valuable tool for large-scale segmentation of 3-D medical image volumes. Regarding the lack of reference data in the domain of medical image analysis,² crowdsourcing has high potential to evolve to a state-of-the-art method to create reference segmentations in 3-D medical image volumes. Due to the nature of 3-D medical image volumes consisting of several successive slices, hybrid crowd-algorithm approaches are mandatory in the context of image segmentation as they can drastically reduce the amount of processed data and therefore the required time and related costs.

Disclosures

The authors declare that there are no conflicts of interest related to this paper.

Acknowledgments

This work has been financed by the Klaus Tschira Foundation (project: “Endoskopie meets Informatik–Präzise Navigation für die minimalinvasive Chirurgie”) and SFB/TRR 125-Cognition-Guided Surgery.

References

1. T. Heimann and H.-P. Meinzer, “Statistical shape models for 3D medical image segmentation: a review,” *Med. Image Anal.* **13**(4), 543–563 (2009).
2. H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique,” *IEEE Trans. Med. Imaging* **35**(5), 1153–1159 (2016).
3. R. Cuingnet et al., “Automatic detection and segmentation of kidneys in 3D CT images using random forests,” *Lect. Notes Comput. Sci.* **7512**, 66–74 (2012).
4. H. R. Roth et al., “DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation,” *Lect. Notes Comput. Sci.* **9349**, 556–564 (2015).
5. A. Montillo et al., “Entangled decision forests and their application for semantic segmentation of CT images,” *Lect. Notes Comput. Sci.* **6801**, 184–196 (2011).

6. Z. Yan et al., "Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition," *IEEE Trans. Med. Imaging* **35**, 1332–1343 (2016).
7. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012).
8. V. C. Raykar et al., "Learning from crowds," *J. Mach. Learn. Res.* **11**, 1297–1322 (2010).
9. L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 319–326 (2004).
10. B. C. Russell et al., "LabelMe: a database and web-based tool for image annotation," *Int. J. Comput. Vision* **77**, 157–173 (2008).
11. T.-Y. Lin et al., "Microsoft COCO: common objects in context," *Lect. Notes Comput. Sci.* **8693**, 740–755 (2014).
12. B. L. Ranard et al., "Crowdsourcing-harnessing the masses to advance health and medicine, a systematic review," *J. Gen. Intern. Med.* **29**(1), 187–203 (2014).
13. C. F. Nodine et al., "How experience and training influence mammography expertise," *Acad. Radiol.* **6**(10), 575–585 (1999).
14. C. F. Nodine and C. Mello-Thoms, "The nature of expertise in radiology," in *Handbook of Medical Imaging*, J. Beutel et al., Eds., pp. 859–894, SPIE Press, Bellingham, Washington (2000).
15. T. Donovan and D. Litchfield, "Looking for cancer: expertise related differences in searching and decision making," *Appl. Cognit. Psychol.* **27**(1), 43–49 (2013).
16. D. Gurari et al., "How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms," in *Proc. IEEE Winter Conf. on Applications of Computer Vision*, pp. 1169–1176 (2015).
17. S. Feng et al., "A game-based crowdsourcing platform for rapidly training middle and high school students to perform biomedical image analysis," *Proc. SPIE* **9699**, 96990T (2016).
18. M. T. McKenna et al., "Strategies for improved interpretation of computer-aided detections for CT colonography utilizing distributed human intelligence," *Med. Image Anal.* **16**(6), 1280–1292 (2012).
19. J. H. Park et al., "Crowdsourcing for identification of polyp-free segments in virtual colonoscopy videos," *Proc. SPIE* **10138**, 101380V (2017).
20. E. Heim et al., "Crowdgestützte Organsegmentierung: Möglichkeiten und Grenzen," in *14. Jahrestagung der Deutschen Gesellschaft für Computer- und Roboterassistierte Chirurgie*, Bremen, Germany, pp. 37–42 (2015).
21. L. Maier-Hein et al., "Can masses of non-experts train highly accurate image classifiers?" *Lect. Notes Comput. Sci.* **8674**, 438–445 (2014).
22. L. Maier-Hein et al., "Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence," *Lect. Notes Comput. Sci.* **9901**, 616–623 (2016).
23. D. Žikić, B. Glocker, and A. Criminisi, "Encoding atlases by randomized classification forests for efficient multi-atlas label propagation," *Med. Image Anal.* **18**(8), 1262–1273 (2014).
24. S. Albarqouni et al., "AggNet: deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Trans. Med. Imaging* **35**(5), 1313–1321 (2016).
25. S. Bittel et al., "How to create the largest in-vivo endoscopic dataset," in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (2017).
26. F. J. C. dos Reis et al., "Crowdsourcing the general public for large scale molecular pathology studies in cancer," *EBioMedicine* **2**(7), 681–689 (2015).
27. H. Irshad et al., "Crowdsourcing scoring of immunohistochemistry images: evaluating performance of the crowd and an automated computational method," *Sci. Rep.* **7**, 43286 (2017).
28. D. Holst et al., "Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds," *J. Endourol.* **29**(10), 1183–1188 (2015).
29. S. N. Ørting et al., "Crowdsourced emphysema assessment," *Lect. Notes Comput. Sci.* **10552**, 126–135 (2017).
30. A. Chávez-Aragón, W.-S. Lee, and A. Vyas, "A crowdsourcing web platform-hip joint segmentation by non-expert contributors," in *IEEE Int. Symp. on Medical Measurements and Applications Proc. (MeMeA)*, pp. 350–354, IEEE (2013).
31. V. Cheplygina et al., "Early experiences with crowdsourcing airway annotations in chest CT," *Lect. Notes Comput. Sci.* **10008**, 209–218 (2016).
32. M. Rajchl et al., "Employing weak annotations for medical image analysis problems," arXiv:1708.06297 (2017).
33. A. Q. O'Neil et al., "Crowdsourcing labels for pathological patterns in CT lung scans: can non-experts contribute expert-quality ground truth?" *Lect. Notes Comput. Sci.* **10552**, 96–105 (2017).
34. S. Mavandadi et al., "Crowd-sourced biogames: managing the big data problem for next-generation lab-on-a-chip platforms," *Lab Chip* **12**(20), 4102–4106 (2012).
35. L. Maier-Hein et al., "Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences," *Int. J. Comput. Assist. Radiol. Surg.* **10**(8), 1201–1212 (2015).
36. A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical Turk," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 453–456, ACM (2008).
37. G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon mechanical Turk," *Judgment Decis. Making* **5**, 411–419 (2010).
38. G. Roelofs, *PNG: The Definitive Guide*, O'Reilly and Associates, Inc., Sebastopol, California (1999).
39. I. Wolf et al., "The medical imaging interaction toolkit," *Med. Image Anal.* **9**(6), 594–604 (2005).
40. S. Pieper, M. Halle, and R. Kikinis, "3D slicer," in *IEEE Int. Symp. on Biomedical Imaging: Nano to Macro*, pp. 632–635, IEEE (2004).
41. A. Rosset, L. Spadola, and O. Ratib, "OsiriX: an open-source software for navigating in multidimensional DICOM images," *J. Digital Imaging* **17**(3), 205–216 (2004).
42. D. Haehn et al., "Neuroimaging in the browser using the x toolkit," *Front. Neuroinf.* **101**, 1 (2014).
43. J. L. Bernal-Rusiel et al., "Reusable client-side javascript modules for immersive web-based real-time collaborative neuroimage visualization," *Front. Neuroinf.* **11**, 32 (2017).
44. N. Rannou et al., "Medical imaging in the browser with the A* medical imaging (AMI) toolkit," in *34th Annual Scientific Meeting European Society for Magnetic Resonance in Medicine and Biology* (2017).
45. T. Urban et al., "Lesiontracker: extensible open-source zero-footprint web viewer for cancer imaging research and clinical trials," *Cancer Res.* **77**(21), e119–e122 (2017).
46. J. Ross et al., "Who are the crowdworkers?: shifting demographics in mechanical Turk," in *CHI Extended Abstracts on Human Factors in Computing Systems*, pp. 2863–2872, ACM (2010).
47. M. Nolden et al., "The medical imaging interaction toolkit: challenges and advances," *Int. J. Comput. Assist. Radiol. Surg.* **8**(4), 607–620 (2013).
48. <http://aws.amazon.com/sdk-for-python> (August 2018).
49. <http://cppmicroservices.org> (August 2018).
50. M. Fowler, *Patterns of Enterprise Application Architecture*, Addison-Wesley Longman Publishing Co., Inc., Boston, Massachusetts (2002).
51. <http://php.net> (August 2018).
52. <http://mysql.com> (August 2018).
53. J. J. Chen et al., "Opportunities for crowdsourcing research on Amazon mechanical Turk," *Interfaces* **5**(3) (2011).
54. R. T. Fielding, "Architectural styles and the design of network-based software architectures," PhD Thesis (2000).
55. T. Heimann et al., "A shape-guided deformable model with evolutionary algorithm initialization for 3D soft tissue segmentation," *Lect. Notes Comput. Sci.* **4584**, 1–12 (2007).
56. W. E. Lorensen and H. E. Cline, "Marching cubes: a high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graphics* **21**(4), 163–169 (1987).
57. D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica* **10**(2), 112–122 (1973).
58. L. Ibanez et al., *The ITK Software Guide*, pp. 131–160, 2nd ed., ch. Gemoetric Transformations, Kitware, Incorporated, Clifton Park, New York (2005).
59. S. M. Pomerantz et al., "Liver and bone window settings for soft-copy interpretation of chest and abdominal CT," *Am. J. Roentgenol.* **174**(2), 311–314 (2000).

60. E. Kuntz and H.-D. Kuntz, *Hepatology, Principles and Practice: History, Morphology, Biochemistry, Diagnostics, Clinic, Therapy*, pp. 171–171, Springer Science and Business Media, Berlin (2006).
61. R. Lamba et al., “CT Hounsfield numbers of soft tissues on unenhanced abdominal CT scans: variability between two different manufacturer’s MDCT scanners,” *Am. J. Roentgenol.* **203**(5), 1013–1020 (2014).
62. T. Boutell, “PNG (portable network graphics) specification version 1.0” RFC 2083, <https://www.rfc-editor.org/info/rfc2083> (March 1997).
63. <http://leafletjs.com> (August 2018).
64. J. D. Foley et al., *Introduction to Computer Graphics*, Vol. **55**, Addison-Wesley Reading, Boston (1994).
65. T. Heimann et al., “Comparison and evaluation of methods for liver segmentation from CT datasets,” *IEEE Trans. Med. Imaging* **28**(8), 1251–1265 (2009).
66. O. Commowick and S. K. Warfield, “Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE,” *Lect. Notes Comput. Sci.* **6363**, 25–32 (2010).
67. L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology* **26**(3), 297–302 (1945).
68. F. Wilcoxon, “Individual comparisons by ranking methods,” *Biom. Bull.* **1**(6), 80–83 (1945).
69. S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Stat.* **6**, 65–70 (1979).
70. W. Thong et al., “Convolutional networks for kidney segmentation in contrast-enhanced CT scans,” *Comput. Meth. Biomech. Biomed. Eng.* **6**(3), 277–282 (2018).
71. Y. Zheng et al., “Deep learning based automatic segmentation of pathological kidney in CT: local versus global image context,” in *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, L. Lu et al., eds., pp. 241–255, Springer, Cham (2017).
72. L. Gottlieb et al., “Pushing the limits of mechanical Turk: qualifying the crowd for video geo-location,” in *Proc. ACM Multimedia Workshop on Crowdsourcing for Multimedia*, pp. 23–28, ACM (2012).
73. C. Van Pelt and A. Sorokin, “Designing a scalable crowdsourcing platform,” in *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pp. 765–766, ACM (2012).
74. A. Kittur et al., “The future of crowd work,” in *Proc. of the Conf. on Computer Supported Cooperative Work*, pp. 1301–1318, ACM (2013).

Eric Heim received his PhD in computer science from the University of Heidelberg in 2018 and received his MSc degree in applied computer science from the University of Heidelberg (2013). His research in the division of Computer Assisted Medical Interventions of the German Cancer Research Center Heidelberg (DKFZ) focuses on crowdsourcing-enhanced algorithms in the context of medical procedures.

Tobias Roß received his MSc degree in medical informatics from the University of Heidelberg/Heilbronn. His masters thesis was about crowd-algorithm collaboration for large-scale image-annotation. Currently, he is working in the Division of Computer Assisted Medical Interventions at the German Cancer Research Center (DKFZ) in Heidelberg as a PhD student. His research focuses on surgical data science and how to make use of the huge amount of unlabeled health care data for machine learning algorithms.

Alexander Seitel received his doctorate in medical informatics from the University of Heidelberg in 2012 and received his diploma in computer science from the Karlsruhe Institute of Technology (2007). He conducted a two-year postdoctoral fellowship at the University of British Columbia (Canada) and is currently working as a scientist in the Division of Computer Assisted Medical Interventions at the

German Cancer Research Center (DKFZ), leading a translational project on ultrasound navigated percutaneous needle insertions.

Matthias Eisenmann is a scientist at the German Cancer Research Center (DKFZ), Heidelberg, working in the Division of Computer Assisted Medical Interventions. He received his MSc in informatics from Karlsruhe Institute of Technology, Germany, in 2015. His research focuses on good scientific practices in biomedical image analysis competitions as well as clinical translation of a concept for assisting ultrasound-guided needle insertions following a quality-centered process.

Jasmin Metzger is a scientist in the Division of Medical Image Computing at the German Cancer Research Center (DKFZ), Heidelberg. She received her diploma in medical informatics from the University of Heidelberg in 2012. Her research focuses on standardization and interoperability of data management and processing. She is engaged in the development and distribution of the Medical Imaging and Interaction Toolkit—a research tool for medical image processing.

Gregor Sommer received his doctorate in medicine from the University of Freiburg in 2008 and received his diploma degree in physics from the University of Freiburg (2005). He is a board certified radiologist and nuclear medicine physician and currently working as a staff radiologist and assistant division head for cardiac and thoracic imaging at the University Hospital Basel, Switzerland. His research is focused on MRI of the lungs and quantitative techniques for cardiac and thoracic imaging.

Alexander W. Sauter completed his radiology residency at the University Hospital Tübingen in 2013. During this time, he was also a group leader for translational imaging at the Werner Siemens Imaging Center. In 2017, he finalized a residency in nuclear medicine at the University Hospital of Basel. Currently, he is a senior physician for cardiothoracic imaging. His research activities focus on multiparametric imaging and artificial intelligence applications.

Fides Regina Schwartz received her doctorate in medicine from the University of Basel in 2018, and received her MD degree from the University of Heidelberg (2013). She completed the Swiss radiology board exams in 2017 and her work as a research fellow at Duke University (Durham, NC, USA) is focused on computer tomography of the vascular system.

Hannes Götz Kenngott received his approbation in medicine 2008. He has been working as a surgeon for nine years and leads the surgical research group New Technologies and Data Science. His expertise is computer-assisted surgery, clinical decision support systems, medical robotics, cyber-physical systems in surgery, and machine learning on medical images and data. He is also part of international consortia in the area of surgical data science and surgical ontology standardization.

Lena Maier-Hein received her PhD from Karlsruhe Institute of Technology with distinction in 2009, and conducted her postdoctoral research in the Division of Medical and Biological Informatics at the German Cancer Research Center (DKFZ) and at the Hamlyn Centre for Robotics Surgery at Imperial College London. As a full professor at the DKFZ, she is now working in the field of computer-assisted medical interventions, focusing on multimodal image processing, surgical data science, and computational biophotonics. She has and continues to fulfill the role of (co-) principal investigator on a number of national and international grants including a European Research Council starting grant 2014.

Biographies for the other authors are not available.