Computer Methods and Programs in Biomedicine 182 (2019) 105102

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Review

A review on multiplatform evaluations of semi-automatic open-source based image segmentation for cranio-maxillofacial surgery ☆

Jürgen Wallner[a,b,*], Michael Schwaiger[a,b], Kerstin Hochegger[b,c], Christina Gsaxner[a,b,c], Wolfgang Zemann[a], Jan Egger[a,b,c,d]

[a] Medical University of Graz, Department of Oral and Maxillofacial Surgery, Auenbruggerplatz 5/1, Graz 8036, Austria
[b] Computer Algorithms for Medicine Laboratory, Graz 8010, Austria
[c] Institute for Computer Graphics and Vision, Graz University of Technology, Inffeldgasse 16c/II, Graz 8010, Austria
[d] Shanghai Jiao Tong University, School of Mechanical Engineering, Dong Chuan Road 800, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 2 May 2019
Revised 9 September 2019
Accepted 27 September 2019

Keywords:
Multi-platform
Algorithmic
Bone segmentation
Patient-individualized medicine

ABSTRACT

Background and objectives: Computer-assisted technologies, such as image-based segmentation, play an important role in the diagnosis and treatment support in cranio-maxillofacial surgery. However, although many segmentation software packages exist, their clinical in-house use is often challenging due to constrained technical, human or financial resources. Especially technological solutions or systematic evaluations of open-source based segmentation approaches are lacking. The aim of this contribution is to assess and review the segmentation quality and the potential clinical use of multiple commonly available and license-free segmentation methods on different medical platforms.

Methods: In this contribution, the quality and accuracy of open-source segmentation methods was assessed on different platforms using patient-specific clinical CT-data and reviewed with the literature. The image-based segmentation algorithms GrowCut, Robust Statistics Segmenter, Region Growing 3D, Otsu & Picking, Canny Segmentation and Geodesic Segmenter were investigated in the mandible on the platforms 3D Slicer, MITK and MeVisLab. Comparisons were made between the segmentation algorithms and the ground truth segmentations of the same anatomy performed by two clinical experts ($n = 20$). Assessment parameters were the Dice Score Coefficient (DSC), the Hausdorff Distance (HD), and Pearsons correlation coefficient (r).

Results: The segmentation accuracy was highest with the GrowCut (DSC 85.6%, HD 33.5 voxel) and the Canny (DSC 82.1%, HD 8.5 voxel) algorithm. Statistical differences between the assessment parameters were not significant ($p < 0.05$) and correlation coefficients were close to the value one ($r > 0.94$) for any of the comparison made between the segmentation methods and the ground truth schemes. Functionally stable and time-saving segmentations were observed.

Conclusion: High quality image-based semi-automatic segmentation was provided by the GrowCut and the Canny segmentation method. In the cranio-maxillofacial complex, these segmentation methods provide algorithmic alternatives for image-based segmentation in the clinical practice for e.g. surgical planning or visualization of treatment results and offer advantages through their open-source availability.

This is the first systematic multi-platform comparison that evaluates multiple license-free, open-source segmentation methods based on clinical data for the improvement of algorithms and a potential clinical use in patient-individualized medicine. The results presented are reproducible by others and can be used for clinical and research purposes.

Contents

☆ *Note*: Due to the Editor's suggestion, this manuscript has been submitted as a Review article

* Corresponding author at: Medical University of Graz, Department of Oral and Maxillofacial Surgery, Auenbruggerplatz 5/1, Graz 8036, Austria.
E-mail address: j.wallner@medunigraz.at (J. Wallner).

https://doi.org/10.1016/j.cmpb.2019.105102
0169-2607/

1. Introduction

In the last two decades the discipline of cranio-maxillofacial surgery has undergone a remarkable rate of software based technological innovation. This is especially related to the complex three dimensional (3D) anatomy of the face in combination with the need of surgical precision and an increasing number of requests for morphological 3D visualized surgery and preoperative planning procedures [1,2]. The therefore needed advanced technological computer assistance is mostly based on 3D surface reconstructions or volume renderings of anatomical structures generated by image-based segmentation methods [3,4].

These image-based segmentation methods operate on radiological image data from computed tomography (CT), positron emission tomography (PET/CT) or magnet resonance imaging (MRI) scans [5–8]. In the last years the radiological image data has increasingly been enlarged in the most clinical centers, due to the accuracy of modern image scanners and the low time consumption of image reconstructions which has resulted in a so far ongoing and rapidly growing interest in image-based segmentation and medical 3D image analysis [7–10].

In the cranio-maxillofacial field, image-based segmentation processes constitute an important step in the diagnosis and treatment support in complex surgical cases [9–11]. Such complex surgical cases include, amongst others, the field of orthognathic surgery, complex facial bone trauma surgery, reconstructive posttraumatic or oncological surgery and implantology as well as congenital cranio-maxillofacial deformities [12].

By the use of image segmentation, anatomical structures of interest in the face and skull can virtually be localized, quantified and visualized to simulate 1) interactive treatment plans, 2) complex surgical procedures and/or 3) therapeutic outcomes in three dimensions [13,14]. Further, image-based segmentation can be used to prepare and generate 3D printed models to support diagnosis and treatment pathways [11]. If functionally stable, these computer-based procedures might lead to a precise preoperative representation of treatment goals, a shortened treatment or operation time and a more accurate therapeutic outcome [6,15].

Accordingly, image-based segmentation algorithms constitute an important step in the whole diagnosis and treatment procedure of the patient and are therefore known to be more or less the gold standard in big clinical centers when dealing with complex surgical cases.

However, although many segmentation methods are available for the cranio-maxillofacial complex [8,16–19] and many interactive medical image-based segmentation approaches can be found in the literature as given in an overview by e.g. Zhao and Xie [20], the practical in-house use of these segmentation methods in the clinical centers of head and neck or cranio-maxillofacial surgery is in fact often strongly limited due to technical, financial or human resources. This was also observed by Egger concerning other medical fields [21].

Having these points in mind, little is known about the segmentation quality, the accuracy or a systematic multiplatform evaluation of commonly available open-source based image segmentation methods. This may be the case because many clinical centers usually use commercially available segmentation methods and focus on industry licensed software packages.

Hence, there is certain lack of knowledge in the technological solution of using open-source based software segmentation methods in clinical cranio-maxillofacial surgery. This is especially true for the systematic evaluation of segmentation methods or for multiplatform comparisons of different segmentation approaches.

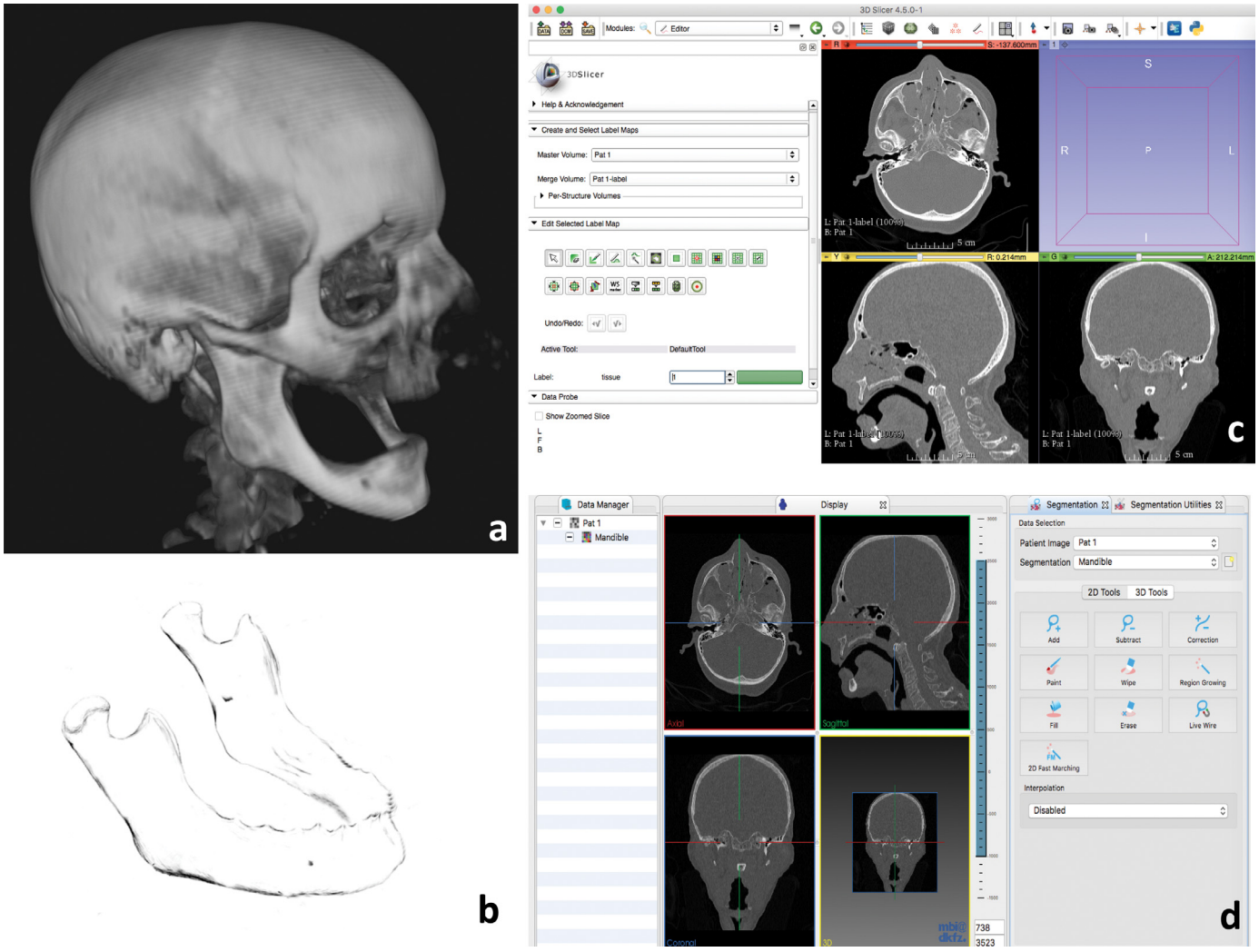


Fig. 1. a-d – A CT-based 3D model of the skull used in this contribution (**a**) and a schematic complete mandibular bone without teeth are shown (**b**). Due to their clinical relevance in trauma and reconstructive surgery the mandible was used as the anatomical structure of choice to investigate multiple open-source based segmentation algorithms on different medical software platforms. Both, non-atrophic and atrophic bones were included in the investigation, but only complete mandibles without teeth were used to form a homogenous and reproducible segmentation control group. The algorithmic segmentations were based on clinical data sets and were carried out on multiple platforms. The medical software platforms 3D slicer in the editor mode (**c**) and MITK in segmentation mode (**d**) are shown.

Thus, the aim of this contribution was to systematically assess the quality and accuracy of several open-source based segmentation algorithms compared on multiple platforms for cranio-maxillofacial surgery including a review of the literature.

Furthermore, the results gathered in this contribution are aimed to allow other groups or labs to understand and reproduce the findings and shared data of this investigation.

2. Material and methods

In this contribution, the segmentation quality and accuracy of several conventionally available and open source-based segmentation algorithms was assessed on multiple medical imaging platforms in comparison to the ground truth of the same anatomy, and reviewed with the literature. The lower jaw (mandible) was chosen as anatomical structure of reference for the segmentation procedures (Fig. 1a and b).

In this contribution, open-source based algorithms, tools and platforms were included according to the following defined points: 1) Easy availability for everybody on an open-source basis; 2) License-free version for end-users; 3) Possible reproducibility of the experiment for the reader.

All algorithms that exist “only” as publications or as pure source code repositories, e.g. on GitHub were excluded, because this would need a reimplementation or at least a compilation from someone with extensive knowledge in software engineering.

Six open-source based algorithms were selected for mandibular bone segmentation and segmentation quality assessment. The evaluated segmentation algorithms were: 1) GrowCut, 2) Robust Statistics Segmenter (RSS), 3) Region Growing 3D, 4) Otsu & Picking, 5) Canny Segmentation and 6) Geodesic Segmenter. These common algorithms were chosen to be known as functionally stable working semi-automatic segmentation methods which are compatible with many software platforms and programs such as graphic editors and are easy to use (Vezhnevets, Konouchine 2005). The six segmentation algorithms were analyzed on multiple imaging platforms in the following order: 1) 3D Slicer (GrowCut and RSS), 2) MITK (3D Region Growing and Otsu & Picking), 3) MeVisLab (Canny Segmentation, Geodesic Segmenter). The chosen platforms, 3D Slicer, MeVisLab and MITK, are easy to download, install and use for everybody. Furthermore, since these platforms are widely used in medical imaging, they are very extensive and offer good documentation and user support, also from the community.

The six open-source based algorithms were carried out on three medical imaging platforms using semi-automatic bone contouring on medical data sets of the mandible that originated from clinical routine (Fig. 1a and b).

2.1. Data selection

For the segmentation process 45 CT-data sets were provided as DICOM files and collected during the clinical routine at the department of cranio-maxillofacial surgery at the Medical University of Graz, Austria. Only high-resolution data sets (512×512), with slice thicknesses not exceeding 1.5 mm, with 0.25 mm pixel size and physiological, complete mandibular bone structures without teeth were included in the selection process. However, incomplete data sets consisting of mandibular structures altered by iatrogenic or pathological factors or fractured mandibles, as well as data sets showing ostheosynthesis materials in the lower jaw, were excluded. All data sets were acquired within a twelve month period (between 2013 and 2017). The CT scans for the data set collection were medically indicated due to diagnosis and/or treatment reasons during the clinical routine. Only already existing data sets originating from such CT scans were used for this data collection.

According to the inclusion criteria 20 CT-data sets were selected and 25 were excluded during the selection process. From the 20 CT-data sets, 10 data sets ($n = 10$, 6 male, 4 female) were further selected in a randomization process performed by a computer program (Randomizer®; https://www.randomizer.at; randomization for clinical and non-clinical trials; Graz, Austria), to form an experimental segmentation group for the algorithmic segmentations.

The control group consisted of objectively created bone structure volumes of the lower jaw (ground truth) according to the selected CT-data sets. To create these ground truth volumes for a comparative assessment, manual slice-by-slice segmentation of the randomly selected lower jaw data sets was carried out twice by two clinical experts (A, and B). The clinical experts were both specialized maxillofacial surgeons with more than ten years of clinical experience. More precisely, each data set was segmented manually by clinical expert A and B to create two independent ground truth segmentations (Ground truth A and B) of each data set. To ensure the generation of high quality ground truth data, only physiologic data sets with clear bone contours and anatomical structures without artifacts were used in this contribution, according to the mentioned inclusion criteria.

All data sets were completely anonymized by the authors before their use. Any patient specific information from the medical records was deleted during the anonymization process. Only de-identified data was used in this contribution. The data sets were provided as NRRD files.

2.2. Segmentation process

Semi-automatic segmentation with the algorithms 1) GrowCut, 2) RSS, 3) 3D Region Growing 3D, 4) Otsu & Picking, 5) Canny Segmentation and 6) Geodesic Segmenter was carried out on the selected data sets using commonly known medical image computing and scientific visualization platforms 3D Slicer, MITK and MeVisLab, which can be used in a variety of medical image-based applications [22,23] (Fig. 1c, d and 2 a). Each segmentation algorithm was applied to each data set. For each segmentation algorithm, the individual workflow provided by the medical platform was used for initialization. These medical software platforms are also conventionally and freely available and offer many additional options for medical image-based analysis such as 3D reconstructions, complex visualizations or preparations of 3D printable models. For the algorithmic segmentation process, the data set was loaded into the medical platform and the segmentation algorithm was initialized according to the algorithm's specific user-guidance. The focus was set on semi-automatic algorithms requiring a minimum of users' interaction. The only task to be fulfilled by the user was initialization. This was done by drawing seed points or lines on the medical image (CT), or setting other parameters required in the program in order to obtain a segmentation. The study participants have had five minutes of training before they started the semi-automatic segmentation. The segmentation was carried out for the complete mandibular bone without modification of the data set, post-processing or the skipping of slices. The segmentation results were saved as a 3D mask and set in comparison to the manual generated ground truth segmentations from the two clinical experts (A, B) using standard assessment parameters in image-based segmentation (Fig. 6a–d).

Slice-by-slice segmentation for the ground truth data generation was carried out on the scientific medical prototyping platform MeVisLab (Fig. 4c). According to the software's function, an individually created modular framework was integrated in the software platform for the ground truth generation. The MeVisLab software was used by the two clinical experts (A, B) to outline the complete mandibular bone in the selected CT-data sets in axial directions. The selected data sets were loaded into the platform and were successively segmented manually on a slice-by-slice basis and saved as 3D ground truth masks (Fig. 5).

After all segmentation processes were finished, 10 ground truth data sets from clinical expert A, 10 ground truth data sets from clinical expert B (control group, ground truth) and 10 semi-automatic segmentation data sets from each of the six algorithms created by a non-medical user were available (segmentation group). The data sets were compared among themselves by defined standard assessment parameters as follows: Algorithmic segmentation vs. Ground truth segmentation A, Algorithmic segmentation vs. Ground truth segmentation B, Ground truth segmentation A vs. Ground truth segmentation B. The comparison was done for each algorithmic segmentation method: 1) GrowCut, 2) RSS, 3) 3D Region Growing 3D, 4) Otsu & Picking, 5) Canny Segmentation and 6) Geodesic Segmenter on their associated medical platforms 3D Slicer (Grow Cut, RSS), MITK (3D Region Growing, Otsu and Picking) and MeVisLab (Canny and Geodesic).

In every case, all patient-specific CT-data of the complete lower jaw were included. Slices were neither deleted, nor skipped or modified, only the original and complete dataset of the whole mandible was used.

A more precise description about the used segmentation algorithms, algorithmic platforms and the ground truth references is provided in the section "Theory and detailed information".

2.3. Assessment criteria

To assess the segmentation quality and a potential clinical use of the segmentation algorithms, a user had to initialize the semi-automatic segmentation approach by marking parts of the mandibular bone and the background in axial, sagittal and coronal slices respectively. Each data set was segmented only once on the according platform by each semi-automatic algorithm.

After the segmentations were finished, the segmentation quality and accuracy of the semi-automatic segmentations was assessed by defined standard parameters. The segmentation quality was assessed through the overlap between the semi-automatic open source segmentation and the ground truth of the same anatomy by using the DICE Score coefficient (DSC,%) [33] and the Hausdorff Distance (HD, voxel) [34].

In more detail, the DSC is the agreement between two binary volumes and is calculated as follows. It measures the relative volume overlap between *R1* and *R2*, where *R1* and *R2* are the binary

masks from two segmentations. $V(\cdot)$ is the volume (e.g., in mm^3) of voxels inside a binary mask, obtained by counting the number of voxels, then multiplying that value by the voxel size (Eq. (1)).

$$DSC = \frac{2 \cdot V(R1 \cap R2)}{V(R1) + V(R2)} \quad (1)$$

The HD between two binary volumes is defined by the *Euclidean* distance between the boundary voxels of the masks. Given again the sets *R1* and *R2* of two segmentations that consist of the points that correspond to the centres of the segmentation mask boundary voxels in the two images, the directed *HD h(A,R)* is defined as the minimum *Euclidean* distance from any of the points in the first set to the second set, and the *HD* between the two sets *H(A,R)* is the maximum of these distances (Eq. (2)).

$$h(A,R) = \max_{a \in R}(d(a,R)), \ \textit{where}\ d(a,R) = \min_{r \in R} \|a - r\|$$

$$H(A,R) = \max(h(A,R), h(R,A)) \quad (2)$$

Both, DSC and HD are commonly used standard parameters in the evaluation of various techniques in volume and image rendering 7 [33,34].

The agreement between the manual slice-by-slice segmentations (ground truth A, B) was, additionally to DSC and HD, determined by volume and voxel values. This was done for an even more precise comparative assessment of the manually generated ground truth segmentations (A, B), in order to prove the validity of the used control group.

The mentioned parameters were assessed for each of the 10 data sets and for each segmentation algorithm on the according medical platform.

The semi-automatic segmentations were then directly compared to each of the two ground truth segmentations (A, B) (Fig. 6a and b). Since each ground truth data set consisted of two manual segmentations performed by two clinical experts (A, B), these two manual segmentations were also compared amongst themselves (Fig. 6d) to avoid bias causing variations in the ground truth segmentations.

Additionally, the total time spent by medical experts with the technical applications and their semi-automatic tools to achieve results of ground-truth quality was measured, taking also into account the post-processing.

2.4. Statistical methods

Descriptive statistical calculations were used to summarize the measurements and assessment parameters including minimum, maximum, mean values and standard deviations. This was performed for the comparisons of DSC and HD values between each segmentation approach and the ground truth segmentation models (A, B), as well as between the ground truth segmentations (A, B) amongst each other. Descriptive statistical analysis was also done for the total segmentation times.

Additionally, analytical statistical methods were calculated to compare the ground truth segmentations (A, B) amongst each other. These analytical statistical methods consisted of calculated paired t-tests (p) and Pearson's product-moment correlation coefficients (r) [24,25], boxplots and regression analysis including regression lines through the origin. Probability values were calculated between the ground truth segmentations (A, B). P-values under 0.05 ($p < 0.05$) were assumed to be significant.

Statistical calculations were performed using the R software (R-project®, v. 3.1.2; R Foundation for Statistical Computing, Vienna, Austria).

An applied protocol was used for a systematic comparison in order to screen the eligibility of the clinical data and assess the quality of the segmentation algorithms on the according platform (Table 2).

The complete analysis, segmentation algorithm usage and assessment parameter calculation of the several segmentation algorithms on the multiple platforms was all done on a conventional MacBook Pro (Late 2011) containing a CPU of 2.4 GHz Intel Core i5, a RAM of 4 GB 1333 MHz DDR3 and a graphics card of Intel HD Graphics 3000 384 MB. Slicer and MITK were running under Mac OS X 10.9.5 Mavericks whereas MeVisLab was used under Mac OS X 10.12.5 macOS Sierra.

This contribution was approved by the internal review board (IRB) of the medical university of Graz, Austria (IRB: EK-29–143 ex 16/17).

3. Theory and detailed information

The following section provides detailed background information about the technical content, the medical platforms, reviews the segmentation methods that were evaluated in this contribution and shows their user interfaces.

3.1. Medical imaging platforms

3.1.1. MeVisLab (MeVisLab 2.8.2)

Medical image processing development, research and scientific visualization can be done by the cross-platform modular framework MeVisLab [26] (Fig. 2a and b). A generic framework can be found in the MeVis Image Processing Library, which is object-oriented. As algorithms are self-descriptive modules, fast integration and testing as well as developing new algorithms or clinical application prototypes is possible. Combinations of functional units or modules form complex networks for image processing such as segmentation, volumetry, functional or quantitative morphological analysis or filtering. Surgery planning is one of the software aids that is realized by MeVisLab. Formats supported by MeVisLab are for example TIFF, BMP, RAW, PNG, DICOM and JPG. Fig. 2 shows networks implemented in the open-source platform MeVisLab that were used in this contribution. (www.mevislab.de).

3.1.2. MITK (MITK workbench 2016.11)

The "Medical Imaging Interaction Toolkit" offers a combination of registration, visualization and image segmentation in an application framework for image processing (Fig. 1d). The open-source software unites VTK (Visualization Toolkit) and ITK (Insight Toolkit) and is available for free. Features not covered by ITK or VTK can be realized using MITK by developing own plug-ins [27]. (www.mitk.org).

3.1.3. Slicer (3D slicer 4.5.0)

The open-source software platform 3D Slicer can be used for interactive segmentation and registration, three-dimensional visualization and volume rendering, as well as for image processing. The "Editor Mode" that is used for segmentation is shown in Fig. 1c. (www.slicer.org).

3.2. Segmentation algorithms

Each medical platform offers various algorithms and techniques that can be used for the image-based segmentation of a data set. Table 1 shows the platforms and the algorithms that were chosen in this contribution. Additional step-by-step video tutorials can be found in the section 6.2 (Tutorial availability).

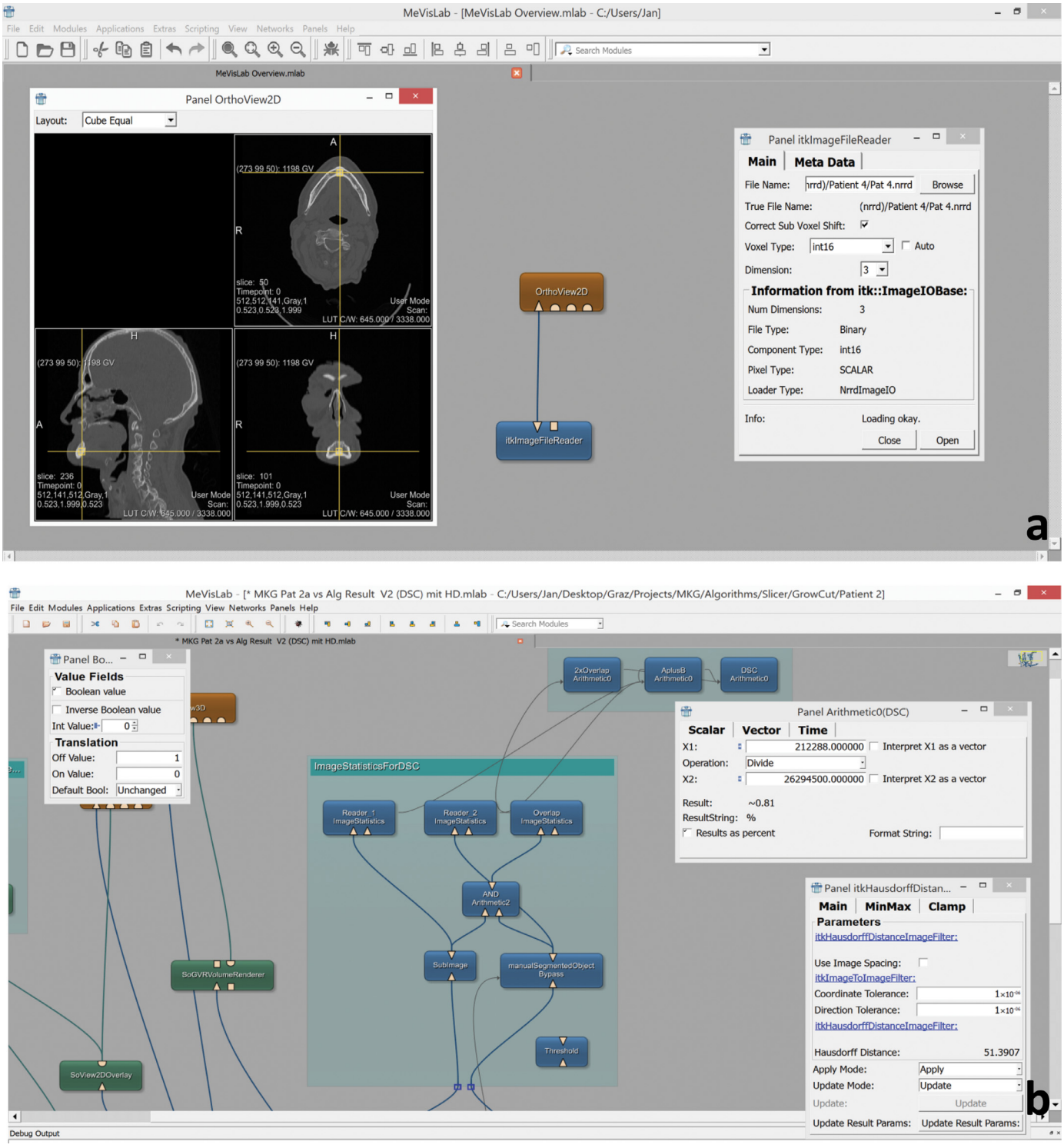


Fig. 2. **a** and **b** – The medical software platform MeVisLab is shown, which was also used to carry out the segmentations of different algorithms (**a**) and determine the Dice Score Coefficients (DSC) and the Hausdorff Distances (HD) of each segmentation and further to compare the algorithmic to the ground truth segmentations and the ground truth segmentations amongst each other. To assess the segmentation quality DSC and HD values were used as standard parameters in an individualized modular framework that was implemented in MeVisLab for this contribution (**b**).

Table 1
Segmentation algorithms and according medical platforms.

Platform	Slicer	MITK	MeVisLab
Segmentation Algorithm	GrowCut Robust Statistics Segmenter (RSS)	Region Growing 3D Otsu & Picking	ITK Canny Segmentation Level Set Image Filter ITK Geodesic Segmenter Active Contour Level Set Image Filter

Table 1: The segmentation algorithms that were evaluated in this contribution are shown according to their medical platforms. All algorithms and platforms are conventionally available, license-free and open-source based.

3.2.1. GrowCut 3.6 (3D slicer 4.5.0)

A set of scribbles inputted by the user is used for foreground and background so that the region growing algorithm works. The competitive algorithm uses cellular automata. Inputted scribbles form the basis are used to automatically compute a ROI (Region of Interest). All scribbles, especially the labels of those scribbled pixels, are used to iteratively label all remaining pixels. As soon as a pixel's label is immutable and all pixels within the ROI are labeled, the algorithm converges. The segmentation can be edited using various options when necessary. https://www.slicer.org/wiki/Modules:GrowCutSegmentation-Documentation-3.6.

3.2.2. RSS 3.6 (3D slicer 4.5.0)

The generic Robust Statistics Segmenter uses a label map to initialize a target object. Hereby, the boundary is extracted by evolving an active contour model. Contour models accurately localize nearby edges. The capture region can be enlarged by using scale-space continuation to surround the feature. The so called "snake" is an energy-minimizing spline that is influenced by image forces and guided by external constraint forces that pull it towards lines and edges [28]. https://www.slicer.org/wiki/Modules:RobustStatisticsSeg-Documentation-3.6.

3.2.3. Otsu 2016.11 (MITK workbench 2016.11)

Otsu is an algorithm based on thresholding which helps to find the ideal threshold for segmentation. There are two classes, separated through a threshold. A probability of occurrence is given for a certain value. The average grey-scale value as well as the average value of each class is known. Each pixel of a grey-scale valued image is colored either black or white when converted into a binary image through a non-linear function depending on whether the pixel is above or below a certain threshold [29]. http://docs.mitk.org/2016.11/org_mitk_views_segmentation.html#org_mitk_gui_qt_segmentationUserManual3DOtsuTool.

3.2.4. Region growing 2016.11 (MITK workbench 2016.11)

The region growing algorithm is initialized by a seed point and the definition of a threshold interval. Afterwards, the segmentation result can interactively be adapted via a slider. Region growing performs best when the region to be extracted has a great contrast to the background [30]. http://docs.mitk.org/2016.11/org_mitk_views_segmentation.html#org_mitk_gui_qt_segmentationUserManual3DRGTool.

3.2.5. ITK canny segmentation level set image filter 5.1.0 (MeVisLab 2.8.2)

A well-known algorithm for edge detection is the canny edge detector. It aims to detect existing edges as accurately as possible and avoids the detection of non-existing edges. Noise reduction is achieved with a two dimensional Gaussian filter. The size of the smoothing operator is adapted to the situation and Laplace or Sobel operators are used on the smoothed image. A thinned line can be realized through tamping down non extrema. Gradients cannot be classified exactly when their slope is average. Hysteresis is used to decide whether the gradient is an edge or not [31]. https://itk.org/Doxygen/html/classitk_1_1CannySegmentationLevelSetImageFilter.html.

3.2.6. ITK geodesic active contour level set image filter 5.1.0 (MeVisLab 2.8.2)

Until shape boundaries are reached, an initial contour evolves outwards (or inwards to form segments structures). A level set speed function is used, which is based on an edge potential map provided by the user [32]. https://itk.org/Doxygen/html/classitk_1_1GeodesicActiveContourLevelSetImageFilter.html.

3.3. Selection of segmentation algorithms

Image-based segmentation algorithms can be classified by their underlying segmentation strategy, such as thresholds, edges or regions. The algorithms in this contribution should cover each group to guarantee a profound evaluation. The choice of the algorithms, tools and platforms for this experiment was done accordingly to cover the four main methods for image segmentation which are a) pixel-value based methods, particularly thresholding using the Otsu algorithm, b) edge based methods such as Canny or Robust Statistics Segmenter c) region based, in particular Region Growing and Grow Cut, and finally d) a numerical method using Level Sets.

It was also important to avoid the use of modules that are based on the same ITK module offered by different open-source based platforms (e.g. MeVisLab and MITK both offer ITK modules).

After intensive testing of the various segmentation facilities available in the platforms Slicer, MeVisLab and MITK prior to the start of this experiment, the following algorithms were determined to provide the most potential for an efficient segmentation in the cranio-maxillofacial complex in our CT images: Grow Cut (3D Slicer), RSS (3D Slicer), Otsu (MITK), Region Growing (MITK), ITK Canny Segmentation Level Set Image Filter (MeVisLab), ITK Geodesic Segmenter Active Counter Level Set Image Filter (MeVisLab).

3.4. Usage of segmentation algorithms

For the usage of the semi-automatic segmentation algorithms, all parameters were chosen manually by the study participants based on their individual experience, as this would also be the case in a practical use. To decrease a potential variance resulting from the human factor in the usage of semi-automatic segmentation tools to a minimum, the study participants have had five minutes of training time before they started the semi-automatic segmentation.

3D Slicer is an intuitive and eminently user-friendly platform, which provides a lot of information and functions to the user. An explanation of how segmentations can be approached is given shortly, but efficiently, on the corresponding website (www.slicer.org). The initialization of GrowCut is done with a few simple steps: First an image has to be loaded using a data set and then the icon "files to add" has to be chosen. The drag down field, first stating 'Welcome to Slicer' offers various options. Here "Editor" should be chosen along with "Generic Anatomy Colours". The paint effect marks first the inner structures of the region to be segmented (positive gesture) – usually green for tissue. Next, the parts of the image that should be segmented have to be marked. Afterwards, the color of the editor (usually orange was chosen for the bone) has to be changed to set a border (negative gestures) and encapsulate the region that has to be segmented. After initialization, the three views of the image are shown and "GrowCut Effect" has to be chosen. Finally, the "Change Label Effect" shows the segmentation.

The workflow of the RSS under Slicer is as follows: after loading the image, "Editor" has to be chosen to set a seed region. RSS uses parameters such as intensity homogeneity (IH), boundary smoothness (BH) and volume to refine the segmentation. In this contribution, the parameters suggested in the RSS Tutorial were taken. In the drag down field "Segmentation – Specialized – Robust Statistics Segmenter" has to be chosen. Then the "Editor" mode has to be chosen to mark regions to be segmented. For RSS only the green color was used. For all data sets the set parameters for the RSS segmentation were the same (Volume 100, IH 0.5, BS 0), except in case 8 which required a volume of 25. This was caused by the few grey tones in the image.

Under MITK (www.mitk.org), the combination of Otsu, a thresholding algorithm, and Picking provides the possibility to first select

a number of regions that are then separated through dividing pixels on the basis of the image histogram. Afterwards, the required region can be selected by Picking. First, a file has to be loaded using "open file". A color as well as a name for the segmentation can be entered using the tool "Segmentation". In "3D Tools" the icon "Otsu" can be found, which divides the image into the number of regions set by the user. Picking then offers the possibility to choose one of these areas by clicking on it. "Confirm segmentation" removes the other areas and only leaves the one required.

To apply 3D Region Growing under MITK, a file is loaded and name and color for the segmentation are chosen. Region Growing 3D can be found in "3D Tools". It requires the definition of a threshold interval through setting a seed point by clicking onto the required area. After running the segmentation a preview of the segmentation is shown. This can then be adapted interactively using "Adapt Region Growing". "Confirm Segmentation" then saves the segmented area.

The example network available in the MeVisLab (www.mevislab.de) "Help" function was used for the application of the ITK Canny Segmentation Level Set Image Filter as well as for the ITK Geodesic Active Contour Level Set Image Filter. It combines matching components necessary for the segmentations. The Load Module was exchanged for an NRRD file reader, a base saver was added together with an NRRD file saver. After opening the image, the marker list was deleted (for not having wrong seed points saved later on) and new seed points were entered. Then, via "save base", every seed point was saved to ensure the accurate position. When setting "seed points" in slices in the region of the upper head, a great part of the skull was segmented as well along with the segmentation "running out".

3.5. Ground truth reference

The segmentation references (ground truth) were obtained by manual slice-by-slice segmentations prepared by two clinical experts (A, B). Each clinical expert segmented the same 10 data sets, in order to compare the algorithmic segmentations. The ground truth segmentation was structured as follows: the manual segmentation was outlined according to the outer compact cortical bone, which is the physiological border of an anatomical bone structure. Although the inner part of every bone contains less compact cancellous bone which has natural cavities and holes due to weight and tension reasons, both compact cortical and less compact cancellous structures form together the whole anatomy of the bone. The proportion and volume between cortical and cancellous bone can vary from one individual to another and depends on the bone quality. However, in any case, the compact cortical bone is physiologically the outer border of the whole anatomical bone structure. Therefore, the mandible was manually outlined according to the compact bone border and the inner cancellous bone structure, including cavities and holes, was "filled".

The output of an algorithmic segmentation was compared to the ground truth segmentations of the clinical experts using 3D masks. This was done for each algorithm and for the ground truth segmentations amongst each other. A modular framework was established in MeVisLab to create the 3D masks from the ground truth references and from the algorithmic segmentations. The 3D masks (3D models) were then compared using the network in Fig. 2b, which was used to calculate the DSC (%) and HD (voxel) values of each 3D mask. Comparative assessments of the segmentation quality between the 3D masks were done using DSC (%) and HD (voxel) values. The accordance between a manual ground truth and the semi-automatic segmentations was important in order to find out if an implemented segmentation algorithm on a medical software platform can be used to replace the ground truth reference performed by a clinical expert. The comparisons were performed between the algorithmic segmentations and the ground truth segmentations, and between ground truth segmentations of clinical expert A and clinical expert B amongst each other. For the ground truth segmentations, volume and voxel values were determined additionally to the DSC and HD values, to compare the ground truth segmentations in more detail. Statistical analyses supported the comparative assessments. The statistical analyses are described in the section statistical methods.

3.6. Post processing

Each medical software platform offers additional possibilities and functions to adapt, edit or manually adjust the segmentation results that the used segmentation algorithms generate. However, the results were not manipulated, filtered, added or otherwise modified by using these options to ensure a genuine reproducibility of this contribution. Furthermore, a functionally stable algorithm that is able to segment a complete bone structure without manual intervention is important for further research concerning this topic. Moreover, the applicability of conventionally available and open-source based algorithms in image-based segmentation might be improved so that these algorithms can potentially directly be used for clinical medical purposes in the future. The time needed for post processing has also been taken into account in the measurements of the total segmentation times spent (Tables 9a and 10). Using the computer hardware of this study, the time for post processing was only about 1 to 2 min for the semi-automatic segmentations, no additional post processing was applied to the manual ground truth segmentations.

4. Results

This section presents all results generated by comparing the segmentation quality and accuracy obtained by each medical software platform (Figs. 1c, d and 2a) and each investigated segmentation algorithm (Table 1) with the ground truth reference performed by two clinical experts (A, B). The steps and procedures of this systematic comparison were performed using an applied protocol (Table 2). For the assessment of the segmentation accuracy standard parameters in the evaluation of image-based segmentation such as the Dice Score Coefficient (DSC,%) [33] and the Hausdorff Distance (HD, voxel) [34] were calculated (Fig. 2b). Furthermore, total segmentation times are shown in Tables 9a and 10. The complete results of the segmentation quality assessments are shown in the Tables 3 to 18.

Table 3 shows the result of comparing the outputs of the algorithmic segmentation with GrowCut (3D Slicer) to the manually performed ground truth segmentations (Fig. 3a). Table 4 shows the result of comparing the outputs of the algorithmic segmentation with Robust Statistics Segmenter (RSS) (3D Slicer) to the manually performed ground truth segmentations (Fig. 3b). Table 5 shows the result of comparing the outputs of the algorithmic segmentation with 3D Region Growing (MITK) to the manually performed ground truth segmentations (Fig. 3c). Table 6 shows the result of comparing the outputs of the algorithmic segmentation with Otsu & Picking (MITK) to the manually performed ground truth segmentations (Fig. 3d). Table 7 shows the result of comparing the outputs of the algorithmic segmentation with Geodesic Segmenter (MeVisLab) to the manually performed ground truth segmentations (Fig. 4a). Finally, Table 8 shows the result of comparing the outputs of the algorithmic segmentation with Canny Segmenter (MeVisLab) to the manually performed ground truth segmentations (Fig. 4b).

In this contribution, the main focus lies on the comparison between open source based semi-automatic algorithmic segmentation on multiple platforms to the manually performed ground truth segmentation from clinical experts. The defined assessment

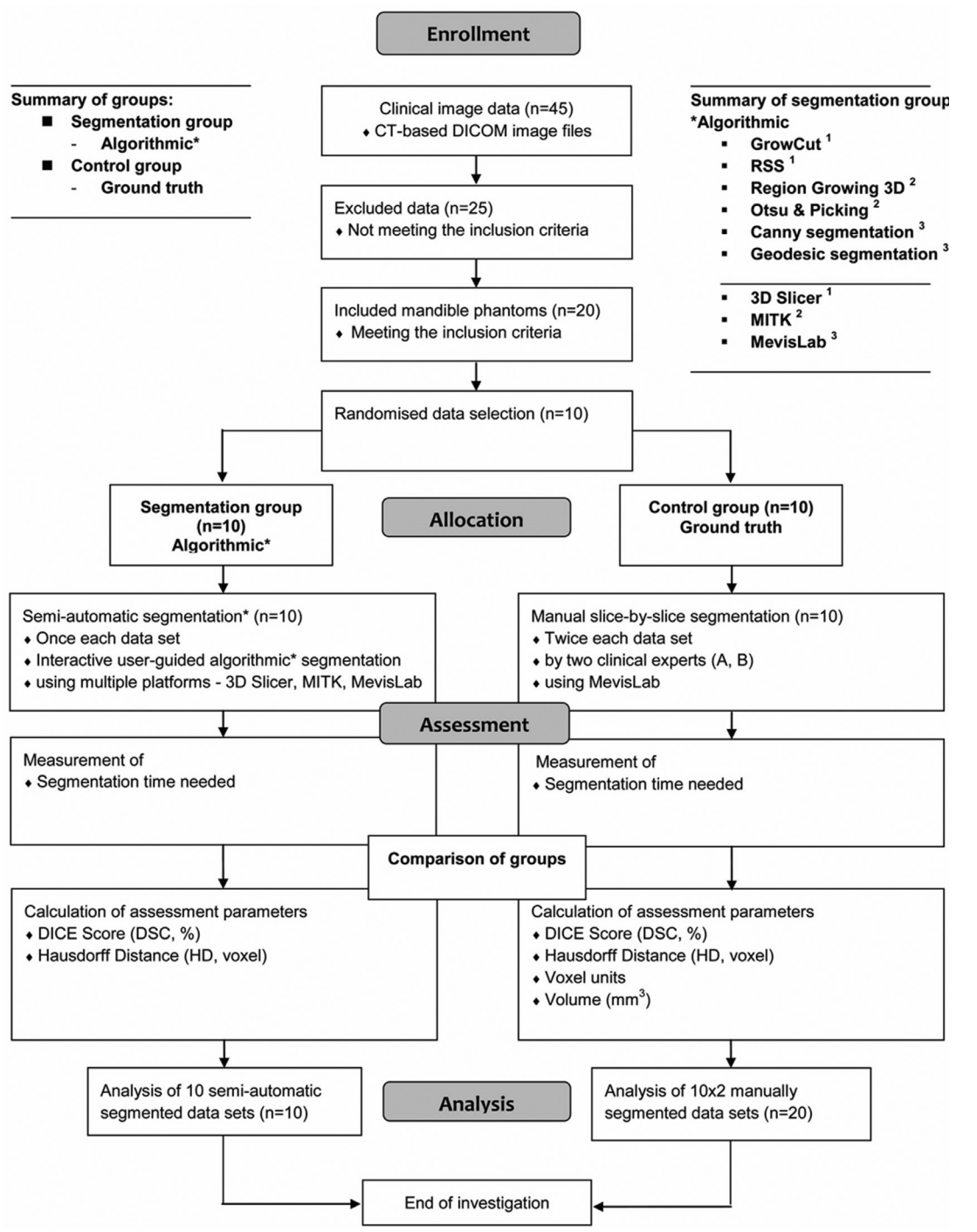


Table 2: An applied protocol was used in this contribution for a systematic comparison in order to screen the eligibility of the clinical data and assess the quality of the segmentation algorithms on the according platform. The contribution procedure and assessment parameters can be seen in the according boxes.

parameters DSC and HD obtained from each segmentation algorithm were summed over 10 patient data sets and then arithmetically averaged to obtain one comparable overall average value. This was done to further assess the segmentation quality and accuracy of the algorithm to the ground truth segmentations. These overall results are shown in Table 9.

The manual ground truth segmentations (A, B) (Figs. 4c and 5) were compared amongst each other for each case (Tables 10 and 11) and proofed for validity as a control group (Fig. 5) (Tables 12–16).

The initialization of the data sets took approximately 1 min, but computation of the algorithmic segmentation went up to 10 – 15 min depending on the platform (Slicer, MeVisLab or MITK) and algorithm. MITK's computation time was very slow, and it took a while to reach the optimum threshold parameters. MeVisLab is a CPU-intensive platform and takes a few seconds for initialization and a few minutes for running. Slicer took the longest time to compute the results (15 min) (Table 9a). Total manual ground truth segmentation times were higher with 38 min on average (Table 10).The overlap agreement between the two ground truth

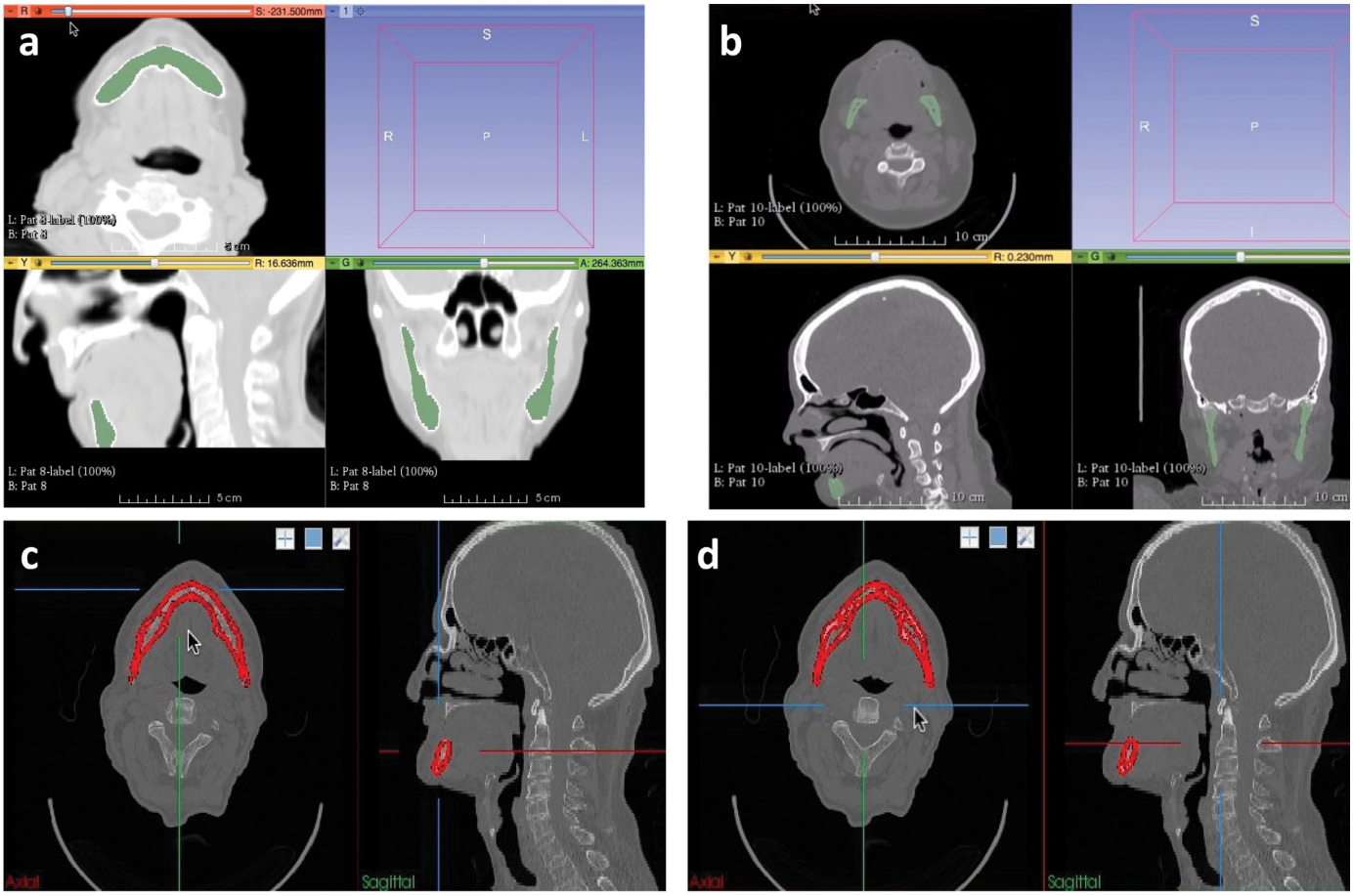


Fig. 3. a-d– The segmentation results of the open-source based algorithms GrowCut (**a**) and RSS (**b**) in 3D slicer are shown, as also the segmentation results of the algorithms Region Growing 3D (**c**) and Otsu and Picking (**d**) in MITK.

Table 3
Segmentation results: GrowCut.

Case No.	Clinical expert A		Clinical expert B	
#	DSC (%)	HD (voxel)	DSC (%)	HD (voxel)
1	83.26	29.22	83.6	27.91
2	80.73	51.39	80.66	50.96
3	82.73	21.35	83.77	20.71
4	88.42	19.65	88.69	19.34
5	80.81	57.46	80.59	57.46
6	87.80	29.10	88.79	28.86
7	86.00	29.50	86.34	33.65
8	88.27	49.49	87.76	47.84
9	90.33	19.87	89.85	19.34
10	86.28	28.14	87.49	28.25
Mean	85.46	33.51	85.75	33.43
Min	80.73	19.65	80.59	19.34
Max	90.33	57.46	89.85	57.46
SD	3.38	13.98	3.39	13.86

Table 3: The segmentation results of the algorithm GrowCut in comparison to the two ground truth segmentations of clinical expert A and B (control group) are shown using the standard parameters DSC (%) and HD (voxel). The table gives an overview about the congruence between the segmentation with GrowCut and the ground truth segmentations by clinical expert A and clinical expert B.

Table 4
Segmentation results: Robust Statistics Segmenter (RSS).

Case No.	Clinical expert A		Clinical expert B	
#	DSC (%)	HD (voxel)	DSC (%)	HD (voxel)
1	68.10	18.14	66.37	18.00
2	69.34	15.56	71.52	15.00
3	68.61	11.87	70.57	11.87
4	73.92	12.21	75.84	12.04
5	78.97	19.24	78.94	19.24
6	76.70	12.57	78.04	13.45
7	76.99	12.00	76.93	11.87
8	74.90	14.46	75.74	16.19
9	72.78	14.04	73.59	14.35
10	79.76	8.37	77.80	8.37
Mean	74.00	13.84	74.53	14.04
Min	68.1	8.37	66.37	8.37
Max	79.76	19.24	78.94	19.24
SD	4.24	3.21	3.99	3.24

Table 4: The segmentation results of the algorithm RSS in comparison to the two ground truth segmentations of clinical expert A and B (control group) are shown using the standard parameters DSC (%) and HD (voxel). The table gives an overview about the congruence between the segmentation with RSS and the ground truth segmentations by clinical expert A and clinical expert B.

segmentations (A, B) yielded to an average DSC of 94.09±1.17% and the average HD was 3.87 ± 1.21 voxel units (Table 10). The measurement values of the ground truth segmentations' volumes (Tables 12 and 13) and voxels (Tables 14 and 15) in the regression models were localized closely along the regression lines (Tables 13 a and 15 a). Especially the created boxplots were similar for volume and voxel values between the ground truth segmentations (A, B) (Tables 13b and 15b).

The calculated probability values between the volume and voxel values of the manual ground truth segmentations were not signif-

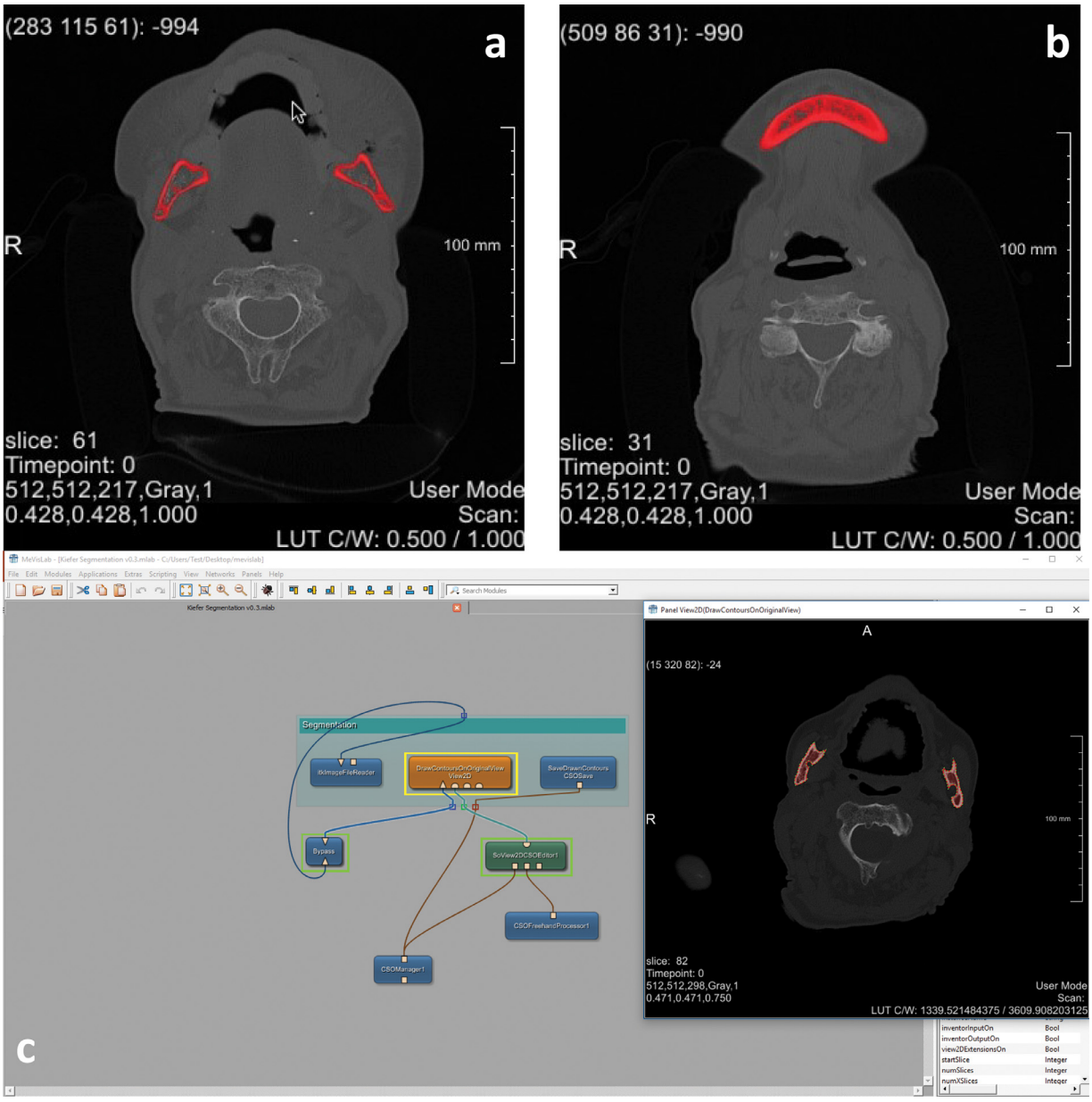


Fig. 4. a-c– The segmentation results of the open-source based algorithms Geodesic Segmenter (**a**) and Canny Segmentation (**b**) in MeVisLab are shown. Moreover, a ground truth segmentation by clinical expert A which was also carried out in MeVisLab is shown including the therefore individualized and implemented modular framework (**c**).

icant ($p > 0.05$) (Table 16). Thus, the ground truth segmentations were not significantly different from each other. Furthermore, the product-moment correlation coefficient (Pearson, r) of volume and voxel values was close to the value one ($r > 0.99$) when comparing the ground truth segmentations (Table 16).

For a visual assessment of the performed segmentations, Fig. 6 presents the overlap of an algorithmic segmentation (Fig. 6a) and a manual ground truth segmentation (Fig. 6b). When comparing these figures, a clear and more sensitive surface visualization could be achieved by the algorithmic segmentation (Fig. 6a), since the ground truth schemes were created manually according to the number of image slices (Fig. 6b). The slice-by-slice ground truth segmentation is truly visualized as a stepwise surface contouring, meaning each visualized step defines one slice, without skipping, deleting or modifying slices from the original dataset. The algorithmic segmentation is further superimposed by an anatomical 3D reconstruction, including facial bone structures (Fig. 6c). Furthermore, two ground truth segmentations performed by clinical expert A and B of one patient case were also superimposed in a 3D visualization to give an example of the coincidence of the ground truth segmentations (Fig. 6d).

5. Discussion

Computer-assisted technologies based on algorithmic software segmentation are a massively increasing topic in the medical domain [15]. This is especially valid for complex surgical cases where 3D visualization of anatomical structures and preoperative planning of surgical procedures is important in order to reduce the diagnosis and treatment time and to improve the therapeutic outcome [35].

However, in most medical fields, computer assisted technologies such as medical image processing and image-based segmentation have only just entered clinical practice within the last decade, and many of them are still in an ongoing research or development stage [36]. Hence, although an already published work investi-

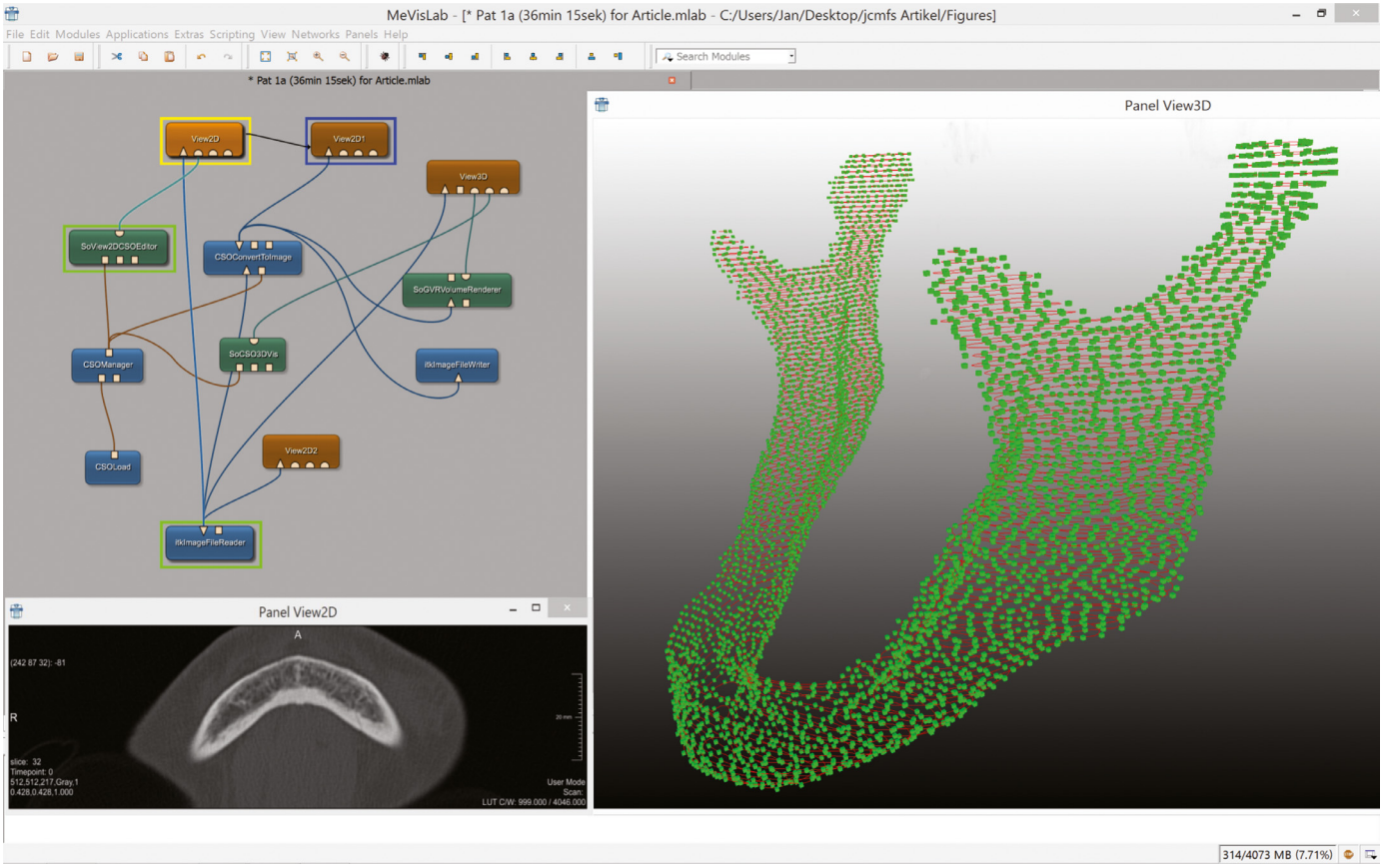


Fig. 5. Ground truth segmentation: The ground truth segmentations were achieved by manual slice-by-slice segmentations by two clinical experts under MeVisLab. The screenshot shows the MeVisLab network and its modules and connections (upper left), an axial slice to draw a contour manually (lower left) and the completely segmented mandibular bone (green) in a 3D visualization (right). The single contours have been used to generate a solid 3D mask to evaluate and compare the semi-automatic segmentations. The ground truth segmentations have been used as control group to assess the segmentation quality of the algorithmic segmentations.

Table 5
Segmentation results: 3D region growing.

Case No.	Clinical expert A		Clinical expert B	
#	DSC (%)	HD (voxel)	DSC (%)	HD (voxel)
1	34.26	22.23	34.45	22.11
2	56.73	14.87	56.84	13.49
3	63.89	7.62	64.62	7.81
4	75.42	8.25	76.07	6.71
5	71.23	6.78	71.15	7.21
6	65.02	10.05	65.17	9.27
7	68.41	8.25	68.81	8.25
8	46.65	12.37	45.81	12.33
9	33.91	29.88	33.69	30.74
10	67.36	8.37	69.46	9.11
Mean	58.29	12.87	58.61	12.70
Min	33.91	6.78	33.69	6.71
Max	75.42	29.88	76.07	30.74
SD	15.03	7.58	15.4	7.81

Table 5: The segmentation results of the algorithm 3D Region Growing in comparison to the two ground truth segmentations of clinical expert A and B (control group) are shown using the standard parameters DSC (%) and HD (voxel). The table gives an overview about the congruence between the segmentation with 3D Region Growing and the ground truth segmentations by clinical expert A and clinical expert B.

Table 6
Segmentation results: Otsu & Picking.

Case No.	Clinical expert A		Clinical expert B	
#	DSC (%)	HD (voxel)	DSC (%)	HD (voxel)
1	81.77	14.59	80.58	12.88
2	59.40	13.40	59.63	12.73
3	63.94	7.62	64.65	7.81
4	67.01	8.54	67.43	7.48
5	56.79	11.58	56.86	12.08
6	69.63	8.6	69.93	8.6
7	53.48	12.57	53.48	12.57
8	43.99	12.96	43.18	12.96
9	59.71	12.45	58.80	12.73
10	67.92	8.37	70.03	9.11
Mean	62.36	11.06	62.46	10.9
Min	43.99	7.62	43.18	7.48
Max	81.77	14.59	80.58	12.96
SD	10.26	2.53	10.41	2.33

Table 6: The segmentation results of the algorithm Otsu & Picking in comparison to the two ground truth segmentations of clinical expert A and B (control group) are shown using the standard parameters DSC (%) and HD (voxel). The table gives an overview about the congruence between the segmentation with Otsu & Picking and the ground truth segmentations by clinical expert A and clinical expert B.

gates the single outcome of the open-source segmentation method GrowCut [37], there has been a lack of research in terms of quality and accuracy and in terms of systematic comparisons of multiple segmentation algorithms and platforms. This is especially true for the challenging evaluation of license-free, open-source based segmentation approaches on a controlled clinical basis, although these open-source based segmentation approaches are easily available and can be independently used by many centers and groups for both clinical and research purposes.

Table 7
Segmentation results: Geodesic segmenter.

Case No.	Clinical expert A		Clinical expert B	
#	DSC (%)	HD (voxel)	DSC (%)	HD (voxel)
1	75.71	15.00	74.43	13.75
2	71.61	14.87	73.86	14.28
3	81.23	9.17	83.42	9.90
4	85.56	11.23	85.56	11.23
5	79.62	8.60	79.62	8.60
6	76.37	7.87	76.37	7.87
7	80.44	7.28	80.45	7.68
8	80.15	7.68	81.26	8.06
9	75.88	8.67	76.63	8.66
10	76.34	11.61	73.76	10.86
Mean	78.29	10.19	78.54	10.08
Min	71.61	7.28	73.76	7.68
Max	85.56	15	85.56	14.28
SD	3.89	2.87	4.15	2.4

Table 8: The segmentation results of the algorithm Geodesic Segmenter in comparison to the two ground truth segmentations of clinical expert A and B (control group) are shown using the standard parameters DSC (%) and HD (voxel). The table gives an overview about the congruence between the segmentation with Geodesic Segmenter and the ground truth segmentations by clinical expert A and clinical expert B.

Table 8
Segmentation results: Canny segmentation.

Case No.	Clinical expert A		Clinical expert B	
#	DSC (%)	HD (voxel)	DSC (%)	HD (voxel)
1	78.72	9.95	77.31	7.87
2	72.88	9.11	74.96	9.27
3	82.20	8.31	84.67	8.77
4	86.41	7.62	87.42	6.48
5	81.79	5.48	81.62	5.92
6	86.02	7.55	86.02	7.55
7	81.56	8.31	81.59	8.12
8	86.33	8.77	87.18	9.11
9	75.39	13.08	75.59	13.15
10	88.21	7.55	86.66	7.55
Mean	81.95	8.57	82.30	8.37
Min	72.88	5.48	74.96	5.92
Max	88.21	13.08	87.42	13.15
SD	5.06	1.98	4.87	1.99

Table 7: The segmentation results of the algorithm Canny Segmentation in comparison to the two ground truth segmentations of clinical expert A and B (control group) are shown using the standard parameters DSC (%) and HD (voxel). The table gives an overview about the congruence between the segmentation with Canny Segmentation and the ground truth segmentations by clinical expert A and clinical expert B.

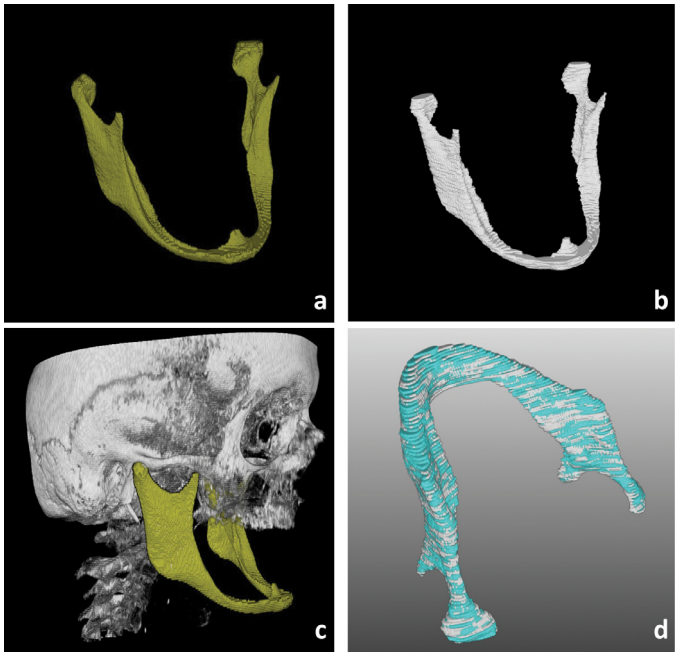


Fig. 6. a-d. The overlap and accuracy of an algorithmic segmentation (gold) (**a**) and a manual ground truth segmentation (white) (**b**) is shown. When comparing these figures a clear and more sensitive surface visualization could be achieved by the algorithmic segmentation (gold) (**a**), since the ground truth schemes were created manually according to the number of image slices (white) (**b**). The slice-by-slice ground truth segmentation is truly visualized as a stepwise surface countering, meaning each visualized step defines one slice, however without skipping, deleting or modifying slices from the original dataset. The algorithmic segmentations are further superimposed in an anatomical 3D reconstruction including facial bone structures **(c)**. Further two ground truth segmentations performed by clinical expert A (turquoise) and by clinical expert B (grey) of one patient case were superimposed in one 3D visualization to give an example about the coincidence of the ground truth segmentations (**d**). *Note: The segmented mandible is – due to missing teeth – strongly atrophied. The thin bone is well visualized by the algorithmic segmentation.*

In this investigation, selected data sets of the mandible originating from the clinical routine have been segmented manually on a slice-by-slice basis by two clinical experts (ground truth control group) and semi-automatically with multiple open-source based segmentation algorithms (interventional segmentation group). The image-based segmentations were compared successively to assess the segmentation quality and accuracy by defined assessment parameters (DSC, HD etc.) and compared with the literature. Additionally, pure generated ground truth segmentations have been compared amongst each other to prove the validity of the used control group.

DSC and HD values are known to be valid parameters for assessing the overlap and agreement of two segmented volumes [33,34,38] and were already used by others to assess interactive segmentation processes of e.g. glioblastoma multiforma in the brain [39].

Regarding to the literature, there are some articles that evaluate the accuracy of image-based software segmentation of 3D models or computer-aided 3D surface reconstructions from CT-based DICOM image files [40–46]

Concerning the cranio-maxillofacial field there are just a few articles dealing directly with the assessment of image-based software segmentation, when those only focusing on the dental medical field are excluded [33,47].

In the cranio-maxillofacial field image-based software segmentation was also done by Szymor et al. and Yan-Hui Sang et al. to assess the accuracy of software segmentation and 3D surface reconstructions in their studies [34,47]. Szymor et al. evaluated the segmentation accuracy of parts of the inner orbital wall by comparing the segmentation approach with 3D printed models of the same structure [47].

However, none of these existing articles evaluated multiple image-based software segmentation methods on different medical platforms by using clinical ground truth data or volumes of the same anatomy.

In medical visualization, ground truth data or volumes can be used to assess the quality and accuracy of an image-based software process very precisely, since a direct comparison to the real visualized volume size is feasible and can be objectively measured. The Ground truth is known as a segmented image-based virtual model that has the real visualized size or volume of the structure of interest. Therefore, the ground truth can be compared to a direct 3D cartography (mapmaking) of a structure. This was considered in our contribution to show an occurring variability in the image-based segmentation process and to provide accurate results in the assessment procedure.

The difficulty and high effort in the creation of ground truth data may be a reason for the low existence of these volumes as a controlled data sample group for the comparison of segmentation approaches and computer-aided image-based software processes. If used, the creation of more or less true and objectively valid ground truth segmentations must be ensured to avoid invalid subjectively created volumes. Objective, valid ground truth volumes are important because otherwise, an extensive bias in the comparative assessment procedure may occur.

In this contribution, the mandible (Fig. 1a and b) was chosen for segmentation, being the biggest and strongest bone in the maxillofacial complex that consists of solid biological bone structure [48]. Clinically, the mandible is often involved in trauma injury due to traffic and sport accidents or violent crimes [10,49]. Therefore, the lower jaw is clinically relevant [36] representing with about 40% the highest occurrence of all facial fractures in the craniomaxillofacial field.

The performed data-selection process in this contribution was done for the following reasons: First, the frequent involvement of the lower jaw in potentially time-consuming trauma cases and second, the solid anatomy of the bone. These reasons both lead to 1) the need of frequent clinically relevant surgical interventions such as complex osteosynthesis and lower jaw reconstructions [37], 2) the opportunity of an objective comparison between a segmentation method and the ground truth data of the same anatomical structure and 3) the opportunity of forming a homogenous control data group according to defined inclusion and exclusion criteria.

In this investigation, we selected open-source based and license free segmentation algorithms and medical platforms because these software packages are 1) easily available, 2) can be used in many centers, 3) do not create additional financial costs, 4) are independent from third parties or the industry, 5) are reproducible by others 6) offer multiple functions for image processing, 7) can be further developed and 8) were not - in contrast to some other commercial image processing packages - systematically evaluated yet.

According to our results, functionally stable segmentation outcomes could be achieved within this contribution. Overall, the results show that the GrowCut, an algorithm based on region growing, achieved the highest DSC results in this contribution matching the ground truth segmentation to 85.6% on average (Table 9a and b). The result is shown in Fig. 3a. However, although the DSC values were highest with the GrowCut segmentation, the algorithm also provided high HD values over 33 voxel, which is far beyond the values of other algorithms, meaning that there were some outliers.

Segmentations with Region Growing 3D could not provide the same segmentation accuracy as GrowCut, since Region Growing 3D only segments the outer bone border while leaving the inner cancellous cavities “empty”, which impairs the actual DSC of 58.4%, but delivers a lower HD of 12.8 voxel. Both algorithms GrowCut and Region Growing 3D are based on region growing.

The results of Otsu & Picking in Fig. 3d look similar to the result of Region Growing 3D in Fig. 3c. Table 9a and b shows that the DSC values of Region Growing 3D (58.4%) and Otsu & Picking (62.4%) are also related to each other although Otsu & Picking is an algorithm based on thresholds and Region Growing 3D is a region growing algorithm. This similarity occurs because both algorithms focus on edges of the bone without filling cancellous bone structure.

Otsu & Picking (Table 9 a and b) has a lower DSC compared to GrowCut, as it only segments the edges of the bone. While GrowCut “fills” the inner cancellous cavities and holes in the mandible, Otsu & Picking generates a model only consisting of actual osseous matter, which means cancellous cavities are left unfilled. This is shown in Fig. 3d. However, the HD of Otsu & Picking provides better results compared to the GrowCut algorithm being 10.9 respectively 33.5 voxel (Table 9 a and c).

The DSC results obtained by Geodesic Segmentation with 78.4% (Fig. 4a) and RSS with 74.3% (Fig. 3b) are similar. Both use active contour models. The deviation occurs because of seed points being placed individually, according to the platform performing the algorithms.

Canny Segmentation is the only edge-based algorithm, achieving good accordance to the clinical experts’ ground truth segmentation of a high DSC with 82.1% and a low HD of 8.5 voxel (Table 9 a–c). Although GrowCut offers the best segmentation result according to the DSC value for segmentation, Canny Segmentation is not as labor-intensive as GrowCut with regards to initializations and handling.

Initialization of segmentation parameters is the most influencing factor concerning the outcomes of the algorithms. Therefore, an appropriate period of time should be taken to gather experience using the algorithms, especially for the initialization.

However, taking both, DSC and HD into consideration, the Canny Segmentation performs best and therefore provides the best segmentation quality according to the used assessment parameters.

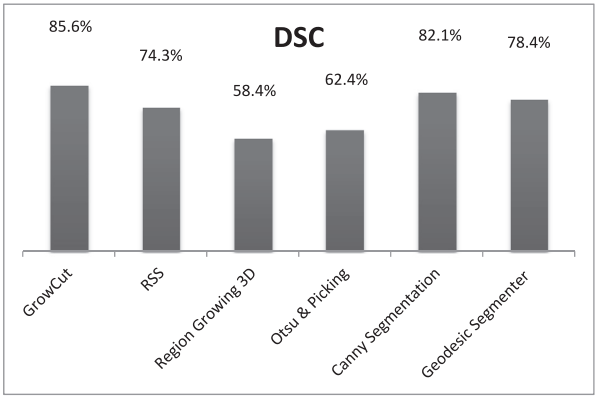
According to the clinical experience, DSC values of over 80% are in general accurate enough in cranio-maxillofacial surgery for an adequate clinically relevant use. When reviewing the literature, this experience was also observed in semi-automatic segmentation processes with GrowCut in the assessment of glioblastoma mulitforma volumetries in the neurosurgical field [27]. In the craniomaxillofacial field, DSC values of over 80% are described to be clinically mostly acceptable for 3D visualization, 3D printable model preparation or 3D template design for the preoperative orientation of osteosynthesis materials or surgical implants [37,47]. These findings are in accordance with the DSC values of the Grow Cut and the Canny segmentation method in this contribution. Regarding to the literature, these segmentation methods can both provide enough segmentation quality and accuracy to represent technological solutions in open-sourced based algorithmic image segmentation, at least when used in solid bone structures of the skull face [27,37,47].

When comparing the two ground segmentations (A, B) amongst each other, DSC values were high with $94.09 \pm 1.17\%$ and HD values were small with 3.87 ± 1.21, which shows a high coincidence (Tables 10 and 11). Also, the created regression models support these findings showing visually that the volume and voxel values were closely related to the constructed regression lines (Tables 13 and 15). Furthermore, neither volume values nor voxel units of these manual segmentations were significantly different from each other ($p > 0.05$). Moreover high direct positive correlation near the value one of both volume and voxel values between the ground truth segmentations could be observed ($r > 0.99$) (Tables 12, 14 and 16). These statistical calculations show a high similarity between the compared ground truth volumes. These results show that the used ground truth schemes were nearly identical by achieving a very high degree of segmentation overlap, although they were generated independently by two clinical experts (A, B). Therefore, the ground-truth schemes could be used as an objective, valid control group data sample without bias, not causing significant variability or comparison inaccuracy.

Recapitulating the total segmentation times spent, a segmentation can be done within 10 min semi-automatically, depending on the computer used (CPU size, age) and the used algorithm. Comparing the total average semi-automatic segmentation times (Table 9a) with the total average segmentation times spent for the ground truth reference (Table 10), the semi-automatic segmentation clearly saves time. The most time consuming part when using

Table 9
a–c Overall average results of multiplatform segmentation algorithms.

Overall Results												
	Slicer				MITK				MeVisLab			
	GrowCut		RSS		Region Growing 3D		Otsu & Picking		Canny Segmentation		Geodesic Segmenter	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD
	%	voxel	%	voxel	%	voxel	%	voxel	%	voxel	%	voxel
Mean	85.6	33.5	74.3	13.9	58.4	12.8	62.4	10.9	82.1	8.5	78.4	10.1
Min	80.59	19.34	66.37	8.37	33.69	6.71	43.18	7.48	72.88	5.48	71.61	7.28
Max	90.33	57.46	79.76	19.24	76.07	30.74	81.77	14.59	88.21	13.15	85.56	15
	Total Time of Segmentation											
	minutes											
Mean±SD	13±0.5		15±0.25		5±0.25		4±0.75		10±0.5		9±0.75	



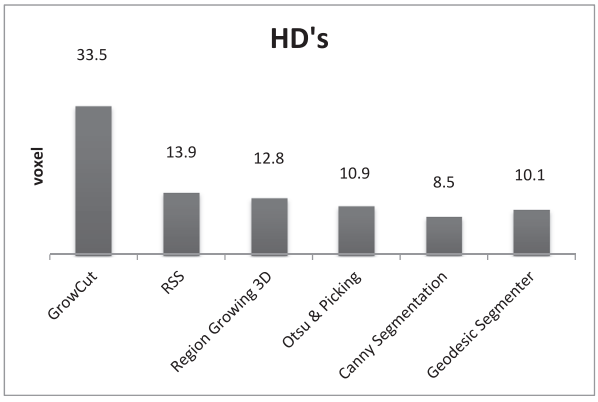


Table 9a-c: The main focus of this contribution lies on the comparison of the manual ground truth to the semi-automatic algorithmic segmentations. Both DSC (%) and HD values (voxel) (meaning result of clinical expert A and clinical expert B) of each algorithm were added, then arithmetically averaged to obtain one overall average value comparing the segmentation quality and accuracy of the algorithms to a manual ground truth segmentation. The results are descriptively shown in Table 9**a** using the standard parameters DSC (%) and HD (voxel) (**a**)). Table 9**a** further shows the average of the total segmentation time (total time) that medical experts spent with the semi-automatic tools to achieve qualitative segmentation results including post processing (*SD: standard deviation*). The **tables b** and **c** give a graphical overview about the DSC (**b**) and HD (**c**) values of these overall average results. The GrowCut and the Canny segmentation algorithm show the highest DSC (%) (**b**) values, providing the high congruence to the compared ground truth control segmentations. The Canny segmentation provides further the lowest HD (voxel) (**c**) values, also providing high congruence to the compared ground truth control segmentations. In order due to high DSC and low HD values the Canny segmentation shows the highest segmentation accuracy and offers the highest segmentation quality.

semi-automatic segmentation is the computation time of the algorithms, which requires no user input and can run in the background, while other tasks can be performed by the user. As all results were generated on the same computer (see Material and Methods), it highly depends on the age and RAM of the computer used. In every case, the segmentation on clinical data sets using semi-automatic algorithms took less time than the manual segmentation (ground truth segmentation) when high segmentation quality should be achieved.

Despite these results, we are aware of some limitations concerning this contribution: First, the segmentation methods and medical platforms assessed have not been specifically developed for cranio- maxillofacial surgery. Some of them have been used in other analysis concerning software applications [50]. Second, experimental lower jaw segmentation has, amongst others, partly been previously carried out by other groups including image-based mandibular nerve extraction [51] and has been described within a pilot project in combination with a computer-aided trauma

Table 10
Descriptive ground truth segmentation results.

Case No. #	Ground truth A DSC (%) %	Ground truth B HD (voxel) voxel	Ground truth A Total Time minutes	Ground truth B Total Time minutes
1	94.33	3.16	36	40
2	91.72	5.20	46	40
3	92.65	3.16	38	39
4	94.66	6.32	38	38
5	93.68	3.32	37	35
6	94.48	4.12	43	40
7	94.11	4.69	38	42
8	94.23	2.24	36	37
9	92.53	4.24	38	38
10	93.73	3.46	36	35
Mean	94.09	3.87	38.6	38.4
Min	91.72	2.24	36	35
Max	95.55	6.32	46	42
SD	1.17	1.21	3.31	2.27

Table 10: The segmentation results of the algorithm manually generated two ground truth segmentations of clinical expert A and B (control group) are shown in comparison amongst each other using the standard parameters DSC (%) and HD (voxel). Table 9 gives a descriptive overview about the congruence between the segmentation of clinical expert A and clinical expert B. Additionally the total segmentation times (total time) spent by the medical experts A and B with the applications is given in the times column to achieve results of ground truth quality.

Table 11
a and b Graphical ground truth segmentation results.

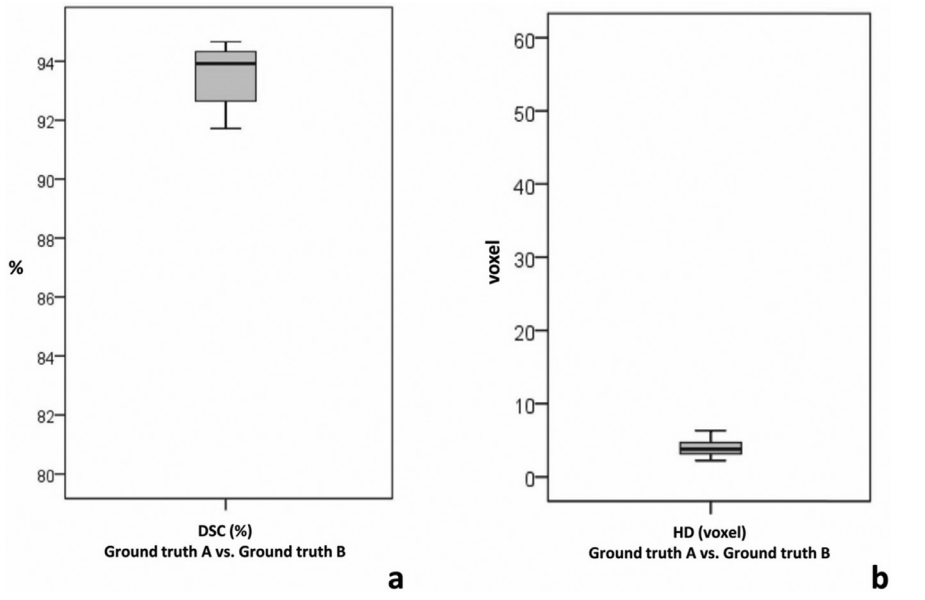


Table 11 a and b: The segmentation results of the algorithm manually generated two ground truth segmentations of clinical expert A and B (control group) are shown in comparison amongst each other using the standard parameters DSC (%) (**a**) and HD (voxel) (**b**). Table 10 gives a graphical overview about the congruence between the segmentation of clinical expert A and clinical expert B using boxplot diagrams. Between the ground truth segmentations A and B DSC (%) values are high and HD (voxel) values are low, which shows a high congruence.

simulation system by haptic feedback [10]. However, there is a missing comparative assessment of the segmentation accuracy in these previous reports. Third, although the data sets used in this contribution were randomly selected from the clinical routine, a higher amount of data-samples would probably have more impact in assessing the segmentation quality and a potential practical use. Fourth, although the ground truth generation was performed as valid as possible by two clinical experts and was proofed as a valid control group by analytical statistical calculations, a real image-based ground truth scheme is in fact impossible to create. This is the case because every segmentation method has to rely on certain image-based landmarks that have to be set at the beginning. Fifth, some difficulty occurred when segmenting the lower jaw's condoyle, since this region is physiologically overlapped by the skull base and strongly interferes with other anatomical structures. Sixth, we deliberately did not include the segmentation of teeth within our medical data sets and only used complete physiological mandibular data sets, because metal material from dental restorations often lead to strong image-based artifacts and generate incomplete or inaccurate CT-scans. Such artifacts could have interfered with the algorithmic segmentations and would have limited a multiple assessment on different platforms. Although artifacts can probably be more or less compensated in the image slices by modern CT-scanner machines after the scan is

Table 12
Ground truth volume results.

Case No.	Ground truth Volumes (mm^3)	
	Ground truth A	Ground truth B
1	30507.8	29413.4
2	17333	17730.4
3	19356.9	20067.2
4	46506.9	47508.8
5	39813.6	39733
6	30861.2	31283.1
7	45792.7	45492.8
8	31525.1	32288.9
9	18150.5	18686.3
10	32951.8	31296.5
Mean	31.28	31.35
Min	17.33	17.73
Max	46.51	47.51
SD	10.69	10.59

Table 12: The volumes measurements of the ground truth segmentations A and B are shown in mm^3 for every case and summarized in cm^3.

Table 14
Ground truth voxel results.

Case No.	Ground truth Voxel	
	Ground truth A	Ground truth B
1	166749	160767
2	118277	120989
3	54887	56901
4	84897	86726
5	153211	152901
6	96836	98160
7	211925	210537
8	77436	79312
9	123856	127,512
10	103396	98202
Mean	119147	119200.7
Min	54887	56901
Max	211925	210537
SD	± 46957.5	± 45568.9

Table 14: The voxel measurements of the ground truth segmentations A and B are for every case.

completed, the data set would have been modified from the original. Hence, CT-scans with such artifacts were excluded in the data selection process to obtain a qualitative comparative assessment of the investigated segmentation methods. Moreover, incomplete data sets, data sets including pathological or cystic lesions or data sets with artifacts, missing or damaged slices would have affected the ground truth generation due to strong occurring subjectivity in the manual segmentation process. For an adequate objectivity in the assessment of the investigated segmentation methods, we tried to create an accurate and clearly generated valid control group of the same anatomy (ground truth). Seventh, we did not segment the complete skull in this contribution, since we wanted to perform this multiplatform analysis of different segmentation algorithms on one defined anatomical structure. Eight, we want to state that the semi-automatic segmentation results may probably vary if compared to differently generated ground truth models than it was done in this contribution. However, the manual slice-by-slice outlining of the mandibular bone provided ground truth models that were accurate enough and statically valid to form a representative ground truth control group.

Ninth, we want to state that no medical CE-marked software (e.g. Brainlab, Materialise CMF Module, Maxilim, IPS Case Designer) was evaluated in this contribution, although CE-marked software is - like other computer assisted technologies - becoming increasingly important in software-based surgical planning procedures. However, with this investigation we did not want to create a new gold standard by evaluating multiple open-source based software segmentation approaches. Rather, we wanted to provide a first systematic analysis of multiple segmentation approaches on several medical platforms that are independently available, reproducible and show a different course of action in image-based segmentation without commercial and license based complex software packages. In any case, CE-marked software is probably the gold standard in some head and neck surgery departments, but only if these software packages are fully available and supported at the clinical center. Still, numerous departments do not routinely work with these CE-marked software packages since the packages are functionally complex, highly expensive and usually need additional human or technical resources for their use. This may be the reason why software packages, especially without CE-marks, are still investigated scientifically in other studies for a potential clinical use [47].

Table 13
a and b Regression analysis of ground truth volume measurements.

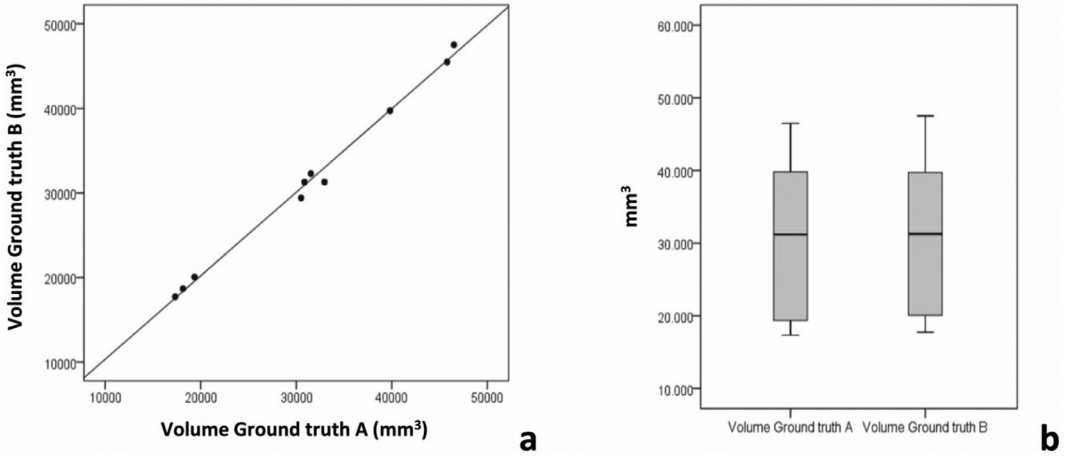


Table 13: Ground truth (A, B) volume measurements (mm^3) are graphically compared in a regression model (**a**) and a boxplot diagram (**b**). The volume measurement distributions are similar along the regression line between the ground truth segmentations A and B (**a**). Moreover volume measurement distributions are closely located along their regression line (**a**). Volume measurements of the segmentations are further shown in a boxplot diagram (**b**), providing nearly equally ground truth segmentations.

Table 15
a and b Regression analysis of ground truth voxel measurements.

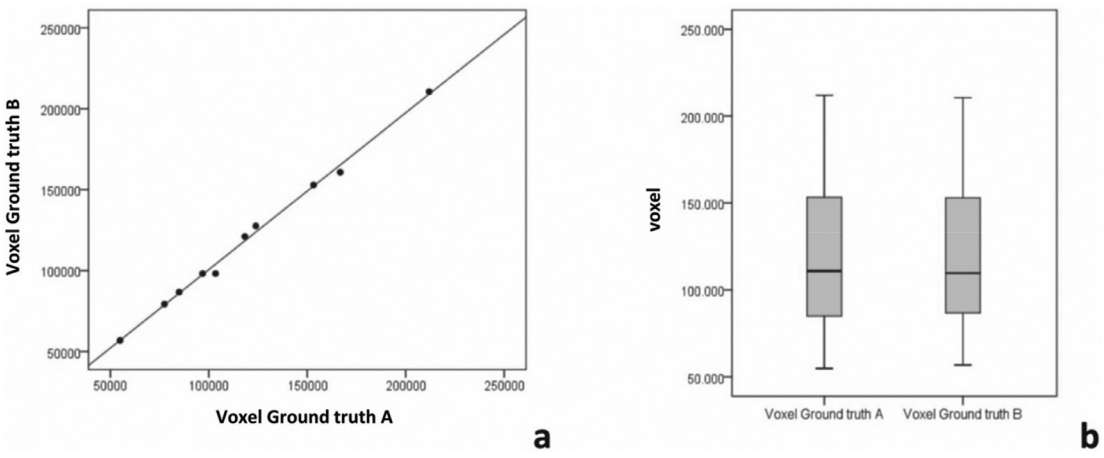


Table 15: Ground truth (A, B) voxel measurements are graphically compared in a regression model (**a**) and a boxplot diagram (**b**). The voxel measurement distributions are similar along the regression line between the ground truth segmentations A and B (**a**). Moreover voxel measurement distributions are closely located along their regression line (**a**). Voxel measurements of the segmentations are further shown in a boxplot diagram (**b**), providing nearly equally ground truth segmentations.

Table 16
Comparison of ground truth segmentations.

Comparison of Ground truth segmentations (A, B)		
Volume	Significance (p)	Coefficient (r)
Ground truth A vs. Ground truth B	$p = 0.803$	$r = 0.997$
Voxel	Significance (p)	Coefficient (r)
Ground truth A vs. Ground truth B	$p = 0.960$	$r = 0.998$

Table 16: Probability values (pared *t*-test, p) and Product-moment correlation (Pearson, r) for volume and voxel comparisons of the manual ground truth segmentations (A, B) are shown to assess their difference. No statistical significance ($p<0.05$) was observed between the segmented volumes or number of voxels. A high direct proportional correlation (r) close to the value one can be seen between the segmented ground truth segmentations. This both show that the used ground truth segmentations were not significantly different from each other and also reached a very high degree in their correlation to each other. In order these manually generated segmentations by two clinical experts (A, B) can be seen as valid ground truth control samples in this contribution. *Note: The probability (p) and product-moment correlation (Pearson, r) calculations were based on the volume and voxel values that are shown in the Tables 12–15.*

According to the results of this contribution, the analyzed segmentation methods have all been functionally stable. The GrowCut and especially the Canny segmentation algorithm might be relevant for the segmentation processes in patient-individualized medicine, which has recently enlarged to a high potential topic in today's clinical and theoretical-experimental medicine [52,53]. The high relevance in research concerning this fact and the importance of ongoing developments and assessments in 3D image-based processes such as segmentation-based 3D reconstruction of image acquisitions can also be found in the literature [41,42,54,55].

In that context, the image-based segmentation processes of both the GrowCut and the Canny algorithm can clinically be valid for 3D visualizations, preparation of 3D printable models, 3D printing and/or patient-specific 3D template design for osteosynthesis materials or surgical implant adaption for both research and clinical purposes.

The investigated open-source based segmentation approaches could be relevant for head and neck departments that try to avoid additional financial costs, which strongly occur in medical image-based analysis and image-based software processing due to monetary and licensed-based services. For example, only the 3D printing industry's revenue worldwide from products and services is over US$ 4 billion and fast growing, with 13.1% of the industry attributed to the medical sector [56].

Taking these points and the results of this contribution into account, the investigated GrowCut and Canny segmentation algorithms may provide accurate segmentation results on the used medical platforms and adequate segmentation quality when clinically used as mentioned above. Similar, this can also be found when reviewing the literature, since the results presented in this article are in some accordance to already existing works that focus on the assessment of segmentation approaches based on manufactured 3D model production [29,30–32,35] or the segmentation of tumors and cDNA in other medical and surgical fields [27,57–59]. Further the findings of this contribution are also in correlation with the findings of Szymor et al., that an open-source image-based software segmentation method is adequate enough for a potential clinical use [47]. According to Szymor et al. we can also suggest the use of open-source based software packages in the clinical practice, especially regarding the GrowCut and the Canny segmentation approach.

Concerning the Pro and Cons of the technical methods used in this contribution, the following can be stated (Table 17):

Slicer: Slicer is an intuitive and self-explaining platform, nevertheless application and handling of the algorithms demands experience.

As region growing works best when there is a strong contrast between the region being extracted and the background, images consisting of only white and light grey need special procedures to obtain the best result.

RSS using active contour models react sensitively to numerousness of seed points, while lines or encircled lines either produce fragmentary or leaking segmentations. Although computation time took the longest using Slicer, about 10 – 15 min, it has proven to run stably and achieve the best congruence between manual and automatic segmentations. Computation time varies using different hardware components, as it is hardware dependant.

MITK: Otsu and Picking is the easiest and fastest way to conduct segmentations as there are no parameters to be set, and only a number of regions is selected.

The DSCs and HDs appear not to be as high, because this method did not fill cavities. Additional editing might enhance the results. This also extends to 3D Region Growing as it only segments bone, not filling cavities. Therefore the DSC is not as accurate as for

algorithms segmenting bone and cavities. It is not as applicable for clinical uses as GrowCut might be if cavities are needed.

Results highly depend on the mandibular form. Selection of parameters is difficult as there is no common theme to be followed. The thresholds need not be selected as the seed point defines the upper and lower threshold.

Due to the interactive adaption to region growing, the segmentation time is not prolonged.

Otsu & Picking segments only the bone and takes no longer than 5 min. There are no parameters to choose which makes Otsu & Picking a fast and easy method for segmentation. Otsu allows selecting a number of regions (from 4 to 6). The head is divided into these regions, so that the required region can then be chosen with Picking. The other regions disappear. The division into regions depends on the grey tones of the image and significant borders between bone and tissue.

MeVisLab: Trials applying the ROI module for both algorithms showed that the segmentation is more accurate, even in upper mandibular parts, since the skull is not segmented. The inclusion of more seed points in superior slices is enabled for that reason.

Unfortunately the application of ROI resulted in a reformatting of the image, thus the comparability of the segmentation was lost. As a consequence thereof the idea of using ROI was discarded.

The algorithms achieved similar results when considering the average, lowest and roughly the highest DSCs and HDs. The use was straightforward after dealing with modules, applications and principles of the algorithms. The example network (Fig. 7) was used to create the segmentations. Slight modifications concerning the process of image saving and reading were made in order to allow the use of NRRD files. The View 2D module was opened to directly set seed points on the image. The first run provides a segmentation shown in red which can be modified by setting more seed points.

More seed points in upper regions can mean a better segmentation which extracts more of the lower jawbone, as the active contour model evolves including the temporomandibular joint region without picking parts of the skull. The active contour model could also evolve including the skull, together with blemishes around the temporomandibular joint region.

Although several modules in the network provide parameters to be selected individually, the default settings were used, except for the Sigmoid Filter. Beta was increased to allow a segmentation closer to the edges of the mandible (front U shape) as well as to erase points left out. It was not advisable to set Beta very high as the result did not improve. The mandibular front was segmented along with fuzzy boundaries. However, when Beta is set too low the segmentation runs out and leaves blemishes. Beta can be adapted interactively and the segmentation is computed within a few seconds.

In summary, complete functionally stable and time saving image-based segmentation could be performed by the algorithmic segmentation methods. At least, the semi-automatic segmentation quality performed by the GrowCut and especially by the Canny algorithm might be accurate enough for a potential clinical use. The segmentation outcome provided by the Canny algorithm was also quite close to the clinical experts' ground truth segmentations, although the algorithm could not fully replace the clinical experts' segmentation accuracy. Additionally, advantages of the investigated segmentation methods are 1) a free access to the segmentation software and the used platform, 2) a conventional instead of commercial segmentation approach (meaning the avoidance of license-based monetary services e.g. outsourced services or acquisition of monetary software) and 3) a clinical relevant use due to accurate and valid segmentation results that is generally not compulsory limited to the mandible as long as solid bone structures are used.

Thus, supporting tasks for surgical diagnosis and treatment procedures in patient-individualized medicine [40,41,52] such as 1) 3D visualization, 2) preparation of 3D printable models, 3) 3D printing and/or 4) patient-specific surgical 3D template design for osteosynthesis material or surgical implant adaption in the clinical practice can be directly performed in-house on an open-source basis. This can be valid for both complex surgical cases and the clinical routine.

Table 17
Pro and Cons of segmentation algorithms and according medical platforms.

Pros of Technical Methods			
Platform	Slicer	MITK	MeVisLab
Segmentation Algorithm	GrowCut	Region Growing 3D	Canny Segmentation
	-Achieved best segmentation results	-Combination of placing seed points and using the Adapt Region Growing slider, which gives instant feedback of the segmentation results	-Needs only very few seed points (around three)
		-Unfilled cavities in the cancellous bone*	-Mostly filled cavities in the cancellous bone*
	RSS	Otsu & Picking	Geodesic Segmentation
	-Mostly filled cavities in the cancellous bone*	-No need of precise seed point placement, (only region has to be chosen)	-Needs only few seed points (around five)
		-Unfilled cavities in the cancellous bone*	-Unfilled cavities in the cancellous bone*
Cons of Technical Methods			
Platform	Slicer	MITK	MeVisLab
Segmentation Algorithm	GrowCut	Region Growing 3D	Canny Segmentation
	-Brush-based user input of fore- and background in several slices	-Needs training/experience in placing seed points and using Adapt Region Growing slider	-Needs training/experience in placing seed points
		-Unfilled cavities in the cancellous bone*	-Mostly filled cavities in the cancellous bone*
	RSS	Otsu & Picking	Geodesic Segmentation
	-Several seed points and parameter settings	-No options to influence or enhance the segmentation result	-Needs training/experience in placing seed points
	-Needs training/experience in placing seed points	-Unfilled cavities in the cancellous bone*	-Mostly filled cavities in the cancellous bone*

Table 17: Overview about the Pro-and Cons of the used technical methods. Advantages and disadvantages of the segmentation algorithms and the according medical platforms are shown. * *Can be a pro and a con, depending if a user needs only the "unfilled" cavities or the "filled" cavities and holes in the in the cancellous bone of the lower jaw.*

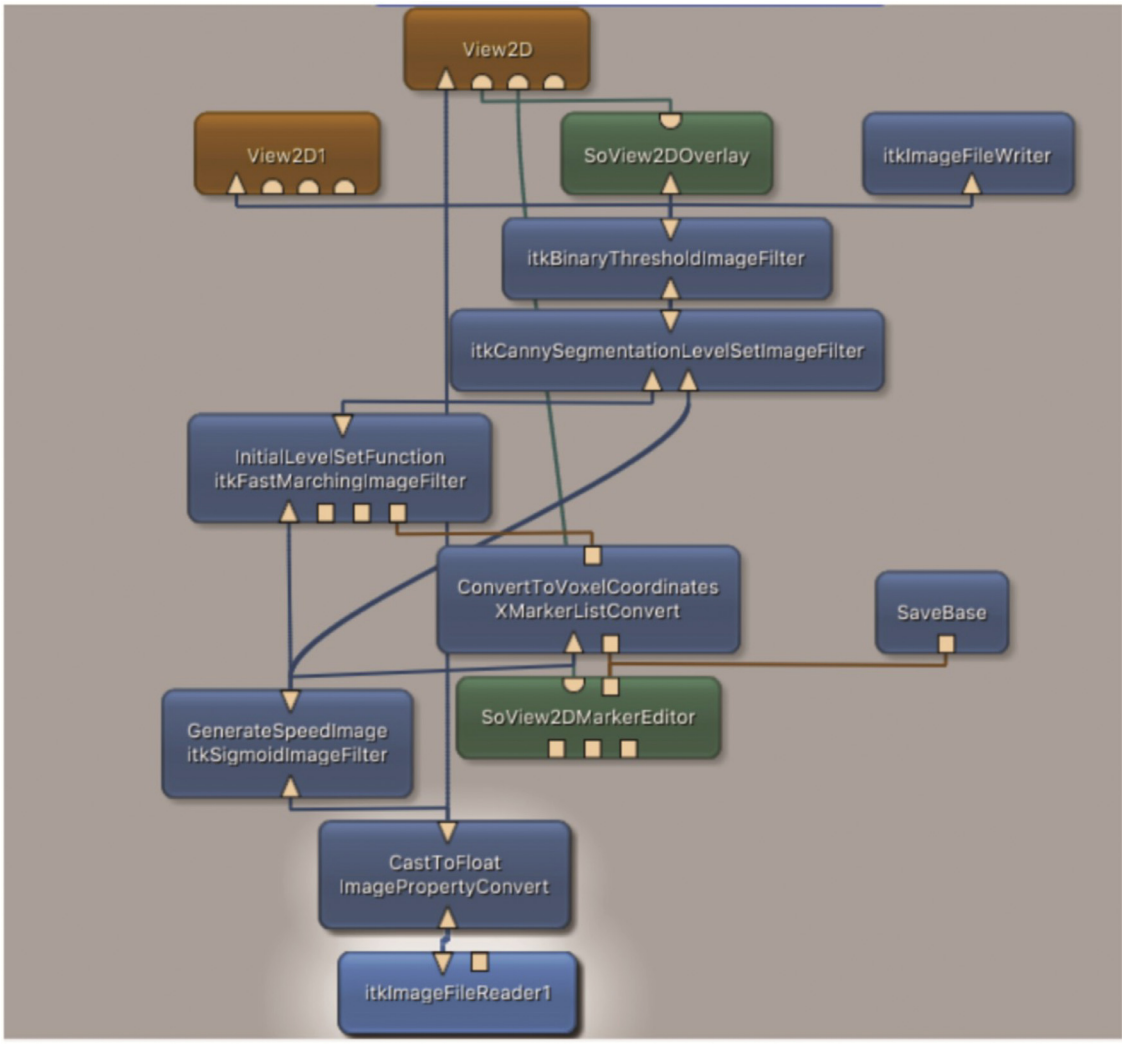


Fig. 7. The example network that was used to create the segmentations is shown. In this example the network was used for semi-automatic segmentation with the Canny algorithm (MeVisLab).

Segmentation is typically the first step in a medical image analysis pipeline, therefore incorrect segmentation affects any subsequent steps. Automatic medical image segmentation is known to be one of the most complex problems in image analysis and still under active research. Zhang estimated already in 2006 that there are over 4000 image segmentation algorithms [60]. However, the majority of these algorithms are only available locally to the research groups who developed them and an own usage would need a reimplementation. Therefore the main advantage of this contribution is the availability of the algorithms for end users, which stands in strong contrast to image processing techniques and segmentation algorithms that just exist on papers or as descriptions and must be re-implemented by software developers for usage. Before creating and implementing new algorithms for a certain anatomical structure or part of the body, an analysis of how well already existing tools work is necessary in order to provide information about the existing algorithmic usability.

Therefore, the gathered knowledge in this contribution might help to improve ideas, eradicate deficiencies or maintain efficient strategies in image-based segmentation.

The data and results presented in this investigation are objectively reproducible by others because only conventionally and already existing software platforms have been used that are available for everybody.

For a future work, the segmentation results achieved within this contribution can support the computer-aided reconstruction of facial defects with miniplates or support oral and maxillofacial implantological procedures [61,62]. The 3D reconstructions from the segmentations can be used for a patient-individualized treatment support [40,41,52] in the facial area, where both a functional and an aesthetic outcome is very important for postoperative life quality and rehabilitation. Moreover, the results can be imported into recently released devices such as medical Augmented Reality (AR) systems for surgical navigation [63], Virtual Reality (VR) environments [64] or optical see-through head-mounted displays (HMD) [65] to be used e.g. for the resection of tumors or complex surgical cases in cranio-maxillofacial and head and neck surgery. The results can be imported into AR for intraoperative guided therapy and in VR for a photorealistic preoperative planning.

The ground truth data and segmentations of this contribution can be used for the training procedures of deep learning networks, for a fully-automatic segmentation of pathological lesions such as tumors in CTs or PET/CTs or for a further comparison of CE-marked or ISO certified software packages such as Brainlab, Materialise

CMF Module or other third-party toolkit algorithms.

Platform availability

MeVisLab 2.8.2: www.mevislab.de
ITK Canny Segmentation Level Set Image Filter 5.1.0: https://itk.org/Doxygen/html/classitk_1_1CannySegmentationLevelSetImageFilter.html
ITK Geodesic Active Contour Level Set Image Filter 5.1.0: https://itk.org/Doxygen/html/classitk_1_1GeodesicActiveContourLevelSetImageFilter.html
MITK Workbench 2016.11: www.mitk.org
3D Otsu 2016.11: http://docs.mitk.org/2016.11/org_mitk_views_segmentation.html#org_mitk_gui_qt_segmentationUserManual3DOtsuTool
3D Region Growing 2016.11: http://docs.mitk.org/2016.11/org_mitk_views_segmentation.html#org_mitk_gui_qt_segmentationUserManual3DRGTool
3D Slicer 4.5.0: www.slicer.org
GrowCut 3.6: https://www.slicer.org/wiki/Modules:GrowCutSegmentation-Documentation-3.6
RSS 3.6: https://www.slicer.org/wiki/Modules:RobustStatisticsSeg-Documentation-3.6

Data availability

The CT-data used for the assessments in this investigation can freely be downloaded for own research and/or reproducibility purposes, but we kindly asked to cite our work.

Wallner J, Egger J (2018). Mandibular CT Dataset Collection. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.6167726.v2.

Tutorial availability

Tutorial videos demonstrating the interactive segmentations can be found under the following YouTube-channel:
https://www.youtube.com/c/JanEgger/

Step-by-step tutorial videos for the semi-automatic segmentation algorithms can be found under the following online-weblinks:
GrowCut: https://www.youtube.com/watch?v=WZoEu0z1z3o.
RSS: https://www.youtube.com/watch?v=K2tz3j2wEfc.
Region Growing 3D: https://www.youtube.com/watch?v=Iv4T_GlWapA.
Otsu & Picking: https://www.youtube.com/watch?v=zUN02PD7xHc.
Canny Segmentation: https://www.youtube.com/watch?v=ro3Wl4v880I.
Geodesic Segmentation: https://www.youtube.com/watch?v=uowlox8K2e8.

Author contribution statement

Conceived and designed the experiments: JW JE. Performed the experiments: JW JE. Analyzed the data: JW KH JE. Contributed reagents/materials/analysis tools: JW MS KH WZ JE. Wrote the paper: JW JE.

Additional information

Additional information or data concerning this contribution can be provided by the authors, if requested.

Declaration of Competing Interest

The authors of this paper have no potential conflict of interest. The authors disclose any financial and personal relationships with other people or organizations that inappropriately influence (bias) this work. This disclosure includes employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications/registrations and grants or other funding.

Acknowledgements

This investigation was approved by the Internal Review Board (IRB) of the Medical University of Graz, Austria (IRB: EK-29–143 ex 16/17).

The work received funding from the Austrian Scientific Fund (FWF-KLIF): “enFaced: Virtual and Augmented Reality Training and Navigation Module for 3D-Printed Facial Defect Reconstructions” (KLI-678-B31; P.I.: Jürgen Wallner and Jan Egger) and the TU Graz Lead Project (Mechanics, Modeling and Simulation of Aortic Dissection).

Some of the present data of this collection have partly been used in already published works [37,60,66–68]. In these publications the data was used as testing data for single algorithms or deep learning networks. Some CT datasets were used to test a deep learning network [67] in the mandible for full automatic network-based segmentation and to test the accuracy of the single open-source algorithm GrowCut (www.growcut.com) in the mandible [37]. With this investigation [37], parts of the CT data used for algorithmic testing were initially uploaded as a figshare repository for reproducibility reasons because of the publication requirements of the journals. The overall HD and DSC results presented in this publication were partly presented within a conference paper [66]. One CT-dataset was used in .stl (standard triangle language) file format as a surface model without segmentation to evaluate a software tool for computer-aided positioning planning of miniplates for oral & maxillofacial surgery [60]. Further, a collection of the compared manual segmented ground truth models was made available for end users out of reproducibility reasons [67].

However, only this manuscript includes the complete multiplatform comparison and systematic evaluation of the multiple segmentation methods, the detailed results and the full amount of all data. This especially includes the overall results and data of all compared segmentation algorithms and the different platforms which are only within this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2019.105102.

References

[1] S. Mazzoni, G. Badiali, L. Lancellotti, L. Babbi, A. Bianchi, C Marchetti, Simulation-guided navigation: a new approach to improve intraoperative three-dimensional reproducibility during orthognathic surgery, J. Craniofac. Surg. 21 (6) (2010) 1698–1705.
[2] M.J. Zinser, H.F. Sailer, L. Ritter, B. Braumann, M. Maegele, J.E. Zoller, A paradigm shift in orthognathic surgery? A comparison of navigation, computer-aided designed/computer-aided manufactured splints, and "classic" intermaxillary splints to surgical transfer of virtual orthognathic planning, J. Oral Maxillofac. Surg. 71 (12) (2013) 2151.e1–2151.e21.
[3] C. Chu, D.L. Belavý, G. Armbrecht, M. Bansmann, D. Felsenberg, G. Zheng, Fully automatic localization and segmentation of 3D vertebral bodies from CT/MR images via a learning-based method, PLoS ONE 10 (11) (2015) e0143327.
[4] D. Zukic, A. Vlasák, J. Egger, D. Horinek, C. Nimsky, A. Kolb, Robust detection and segmentation for diagnosis of vertebral diseases using routine MR images, Comp. Graph. Forum 33 (6) (2014) 190–204.
[5] S.D. Olabarriaga, A.W. Smeulders, Interaction in the segmentation of medical images: a survey, Med. Image Anal. 5 (2) (2001) 127–142.
[6] G. Orentlicher, D. Goldsmith, A. Horowitz, Applications of 3-dimensional virtual computerized tomography technology in oral and maxillofacial surgery: current therapy, J. Oral Maxillofac. Surg. 68 (8) (2010) 1933–1959.
[7] J. Egger, R.R. Colen, B. Freisleben, C. Nimsky, Manual refinement system for graph-based segmentation results in the medical domain, J. Med. Syst. 36 (5) (2012) 2829–2839.

[8] L. Wang, K.C. Chen, Y. Gao, F. Shi, S. Liao, G. Li, S.G. Shen, J. Yan, P.K. Lee, B. Chow, N.X. Liu, J.J. Xia, D. Shen, Automated bone segmentation from dental CBCT images using patch-based sparse representation and convex optimization, Med. Phys. 41 (4) (2014) 043503.
[9] S.C. Schvartzman, R. Silva, K. Salisbury, D. Gaudilliere, S. Girod, Computer-aided trauma simulation system with haptic feedback is easy and fast for oral-maxillofacial surgeons to learn and use, J. Oral Maxillofac. Surg. 72 (10) (2014) 1984–1993.
[10] M.T. McCann, M. Nilchian, M. Stampanoni, M. Unser, Fast 3D reconstruction method for differential phase contrast X-ray CT, Opt. Express 24 (13) (2016) 14564–14581.
[11] S. Raith, S. Wolff, T. Steiner, A. Modabber, M. Weber, F. Hölzle, H. Fischer, Planning of mandibular reconstructions based on statistical shape models, Int. J. Comput. Assist. Radiol. Surg 12 (1) (2017) 99–112.
[12] R. Olszewski, Three-dimensional rapid prototyping models in cranio-maxillofacial surgery: systematic review and new clinical applications, Proc. Belgian R. Acad. Med. 2 (43) (2013) e77.
[13] M. Poon, G. Hamarneh, R. Abugharbieh, Efficient interactive 3D Livewire segmentation of complex objects with arbitrary topology, Comp. Med. Imag. Graph. 32 (8) (2008) 639–650.
[14] G. Badiali, V. Ferrari, F. Cutolo, C. Freschi, D. Caramella, A. Bianchi, C. Marchetti, Augmented reality as an aid in maxillofacial surgery: validation of a wearable system allowing maxillary repositioning, J. Cranio-Maxillo-Fac. Surg. 42 (8) (2014) 1970–1976.
[15] S. Tucker, L.H. Cevidanes, M. Styner, H. Kim, M. Reyes, W. Proffit, T. Turvey, Comparison of actual surgical outcomes and 3-dimensional surgical simulations, J. Oral Maxillofac. Surg. 68 (10) (2010) 2412–2421.
[16] D. Terzopoulos, T. McInerney, Deformable models and the analysis of medical images, Stud. Health Technol. Inform. 39 (1997) 369–378.
[17] T. McInerney, G. Hamarneh, M. Shenton, D. Terzopoulos, Deformable organisms for automatic medical image analysis, Med. Image Anal. 6 (3) (2002) 251–266.
[18] Y. Kang, K. Engelke, W.A. Kalender, Interactive 3D editing tools for image segmentation, Med. Image Anal. 8 (1) (2004) 35–46.
[19] D. Kainmueller, H. Lamecker, H. Seim, M. Zinser, S. Zachow, Automatic extraction of mandibular nerve and bone from cone-beam CT data, Med. Image Comput. Comput. Assist. Interv. 12 (2) (2009) 76–83.
[20] F. Zhao, X. Xie, An overview of interactive medical image segmenation, Ann. BMVA 2013 (7) (2013) 1–22.
[21] J. Egger, Refinement-cut: user-guided segmentation algorithm for translational science, Sci. Rep. 4 (2014) 5164.
[22] N. Archip, O. Clatz, S. Whalen, D. Kacher, A. Fedorov, A. Kot, N. Chrisochoides, F. Jolesz, A. Golby, P.M. Black, S.K. Warfield, Non-rigid alignment of pre-operative MRI, fMRI, and DT-MRI with intra-operative MRI for enhanced visualization and navigation in image-guided neurosurgery, Neuroimage 35 (2) (2007) 609–624.
[23] Y. Hirayasu, M.E. Shenton, D.F. Salisbury, C.C. Dickey, I.A. Fischer, P. Mazzoni, T. Kisler, H. Arakaki, J.S. Kwon, J.E. Anderson, D. Yurgelun-Todd, M. Tohen, R.W. McCarley, Lower left temporal lobe MRI volumes in patients with first-episode schizophrenia compared with psychotic patients with first-episode affective disorder and normal subjects, Am. J. Psychiatry 155 (10) (1998) 1384–1391.
[24] R. Fisher, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, Biometrika 10 (4) (1915) 507–521.
[25] J. Rodgers, W. Nicewander, Thirteen ways to look at the correlation coefficient, Am. Stat. 42 (1) (1988) 59–66.
[26] J. Egger, J. Tokuda, L. Chauvin, B. Freisleben, C. Nimsky, T. Kapur, W. Wells, Integration of the OpenIGTLink network protocol for image-guided therapy with the medical platform MeVisLab, Int. J. Med. Robot. Comp. Assist. Surg. 8 (3) (2012) 282–290.
[27] I. Wolf, M. Vetter, I. Wegner, T. Böttger, M. Nolden, M. Schöbinger, M. Hastenteufel, T. Kunert, H.P. Meinzer, The medical imaging interaction toolkit, Med. Image Anal. 9 (6) (2005) 594–604.
[28] Y. Gao, R. Kikinis, S. Bouix, M. Shenton, A. Tannenbaum, A 3D interactive multi-object segmentation tool using local robust statistics driven active contours, Med. Image Anal 16 (6) (2012) 1216–1227.
[29] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Sys., Man. Cyber 9 (1) (1979) 62–66.
[30] R. Haralick, L Shapiro, Image segmentation techniques, Comp. Vision Graph. Image Process. 29 (1985) 100–132.
[31] J Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intellig. 8 (6) (1986) 679–698.
[32] V. Caselles, R. Kimmel, G. Sapiro, Geodesic active contours, Int. J. Comp. Vision 22 (1) (1997) 61–97.
[33] M.P. Sampat, Measuring intra- and inter-observer agreement in identifying and localizing structures in medical images, IEEE Int. Conf. Image Process. 2006 (2006) 1–4.
[34] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, Comparing images using the Hausdorff distance, IEEE Trans. Pattern Anal. Mach. Intell. 15 (9) (1993) 850–863.
[35] N. Byrne, M. Velasco Forte, A. Tandon, I. Valverde, T. Hussain, A systematic review of image segmentation methodology, used in the additive manufacture of patient-specific 3D printed models of the cardiovascular system, JRSM Cardiovasc. Dis. 5 (2016) e2048004016645467.
[36] J.S. Mulford, S. Babazadeh, N. Mackay, Three-dimensional printing in orthopaedic surgery: review of current and future applications, ANZ J. Surg. 86 (9) (2016) 648–653.
[37] J Wallner, et al., Clinical evaluation of semi-automatic open-source algorithmic software segmentation of the mandibular bone: practical feasibility and assessment of a new course of action, PLoS ONE 13 (5) (2018) e0196378.
[38] K.H. Zou, S.K. Warfield, A. Bharatha, C.M. Tempany, M.R. Kaus, S.J. Haker, W.M. Wells, F.A. Jolesz, R. Kikinis, Statistical validation of image segmentation quality based on a spatial overlap index, Acad. Radiol. 11 (2) (2004) 178–189.
[39] J. Egger, T. Kapur, A. Fedorov, S. Pieper, J.V. Miller, H. Veeraraghavan, B. Freisleben, A.J. Golby, C. Nimsky, R. Kikinis, GBM volumetry using the 3D Slicer medical image computing platform, Sci. Rep. 3 (2013) 1364.
[40] M. Shahbazian, R. Jacobs, J. Wyatt, G. Willems, V. Pattijn, E. Dhoore, C. VAN Lierde, F. Vinckier, Accuracy and surgical feasibility of a CBCT-based stereolithographic surgical guide aiding autotransplantation of teeth: in vitro validation, J. Oral. Rehabil. 37 (11) (2010) 854–859.
[41] Z. Fourie, J. Damstra, R.H. Schepers, P.O. Gerrits, Y. Ren, Segmentation process significantly influences the accuracy of 3D surface models derived from cone beam computed tomography, Eur. J. Radiol. 81 (4) (2011) e524–e530.
[42] S. Akyalcin, D.J. Dyer, J.D. English, C. Sar, Comparison of 3-dimensional dental models from different sources: diagnostic accuracy and surface registration analysis, Am. J. Orthodont. Dentofac. Orthoped. 144 (6) (2013) 831–837.
[43] W.P. Engelbrecht, Z. Fourie, J. Damstra, P.O. Gerrits, Y. Ren, The influence of the segmentation process on 3D measurements from cone beam computed tomography-derived surface models, Clin. Oral Investig. 17 (8) (2013) 1919–1927.
[44] M. Martorelli, P. Ausiello, R. Morrone, A new method to assess the accuracy of a Cone Beam Computed Tomography scanner by using a non-contact reverse engineering technique, J. Dent. 42 (4) (2014) 460–465.
[45] E. Huotilainen, R. Jaanimets, J. Valasek, P. Marcian, M. Salmi, J. Tuomi, A. Makitie, J. Wolff, Inaccuracies in additive manufactured medical skull models caused by the DICOM to STL conversion process, J. Cranio-Maxillo-Fac. Surg. 42 (5) (2014) e259–ee65.
[46] Y.H. Sang, H.C. Hu, S.H. Lu, Y.W. Wu, W.R. Li, Z.H. Tang, Accuracy assessment of three-dimensional surface reconstructions of in vivo teeth from cone-beam computed tomography, Chin. Med. J. 129 (12) (2016) 1464–1470.
[47] P. Szymor, M. Kozakiewicz, R. Olszewski, Accuracy of open-source software segmentation and paper-based printed three-dimensional models, J. Cranio–Maxillo-Fac. Surg. 44 (2) (2016) 202–209.
[48] M. Khouri, M. Champy, Results of mandibular osteosynthesis with miniaturized screwed plates. Apropos of 800 fractures treated over a 10-year period, Ann. Chir. Plast. Esthet. 32 (3) (1987) 262–266.
[49] B. Chrcanovic, Fixation of mandibular angle fractures: in vitro biomechanical assessments and computer-based studies, Oral Maxillofac. Surg. 17 (4) (2013) 251–268.
[50] J. Egger, T. Kapur, C. Nimsky, R. Kikinis, Pituitary adenoma volumetry with 3D Slicer, PLoS ONE 7 (12) (2012) e51788.
[51] D. Kainmueller, H. Lamecker, H. Seim, M. Zinser, S. Zachow, Automatic extraction of mandibular nerve and bone from cone-beam CT data, Med. Image Comput. Comput. Assist. Interv. 12 (2) (2009) 76–83.
[52] Z. Yaari, D. da Silva, A. Zinger, E. Goldman, A. Kajal, R. Tshuva, E. Barak, N. Dahan, D. Hershkovitz, M. Goldfeder, J.S. Roitman, A. Schroeder, Theranostic barcoded nanoparticles for personalized cancer medicine, Nat. Commun. 7 (2016) 13325.
[53] S. Chia, J.L. Low, X. Zhang, X.L. Kwang, F.T. Chong, A. Sharma, D. Bertrand, S.Y. Toh, H.S. Leong, M.T. Thangavelu, J.S.G. Hwang, K.H. Lim, T. Skanthakumar, H.K. Tan, Y. Su, S. Hui Choo, H. Hentze, I.B.H. Tan, A. Lezhava, P. Tan, D.S.W. Tan, G. Periyasamy, J.L.Y. Koh, N. Gopalakrishna Iyer, R. Das Gupta, Phenotype-driven precision oncology as a guide for clinical decisions one patient at a time, Nat. Commun. 8 (1) (2017) 435.
[54] C.J. Niedworok, A.P. Brown, M. Jorge Cardoso, P. Osten, S. Ourselin, M. Modat, T.W. Margrie, aMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data, Nat. Commun. 7 (2016) 11879.
[55] E. Faure, T. Savy, B. Rizzi, C. Melani, O. Stašová, D. Fabrèges, R. Špir, M. Hammons, R. Čúnderlík, G. Recher, B. Lombardot, L. Duloquin, I. Colin, J. Kollár, S. Desnoulez, P. Affaticati, B. Maury, A. Boyreau, J.Y. Nief, P. Calvat, P. Vernier, M. Frain, G. Lutfalla, Y. Kergosien, P. Suret, M. Remešíková, R. Doursat, A. Sarti, K. Mikula, N. Peyriéras, P. Bourgine, A workflow to process 3D+time microscopy images of developing organisms and reconstruct their cell lineage, Nat. Commun. 7 (2016) 8674.
[56] T. Caffrey, T. Wohlers, Additive manufacturing state of the industry, Manuf. Eng. 154 (2015) 67–78.
[57] A. Hamamci, G. Unal, N. Kucuk, K. Engin, Cellular automata segmentation of brain tumors on post contrast MR images, Med. Image Comput. Comput. Assist. Interv. 13 (3) (2010) 137–146.
[58] S. Katsigiannis, E. Zacharia, D Maroulis, Grow-cut based automatic cDNA microarray image segmentation, IEEE Trans. Nanobiosci. 14 (1) (2015) 138–145.
[59] E. Kostopoulou, S. Katsigiannis, D. Maroulis, A custom grow-cut based scheme for 2D-gel image segmentation, IEEE Eng. Med. Biol. Soc. 2015 (2015) 2407–2410.
[60] Y.-J. Zhang, in: Advances in Image and Video Segmentation, IRM Press, Hershey, PA, 2006, p. 457.
[61] J. Egger, J. Wallner, M. Gall, X. Chen, K. Schwenzer-Zimmerer, K. Reinbacher, D. Schmalstieg, Computer-aided position planning of miniplates to treat facial bone defects, PLoS ONE 12 (8) (2017) e0182839.

[62] X. Chen, L. Xu, Y. Yang, J. Egger, A semi-automatic computer-aided method for surgical template design, Sci. Rep. 6 (2016) 20280.
[63] D. Schmalstieg, T. Höllerer, Augmented Reality: Principles and Practice. 1st ed., Paperback, 528 Pages Edn, Addison-Wesley Professional, 2016 ISBN 978-0321883575.
[64] J. Egger, M. Gall, J. Wallner, P. Boechat, A. Hann, X. Li, X. Chen, D. Schmalstieg, HTC Vive MeVisLab integration via OpenVR for medical applications, PLoS ONE 12 (2017) e0173972.
[65] X. Chen, L. Xu, Y. Wang, H. Wang, F. Wang, X. Zeng, Q. Wang, J. Egger, Development of a surgical navigation system based on augmented reality using an optical see-through head-mounted display, J. Biomed. Inform. 55 (2015) 124–131.
[66] J. Egger, K. Hochegger, M. Gall, X. Chen, K. Reinbacher, K. Schwenzer-Zimmerer, D. Schmalstieg, J. Wallner, Algorithmic evaluation of lower jawbone segmentations, Proc. SPIE Med. Imag. Conf. (2017) 10137 -11.
[67] B. Pfarrkirchner, C. Gsaxner, L. Lindner, N. Jakse, J. Wallner, D. Schmalstieg J. Egger, Lower jawbone data generation for deep learning tools under MeVisLab, Proc. SPIE Med. Imag. Conf. (2018) 10578–10596.
[68] J. Wallner, I. Mischak, J. Egger, Computed tomography data collection of the complete human mandible and valid clinical ground truth models, Sci. Data 6 (2019) 190003.