On the Evaluation of Segmentation Editing Tools

Frank Heckel,[a] Jan H. Moltz,[a] Hans Meine,[a] Benjamin Geisler,[a] Andreas Kießling,[b] Melvin D'Anastasi,[c] Daniel Pinto dos Santos,[d] Ashok Joseph Theruvath,[d] Horst K. Hahn[a]
[a]Fraunhofer MEVIS, Universitaetsallee 29, 28357 Bremen, Germany
[b]Philipps-University Marburg, Department of Diagnostic Radiology, Baldingerstrasse, 35043 Marburg, Germany
[c]University Hospital of Munich, Department of Clinical Radiology, Marchioninistrasse 15, 81377 Munich, Germany
[d]University Hospital Mainz, Department of Diagnostic and Interventional Radiology, Langenbeckstrasse 1, 55131 Mainz, Germany

Abstract. Efficient segmentation editing tools are important components in the segmentation process, as no automatic methods exist that always generate sufficient results. Evaluating segmentation editing algorithms is challenging, because their quality depends on the user's subjective impression. So far, no established methods for an objective, comprehensive evaluation of such tools exist and particularly intermediate segmentation results are not taken into account. We discuss the evaluation of editing algorithms in the context of tumor segmentation in CT. We propose a rating scheme to qualitatively measure the accuracy and efficiency of editing tools in user studies. To objectively summarize the overall quality we propose two scores based on the subjective rating and the quantified segmentation quality over time. Finally, a simulation-based evaluation approach is discussed, which allows a more reproducible evaluation without the need for a user. This automated evaluation complements user studies, allowing a more convincing evaluation, particularly during development, where frequent user studies are not possible. The proposed methods have been used to evaluate two dedicated editing algorithms on 131 representative tumor segmentations. We show how the comparison of editing algorithms benefits from the proposed methods. Our results also show the correlation of the suggested quality score with the qualitative ratings.

Keywords: segmentation editing, interactive segmentation, evaluation, validation, automation, simulation.

Address all correspondence to: Frank Heckel, Fraunhofer MEVIS, Universitaetsallee 29, 28357 Bremen, Germany; E-mail: frank.heckel@mevis.fraunhofer.de

1 Introduction

Segmentation is one of the essential tasks in medical image analysis. For the (semi-)automatic segmentation of objects in 3D medical images, such as computed tomography (CT) or magnetic resonance imaging (MRI), many algorithms have been developed during the past decades for specific purposes.[1,2] Segmentation algorithms can be categorized by the degree of automation into

- *Fully automatic* methods, which do not require any intervention by the user,
- *Semi-automatic* methods, where the user initializes or parameterizes the algorithm, e.g. by appropriately marking the object of interest,

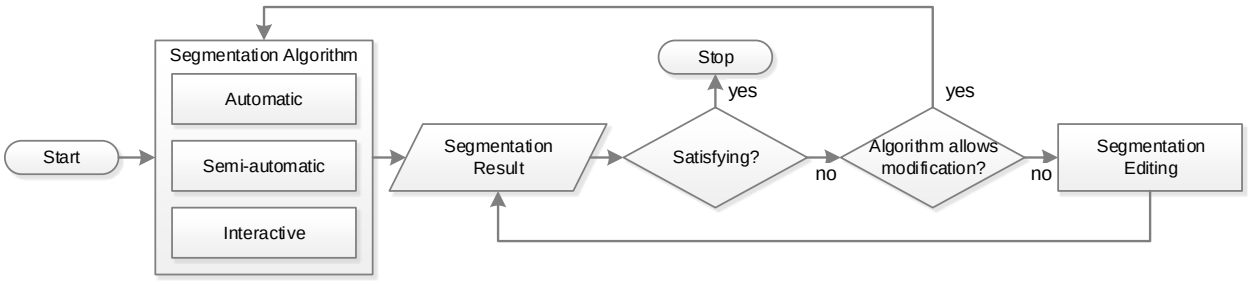


Fig 1 Computer-assisted segmentation process with an optional segmentation editing step. Note that by definition interactive segmentation algorithms always provide the possibility to modify the segmentation result.

- *Interactive* methods, which are based on an *iterative process* in which the user plays a central role by steering and correcting a computer generated segmentation result and

- *Manual* tools, where the object of interest is delineated or "painted" by hand in 2D on each slice of the image for example.

If an automatic segmentation algorithm is not available for a specific task or if it has failed, interactive segmentation algorithms are often used as a fallback. Using interactive tools, a 3D segmentation is typically generated by a set of 2D user inputs on the slices of the 3D image and each interaction immediately modifies the segmentation result. For interactive segmentation in 2D and 3D, several algorithms have been suggested, such as live-wire,[3] SketchSnakes,[4] an interactive watershed transform,[5] graph cuts[6] or random walker[7] for example. In contrast to 2D algorithms, 3D methods generate a new 3D result based on a 2D input.

In cases were a (semi-)automatic algorithm has failed, a different solution to the segmentation problem is a manual correction of the automatically generated insufficient segmentation result as shown in Fig. 1. This *segmentation editing* can be seen as a special case of interactive segmentation. In contrast to general interactive segmentation, segmentation editing typically starts with an *initial segmentation* that the user *locally* corrects in several steps until it matches his or her needs (see Fig. 2). One step can be interpreted as one interaction. In each step, the user reacts on

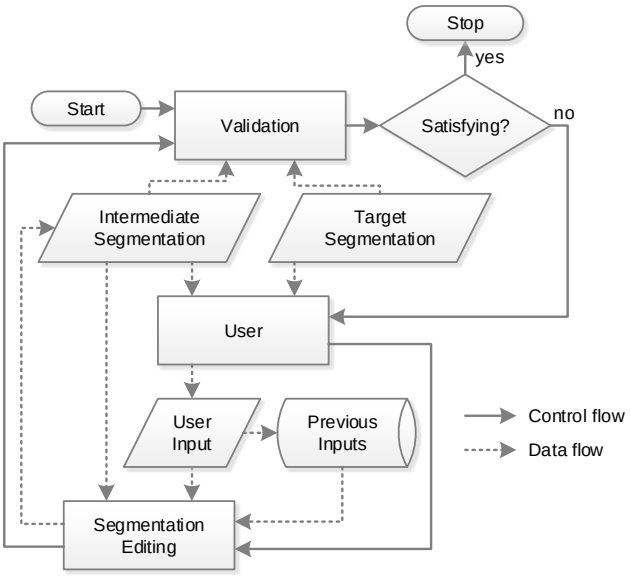


Fig 2 Stepwise segmentation editing process. The validation of the current segmentation result is visually performed by the user with respect to his or her intended result (target segmentation). Note that the target segmentation only exists in the mind of the user.

the current segmentation by a specific input, resulting in a new segmentation. We call these temporary segmentation results of the editing process *intermediate segmentations*, while we refer to the user's intended result as *target segmentation*. Some (semi-)automatic segmentation algorithms provide dedicated editing functionality.[8–10] If the (semi-)automatic segmentation algorithm does not provide the possibility to modify its results, dedicated editing tools exist, which are independent of the initial segmentation algorithm (cp. Fig. 1). By analogy to interactive segmentation algorithms, dedicated editing tools can modify the segmentation result in 2D[11–13] or 3D.[14–21] A detailed overview on segmentation editing for medical imaging is given in.[22]

For assessing the similarity of a *single* segmentation result with respect to a *reference segmentation*, i.e. its *quality*, various measures exist. We refer to this as *static validation*. Common static quality measures include volume-based metrics, like the volume overlap (Jaccard coefficient) and the Dice coefficient, as well as surface-based metrics, like the mean and maximum surface distance (Hausdorff distance),[23] and a combined measure known as the MICCAI score.[24] Reference seg-

mentations are often given by manual delineations generated by domain experts, which are used as a surrogate for the unknown ground truth.[25] An objective quantitative evaluation of interactive segmentation algorithms or algorithms for segmentation editing is more challenging, though, because of their *dynamic* nature and because their quality also depends on the user's subjective impression and intention. Particularly in the context of segmentation editing no established metrics exist for objectively and comprehensively measuring the quality of an algorithm. Most often only the final segmentation result is compared to a reference segmentation. In addition, measures such as the number of interactions or the required editing time are often reported. The quality of intermediate segmentation results is typically not taken into account in order to measure the quality of such tools, even though they are very important for the subjective quality.

In this paper, we discuss the evaluation of segmentation editing algorithms in the context of tumor segmentation for chemotherapy response monitoring, where the volume of a tumor is assessed over time using CT. As an example application of our evaluation approach, we assess two dedicated 3D sketch-based segmentation editing tools that we have proposed earlier.[22] Nevertheless, the evaluation methods proposed in this paper can be used for other scenarios, other imaging modalities and other segmentation editing tools as well, as long as they can be represented by the stepwise process shown in Fig. 2.

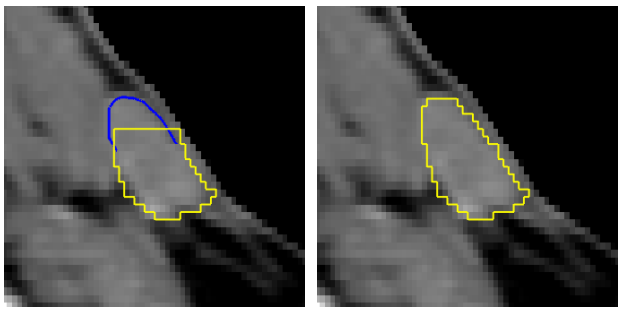
Typically, interactive algorithms are evaluated in terms of user studies. We discuss how user studies should be designed and we propose a qualitative rating scheme for analyzing the subjective quality aspects. We also propose a quality score for segmentation editing tools that accumulates the quality of intermediate segmentation results into a single measure. Based on these methods, results from a study with five radiologists are presented, where two3D editing tools have been utilized for the manual correction of 131 representative tumor segmentations (lung nodules, liver metastases

and lymph nodes). Finally, we compare these results to the results of a simulation-based validation approach, which we have presented previously.[26]

2 Related Work

Udupa et al. have summarized challenges in the evaluation of segmentation algorithms in the context of medical imaging.[25] The authors also propose a general methodology for the evaluation of such algorithms, including requirements, its implementation and performance metrics, i.e. quality measures. However, specific challenges for interactive approaches are not discussed by Udupa et al. Quality measures that compare segmentation results to reference segmentations generated by domain experts have also been discussed by several other authors.[23,27–29] This is known as supervised evaluation. Some authors have focused on the variability of segmentation results in the context of medical imaging and the analysis of segmentation algorithms with respect to multiple reference segmentations.[30–34] A method that combines several complementary quality measures into a single measure has been proposed by Deng et al.[24] A combined measure that additionally considers the common variability of different users has been proposed in context of the MICCAI segmentation challenge 2007[35] and the MICCAI liver tumor segmentation challenge 2008.[36] This measure is known as the MICCAI score.

Zhang et al. give an overview on unsupervised methods for segmentation evaluation,[37] which do not require a reference segmentation for estimating the quality of a segmentation result. Such methods can be used for on-the-fly self-tuning of segmentation algorithms for example. In order to automatically verify the correctness of segmentation results, Frounchi et al. have proposed a framework called Image Segmentation Automated Oracle (ISAO).[38] It uses machine learning in order to distinguish between consistent and inconsistent segmentation results.



(a) (b)

Fig 3 Sketch-based editing example in 2D for a lymph node in CT where a part is added to the segmentation: (a) initial segmentation (yellow), sketch-based user input (blue) and (b) editing result.

Olabarriaga and Smeulders have discussed human-computer interaction in the context of medical image segmentation.[39] The authors also summarize aspects that need to be considered when evaluating interactive algorithms. Other work on the evaluation of interactive segmentation methods focuses on "scribble-based" approaches like graph cuts or random walker, where the user draws foreground and background markers to influence the result. McGuinness and O'Connor have investigated the evaluation of such algorithms for 2D natural images.[40] Later the authors proposed a simulation-based automated evaluation for scribble-based methods in 2D.[41] For scribble-based interactive segmentation of 3D medical images, Moschidis and Graham proposed a simulation-based framework for performance evaluation[42] as well as a systematic comparison of various interactive segmentation methods.[43] For assessing the reproducibility of a graph-cut-based interactive algorithm for follicle segmentation, Haque et al. have used a similar framework that generates interactions from a database of real user inputs, which they refer to as "correct interactions".[44] Nickisch et al. and Kolhi et al. have discussed the use of a simulation model in the context of 2D natural images.[45,46] They call their model the active robot user. It is used for both evaluating and training interactive segmentation algorithms. The above work focuses on the evaluation of interactive segmentation algorithms, but don't cover the evaluation of segmentation editing algorithms.

In our previous work,[22] we have presented two sketch-based 3D segmentation editing algorithms serving as examples for evaluation in this work. Sketching provides an intuitive 2D interface for segmentation editing, where the user modifies a binary segmentation on a slice s by drawing a contour along the correct object border as shown in Fig. 3. In order to capture the user's intent based on this 2D input and modify the segmentation in 3D accordingly, we have developed both an image-based algorithm (which employs gradients), as well as an image-independent algorithm that is solely based on a geometrical object reconstruction approach.

In,[26] we introduced a simulation-based validation approach for sketch-based editing tools, but have not covered their evaluation or compared it with real users.

The present article focuses on the evaluation of sketch-based editing methods, presents results from a corresponding (previously unpublished) user study, and complements that with the simulation-based approach.

3 Evaluation of Segmentation Editing Tools

Using interactive segmentation tools, the final segmentation result is given by a user-driven, dynamic process (cp. Fig. 2). For the evaluation of interactive tools in general and editing algorithms in particular, it is insufficient to assess the final result only. The quality of such tools is influenced by additional factors, like the number of interactions or the algorithm's "reaction" time. Furthermore, their acceptance depends on the user's expectations, making the quality of interactive tools highly subjective. This subjective quality also suffers from bad intermediate results, i.e., the user would not accept results where the segmentation became worse in a specific editing step with respect to his or her intention. Consequently, user studies play the most important role for testing and comparing interactive algorithms or differing versions of the same algorithm.

The evaluation of segmentation algorithms always depends on the specific task. For example, the requirements on a segmentation algorithm differ depending on whether a tumor should be segmented for volumetric chemotherapy follow-up assessment or for surgery planning. The following three criteria are commonly used during the evaluation of interactive segmentation algorithms:[39,40,43]

- *Accuracy*: Indicates how well the segmentation corresponds to the reference segmentation, which can be assessed quantitatively using the common quality measures (e.g., volume overlap or Hausdorff distance) or qualitatively using rating schemes.

- *Efficiency*: Refers to the amount of work necessary for segmenting the object. Indicators can be the total elapsed time or the number of interactions for example. The efficiency strongly depends on the segmentation task and the user. It is affected by the computational and the interactive part of the algorithm as well as the time for assessment of the segmentation result, making it difficult to be objectively measured.

- *Repeatability*: Indicates how well the same result can be generated over different segmentation sessions by the same user or different users for a specific segmentation task.

In the following sections, we first propose a guideline for user studies, then show how such studies can be used to qualitatively rate segmentation editing algorithms, and finally complement this with a score for objectively measuring their accuracy and efficiency. Repeatability is not explicitly discussed. The proposed methods can be used to estimate this criterion as well, though, by comparing the results of different users or segmentation sessions.

3.1 Proposed Guideline for User Study Design

User studies allow an assessment of the subjective quality aspects of interactive algorithms. User studies are inherently subjective,[37] which is particularly true for interactive methods.[39] Results of user studies are typically biased, e.g., depending on the order in which the data is observed.[37] In order to avoid biasing, well designed study guidelines and a large number of data and participants are necessary. In practice this is often difficult, though, due to the high effort of studies focusing on interactive tools.

During the past years, we have composed a guideline for the design of our studies, whose aim is to get as much information as possible out of such studies while remaining practicable:

- *Randomize the order of data* in order to avoid biasing. Use different orders for every study participant.

- *Make clinical routine a constraint*: Tell the participants to use the tool in the same way as they would use it in their daily work. This affects the maximum time that the users accept for the segmentation process for example.

- *Acquire multiple results per case*: Let several participants perform the same segmentation task on the same data in order to find issues caused by the data or the segmentation tool.

- *Have heterogeneous users*: If possible, include experienced and inexperienced participants or experts from different fields.

- *Train users*: Explain all aspects of the evaluation and the used tool to the participants and give them at least one example data set. Tell them exactly what the segmentation task is (e.g., volumetric tumor follow-up assessment or surgery planning). For some studies, it can

be important not to tell the users anything about the algorithm and to skip the training step, though, e.g., if the intuitivity of the segmentation tool should be assessed.

- *Record information*: During the study, record any information that might be interesting for future analysis, like all user inputs and intermediate segmentation results, CT window settings, time measurements and user comments for example.

- *Observe the users*: As user studies are subjective, not all information can be recorded automatically. Each participant should be observed by a researcher or domain expert, who should frequently ask the participant to explain what he or she intended by a specific interaction and what result he or she expected from it for example.

- *Build "bulletproof" evaluation tools*: As we typically evaluate an algorithm that is work in progress we need to make sure that the evaluation tool is robust to crashes and malfunctions. This includes for example that no data is lost and that the participant can continue seamlessly with the last processed case after a crash. In addition, a clear workflow needs to be implemented that forces the user to give all necessary information (like comments or ratings) and that avoids an incorrect usage.

This checklist, which can be considered to be good scientific practice, has proven to be a valuable tool for planning and realization of our studies and workshops.

If different algorithms should be compared in a user study, biasing is an even more important issue, because the participants typically remember the data and what interactions they performed in the previous session. The best solution to this would be to delay the evaluation of the other algorithm by some days or even weeks, which often is not practicable, though. A compromise could be to randomize the order of the algorithms.

Table 1 Qualitative 5-point rating schemes for the segmentation editing algorithm and the final segmentation result in the context of volumetric tumor follow-up assessment. The rating is given by the user after one editing run for a specific object (i.e. per *case*). Results within one of the "sufficient" categories can be considered as appropriate for clinical practice.

	Rating		**Meaning for segmentation editing algorithm**	**Meaning for segmentation result**
Sufficient	$++$	Perfect	Errors could be corrected quickly with a few steps and according to the user's expectation.	No (more) manual correction required.
	$+$	Good	Errors could be corrected according to the user's expectation with slightly more effort, though.	Only minor errors that don't affect the volume and that don't have to be corrected.
	0	Acceptable	Errors could be corrected sufficiently at reasonable expense that would be acceptable in clinical routine. A few intermediate results were unexpected, though.	Small errors that only slightly affect the measured volume. In clinical routine, these errors wouldn't be corrected, though.
Insufficient	$-$	Bad	Correcting errors was complicated and took high effort. Several intermediate results were unexpected.	Significant errors that need correction.
	$--$	Unacceptable	Errors could not be corrected or only with far too much effort. Many intermediate results were unexpected.	Unusable segmentation even with (more) manual correction.

3.2 Qualitative Evaluation: The Editing Rating Score

In order to assess the subjective quality of segmentation editing tools, we have designed a 5-point scheme, which assembles both the accuracy and the efficiency into one rating. This rating allows drawing conclusions on how suitable an editing tool is for clinical practice with respect to the given segmentation task. This scheme is summarized in Tab. 1. Ratings of acceptable, good or perfect can be considered as sufficient for clinical practice, while bad or unacceptable ratings are insufficient.

Often there is no reference segmentation available at the beginning of the study that allows

to quantitatively measure the quality of the segmentation result. In addition, we cannot expect the final segmentation result to be a reference segmentation, because we want to assess the tool's applicability to clinical routine, which means that the user stops editing at some point, or because the algorithm fails. Therefore, our rating scheme distinguishes between the subjective quality of the editing tool and the subjective quality of the final segmentation result.

In order to quantitatively measure the overall subjective quality of a segmentation editing algorithm based on all ratings, we define the *editing rating score* for an editing tool as the average of all ratings (for many cases, by multiple readers) after mapping the ratings ($--$ to $++$) onto the interval $[0 \ldots 1]$:

$$r_{\text{edit}} = \frac{1}{N}(0.0r_{--} + 0.25r_{-} + 0.5r_{0} + 0.75r_{+} + 1.0r_{++}) \tag{1}$$

with N being the number of cases and with the quality counts

$$\begin{aligned} r_{--} &= |\{i : q(i) = --\}| \\ r_{-} &= |\{i : q(i) = -\}| \\ r_{0} &= |\{i : q(i) = 0\}| \\ r_{+} &= |\{i : q(i) = +\}| \\ r_{++} &= |\{i : q(i) = ++\}|\,, \end{aligned} \tag{2}$$

where $q(i)$ is the qualitative rating the reader assigned to the editing algorithm after finishing a specific case i, and $|\cdot|$ denotes set cardinality, i.e., the number of cases with a specific rating.

Ratings of subjective quality have a high variability due to the user's individual standards for

assessing quality and his or her requirements and expectations on the tool and the final segmentation result.[37] These standards might even change during the study, which could influence the rating. For example the expectations on the result and the editing tool change if previous similar cases worked well or badly. Moreover, the participants are typically less concentrated and more inaccurate by the end of the study. As a result, not all slices of the object might be inspected, which could bias the ratings. This makes a good study design crucial for reliable qualitative results, e.g. by randomization and the acquisition of multiple results for each case.

3.3 Quantitative Evaluation: The Editing Quality Score

Given a reference segmentation, a variety of well-known measures exists that can be used to assess the static quality of each intermediate result, such as the volume overlap, the Hausdorff distance or the MICCAI score for example. The result is a plot of the quality over time as shown in Fig. 5. Based on the qualitative rating of the final segmentation result, we can assume to have a reference segmentation for cases whose result has been rated at least as acceptable. Note, however, that reference segmentations generated this way are biased toward the tools by which they have been created, i.e., in our case the initial segmentation algorithm as well as the specific segmentation editing approach. Using it as a reference for a different tool, the quality of each step will certainly be worse and a perfect match will not be achievable.

3.3.1 A Measure for Quantitative Evaluation

The goal of a segmentation editing tool is to finish a given segmentation with as few steps as possible. Therefore, we can assume an editing tool to be "better" if it produces higher quality results with fewer steps.

In order to objectively measure the overall quality of a segmentation editing algorithm with respect to its dynamic nature, we define the *editing quality score*

$$m_{\text{edit},S_{\text{max}}} = \frac{1}{S_{\text{max}}} \left(\sum_{j=1}^{\min(S,S_{\text{max}})} m_j + \hat{S} \cdot m_S \right)$$
$$\hat{S} = \begin{cases} S_{\text{max}} - S & \text{if } S_{\text{max}} > S \\ 0 & \text{if } S_{\text{max}} \leq S \end{cases}, \tag{3}$$

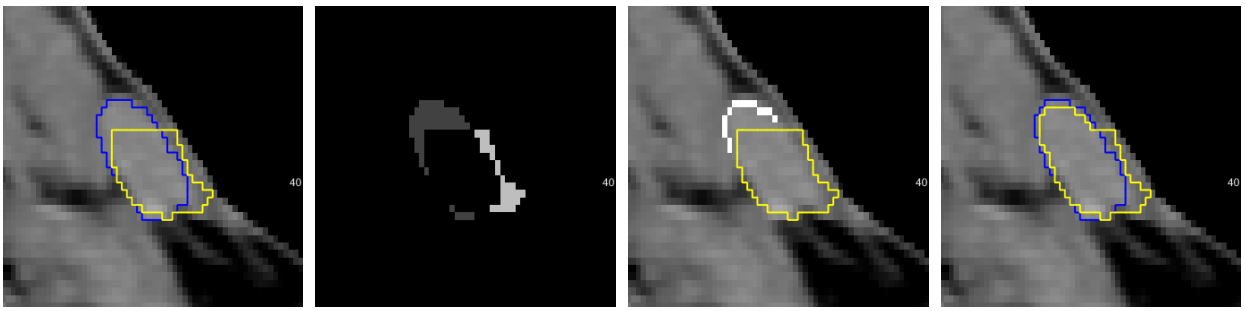
with S being the real number of editing steps and S_{max} being the maximum number of acceptable steps for the specific segmentation task. m_j is the MICCAI tumor segmentation score in a specific step j with respect to the final (reference) segmentation as defined by Deng and Du as well as Heimann et al.,[23,36] which gives $m_{\text{edit},S_{\text{max}}} \in [0, 100]$. m_j could also be replaced by any other static quality measure, though. m_S is the quality of the final segmentation result, which typically, but not necessarily, equals the maximum of the quality measure. $m_{\text{edit},S_{\text{max}}}$ equals the area under the quality curve within $[1, S_{\text{max}}]$, skipping the initial segmentation at $i = 0$ (cp. Fig. 5). Making the editing score dependent on a maximum number of acceptable editing steps has two advantages: First, it rewards algorithms that need fewer steps while it penalizes algorithms that do not allow a correction with an acceptable amount of work. Second, it keeps the measure comparable over different algorithms that differ in the real number of editing steps S, as long as S_{max} remains the same. S_{max} needs to be defined once in advance for the specific segmentation task. From various interviews with radiologists from different clinics, we got the feedback that due to time constraints, a maximum of five correction steps would be performed in clinical practice, at least in the context of chemotherapy response monitoring, giving $S_{\text{max}} = 5$ in our domain. Note that undo operations

are counted as steps as well.

3.3.2 Measuring the Influence of Algorithmic Changes

In particular during algorithm development, it is virtually impossible and impractical to conduct frequent user studies. Nevertheless, it is mandatory to measure the influence of algorithmic changes on its quality. This is equally important for optimizing parameters of the algorithm. One solution to this could be to reuse the recorded interactions and intermediate results from a user study for evaluation. In this setting, the modified algorithm is applied using the "old" inputs in every step. However, both the inputs and the intermediate results strongly depend on each other. Each user input directly depends on the current segmentation state, which indirectly depends on *all* previous intermediate results and user inputs. Therefore, changes in the underlying segmentation algorithm not only change the intermediate results but also the user inputs to the algorithm that are necessary to converge to the user's intended result. As a consequence, the old user inputs become suboptimal or even invalid with respect to the modified algorithm.

Nevertheless, the stored intermediate results can be useful for quantitative assessment of the segmentation quality during development and optimization. Considering each step as a single modification that is independent of the previous ones, we can apply one editing step using the modified algorithm and measure the static quality of its result in order to compare it to the quality of the old result. This, however, does not allow an evaluation of the dynamic aspects of the editing tool, for example whether fewer steps are needed after algorithmic optimizations.



(a) (b) (c) (d)

Fig 4 Simulation example for the exemplary lymph node from Fig. 3: (a) initial segmentation (yellow) and reference segmentation (blue), (b) errors in the initial segmentation (encoded by different gray values), (c) generated correction contour (white voxels) and (d) result after applying the image-based segmentation editing algorithm.[22]

Table 2 Overview on the data used in the study. Cases which have been edited with both editing algorithms are included in the "number of edited lesions". Cases in which the final segmentation result has been rated at least as acceptable for both editing algorithm are included in the "number of lesions with reference segmentations". Note that the study time was restricted to one hour and cases could be skipped.

Dataset	Number of lesions	Processed by	Number of edited lesions	Number of lesions with reference segmentation
List 1	96	User 1 (no experience)	25	20
		User 2 (high experience)	32	29
List 2	95	User 3 (medium experience)	27	21
		User 4 (medium experience)	29	24
List 3	95	User 5 (medium experience)	18	16
	286		131	110

3.4 Simulation-Based Evaluation

As discussed in Sec. 3.3.2, new user studies are mandatory after algorithmic changes for evaluating the dynamic aspects of an editing tool. However, they require a relatively high effort and cannot be performed after each small parameter change. In addition, user studies suffer from an intrinsic inter- and intra-observer variability, limiting their reproducibility. Even if the same object is segmented twice by the same user using the same interactive tool, the results inevitably differ due to different use of the editing tool (i.e. different user inputs) or due to a different judgment (i.e. the user considers the border between object and background to be located at different positions in

different segmentation sessions).[39]

In order to allow an objective, reproducible evaluation and comparison of 3D segmentation editing tools, without the necessity of the user, replacing the user by a simulation has shown to be an appropriate solution.[41,42,44–46] In the context of sketch-based 3D editing, we have previously proposed such a simulation-based validation approach, where plausible user inputs are iteratively generated based on the current (intermediate) and the given reference segmentation.[26]

In each iteration of the editing simulation, we first determine all *errors* in the current segmentation result (i.e. parts that are missing or which are unintentionally included with respect to the reference segmentation) as shown in Fig. 4b. From all errors in 3D, the most probably corrected one is selected based on its volume and compactness and similarly, the view (axial, sagittal or coronal) and slice s most probably used for drawing the 2D contour are selected. Finally, the sketch-based user input on s is generated, defined as the intersection of s with the surface of the error region (minus the surface of the current segmentation) as shown in Fig. 4c.

4 Results

We have used the proposed methods in order to evaluate and compare our image-based and our image-independent sketch-based 3D segmentation editing tools,[22] to which we refer as algorithm 1 and algorithm 2 in the following. Both a user study and a simulation-based evaluation have been performed, using a dataset of a total number of 286 representative tumor segmentations (lung nodules, liver metastases and lymph nodes) whose initial segmentations were rated as insufficient by clinicians in previous studies. The initial segmentations were generated with the dedicated semi-automatic segmentation algorithms by Moltz et al.[47]

Fig. 5 shows some examples of this dataset. The presented data shows the results that are

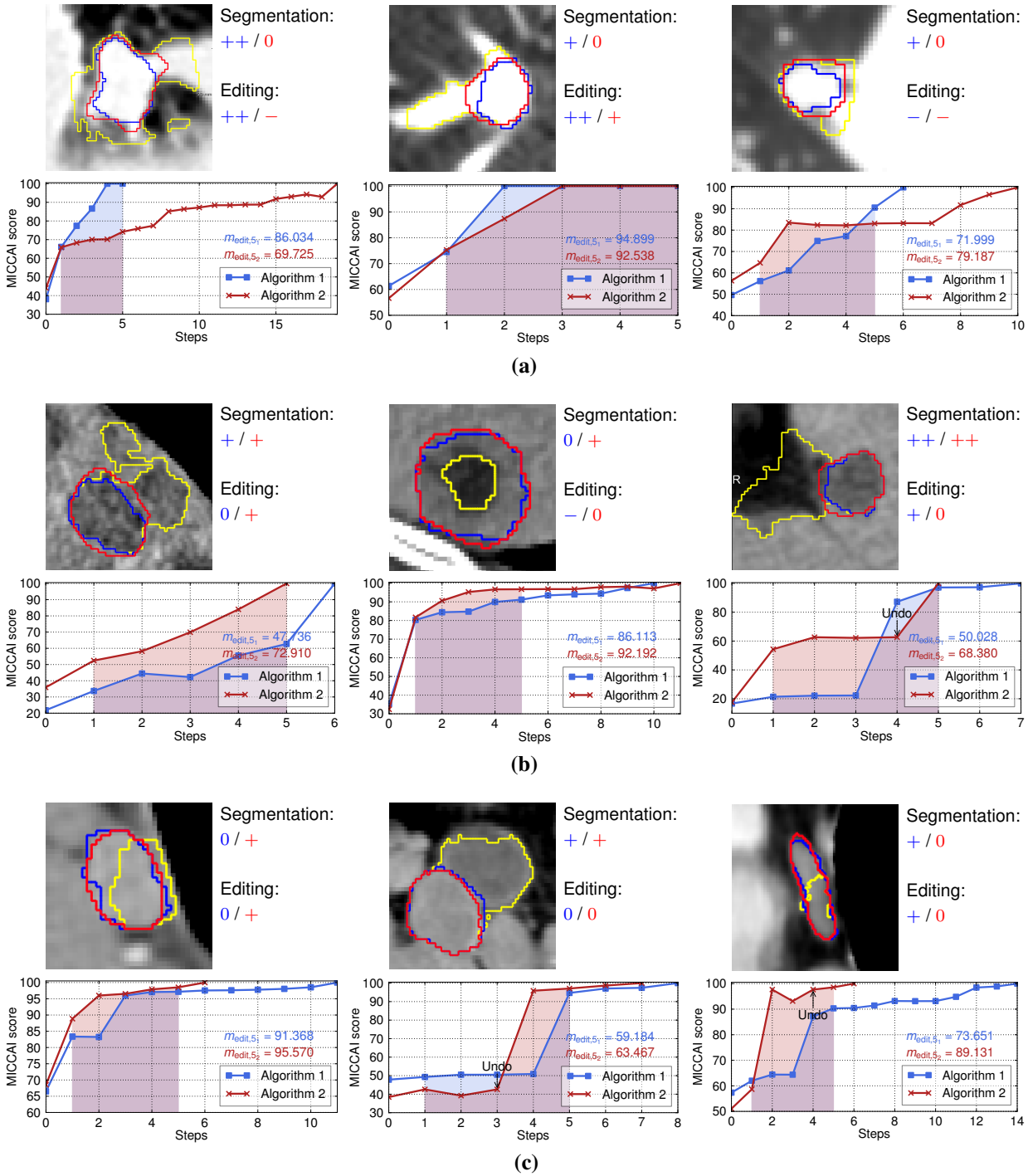


Fig 5 Examples from the study data: (a) lung nodules, (b) liver metastases and (c) lymph nodes. The initial segmentation is visualized in yellow. The blue and red contours show the final segmentation after editing with algorithm 1 and algorithm 2, respectively. The ratings refer to the subjective quality with respect to algorithm 1 / 2 (cp. Tab. 1). The highlighted areas under the curves indicate $m_{\text{edit},5}$. Note that the curves have been extended to step $\max(S, 5)$.

generated by the proposed methods and how the comparison of editing algorithms benefits from them. In addition, commonly used criteria such as time measurements and the number of steps are reported.

4.1 User Study

4.1.1 Study Design

Five radiologists participated in the study. The participants had different levels of experience with respect to the editing tools, which was determined by a questionnaire in advance. The study was conducted according to the guideline presented in Sec. 3.1. The 286 tumors were subdivided into three lists, each of which was processed by up to two clinicians in a randomized order as shown in Tab. 2. Each participant started with a different editing algorithm and got a short introduction to the software. The clinicians were observed by technical experts, who were advised not to influence the clinicians. All user inputs and intermediate results were stored by the software.

The total time for the study was restricted to one hour for each algorithm. In addition, the user had the option to skip a case, e.g., if he or she judged it as irrelevant for chemotherapy response assessment, resulting in 131 edited tumor segmentations (cp. Tab. 2). Last but not least, the users were told to only spend as much time per case as they would accept in routine. The final segmentations and the editing algorithms were rated according to the scheme shown in Tab. 1, which was shown throughout the study via a video projector. The study was performed on an Intel Core i7-2600 (3.4 GHz) with 16 GB RAM running Windows 7 64-bit.

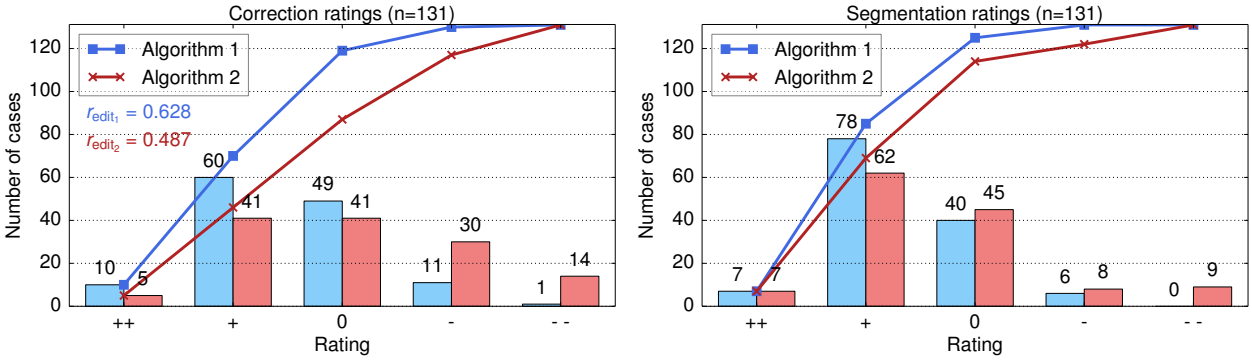


Fig 6 Qualitative results for the editing algorithms as well as the final segmentation results. The curves show the cumulative ratings, i.e., the number of cases for which the rating is $\geq \{++ \mid + \mid |0| \mid - \mid --\}$.

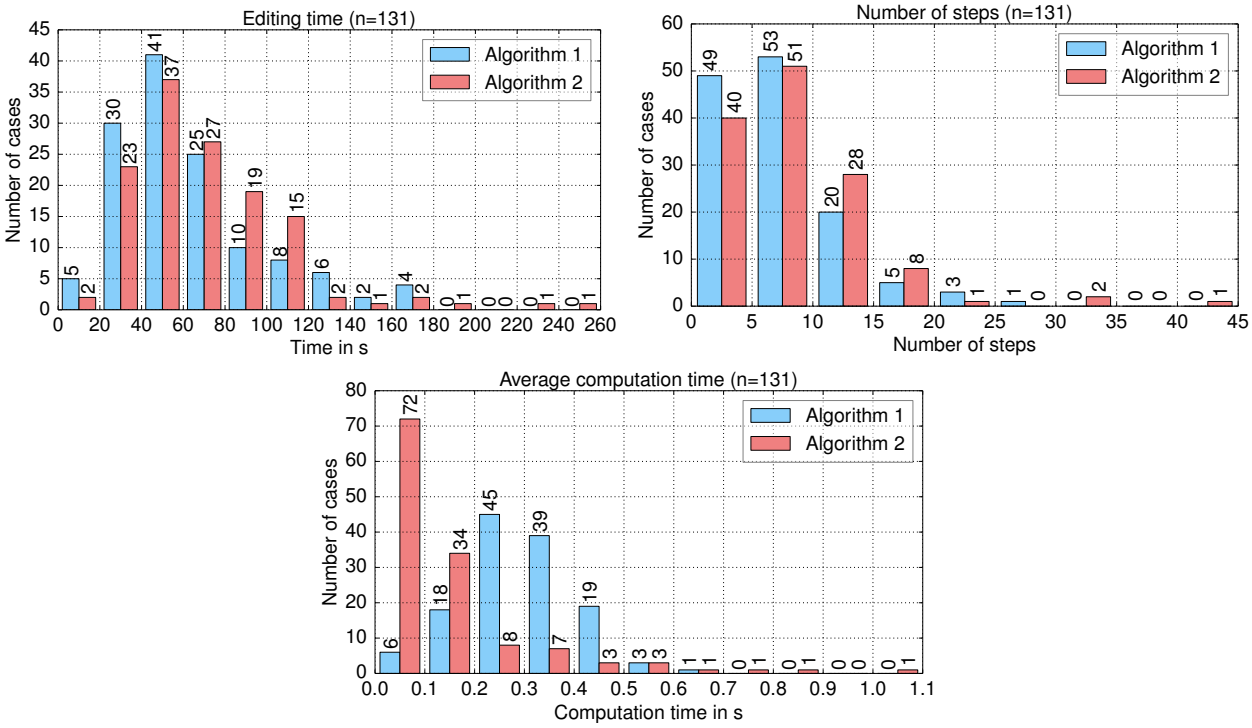


Fig 7 Histogram plots of the editing time (bin size 20), the number of editing steps (bin size 5) and the average computation time of a single step (bin size 0.1).

4.1.2 Qualitative Results

Fig. 6 and Fig. 7 summarize the qualitative results of the study. With algorithm 1, the clinicians needed a median time of 52 s in order to finish the segmentation. Using algorithm 2, the median time was 63 s. Note that these times include the assessment of each intermediate segmentation and the final segmentation result. With both editing methods, the median number of editing steps was

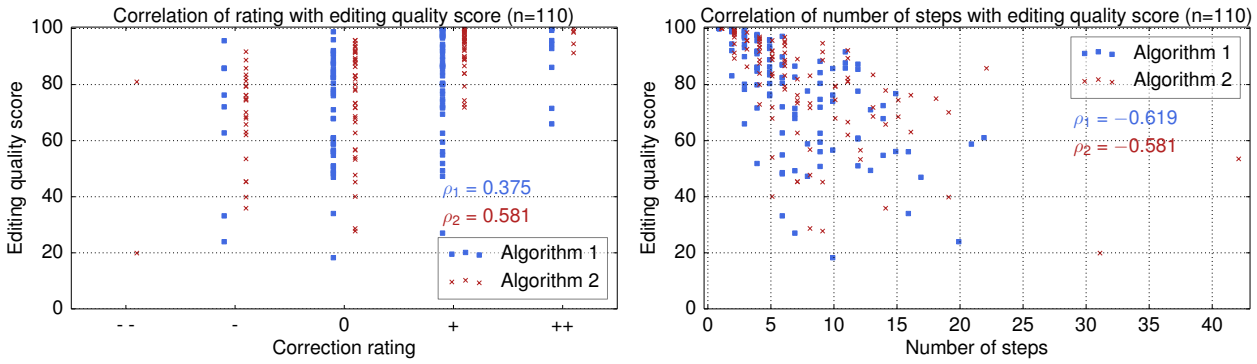


Fig 8 Correlation of the the editing quality score $m_{\text{edit},5}$ with the qualitative rating and the the number of steps. ρ refers to Pearson's correlation coefficient.

seven. A closer look at the distribution of the number of editing steps reveals that there are significantly more cases where the clinicians needed more than 10 steps with algorithm 2 (see Fig. 7). The maximum number of steps was 25 for algorithm 1 and 42 for algorithm 2. Consequently, the rating for algorithm 2 (r_{edit}=0.487) was worse compared to algorithm 1 (r_{edit}=0.628, see Fig. 6). Concerning the computation time of a single editing step, algorithm 2 is significantly faster than algorithm 1, with median computation times of 0.09 s and 0.28 s, respectively.

4.1.3 Quantitative Results

In order to quantitatively analyze the results as discussed in Sec. 3.3, we have selected 110 cases for which the final segmentation was rated as sufficient for both algorithm 1 and 2. Based on these reference segmentations, the MICCAI score has been computed for each intermediate segmentation result from which the editing quality score $m_{\text{edit},5}$ has been computed (cp. Eq. 3). As shown in Fig. 8, $m_{\text{edit},5}$ correlates with both the the qualitative rating and the number of editing steps.

Fig. 9 summarizes the quality of the intermediate results for both editing algorithms. In the first five steps algorithm 2 performs slightly better than algorithm 1, giving a better result with respect to $m_{\text{edit},5}$ (87.995 vs. 84.409). However, in the following steps, the intermediate segmentation results of algorithm 1 show better MICCAI scores. In addition, the intermediate results of algorithm 1

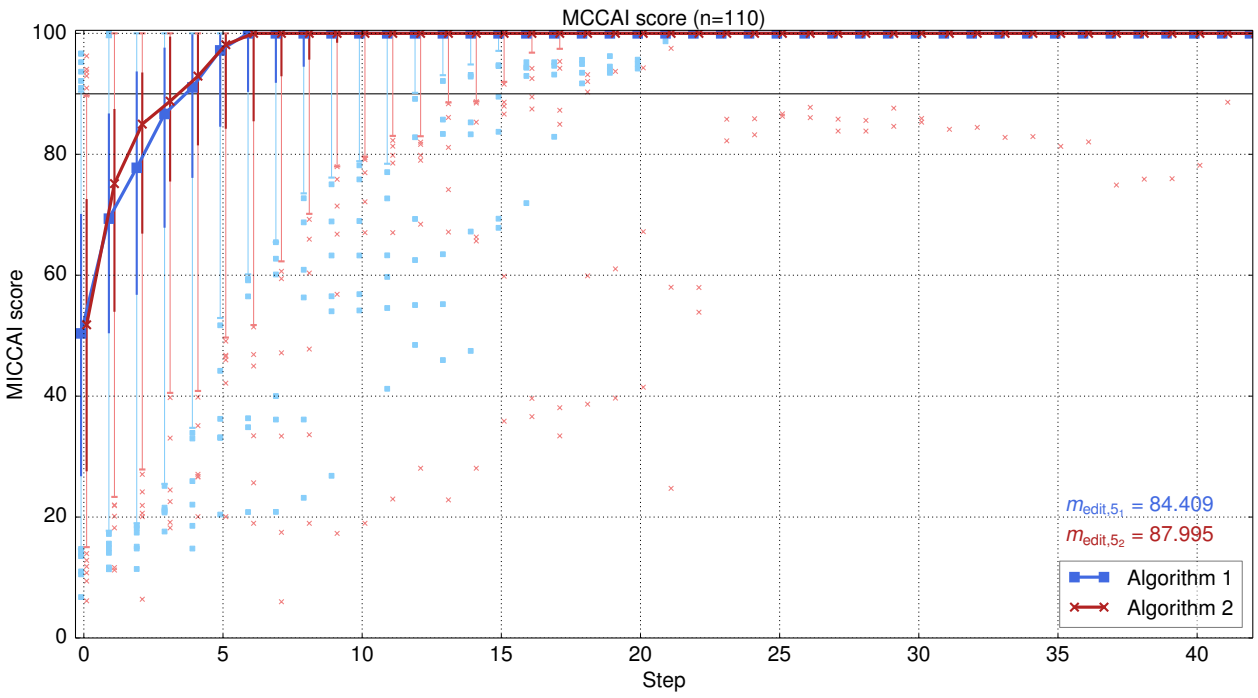


Fig 9 Quality of intermediate segmentation results in the user study. The curves show the median MICCAI score in each step, from which $m_{\text{edit},5}$ has been computed. The thick vertical lines indicate the 25- and 75%-quantiles, while the error bars show the 5- and 95%-quantiles. The dots indicate outliers. The horizontal line at 90 indicates the typical variability between different readers as defined in.[36]

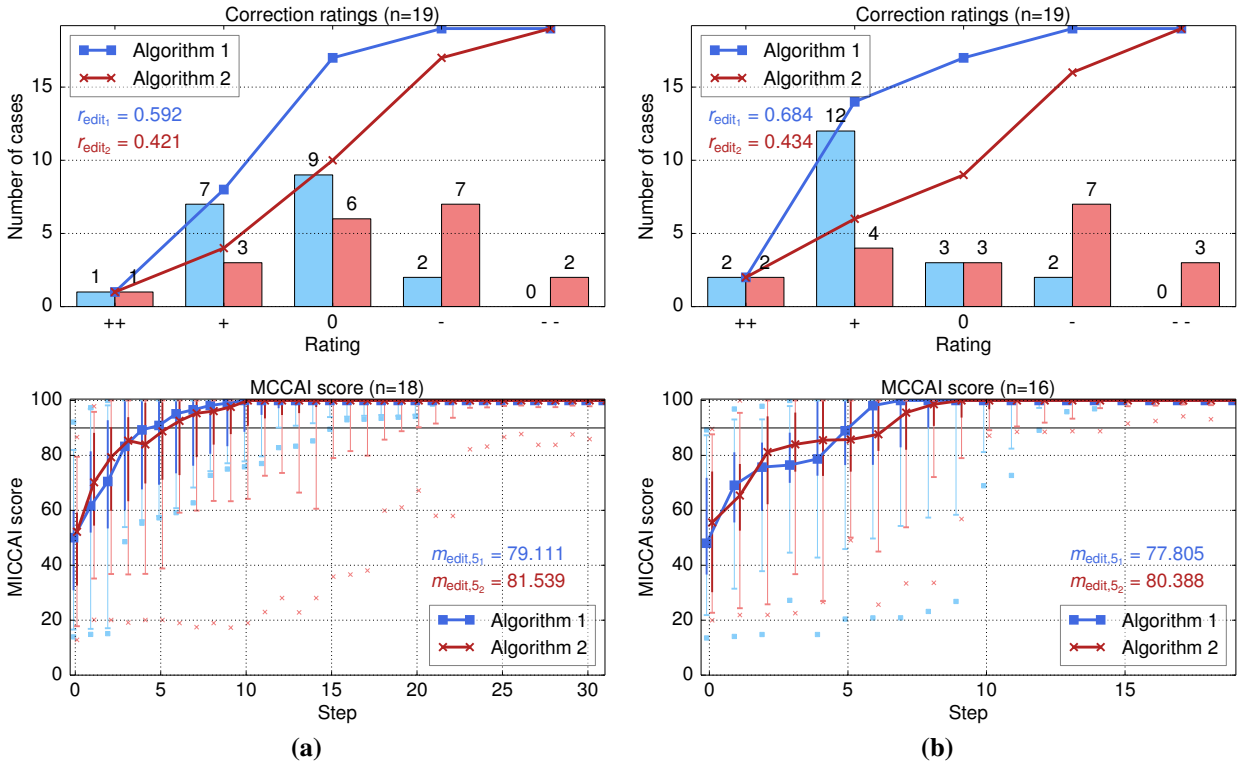


Fig 10 Results of the experienced user (a) compared to the results of the inexperienced user (b) on the same cases with the same initial segmentations.

show a lower variability and smaller outliers starting from step six.

After five editing steps, the median segmentation quality of both editing tools is within the typical variability between different readers as defined in.[36] The difference in the average quality in step 0 is due to the fact that the same initial segmentation is compared to different reference segmentations generated by the specific editing tools.

4.1.4 Influence of the Level of Experience

In order to compare the results of the experienced and the inexperienced user, only 19 cases are considered which have been processed by both users. The results show that ratings by the inexperienced user are slightly better (see Fig. 10). However, the experienced user was able to generate sufficient segmentation results in two more cases (18 vs. 16 cases with at least acceptable final segmentations). The quality of the intermediate results by the experienced user shows a lower variability, in particular for algorithm 1. Moreover, the experienced user was able to better improve the segmentation quality within the first five steps but he also invested more time in the editing process. Both users gave better ratings for algorithm 1.

4.1.5 Observations and Subjective Feedback

The observation of the users revealed two important facts that are difficult to see in the qualitative and the quantitative results. First, some participants tried to perform interactions that are not supported, namely splitting and merging of separate objects. Second, the undo functionality was rarely used. Instead, the participants tried to revoke unintended or erroneous modifications via additional contour-based editing steps.

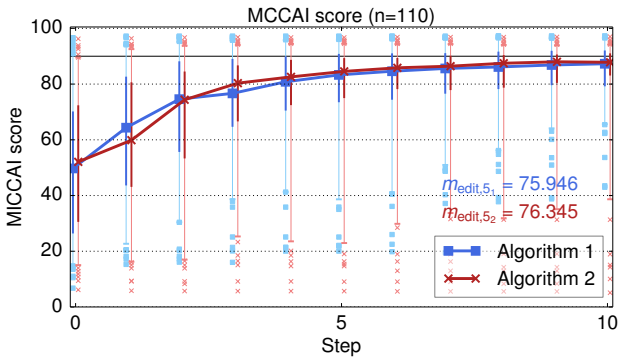


Fig 11 Quality of intermediate segmentation results using the simulation-based evaluation.

In an interview after the study, we asked the clinicians about their preferred tool. Three clinicians preferred algorithm 1 (including the most experienced user) while only one (the inexperienced user) preferred algorithm 2. One clinician did not prefer one algorithm over the other. Particularly for algorithm 2, it was criticized that the results become worse with an increasing number of editing steps, which is also visible in terms of the variability and the outliers in our quantitative analysis (cp. Fig. 9).

4.2 Simulation-Based Evaluation

We have applied our simulation-based evaluation method[26] on the 110 cases for which a reference segmentation was available. For each algorithm, the reference segmentation generated using the same algorithm was used. A maximum of ten steps was performed on each case. With ten steps, the simulation of one case currently takes about 2.5 minutes on average on an Intel Core i7-2620M (2.7 GHz) with 8 GB RAM running Windows 7 64-bit.

The quality of the intermediate segmentation results are shown in Fig. 11. Using the simulation, algorithm 2 shows again slightly better results compared to algorithm 1 with respect to $m_{\text{edit},5}$. (76.345 vs. 75.946, see Fig. 11). However, similar to the results of the user study (cp. Fig. 9), algorithm 2 also shows a higher variability and larger outliers. The segmentation quality achieved

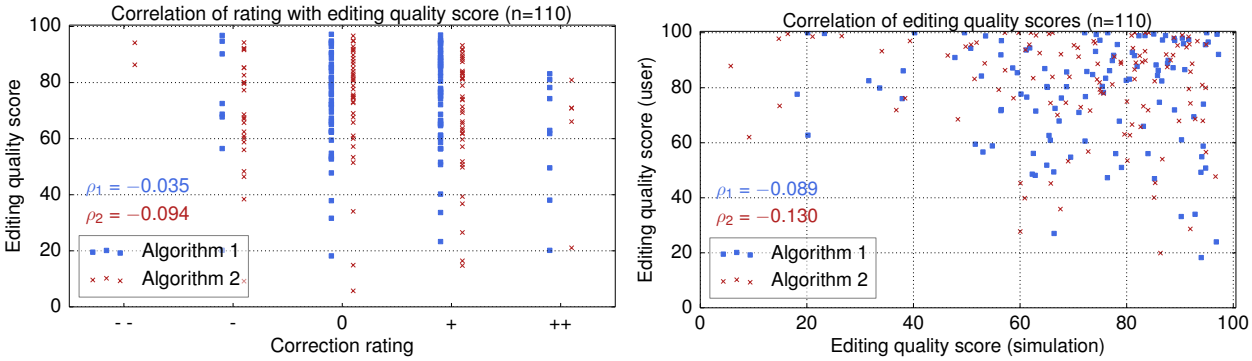


Fig 12 Scatter plots showing no correlation of the simulation-based editing quality score $m_{\text{edit},5}$ with the qualitative results of the user study. ρ refers to Pearson's correlation coefficient.

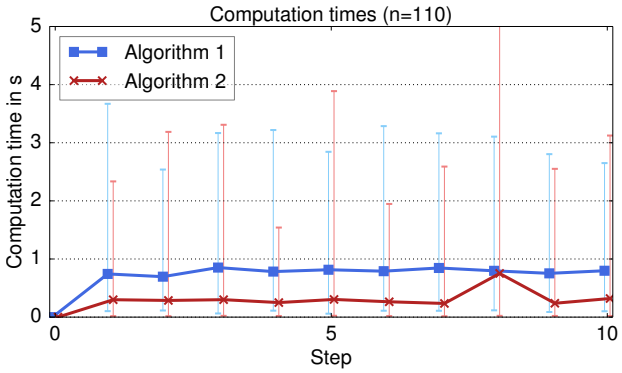


Fig 13 Average computation time in each step during the simulation-based evaluation. The error bars indicate the minimum and maximum computation time in each step.

by the simulation is currently not as good as the results by real users, though (cp. Fig. 9). The quality grows more slowly, resulting in a smaller value for $m_{\text{edit},5}$, and the variability of the results is larger. Overall, the simulation-based results do neither correlate to the qualitative ratings nor to the derived editing quality score from the user study (see Fig. 12).

Finally, we evaluated the average computation time per step, which was much lower for algorithm 2 (see Fig. 13). Although both editing algorithms consider all previous user-inputs in order to perform a specific editing step, the computation time roughly stays the same with an increasing number of steps. The analysis of the computation time also revealed an issue of algorithm 2, where in one case, a single correction in step eight took 44 s.

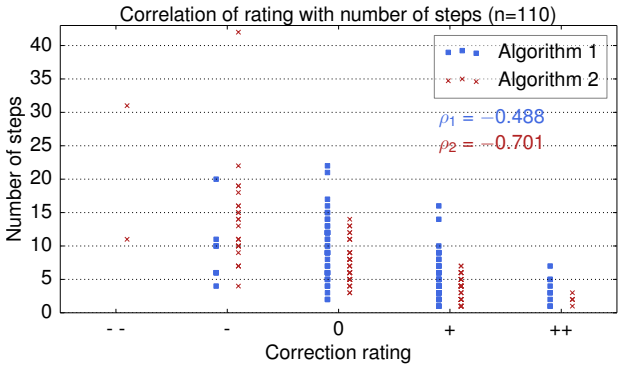


Fig 14 Correlation of the qualitative rating with the number of steps. ρ refers to Pearson's correlation coefficient.

5 Discussion

Segmentation editing is an indispensable step in the segmentation process. In clinical routine, an efficient editing tool is not optional, but a mandatory feature as stated by Heimann and Meinzer,[48] for instance. The lack of such a tool might even limit the acceptance of segmentation-based quantification methods, like measuring the volume of a tumor in the context of chemotherapy follow-up assessment. Even though some algorithms for segmentation editing have been proposed in the last years, their evaluation has been of subordinate significance and established methods for an objective and comprehensive evaluation of such tools are missing so far. This makes it difficult to compare editing tools and to assess their suitability for clinical routine. The methods proposed in this paper aim at filling this gap.

5.1 Qualitative and Quantitative Evaluation

Well-designed user studies in combination with qualitative ratings have shown to be important for the evaluation of segmentation editing tools. Based on the results of those user studies, editing tools can be evaluated quantitatively with respect to the segmentation quality and their dynamic properties, such as the number of editing steps and the computation time of each step.

The proposed editing rating score r_{edit} and the editing quality score $m_{\text{edit},S_{\max}}$ summarize the

qualitative and the quantitative results, respectively, which allows an objective comparison of different algorithms. Quantitative measures give information about the dynamic behavior of an editing algorithm. Even though the proposed editing quality score correlates with the qualitative rating (cp. Fig. 8), it cannot fully replace a subjective, qualitative assessment in the context of segmentation editing.

As already discussed in Sec. 3 and as shown in the examples in Fig. 5, the quality of a segmentation editing algorithm typically suffers from bad intermediate results. As a consequence, participants of the study gave worse ratings if the algorithm showed issues (i.e., if an editing step failed or gave an unexpected result) or if the editing took more effort (i.e. steps) than expected as shown in Fig. 14. Both is only indirectly measured by $m_{\text{edit},S_{\max}}$. Bad results can be identified by a decrease in the static quality in a specific step and by checking for undo operations.

The acceptable effort depends on the specific segmentation task. For example, a higher effort is acceptable for complex objects with low contrast where the initial segmentation shows many errors. In contrast, only a few editing steps are accepted if the segmentation problem looks rather easy, which is difficult to include in a quantitative measure. It is therefore important to analyze the progression of the quality over time for each case in order to detect issues of the editing algorithm.

The results of our user study support the feedback from our clinical partners that five editing steps are a good compromise for clinical practice in the context of volumetric tumor size assessment. After five steps, the average segmentation quality was within the expected variability between different readers. The median number of editing steps was seven and there was a significant number of cases with more than ten editing steps, though. One reason for this could be that, during the user study, the participants were willing to do more steps than in clinical routine in order to explore the limits of the editing algorithms. In addition, the dataset used in the study is not nec-

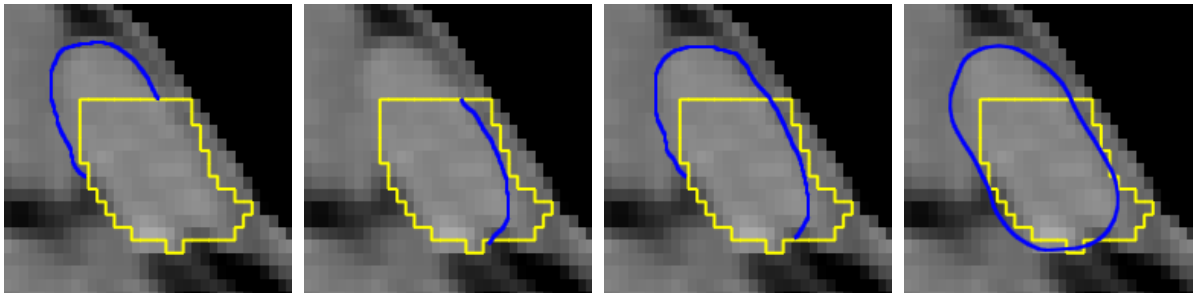


Fig 15 Possible sketch-based corrections (blue): (a) add, (b) remove, (c) add + remove or (d) replace.

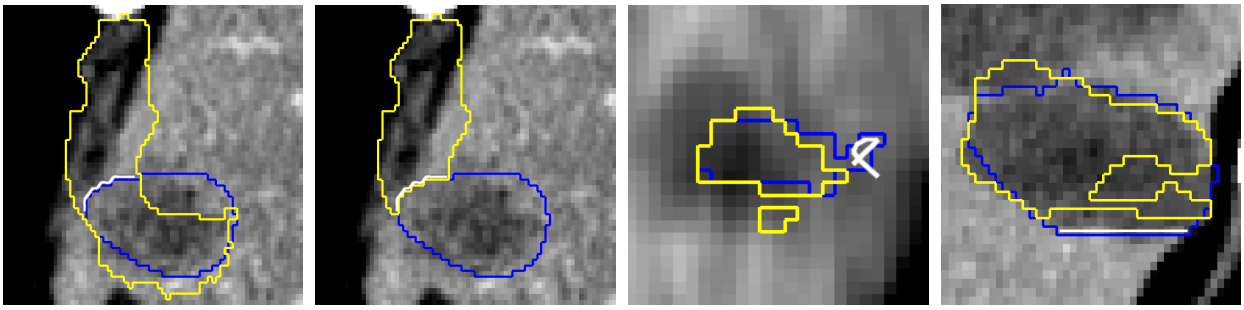


Fig 16 Simulation issues (yellow: intermediate segmentation, blue: reference segmentation, white: generated user input): (a/b) wrong part of the segmentation is kept after a remove operation which the simulation is currently not able to detect, (c) invalid user input interpreted as replace by the editing and (d) invalid user input which is ignored because the view in which the editing has been performed is ambiguous for a straight line.

essarily representative for the majority of cases occurring in routine, as it only contained bad or unacceptable segmentations that needed a significant amount of correction. This, however, only applies to 8–19% of tumor segmentations.[47]

5.2 Simulation-Based Evaluation

Simulating the user in order to provide an automated evaluation has proven to be a useful additional tool for objectively comparing different segmentation editing algorithms or different versions of the same algorithm. The segmentation results of our simulation-based evaluation are currently not as good as the results generated by human experts, though. This is mainly caused by four facts:

1. The simulation is designed to correct one unique error at once, i.e. one 3D connected com-

ponent, while a real user is allowed to correct several errors in one step, e.g. by adding and removing something in one step or by replacing as shown in Fig. 15.

2. The generation of the sketch-based user input might be shifted by up to one voxel "layer" due to their computation in image-space as can be seen in Fig. 4c.

3. The automation is not able to detect errors of the editing tool. For example if the wrong part of the segmentation is kept after removing something from the segmentation as shown in Fig. 16a and Fig. 16b, a real user would undo this step and try it differently. The simulation, however, tries to correct such results in the following by a sequence of sketches.

4. The simulation sometimes generates invalid user inputs (see Fig. 16c and Fig. 16d).

Consequently, the simulation-based results show no correlation to the results of the user study (cp. Fig. 12). Therefore, our simulation approach does not render user studies unnecessary, but they give additional valuable information, particularly in situations where new user studies are not possible. An automated evaluation approach also allows a flexible adaptation, e.g., if additional aspects should be considered that can only be measured at runtime, like the computation time of an editing step. The most important contribution of a simulation-based evaluation is the objective assessment of the influence of algorithmic changes and parameter optimization. In both scenarios the data from previous user studies quickly becomes suboptimal or invalid and new user inputs become necessary as described in Sec. 3.3.2.

5.3 Editing Algorithms

Comparing our two editing algorithms[22] based on the evaluation methods proposed in this work shows that, overall, algorithm 1 (the image-based variant) is better suited for this specific seg-

mentation task. Both algorithms allow the user to efficiently and accurately correct insufficient segmentation results with only a low influence of the level of experience, which is evidence of the intuitivity of the editing tools. However, the segmentation results were not always rated sufficient. As discussed in,[22] this is caused by the applied heuristics, which sometimes fail and give unexpected results, as well as cases were the assumptions of the editing algorithms are not met.

5.4 Generalizability and Comparability

Most of the proposed methods can be applied to other dedicated segmentation editing algorithms and tasks. The rating scheme proposed in Tab. 1, however, has been specifically designed for tumor segmentation in the context of volumetric tumor size assessment and would need to be adapted.

Using the proposed measures, the comparability of editing algorithms is limited by several aspects. The editing quality score $m_{\text{edit},S_{\text{max}}}$ depends on the maximum number of acceptable editing steps S_{max}, which is application specific. S_{max} might also be influenced by the editing algorithm itself. If the computation time of intermediate segmentation results is higher, less editing steps might be accepted by its user for example. In addition, the shape and the size of the specific object of interest determines the absolute value of the quality measure of each intermediate result and hence $m_{\text{edit},S_{\text{max}}}$. The intermediate segmentation results depend on the inputs by the specific user, whose intentions and expectations influence the qualitative rating and therefore the editing rating score r_{edit}, as well.

A segmentation editing challenge, where different tools are applied to the same problems by the same users, could help to overcome these limitations. In such a scenario, the evaluation methods proposed in this paper could help to objectively compare the results of different editing algorithms.

6 Conclusion

The objective evaluation of segmentation editing tools is a complex task. We have discussed aspects to be considered in the evaluation of such tools and we have presented guidelines and methodologies for the qualitative and quantitative evaluation of segmentation editing algorithms.

Due to the dynamic nature of editing tools and because their quality depends on the user's expectation and subjective impression, user studies are the most important instrument for the evaluation and comparison of editing algorithms. In order to objectively summarize qualitative and quantitative results, we have proposed two measures: the editing rating score r_{edit}, which summarizes the subjective quality based on a rating scheme, and the editing quality score $m_{\text{edit},S_{\max}}$, which captures the objectively measurable quality of intermediate segmentation results. This is complemented by a reproducible evaluation without the need for a user, where plausible interactions are simulated, for which we have shown to provide a useful tool.

As a real-life application of the proposed evaluation methods, we have compared two editing algorithms in the context of volumetric tumor size assessment for chemotherapy response monitoring. Our results show the correlation of $m_{\text{edit},S_{\max}}$ with the qualitative ratings as well as the complementary benefits of qualitative, quantitative and simulation-based evaluations, allowing an objective and comprehensive assessment of the quality of segmentation editing tools. Although our discussions focus on tumor segmentation in CT, the presented guidelines and measures can be applied to other segmentation editing tools as well.

7 Future Work

Future work could focus on the evaluation of the repeatability, which was out of our scope so far. For example, our simulation-based evaluation approach could be extended by the simulation

of the inaccuracy and variability of user interactions, which would allow drawing conclusions on the robustness of the editing algorithm to varying inputs. This could serve as a measure for reproducibility. In addition, the simulation-based evaluation needs to be improved so that it better correlates with real users. It could also be investigated whether the editing quality score further benefits from additional measures, such as the number of editing steps, the number of undo operations or the computation time per step for example.

So far, all evaluations have been performed using a computer mouse. However, we suppose the human-computer interface to be an important factor for the performance of a segmentation editing algorithm. For example, we expect more accurate results and a higher efficiency for direct human-computer interfaces like digitizers. Evaluations in this direction would be of high value for the development of efficient editing tools for clinical routine.

Acknowledgments

Parts of this work were funded by Siemens AG, Healthcare Sector, Imaging & Therapy Division, Computed Tomography, Forchheim, Germany.

References

1 D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual Review of Biomedical Engineering* **2**(1), 315–337 (2000).

2 D. J. Withey and Z. J. Koles, "Medical image segmentation: Methods and software," in *International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging*, 140–143 (2007).

3 W. A. Barrett and E. N. Mortensen, "Interactive live-wire boundary extraction," *Medical Image Analysis* **1**, 331–341 (1997).

4 T. McInerney, "SketchSnakes: Sketch-line initialized snakes for efficient interactive medical image segmentation," *Computerized Medical Imaging and Graphics* **32**(5), 331–352 (2008).

5 H. K. Hahn and H.-O. Peitgen, "IWT - interactive watershed transform: A hierarchical method for efficient interactive and automated segmentation of multidimensional gray-scale images," in *SPIE Medical Imaging: Image Processing*, **5032**(1), 643–653, SPIE (2003).

6 Y. Boykov and G. Funka-Lea, "Graph cuts and efficient n-d image segmentation," *International Journal of Computer Vision* **70**, 109–131 (2006).

7 L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(11), 1768–1783 (2006).

8 J. Egger, T. Kapur, T. Dukatz, M. Kolodziej, D. Zukić, B. Freisleben, and C. Nimsky, "Square-Cut: A segmentation algorithm on the basis of a rectangle shape," *PLoS One* **7**, e31064 (2012).

9 S. Steger and G. Sakas, "FIST: Fast interactive segmentation of tumors," in *Abdominal Imaging. Computational and Clinical Applications*, H. Yoshida, G. Sakas, and M. Linguraru, Eds., *Lecture Notes in Computer Science* **7029**, 125–132, Springer (2012).

10 T. Shepherd, S. J. D. Prince, and D. C. Alexander, "Interactive lesion segmentation with shape priors from offline and online learning," *IEEE Transactions on Medical Imaging* **31**, 1698–1712 (2012).

11 P. J. Elliott, J. M. Knapman, and W. Schlegel, "Interactive image segmentation for radiation treatment planning," *IBM Systems Journal* **31**(4), 620 –634 (1992).

12 A. Neumann and C. Lorenz, "Statistical shape model based segmentation of medical images," *Computerized Medical Imaging and Graphics* **22**(2), 133–143 (1998).

13 B. van Ginneken, M. de Bruijne, M. Loog, and M. A. Viergever, "Interactive shape models," in *SPIE Medical Imaging: Image Processing*, 1206–1216, SPIE (2003).

14 Y. Kang, K. Engelke, and W. A. Kalender, "Interactive 3D editing tools for image segmentation," *Medical Image Analysis* **8**(1), 35–46 (2004).

15 L. Grady and G. Funka-Lea, "An energy minimization approach to the data driven editing of presegmented images/volumes," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, **2**, 888–895, Springer (2006).

16 T. Ijiri and H. Yokota, "Contour-based interface for refining volume segmentation," in *Pacific Graphics*, *Computer Graphics Forum* **29**(7), 2153–2160, Blackwell Publishing Ltd (2010).

17 P. A. V. Miranda, A. X. Falcão, and G. C. S. Ruppert, "How to complete any segmentation process interactively via image foresting transform," *SIBGRAPI Conference on Graphics, Patterns and Images* , 309–316 (2010).

18 S. Silva, B. S. Santos, J. Madeira, and A. Silva, "A 3D tool for left ventricle segmentation editing," in *International Conference on Image Analysis and Recognition*, **6112**, 79–88, Springer (2010).

19 A. Karimov, G. Mistelbauer, J. Schmidt, P. Mindek, E. Schmidt, T. Sharipov, S. Bruckner, and M. E. Gröller, "Vivisection: Skeleton-based volume editing," in *Eurographics Conference on Visualization*, **32**(3), 461–470 (2013).

20 A. P. Harrison, N. Birkbeck, and M. Sofka, "IntellEditS: Intelligent learning-based editor of segmentations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **8151**, 235–242, Springer Berlin Heidelberg (2013).

21 A. Kronman and L. Joskowicz, "Image segmentation errors correction by mesh segmentation

and deformation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **8150**, 206–213, Springer Berlin Heidelberg (2013).

22 F. Heckel, J. H. Moltz, C. Tietjen, and H. K. Hahn, "Sketch-based editing tools for tumour segmentation in 3D medical images," *Computer Graphics Forum* **32**(8), 144–157 (2013). online first 08/2013.

23 T. Heimann, B. van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. M. M. Cashman, Y. Chi, A. Córdova, B. M. Dawant, M. Fidrich, J. D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmuller, R. I. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Németh, D. S. Raicu, A.-M. Rau, E. M. van Rikxoort, M. Rousson, L. Ruskó, K. A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. M. Waite, A. Wimmer, and I. Wolf, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Transactions on Medical Imaging* **28**(8), 1251–265 (2009).

24 X. Deng, L. Zhu, Y. Sun, C. Xu, L. Song, J. Chen, R. D. Merges, M.-P. Jolly, M. Suehling, and X. Xu, "On simulating subjective evaluation using combined objective metrics for validation of 3D tumor segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, *MICCAI'07*, 977–984, Springer-Verlag, (Berlin, Heidelberg) (2007).

25 J. K. Udupa, V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn, "A framework for evaluating image segmentation algorithms," *Computerized Medical Imaging and Graphics* **30**(2), 75–87 (2006).

26 F. Heckel, M. I. Ivanov, J. H. Moltz, and H. K. Hahn, "Toward automated validation of sketch-based 3D segmentation editing tools," in *Scandinavian Conferences on Image Analysis*, *Lecture Notes in Computer Science* **7944**, 256–265 (2013).

27 Q. Huang and D. Byron, "Quantitative methods of evaluating image segmentation," in *International Conference on Image Processing*, **3**, 53–56 (1995).

28 Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition* **29**(8), 1335–1346 (1996).

29 A. Fenster and B. Chiu, "Evaluation of segmentation algorithms for medical imaging," in *IEEE Conference on Engineering in Medicine and Biology*, 7186–7189 (2005).

30 V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Transactions on Medical Imaging* **16**(5), 642–652 (1997).

31 W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Transactions on Medical Imaging* **25**(11), 1451–1461 (2006).

32 R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(6), 929–944 (2007).

33 J. H. Moltz, S. Braunewell, J. Rühaak, F. Heckel, S. Barbieri, L. Tautz, H. K. Hahn, and H.-O. Peitgen, "Analysis of variability in manual liver tumor delineations in CT scans," in *IEEE International Symposium on Biomedical Imaging*, 1974–1977 (2011).

34 J. H. Moltz, J. Rühaak, H. K. Hahn, and H.-O. Peitgen, "A novel adaptive scoring system for

segmentation validation with multiple reference masks," in *SPIE Medical Imaging: Image Processing*, **7962**(1), 796214–1–10, SPIE (2011).

35 B. van Ginneken, T. Heimann, and M. Styner, "3d segmentation in the clinic: A grand challenge," in *3D Segmentation in The Clinic: A Grand Challenge, MICCAI Workshop*, 7–15, Springer (2007).

36 X. Deng and G. Du, "Editorial: 3D segmentation in the clinic: A grand challenge II - liver tumor segmentation," (2008).

37 H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding* **110**(2), 260–280 (2008).

38 K. Frounchi, L. C. Briand, L. Grady, Y. Labiche, and R. Subramanyan, "Automating image segmentation verification and validation by learning test oracles," *Information and Software Technology* **53**(12), 1337–1348 (2011).

39 S. D. Olabarriaga and A. W. M. Smeulders, "Interaction in the segmentation of medical images: A survey," *Medical Image Analysis* **5**, 127–142 (2001).

40 K. McGuinness and N. E. O'Connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognition* **43**(2), 434–444 (2010).

41 K. McGuinness and N. E. O'Connor, "Toward automated evaluation of interactive segmentation," *Computer Vision and Image Understanding* **115**(6), 868–884 (2011).

42 E. Moschidis and J. Graham, "Simulation of user interaction for performance evaluation of interactive image segmentation methods," in *Medical Image Understanding and Analysis*, 209–213 (2009).

43 E. Moschidis and J. Graham, "A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction," in *IEEE International Symposium on Biomedical Imaging*, 928–931 (2010).

44 S. M. R. Haque, M. G. Eramian, and K. A. Schneider, "Evaluation of interactive segmentation algorithms using densely sampled correct interactions," in *Image Analysis and Processing*, *Lecture Notes in Computer Science* **8156**, 191–200, Springer Berlin Heidelberg (2013).

45 H. Nickisch, C. Rother, P. Kohli, and C. Rhemann, "Learning an interactive segmentation system," in *Indian Conference on Computer Vision, Graphics and Image Processing*, 274–281, ACM (2010).

46 P. Kohli, H. Nickisch, C. Rother, and C. Rhemann, "User-centric learning and evaluation of interactive segmentation systems," *International Journal of Computer Vision* **100**(3), 261–274 (2012).

47 J. H. Moltz, L. Bornemann, J.-M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte, M. Fabel, H.-C. Bauknecht, M. Hittinger, A. Kießling, M. Püsken, and H.-O. Peitgen, "Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans," *IEEE Journal of Selected Topics in Signal Processing* **3**(1), 122–134 (2009).

48 T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Medical Image Analysis* **13**(4), 543–563 (2009).

Frank Heckel is scientific staff member at the Fraunhofer Institute for Medical Image Computing MEVIS, Bremen, Germany. He received a diploma in applied computer science from the Chemnitz University of Technology, Germany, in 2007 and is currently pursuing his PhD at Jacobs Univer-

sity Bremen, Germany. His current research interests include medical image analysis, interactive segmentation and segmentation editing.

Biographies and photographs of the other authors are not available.

List of Figures

List of Tables