Automatic model-based teeth segmentation, numbering and 3-D reconstruction using routinely collected images

vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

DISSERTATION

zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
von

Andreas Wirtz. M.Sc.

geboren in Bonn, Deutschland

Referenten der Arbeit:	Prof. Dr. Arjan Kuijper Technische Universität Darmstadt
	Prof. Dr. techn. Dr.-Ing. eh. Dieter W. Fellner Technische Universität Darmstadt
	Prof. Dr. Reinhard Klein Rheinische Friedrich-Wilhelms-Universität Bonn
Tag der Einreichung:	04.11.2021
Tag der mündlichen Prüfung:	20.12.2021

Darmstadt 2021

Darmstädter Dissertation
D 17

Wirtz, Andreas: Automatic model-based teeth segmentation, numbering and 3-D reconstruction using routinely collected images

Darmstadt, Technische Universität Darmstadt

Jahr der Veröffentlichung der Dissertation auf TUprints: 2022

URN: urn:nbn:de:tuda-tuprints-207272

Tag der mündlichen Prüfung: 20.12.2021

Veröffentlicht unter CC BY-ND 4.0 International

https://creativecommons.org/licenses/by-nd/4.0/

Erklärung zur Dissertation

Hiermit versichere ich die vorliegende Dissertation selbständig nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 04.11.2021 Andreas Wirtz

Acknowledgments

Throughout my research and the writing of this dissertation I have received a great deal of support and assistance.

First of all, I would like to thank my supervisor, Prof. Arjan Kuijper, for his excellent supervision and guidance as well as the constructive feedback regarding my dissertation. I would like to thank Prof. Dieter Fellner for reviewing this work and for giving me the opportunity to pursue my interest for computer vision and medical computing. I would also like to thank Prof. Reinhard Klein for his efforts as reviewer.

I would like to thank Dr. Stefan Wesarg and my colleagues at the department of Visual Healthcare Technologies at Fraunhofer IGD for the excellent working atmosphere, the fruitful scientific discussions and the constructive feedback. In addition, I would like to thank the students I supervised who contributed to my research.

My very special thanks to Cristina for her immeasurable support, understanding and consideration during the time of writing this thesis, as well as the honest feedback and constructive discussions.

Abstract

Oral health issues like tooth decay affect billions of people worldwide. The early detection as well as the treatment of those problems is an important part of today's healthcare. Since the teeth cannot be assessed by visual inspection only as they are partly embedded in soft tissue and bones, medical imaging techniques are employed to provide the required information. The manual analysis of such images is time consuming and prone to inter- and intra-observer variability. For this reason, computer aided systems which automate the extraction of clinically relevant information can be of great benefit to medical professionals. A mandatory step to provide such systems is the segmentation and numbering of individual teeth in the digital images. However, challenges like image quality and characteristics of dental radiographs, patient-specific variations, and the fact that the 32 teeth only belong to 4 different types makes it hard to accurately detect teeth boundaries and difficult to label an individual tooth.

In this context, this thesis focuses on the research questions of *teeth segmentation and numbering in panoramic X-ray images*, and *image-based 3-D reconstruction of the teeth from five colored photographs* which relies on accurate object-level segmentation and numbering of the 2-D teeth outlines (in the photographs). As both topics share the common aspect of segmentation and numbering, a general concept is presented which is also applied in a third application to localize landmarks for the analysis of dental cephalometric images.

This thesis proposes to solve the segmentation and numbering in 2-D by encoding prior knowledge about the teeth shapes and spatial relations in a coupled shape model. Initial placement of the model is performed by exploiting the semantic segmentation performance of neural networks while dynamic adaptation strategies increase the robustness of fitting to model to unseen images. This enables the extraction of teeth contours from both panoramic radiographs and colored photographs. The proposed image-based 3-D teeth reconstruction utilizes the numbered teeth contours from the photographs to deform a mean model of the teeth by minimizing a silhouette-based loss. It is the first fully-automatic image-based teeth reconstruction that aims to reconstruct the majority of teeth and the first approach to perform a reconstruction only on the five photographs of orthodontic records. The landmark localization utilizes the segmentation and numbering concept to predict the location of 19 landmarks by exploiting the spatial relation between landmarks and other structures and refines those predictions using landmark-specific Hough Forests.

The teeth segmentation and numbering of 28 individual teeth in panoramic radiographs achieves average F1 scores of 0.823 ± 0.189 and 0.833 ± 0.108 on two different test sets. The image-based 3-D reconstruction of 24 teeth from five photographs achieves an average symmetric surface distance of 0.807 ± 0.379 mm. The landmark localization in cephalometric images reaches a success detection rate of 76.04 % in the clinically relevant 2.0 mm accuracy range.

Zusammenfassung

Die Mundhygiene ist ein essenzieller Bestandteil der täglichen Gesundheitsroutine und die Mund- und Zahnmedizin ein wichtiger Bereich des Gesundheitswesens. Erkrankungen des Mundes und der Zähne, wie beispielsweise Karies, betreffen weltweit Milliarden von Menschen. Aufgrund der hohen Prävalenz und der kostspieligen Behandlung ist die frühzeitige Erkennung solcher Erkrankungen ein wichtiger Bestandteil der heutigen Gesundheitsversorgung. Hierfür werden medizinische Bildgebungsverfahren eingesetzt, da die Zähne teilweise in Weichgewebe und Knochen eingebettet sind und ihr Gesundheitszustand daher nicht allein durch eine visuelle Inspektion beurteilt werden kann. Die manuelle Analyse der generierten Bilder ist allerdings zeitaufwändig, stark von der Erfahrung des behandelnden Experten abhängig und leidet unter einer hohen Intra- und Interobserver-Variabilität. Aus diesem Grund können computergestützte Systeme, welche die Extraktion klinisch relevanter Informationen automatisieren, von großem Nutzen für das medizinische Fachpersonal sein. Ein notwendiger Verarbeitungsschritt zur Bereitstellung solcher Systeme für die Zahnmedizin ist die Segmentierung und Nummerierung der einzelnen Zähne in den digitalen Bildern. Herausforderungen wie die Bildqualität und -eigenschaften zahnmedizinischer Röntgenbilder sowie die patientenspezifische Variationen des Gebisses erschweren jedoch die genaue Erkennung von Zahngrenzen. Darüber hinaus setzen sich die 32 Zähne des menschlichen Gebisses aus nur vier Zahntypen zusammen, wodurch eine Nummerierung nicht ausschließlich auf Basis der erkannten Form durchgeführt werden kann.

Vor diesem Hintergrund konzentriert sich diese Arbeit auf die Forschungsfragen i) *Zahnsegmentierung und -nummerierung in Panoramaschichtaufnahmen* und ii) *bildbasierte 3-D-Rekonstruktion der Zähne aus fünf Farbfotografien*, welche eine genaue Segmentierung und Nummerierung der individuellen 2-D-Zahnumrisse (in den Fotografien) voraussetzt. Da beide Themen sich mit dem Problem der Segmentierung und Nummerierung beschäftigen, wird zunächst ein allgemeines Konzept für diesen Aspekt vorgestellt. Dieses wird später auch in einer dritten Anwendung zur Lokalisierung von Landmarken für die Analyse von zahnmedizinischen Fernröntgenseitenbildern angewendet.

In dieser Arbeit wird vorgeschlagen die Segmentierung und Nummerierung der Zähne in zahnmedizinischen Bilddaten unter Verwendung von Vorwissen über die Form und die räumlichen Lagebeziehung der Zähne durchzuführen. Dieses statistische Vorwissen wird in einem gekoppelten Formmodell kodiert. Das Modell wird in einem Optimierungsverfahren an die Strukturen im Bild angepasst und bildet den optimalen Kompromiss zwischen den lokalen Bildmerkmalen und der erwarteten Zahnform und Lage ab. Die initiale Platzierung des Modells in den Bilddaten erfolgt unter Ausnutzung der semantischen Segmentierungs-Performance Neuronaler Netze. Die Modellanpassung an ungesehene Bilderdaten erfolgt unter Einsatz dynamischer Anpassungsstrategien, um die Robustheit der lokalen

Merkmalssuche zu erhöhen. Dieser Ansatz ermöglicht die Extraktion von Zahnkonturen sowohl aus Panoramaschichtaufnahmen als auch aus Farbfotografien.

Der vorgeschlagene bildbasierte 3-D-Zahnrekonstruktionsansatz nutzt die nummerierten Zahnkonturen (Silhouetten) aus den Fotografien, um ein initiales Modell der Zähne solange zu deformieren bis es die extrahierten Silhouetten optimal widerspiegelt. Es ist die erste vollautomatische bildbasierte Zahnrekonstruktion, die darauf abzielt, die Mehrheit der Zähne zu rekonstruieren. Darüber hinaus ist es der erste Ansatz, der eine Rekonstruktion nur auf Basis der fünf Fotografien durchführt, die im Zuge einer kieferorthopädischen Behandlung standardmäßig aufgenommen werden. Die Landmarkenlokalisierung nutzt das Segmentierungs- und Nummerierungskonzept zur Vorhersage der Lage von 19 Landmarken in zahnmedizinischen Fernröntgenseitenbildern. Hierfür wird die räumliche Lagebeziehung zwischen den Landmarken und zu anderen Strukturen (Schädel, Haut) ausgenutzt. Die Vorhersagen werden anschließend mit spezifischen Hough-Forests für jede Landmarke verfeinert.

Die Zahnsegmentierung und Nummerierung von 28 Einzelzähnen in Panoramaschichtaufnahmen erreicht einen durchschnittlichen F1-Score von 0.823 ± 0.189 und 0.833 ± 0.108 in zwei unterschiedlichen Datensätzen. Die bildbasierte 3-D-Rekonstruktion von 24 Zähnen aus fünf Fotografien erzielt einen durchschnittlichen symmetrischen Oberflächenabstand von 0.807 ± 0.379 mm. Die Lokalisierung der Landmarken in Fernröntgenseitenbildern im klinisch relevanten Genauigkeitsbereich von 2.0 mm ist zu 76.04 % erfolgreich.

Contents

1. Introduction **1**
- 1.1. Medical challenges . . . 4
  - 1.1.1. Teeth assessment . . . 5
  - 1.1.2. Orthodontic assessment and monitoring . . . 6
- 1.2. Technical challenges . . . 7
  - 1.2.1. Teeth segmentation and numbering . . . 8
  - 1.2.2. 3-D teeth reconstruction on five photographs . . . 10
  - 1.2.3. Conclusion . . . 12
- 1.3. Goals . . . 13
- 1.4. Contributions . . . 16
- 1.5. Structure of the thesis . . . 16

2. Fundamentals **19**
- 2.1. The human teeth . . . 19
- 2.2. Dental notation systems . . . 20
- 2.3. Dental imaging . . . 22
  - 2.3.1. Benefits of dental radiographs . . . 23
  - 2.3.2. Acquisition of dental radiographs . . . 24
  - 2.3.3. Types of dental radiographs . . . 24
  - 2.3.4. Safety of dental radiographs . . . 26
  - 2.3.5. Orthodontic records . . . 27

3. Related Work **29**
- 3.1. Dental Image Segmentation . . . 29
  - 3.1.1. Pixel-based . . . 30
  - 3.1.2. Edge-based . . . 35
  - 3.1.3. Region-based . . . 36
  - 3.1.4. Local shape knowledge . . . 37
  - 3.1.5. Global shape knowledge . . . 39
  - 3.1.6. Spatial relations . . . 40
  - 3.1.7. Neural networks . . . 41
  - 3.1.8. Hybrid CNN methods . . . 45
  - 3.1.9. Discussion & Conclusion . . . 47

3.2. Image-based 3-D teeth reconstruction . . . 48
3.2.1. Feature-based methods . . . 50
3.2.2. Shape-from-shading . . . 51
3.2.3. Shape-from-silhouettes . . . 52
3.2.4. Discussion & Conclusion . . . 53

4. Segmentation & numbering framework **55**
4.1. General concept . . . 55
4.2. Coupled Shape Model . . . 59
4.2.1. 2-D deformable model items . . . 60
4.2.2. Spatial relation . . . 62
4.3. Binary mask generation . . . 63
4.4. CSM initialization . . . 65
4.5. CSM adaptation . . . 66
4.6. Conclusion . . . 70

5. Teeth segmentation & numbering in panoramic radiographs **73**
5.1. Data . . . 75
5.2. Metrics . . . 78
5.3. Coupled shape models . . . 80
5.4. U-Net & binary mask generation . . . 82
5.5. Model initialization . . . 87
5.5.1. Evaluation & discussion . . . 88
5.6. Model adaptation . . . 92
5.7. Experiments and results . . . 95
5.8. Discussion . . . 100
5.9. Conclusion . . . 105

6. Image-based 3-D teeth reconstruction **107**
6.1. Teeth segmentation & numbering . . . 110
6.1.1. 2-D CSM designs . . . 111
6.1.2. U-Net networks . . . 114
6.1.3. Pre-processing and mask generation . . . 115
6.1.4. Automatic 2-D CSM initialization . . . 117
6.1.5. 2-D CSM adaptation . . . 119
6.1.6. Experiments and results . . . 121
6.1.7. Discussion . . . 132
6.2. Deformation-based reconstruction . . . 137
6.2.1. 3-D CSM designs . . . 138
6.2.2. Silhouette-based loss . . . 139
6.2.3. Estimation of projection parameters . . . 140

6.2.4. Automatic 3-D CSM initialization . . . 141
6.2.5. Teeth reconstruction . . . 142
6.3. Experiments and Results . . . 144
6.3.1. Data . . . 144
6.3.2. Metrics . . . 145
6.3.3. Offline training: 3-D coupled shape models . . . 147
6.3.4. Evaluation . . . 149
6.4. Discussion . . . 152
6.5. Conclusion . . . 153

7. Cephalometric landmark localization **155**
7.1. Challenges . . . 156
7.2. Related work . . . 157
7.3. Methods . . . 160
7.3.1. CSM design . . . 161
7.3.2. CSM initialization . . . 162
7.3.3. Location prediction . . . 163
7.3.4. Location refinement . . . 164
7.4. Experiments & results . . . 165
7.4.1. Data . . . 165
7.4.2. Metrics . . . 166
7.4.3. Offline training . . . 167
7.4.4. Landmark localization results . . . 169
7.5. Discussion . . . 172
7.6. Conclusion . . . 173

8. Conclusions and Future Work **175**
8.1. Research results . . . 175
8.1.1. Teeth segmentation & numbering in panoramic radiographs . . . 176
8.1.2. Image-based 3-D teeth reconstruction . . . 176
8.1.3. Cephalometric landmark localization . . . 177
8.2. Future work . . . 178
8.2.1. Teeth segmentation & numbering in panoramic radiographs . . . 178
8.2.2. Image-based 3-D teeth reconstruction . . . 178
8.2.3. Cephalometric landmark localization . . . 179

Bibliography **181**

A. Publications and Talks **195**
A.1. Publications . . . 195

B. Supervising Activities **197**
B.1. Master Thesis 197
B.2. Bachelor Thesis 197

1. Introduction

Oral health is an essential aspect of our day-to-day healthcare and an important part of our healthcare systems. According to the Global Burden of Disease Study of the year 2017[1], it is estimated that oral diseases affect nearly 3.5 billion people worldwide [RAA*18]. Untreated tooth decay (dental caries) in permanent teeth (i.e. the second and final set of teeth which replaces the primary teeth) is not only the most common oral health problem but it is also considered to be the most common health problem in general[2]. Tooth decay in permanent teeth affects roughly 2.3 billion people globally and in addition to that more than 530 million children suffer from dental caries of primary teeth [RAA*18]. Apart from tooth decay, the major oral health conditions are periodontal diseases (due to bacteria in the mouth causing infections and inflammation of the gums and bone tissue surrounding the tooth), various forms of oral cancers, oral manifestations of HIV and oro-dental traumata (e.g. as a result of accidents)[2]. Additionally, abnormal position (malposition) of teeth and misalignment of teeth or jaws (malocclusion) can negatively impact the oral health status of a person. The treatment of dental diseases is costly, averaging to about 5 % of total health expenses and 20 % of out-of-pocket health expenditure [OEC17], i.e direct payments made by patients to health care providers at the time of service use. However, according to the World Health Organization, the majority of oral health conditions are considered to be largely preventable and can be treated in their early stages[2]. This requires regular examination of the oral health status by a medical professional to detect potential diseases before they become a problem.

Dentistry is the branch of medicine that covers the areas of oral- and dental-medicine. In dentistry, medical examination of a patient's set of teeth and oral tissues is performed in order to identify many issues related to oral health. This examination, however, cannot be performed by visual inspection only. The lower part of the teeth (the root) as well as bony structures in the jaw and nasal region are not directly visible and therefore simple visual oral inspection is insufficient for judging the complete oral health status of a patient. Instead, medical imaging is typically utilized for such examinations. Medical imaging generally refers to the process of creating visual representations of the interior of the human body via different techniques for clinical analysis and medical intervention. The intention is to reveal the internal structures like hard tissue (bones and teeth) and soft tissue (e.g. muscles, blood vessels, fat and tendons) hidden underneath the skin to enable diagnosis and treatment of diseases and health conditions.

[1] Global Burden of Disease Study 2017 (GBD 2017) Data Resources: http://ghdx.healthdata.org/gbd-2017 (last visited on 13.10.2021)

[2] World Health Organization Fact sheet on Oral Health: https://www.who.int/news-room/fact-sheets/detail/oral-health (last visited on 13.10.2021)

Figure 1.1.: Dental radiographs: (a) dental radiograph (orthopantanogram) on the left and exemplary intensity variations related to dental diseases on the right: (b) tooth decay (image from [TNB*16], modified) and (c) jaw lesion (image from [Als18], modified).

In dentistry, the medical imaging technique of X-ray is employed in the majority of cases to generate the 2-D and 3-D **dental radiographs** (X-ray images) used during clinical image-based diagnosis to identify many oral health related problems like tooth decay, infections and bone abnormalities [SOP18]. X-ray (or X-radiation; also known as Röntgen radiation) refers to a penetrating form of energy. It can pass through softer tissues like gums and cheeks and is absorbed by denser hard tissues like bones and teeth. By measuring the energy absorption levels, a gray-scale (intensity) image representing these levels can be generated, where areas with high absorption appear white, and areas where radiation can pass through appear black. Health conditions can then be identified by looking for abnormal variations in gray-scale values that may indicate certain diseases. Fig. 1.1 depicts an orthopantanogram (i.e. a type of dental radiograph) as well as some exemplary images for oral diseases with abnormal gray-scale values (indicated by red rectangles).

In addition, medical images are also an important tool for treatment and surgery planning (e.g. for dental implants [Gre15]), treatment monitoring (e.g. root movement during orthodontic treatment [LWP*15]) and even forensic procedures (e.g. post-mortem identification [NAM08a]). In the past, film-based dental radiographs and other non-digital images had to be manually inspected by the dentist in order to identify intensity variations which indicate a certain medical condition. This process is not only time-consuming but also heavily dependent on the experience and visual perception of the dentist because of the low image quality of dental x-ray images and the variability in the dental structures in between patients. The results of this process are therefore prone to inter- and intraobserver variability, leading to different treatment approaches [WHL*16]. Nowadays, the availability of sensor-based digital radiography and other digital image modalities like colored photographs facilitates the possibility for computer-aided analysis of medical images.

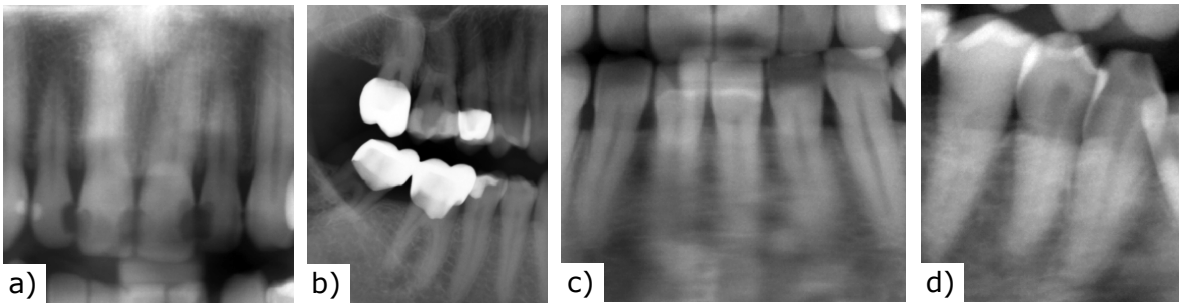


Figure 1.2.: Challenges in dental image segmentation: (a) low contrast between different tissue types, (b) artifacts cause by teeth restorations, dental appliance or diseases, (c) blurry edges of structures, and (d) general image noise.

Computer-aided systems or medical imaging applications which offer interactive or automatic analysis of dental radiographs (or other image types) by extracting clinically relevant information or knowledge, would be of great assistance to dental professionals. These tools could significantly reduce the time required for analysis, enhance the clinical workflow, remove any inter- or intraobserver variability, provide treatment suggestions, assist with treatment planing and simplify the treatment monitoring process. However, automatic or semi-automatic analysis of dental images is a challenging task: low image quality like low contrast as depicted in Fig. 1.2a, artifacts (Fig. 1.2b), blurry edges (Fig. 1.2c), and noise (Fig. 1.2d) make it difficult to clearly identify object boundaries. Due to all these challenges, many different scientific questions need to be solved for different application areas in order to provide the aforementioned computed-aided systems.

The research area dealing with the (automated) extraction of clinically relevant information from medical images (also referred to as medical image analysis) is called **medical image computing** (MIC). It is an interdisciplinary field combining many aspects of different research areas like computer science, mathematics, medicine, information engineering and physics. Researchers in this field focus on the development of computational and mathematical methods for solving challenges concerning medical images and their use for diagnosis, treatment and monitoring. In the dental domain, this includes the extraction of shading information for image-based 3-D teeth reconstruction [AFSEM14], calculation of dental biometrics for human identification [AK13], and teeth contour extraction from dental radiographs [JFR*18]. For almost all of these applications, a segmentation of the relevant dental structures in the image is essential. In addition to that, tooth identification, i.e. the assignment of the appropriate label or number to the detected tooth, is also required. It is also referred to as tooth labeling or tooth numbering, respectively. An example is provided in Fig. 1.3. It shows the segmentation of teeth in a panoramic radiograph.

In this context, this thesis focuses on extracting model-based object-level segmentations and numbering of relevant structures (e.g. teeth) for different scenarios in dentistry and orthodontics as well as performing an image-based 3-D reconstruction of the teeth. These scenarios and their medical chal-

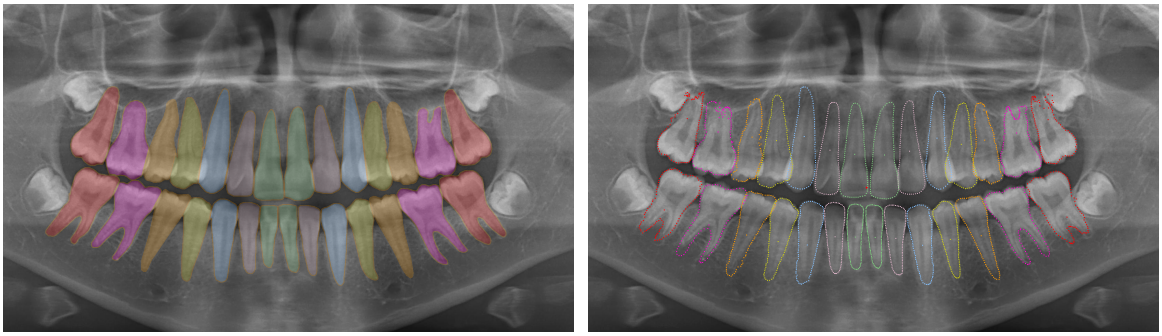


Figure 1.3.: Segmentation of teeth in a dental panoramic radiograph. It is visualized by the contour that represents the individual object boundaries. The left image shows the manually created gold-standard segmentation and the right image depicts the segmentation and assigned labels computed by the framework proposed in this thesis.

lenges will be introduced in the next section, followed by the corresponding technical challenges in Section 1.2. The specific goals of this thesis are explained in Section 1.3 together with a small conclusion. Finally, the overall structure of the thesis is presented at the end of this chapter.

1.1. Medical challenges

Nowadays, medical images are an important tool in dentistry for image-based diagnosis as well as treatment planing and monitoring. They are useful for and used in many different scenarios. As stated above, the most common image modalities used in modern dentistry are 2-D and 3-D dental radiographs (X-ray images). These radiographic images can be categorized into intraoral and extraoral images [WHL*16]. Fig. 1.4 shows a collection of different image types used in dentistry. Intraoral images, like 2-D bitewing- and periapical-radiographs, are obtained inside the patient's mouth and only show specific regions of the set of teeth or individual teeth. They are mostly used to get detailed information about the desired region, e.g. to check for cavities, the health of the tooth root and surrounding bone, or the status of developing teeth. Extraoral images, like 2-D cephalograms and panoramic radiographs or 3-D digital volume tomography (DVT), capture the entire teeth region as well as the surrounding areas like the bones of the nasal and face areas and provide fundamental information about the teeth, jaws and skull of a patient. Therefore, these types of radiographs are used to assess the overall oral health status of a patient, to look for impacted teeth and to monitor growth and development of the jaws in relation to the teeth. In orthodontics, 3-D imprints of the teeth extracted from DVT images or via 3-D scanners are used for treatment planning in case of misaligned teeth. Moreover, colored photographs are also used for documentation purposes. In concrete, they are employed during treatment monitoring for the documentation of the treatment progress with dental braces. More information on dental imaging can be found in Section 2.3.

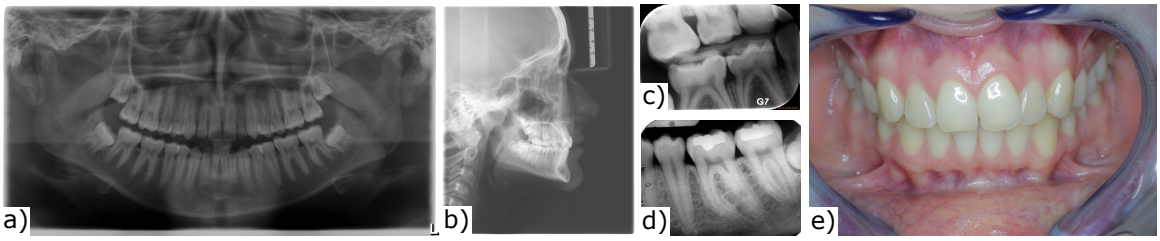
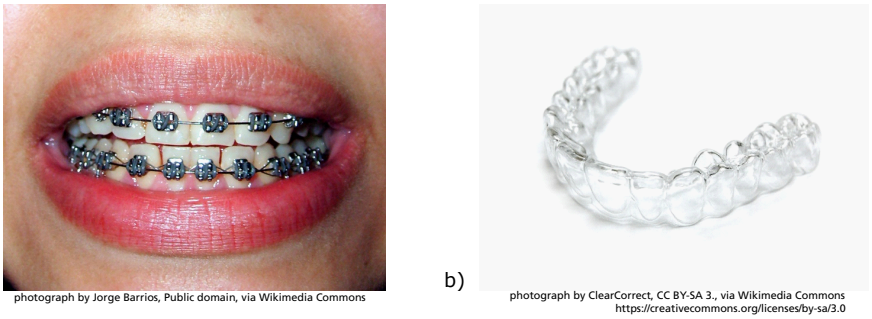


Figure 1.4.: Different image types in dentistry: (a) orthopantomogram, (b) cephalogram, (c) bitewing (image from [WHL*16]), (d) periapical (image from [SFNA04]), (e) photograph.

This thesis focuses on two specific scenarios which are (a) the initial assessment of the patient's set of teeth via panoramic radiographs and (b) the assessment of tooth malposition and malocclusion for orthodontic treatment and treatment monitoring utilizing colored photographs. The two scenarios will be introduced in detail in the next paragraphs together with their medial challenges. Additionally, the technical requirements for corresponding medical image applications will be derived. The technical challenges related to these requirements and how these challenges come together to define the research questions dealt with in this thesis are described in the next section (cf. Section 1.2). An additional scenario – the analysis of cephalometric radiographs for treatment planing – will be introduced in a later chapter of the thesis and demonstrates the flexibility of the approach proposed for solving the research questions.

1.1.1. Teeth assessment

Teeth assessment refers to the process of determining the overall (health) status of all teeth. The dentist will examine all teeth one by one (via a dental radiograph) in order to establish whether the tooth is present, impacted or missing, and to identify any dental diseases like tooth decay, root exposure or infections. In case of a missing or incomplete documentation, teeth assessment can also be used to determine which teeth already underwent treatment of dental diseases (e.g. have fillings or artificial crowns). This examination is typically performed on extraoral panoramic radiographs (cf. Fig. 1.4a). These images depict the entire mouth region which includes all teeth of the upper jaw (maxilla) and lower jaw (mandible) as well as the jaw bone and part of the nasal area. Because of this, they enable the dentist to check the status of the teeth and perform an initial assessment of the dental status of a patient. The identification of the individual teeth relies on the dentist's knowledge of their position in the image as well as their relative position to neighboring teeth. Since the 32 teeth are only separated into four classes (incisor, canine, pre-molar and molar), a numbering/labeling based on their shape alone is almost impossible. Dental diseases are detectable by checking for specific intensity patterns or variations, or abnormal intensity values in certain areas. However, tooth assessment can be a challenging task due to general intensity variation, noise and low contrast in the panoramic radiograph. As a result, the task highly depends on the dentist's visual perception and experience [WHL*16] in dealing with the radiographic image quality as well as identifying abnormal variations. This can lead to an inter-observer



photograph by Jorge Barrios, Public domain, via Wikimedia Commons

photograph by ClearCorrect, CC BY-SA 3., via Wikimedia Commons
https://creativecommons.org/licenses/by-sa/3.0

Figure 1.5.: Two different types of braces used to treat malpositioned teeth and malocclusions: (a) non-removable (metal-)wired braces, (b) removable clear aligners.

variability in the final assessment result and consequently different treatment approaches. Additionally, checking each one of the 32 teeth individually is a time consuming task. A computer-aided system or medical imaging application to automate or assist with tooth assessment would be of great benefit to dental professionals. A robust and accurate segmentation and numbering of the individual teeth in a radiographic image is an essential prerequisite for enabling such systems [SNFA06]. Teeth segmentation refers to the task of partitioning a dental radiograph into multiple non-overlapping segments which each represent an individual tooth. Typically, in image segmentation, these segments are considered homogeneous according to some (signal) characteristics [PXP00]. Teeth numbering refers to the task of assigning the correct label (tooth number) to each segment. Therefore, the resulting technical prerequisite for providing a computer-aided system to automate or assist with tooth assessment is a robust and accurate automatic **teeth segmentation and numbering** in dental panoramic images.

1.1.2. Orthodontic assessment and monitoring

The dental branch of orthodontics is responsible for the diagnosis and treatment of abnormal teeth position (malposition) and misaligned teeth or jaws (malocclusion). These malpositions and malocclusions of individual teeth or entire jaws (e.g. open- and overbite, or overjet) can negatively affect the oral health status of a patient. They can impact the ability to chew or speak, increase the risk of gum disease or tooth decay and potentially influence a person's (psychological) well-being. The treatment of malpositioned or misaligned teeth is a complex procedure that involves the use of aligners or dental braces in order to apply forces to the crown (i.e. the visible upper part) of the teeth to move them to their appropriate position, correct rotation or tilt of individual teeth, or pull teeth out further from the jaw bone. While aligners are transparent non-adjustable plastic braces worn over the teeth which can be removed at any time, a (metal-)wired brace has brackets which are directly glued to the teeth and connected via an adjustable wire. Fig. 1.5 illustrates both types of dental braces.

As a result, metal braces can exhibit stronger and more precise forces to the teeth compared to aligners and are used in more severe cases of malpositioned teeth or malocclusions. Milder cases are

Figure 1.6.: Exemplary set of documentary photographs to illustrate the different views: a) left view, b) anterior view, c) right view, d) maxilla view, e) mandible view. (All images are sections cropped from the original photo for better visualization.)

treated with aligners as these are less disruptive to a patient's daily life since they can be removed if needed. The aligners need to be exchanged in regular intervals (e.g. every two weeks) to ensure optimal treatment and are produced in batches for several treatment steps in advance. Since the treatment of malpositions or malocclusions usually takes between one and two years, regular monitoring is required to ensure the induced changes are in accordance with the treatment plan.

The treatment planning of orthodontic procedures is based on dental radiographs (3-D DVT images) and 3-D dental impressions or intraoral scans, and is already utilizing computed-aided systems. The initial assessment and treatment monitoring (for aligners), however, is only performed visually and documented according to a standardized procedure using five colored photographs as part of the patient's orthodontic record (cf. Section 2.3.5). These colorized photos depict the patient's teeth from a frontal view, lateral views from the left and right as well as top-down views of maxilla and mandible (captured using a mirror). An exemplary set of photographs is shown in Fig. 1.6. Despite being available in digital form, these images are only used for documentation purposes and not employed further for any kind of computed-aided system.

The utilization of these documentary photos for computed-aided systems could enable an automatic initial assessment of the teeth in terms of severity of malposition/malocclusion to support the treatment decision process, or improve the treatment monitoring process. The required information for these tasks could be best computed based on a 3-D model of the set of teeth. Consequently, the resulting technical challenge to be solved here is how to generate such a model of the teeth by performing a **3-D reconstruction** on the five documentation photographs.

1.2. Technical challenges

For each scenario introduced in the previous section, a technical requirement was identified which needs to be solved in order to provide medical imaging applications that can assist with the associated task. These technical requirements depend on a number of technical challenges which need to be solved. This section provides an overview of these challenges and the corresponding research questions for each requirement. The overlap and interaction between these challenges explain how to arrive to the research questions focused on in this thesis.

1.2.1. Teeth segmentation and numbering

In order to extract relevant information about the teeth from dental images, an accurate segmentation and numbering of the teeth is required. Segmentation refers to the task of partitioning an image into multiple non-overlapping segments. While semantic segmentation assigns the same label to all pixels belonging to the same class, instance segmentation also distinguishes between individual instances within one class. As teeth numbering refers to the task of assigning the correct label (tooth number) to each individual tooth, an instance segmentation of the teeth is required. Ideally, this process should be automated and not require any user interaction to offer the most benefit. However, there are several challenges which need to be solved for an automatic teeth segmentation and labeling/numbering in dental medical images:

- The **image quality and characteristics** of radiographs pose a challenge for any automatic segmentation process of the teeth. A dental radiograph can be roughly divided into three main classes [SS74]: (a) teeth area, with high to medium intensity values, (b) dental structural area, with medium intensity values and (c) background area, with the lowest intensity values. However, this classification does not account for regions with over-exposed intensities in the teeth area which correspond to fillings, implants, bridges, braces or artifacts, or under-exposed regions caused by dental diseases like tooth decay (caries) or peridontal disease (cf. Fig. 1.2b). Additionally, the low contrast between the teeth area and the structural area in certain regions, especially around the roots of the teeth, makes it difficult to determine the exact boundary between both areas (cf. Fig. 1.2a). The generally low quality of the image due to noise and low resolution, scanning errors, etc. are additional factors which contribute further difficulties (cf. Fig. 1.2c). All of these characteristics make teeth segmentation especially challenging [AA15].
- The inter-patient **variations in the set of teeth** can be quite significant and offer a big challenge for segmentation and numbering. Naturally, each person's set of teeth is different and can even be used for postmortem identification. In addition to slight variations in the teeth's position and rotation, this also includes abnormalities like gaps caused by missing or extracted teeth, impacted or partially impacted teeth which have not or not fully broken through the gum, and the presence or absence of wisdom teeth. Moreover, malposition or malocclusion of teeth adds additional variation to the expected position of each tooth. All of these factors increase the difficulty of segmentation.
- The **numbering** of the individual (segmented) teeth is difficult due to the characteristics of the human set of teeth. The total number of 32 teeth only consists of 4 different types of teeth (i.e. incisor, canine, pre-molar and molar). The exact denomination of each tooth within a certain class is only determined by their relative position to other teeth, i.e. central incisor vs. lateral incisor or first pre-molar vs. second pre-molar. It becomes even more difficult in cases where teeth have been extracted or fallen out, or almost impossible (without prior knowledge) if the gap caused by missing teeth has closed over time or by orthodontic treatment.
- **Wisdom teeth** contribute to the challenges of automatic segmentation and numbering as they are significantly different to other teeth [SN12]. Wisdom teeth (also referred to as third molar) are

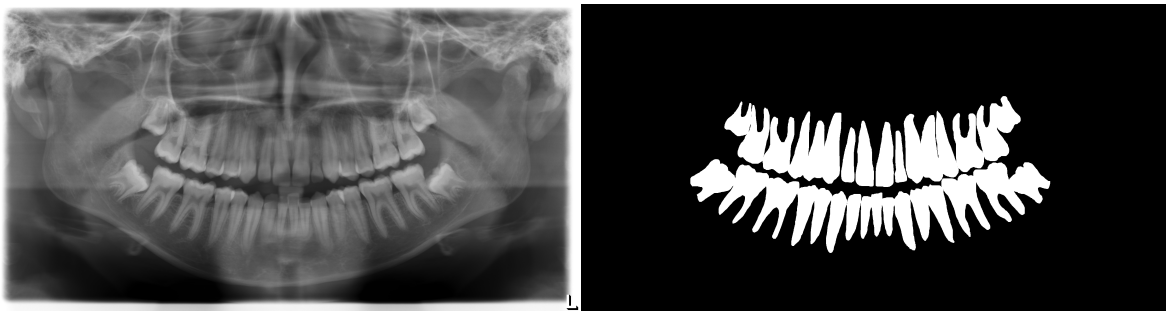


Figure 1.7.: Semantic teeth segmentation: All pixels of the source image (left) are are classified as teeth (white) or background (black), but no separation or labeling of individual teeth is performed.

the most posterior of the three molars. They are less likely to erupt compared to other teeth. The reason being that they do so much later in life compared to other teeth and are therefore much more likely to be affected by impaction due to a lack of space in the dental arch which causes them to get stuck in their position. All these factors contribute to a much larger variation in the position and orientation of the wisdom teeth. Additionally, not all persons develop wisdom teeth and wisdom teeth are more often surgically extracted as impaction can contribute to dental health problems.

Most approaches in the literature over the past decade only relied on image features for dental image segmentation. They utilized pixel-based techniques like global or adaptive thresholding [AA15, IANJ15] and clustering techniques [Als18, ST16, YYMS09] to separate the teeth area from dental structural- and background areas. Others relied on boundary information [GKK*13, Na'17, RAH*14] or region-based methods [LL08, LSSL12, MD11] to handle the difficult image properties of dental radiographs. Ultimately, purely image-based methods proved to be insufficient to accurately extract the teeth boundaries from the image. Moreover, most approaches in this area only focused on a semantic segmentation of the teeth. This means that all teeth are extracted as one single segment instead of a tooth-specific segmentation. If semantic segmentation is performed for a single class label, the resulting segmentation mask indicating the class correspondence of each pixel in the image will only contain the values 0 (background) and 1 (class). It is therefore also referred to as a binary mask. Naturally, no labeling of the teeth is performed in these approaches as the required information about the individual teeth is not available. Fig. 1.7 shows an example of such an image-based segmentation results.

Therefore, relying on image features alone is insufficient in dealing with the challenging image quality and characteristic of dental radiographs. Better performing segmentation approaches integrated prior knowledge about the shape of the teeth in order to better handle low image quality like low contrast and noise as well as the characteristics of dental radiographs like over-exposed intensity values or ill-defined object boundaries. Local shape knowledge was used to penalize jagged or improbable contours in low contrast areas by utilizing information from neighboring areas [LLH10, RRN13,

LHH*14, RRN14, HIHY16]. The usage of global shape knowledge introduced even more prior knowledge into the segmentation process [LGN09]. In recent years, most approaches focus on the use of convolutional neural networks to perform either semantic or instance segmentation, mostly utilizing U-Net (e.g. [NNH*21, KPIB19, SCKM20]) or R-CNN (e.g. [SOP18, JFR*18]) network architectures, respectively.

While segmentation accuracy has greatly improved when employing neural networks, only very few CNN approaches include a tooth numbering [CZL*19, KKJ*20, TTB*19, SPOP20, CLP*21]. Most approaches rely on a post-processing step which is performed after the detection or segmentation of the teeth to assign the teeth numbers and almost always incorporates knowledge about the spatial relation of teeth. Relying only on the extracted contours can be insufficient due to the low amount of teeth classes and similar shapes. Most importantly, only one previous approach combines the tasks of tooth instance segmentation and numbering [SPOP20]. However, the authors only perform a study on the performance of existing network architectures.

In summary, the combined segmentation and numbering of teeth in panoramic images is still open research question. A modeling of the local 2-D shape variation of the individual teeth along with their spatial relation to neighboring teeth is required to handle the challenges encountered with this task of teeth segmentation and numbering. This top-down information (i.e. the prior knowledge encoded in the model) can be used to guide the segmentation process beyond the bounds of the features extracted from the image while constraining it to statistically valid shapes and configurations. More importantly, the explicit modeling inherently includes the number associated with each tooth resulting in an approach that provides both the segmentation and number of each teeth. This thesis presents a novel model-based approach specifically designed for the combined task of teeth segmentation and numbering. Due to the increased shape and spatial variation of wisdom teeth, they will not be covered in this approach.

1.2.2. 3-D teeth reconstruction on five photographs

The aim of image-based reconstruction is to deduce the structure and geometry of 3-D objects from one or multiple 2-D images. It is an ill-posed, inverse problem which tries to recover the lost dimension resulting from the 3-D to 2-D projection process when capturing images. The problem is considered ill-posed since many different shapes can result in the same the 2-D projection. A 3-D image-based reconstruction of the teeth from five colored photographs poses the following challenges:

- The **minimal amount of images** available for the reconstruction poses a significant challenge. Image-based reconstruction already suffers from the fact that depth information is lost during the image acquisition process, making it difficult to reconstruct the 3-D geometry of the object. Typically, image-based reconstruction approaches try to alleviate this problem by using larger amounts of images or fixed camera setups with known configurations to try and recover the lost information. Both approaches are not applicable in the scenario of orthodontic photographs. The number of available images is fixed given the task of using the already established set of five documentary photographs. Moreover, the photographs are captured manually by the dental

professional. While the general viewing direction for each image is predefined according to the documentation protocol, variations in angle, distance and rotation are still common.

- Teeth are **textureless** and their **material properties** result in very challenging appearance properties which makes high quality reconstructions from photographs difficult [AFSEM14, WBG*16]. The thin outer translucent enamel coating of the tooth crown causes the tooth to be extremely specular. The underlying dentin layer however is very diffuse. Both materials exhibit strong subsurface scattering (subsurface light transport), meaning that light which penetrates the surface of the tooth is reflected multiple times at irregular angles before exiting the surface again. Consequently, teeth have very few visible features which makes it very difficult to establish correspondences between different images and therefore very challenging to recover the 3-D information.
- The **simultaneous reconstruction of multiple objects**, as is the case with teeth, poses another challenge. Generally, multi-object reconstruction becomes significantly more difficult (compared to single-object reconstruction) because the objects can suffer from non-trivial occlusions between any number of them. This reduces the overall amount of information for each object that can be extracted from the image(s). In case of the orthodontic images, parts of the teeth remain completely occluded even when using multiple/all perspectives, making a purely image-based reconstruction impossible.
- **Teeth variation**s like missing or malpositioned teeth add additional complexity to the reconstruction problem. They add to the variability of (partial) occlusions and further complicate the correspondence establishment between different views.

Image-based reconstruction is rarely applied in the dental domain, given the aforementioned very challenging appearance properties of the teeth. A typical technique in image-based reconstruction is to use local feature points to establish correspondences between images in order to recover the lost depth information. However, in case of dental photographs, the strongest usable feature is the boundary between individual teeth, or the tooth and the surrounding area [WBG*16]. This means that establishing correspondences between landmark points in different images is not possible since the boundary is not a feature defined by landmarks on the surface of the tooth but instead by the position of the camera relative to the tooth.

Nevertheless, some efforts have been made to reconstruct 3-D tooth models from different image modalities. Abdelrehim et al. [AFSEM14] used shape-from-shading techniques in combination with shape priors to reconstruct the occlusal surface of the crown of the posterior teeth from a single intraoral image. Zhang et al. [ZLS11] utilized the few occlusal features of posterior teeth to apply a 3-D morphing technique and deformed a standard teeth model to reconstruct a missing inlay or crown occlusal surface. In the case of teeth reconstruction from photographs, however, the underlying problem is an image-based multi-object reconstruction in the presence of large occlusions which is still considered as one of the future research directions [HLB21]. Solely relying on information that can be extracted from the images is insufficient to perform a robust and accurate reconstruction, especially in the presence of occlusions. Very few approaches in the literature have performed a reconstruction of multiple teeth, but

none actually use the documentary photographs from orthodontic procedures for the 3-D reconstruction of the teeth. Multi-object reconstruction with occlusions was performed by Farag et al. [FEA*13] and Mostafa et al. [MEA*14]. Both approaches aimed to reconstruct the anterior part of the dental arch from a single colored photograph. They used statistical models of shape and texture in a shape-from-shading approach to model the relationship between 3-D shape and photometric information. The most promising effort was a model-based reconstruction from colored extraoral photographs presented by Wu et al. [WBG*16]. They proposed to use parametric models of the upper and lower dental arch as the basis of their reconstruction. The models were trained to capture the local shape and pose variation of the teeth as well as the global pose and scale variation of the dental arch. As image features, the teeth boundaries were extracted from the images using a Boosted Edge Learning algorithm. The reconstruction was then formulated as a Maximum a Posteriori estimation problem in order to determine the optimal model configuration matching the extracted contours. Their system, however, required a minimum number of 8 photographs to perform the reconstruction. In addition, a manual initialization was also required by manually identifying two teeth per dental arch in each input image as their boundary extraction algorithm did not provide teeth labels/numbers to automatically match teeth.

All in all, there exists no approach which performs an automatic image-based teeth reconstruction from (colored) photographs. Moreover, even existing semi-automatic approaches require more than the proposed five images to perform a successful reconstruction. In order the fill this gap, a fully-automatic, model-based reconstruction approach is presented. It utilizes the boundary information of the teeth (silhouettes) extracted from the five documentary photos to handle the challenging appearance properties. As some posterior teeth, namely second molars and wisdom teeth, are at most visible in a single image, these are not included on the reconstruction as there is simply too less information available to perform a meaningful reconstruction. Moreover, it employs prior domain knowledge about the teeth shape and location in 3-D to enhance and tackle the limited information available from only five photos, the missing surface features and the presence of occlusions. Furthermore, it requires a numbering of the extracted boundaries in order to automatically establish the correspondences between the 2-D boundaries and the 3-D teeth model.

1.2.3. Conclusion

Both scenarios revolve around an accurate object-level teeth segmentation and numbering. While computer-aided systems to automate or assist with tooth assessment depend on segmentation and numbering in panoramic radiographs, the image-based 3-D reconstruction for orthodontic assessment requires a teeth segmentation and numbering in colored photographs. Given the generally predefined viewing directions of the five documentary photographs, a 2-D model-based approach that employs prior domain knowledge by taking into account the shape information of individual teeth in the images as well as their spatial relation to neighboring teeth is applicable for this scenario as well. Therefore, both the panoramic and the photometric teeth segmentation and numbering will utilize the same general concept for this task.

1.3. Goals

This thesis focuses on two main goals or Research Questions (RQ): (a) a teeth segmentation and numbering in dental panoramic radiographs, and (b) an image-based 3-D teeth reconstruction from five orthodontic record photographs (cf. Fig. 1.8).

The first goal is to perform a **teeth segmentation and numbering in dental panoramic radiographs** to support the development of medical imaging applications or computer-aided systems. As discussed in Section 1.2.1, this task requires a modeling of the local 2-D shape variation of the individual teeth along with their spatial relation to neighboring teeth to handle the associated challenges. Medical imaging approaches relying on global shape information often utilize statistical models in the form of active shape model [CTCG95] or active appearance models [CET01]. Techniques combining multiple active shape and appearance models by also utilizing the spatial relations between items are known as coupled (shape) models [TWT*03]. These models have proven useful in a variety of scenarios, like the segmentation of 22 risk structures in the head & neck area in CT images for radiation therapy planning [JSK*14], automatic localization of the sentinel lymph node and additional active nodes on SPECT-CT data [JMDW17], and automated bone segmentation and anatomical landmark localization in Cone-Beam CT [BIV*19]. Since human teeth consist of hard tissue and are therefore not subject to deformations while possessing a well known, characteristic shape, they are well suited for active shape models. The human set of teeth also has a fixed configuration, i.e. the positioning and ordering of the individual teeth inside the mouth, which makes it possible to train coupled shape models without having to worry about overlapping structures. This results in a coupled shape model which combines all teeth shape models into a single (larger) model, enabling the combined segmentation and numbering of all teeth simultaneously. The disadvantage of such models, however, is a sensitivity to the initial placement of the model onto the image. This thesis proposes to use the semantic segmentation performance to encoder-decoder neural networks in combination with dynamic model adaptation to handle the initialization dependency. The approach to teeth segmentation and numbering in dental panoramic radiographs is described in Chapter 5.

The second goal is an **image-based 3-D teeth reconstruction from five photographs** to enable an easier initial assessment of the teeth malposition and a more accurate treatment monitoring. Given the challenges described in Section 1.2.2, deforming an initial guess of the 3-D teeth shapes (deformation-based reconstruction) promises to be the most optimal solution to tackle the technical challenges. The initial guess of the teeth shape depends on prior knowledge and can be best provided by a model-based approach that incorporates the global shape variation of each individual tooth in combination with the spatial relation to other teeth. This model-based top-down information also enhances the deformation-based reconstruction by constraining the performed deformations to valid shapes and providing statistical knowledge about the expected shape of occluded regions. Due to the difficult appearance properties of the teeth in photographs, a silhouette-based loss function is used to compare 2-D projections of the 3-D teeth to their corresponding 2-D contours in the images. This shape-from-silhouette inspired approach will highly depend on an accurate object-level segmentation of the teeth. Additionally, for a fully automatic reconstruction approach, an automated matching between 2-D contours and 3-D model

teeth is required. As concluded in Section 1.2.3, the teeth contour extraction and numbering in photographs is similar to thesis's first research question of teeth segmentation and numbering in dental panoramic radiographs. Therefore, this thesis proposes to utilize the same general approach of coupled shape model, neural network and dynamic adaptation for the task of teeth segmentation and numbering in photographs. The approach to image-based 3-D teeth reconstruction from five photographs is described in Chapter 6.

As both goals/research questions use the same **general concept to teeth segmentation and numbering**, this concept is introduced in a separate chapter, namely Chapter 4. Task specific information about the usage of the concept is provided in the chapter related to the individual research question (cf. Chapters 5 and 6, respectively). Fig. 1.8 illustrates the connections between the different research questions which is reflected in the structure of the thesis (cf. Section 1.5). The flexibility of the presented framework is also demonstrated in a separated application (cf. Chapter 7), which is not directly related to any of the two goals.

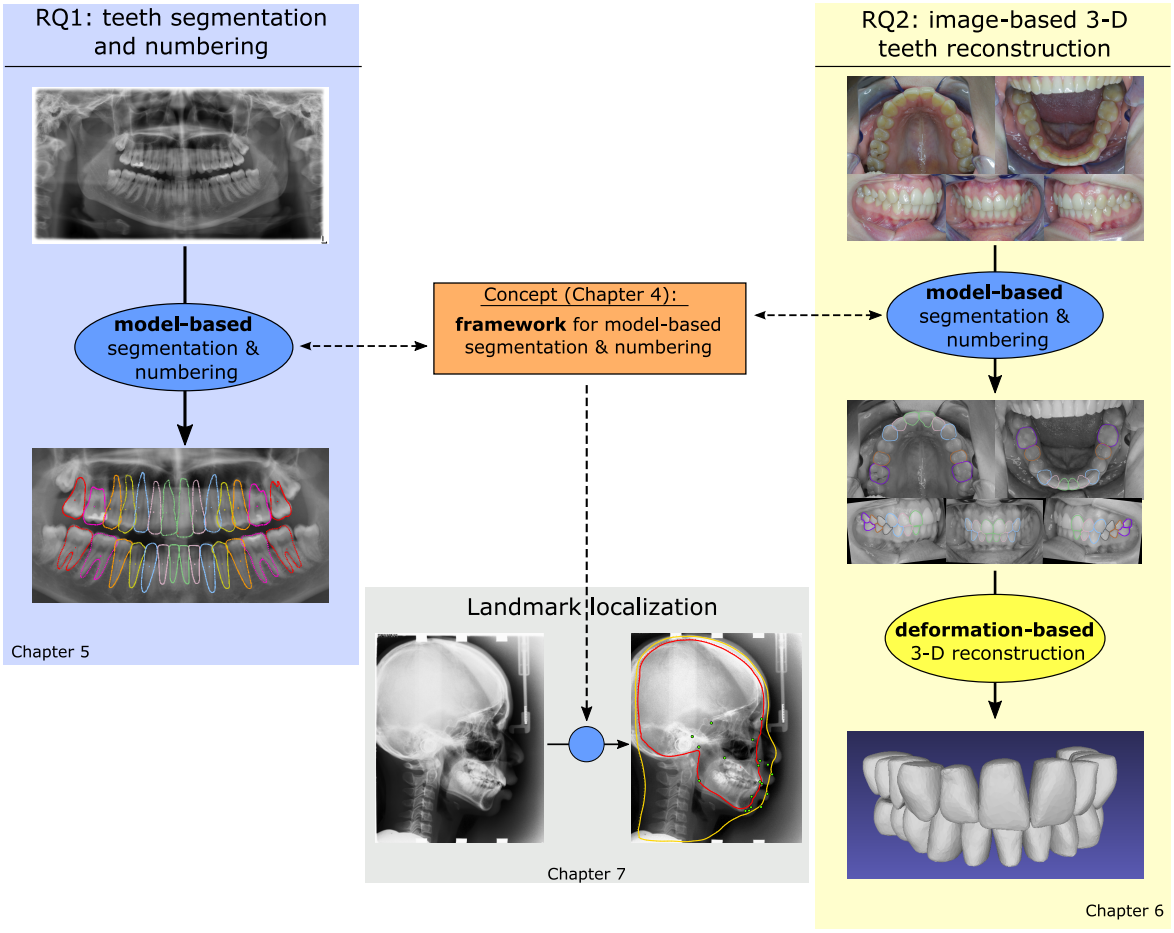


Figure 1.8.: The illustration depicts the two major research questions (RQ) "teeth segmentation and numbering" (blue) and "image-based 3-D teeth reconstruction" (yellow) of this thesis. The common challenge in both topics is a model-based segmentation and numbering of teeth. A common framework (orange) is therefore devised in Chapter 4 and utilized for both tasks. The flexibility of the framework is demonstrated in a third application to perform landmark localization (gray).

1.4. Contributions

The main contributions of this thesis can be summarized as follows:

- It proposes a model-based concept for simultaneous teeth segmentation and numbering. The concept includes tooth-specific statistical shape models to capture the shape variance of each tooth in order to provide prior information to guide the segmentation process while restricting it to valid shapes. In addition, the individual shape models are coupled by their statistical spatial relations to form a coupled shape model. The crucial task of initial positioning of the coupled shape model onto the image is done by employing the semantic segmentation performance of convolutional neural networks and specialized, dynamic adaptation strategies. The concept is described in Chapter 4.
- The model-based segmentation and numbering approach is applied for teeth segmentation and numbering in panoramic radiographs. It is the first approach that explicitly combines spatial and shape domain knowledge for simultaneous teeth segmentation and numbering. It proposes specialized multi-step adaptation strategies for fitting the coupled shape models to the teeth in input images. These strategies control which structures of the model are actively adapted to the image (by utilizing image features) in order to alleviate the limitations of local search algorithms and ensure an optimal segmentation process. The teeth segmentation and numbering in panoramic radiographs in described in Chapter 5.
- It proposes the first fully-automatic image-based 3-D teeth reconstruction approach from colored photographs. It is also the first reconstruction approach which utilizes the documentary photographs typically taken during orthodontic procedures. The approach is deformation-based, where a mean 3-D coupled shape model of the teeth is deformed using a silhouette-based loss function to find the optimal match between model parameters and photographs. Correspondences between 3-D model teeth and extracted silhouettes are established using the numbering information provided by the segmentation framework introduced before (cf. Chapter 4) which exploits the predefined viewing directions of the photographs. The 3-D reconstruction approach is described in Chapter 6.
- It demonstrates the flexibility of the model-based segmentation and numbering concept by applying it to a different type of application, namely the automatic analysis of cephalometric radiographs. The analysis of these types of radiographs relies on the localization of specific landmarks within the image. The capability of the concept to encode domain knowledge is used here to exploit the spatial information between landmark points and other anatomical structures to support the estimation of their location. This application is introduced in Chapter 7.

1.5. Structure of the thesis

The thesis is structured as follows: Chapter 2 provides some fundamental information about selected dental topics. The goal is to facilitate a better understanding of the thesis by introducing knowledge

about the human teeth, dental notation systems and dental imaging. In Chapter 3, the state-of-the-art methods related to the two focus areas of this thesis, namely segmentation (and numbering) of structures in dental images as well as image-based 3-D reconstruction of teeth, are reviewed. While this chapter will mainly focus on the dental domain, it will also include some methods from other medical areas. The concept proposed for teeth segmentation and numbering will be introduced in Chapter 4. It describes the structure and design of the coupled shape model (CSM), the general approach to the initialization of the CSM as well as the dynamic adaptation strategies used to segment the teeth boundaries in the images. The application of this concept for teeth segmentation and numbering in panoramic radiographs is described in Chapter 5. It describes the design of the coupled model as well as application specific methods for model initialization and the adaptation of the model to the input image. The performance of the approach will be evaluated on two different data set and discussed at the end of the chapter. Chapter 6 explains the first fully-automatic photo-based 3-D reconstruction approach of the teeth. It utilizes the five colored photographs from orthodontic records and uses the segmentation and numbering concept to extract the teeth contours from these photographs. The reconstruction approach will be evaluated by comparing the generated 3-D teeth models to gold-standard laser scans. A third application of the framework for automatic cephalometric analysis is introduced in Chapter 7. Finally, Chapter 8 concludes the thesis by providing a summary and discussion of the presented contributions, and suggests possible directions for future work.

2. Fundamentals

This chapter presents some fundamental information from the dental domain which is relevant for the understanding of the topic of this thesis. It describes the human set of teeth, introduces several dental notation systems that are currently used by dental professionals and, most importantly, provides an overview over the imaging techniques used in dentistry. The book 'A Dictionary in Dentistry' by Robert Ireland [Ire10] served as a general reference for most information provided in this chapter. After finishing this chapter, the reader will have sufficient knowledge about the dental domain and imaging techniques used there to be able to follow the rest of the thesis.

2.1. The human teeth

The full human permanent set of teeth consists of 32 individual teeth which serve the function of cutting and grinding pieces of food in preparation for digestion. The teeth are located inside the mouth and are positioned in two *dental arches* on the upper (*maxillary*) and lower (*mandibular*) jaw. Each arch contains 16 teeth, with the maxillary arch being slightly larger than the mandibular one. The permanent set of teeth is the second and final set that humans develop. It replaces the primary set which only consists of 20 teeth, ten on the maxilla and ten on the mandible, during the ages of 6 and 12.

Generally, an individual tooth can be separated into three different parts which are *the crown*, *the neck* and *the root* (cf. Fig. 2.1). The crown is the upper part of a tooth and (under normal circumstances) the only part that is directly visible. It mostly consist of enamel which is the hardest and most highly mineralized substance in the human body. The outer shape of the crown varies between different types of teeth. Located at the core of the crown (i.e. the center of the tooth) is the pulp chamber. It contains the blood vessels and nerves of tooth which enter the tooth through its root. The root part makes up the lower half of the tooth and anchors the tooth in the jaw bone. Teeth have between 1 and 3 roots depending on their type. Each root features a pulp canal through which the blood vessels and nerves run. The area between the crown and the root is called the neck of the tooth. It is considered to be the part where the root exists the jaw bone and is only covered by the gingiva.

There exist four different types of teeth that each have their own specific function. Fig. 2.2 depicts the four types as well as their spatial configuration on a human jaw. The incisor teeth are located at the front (of the teeth arc) and serve the function of cutting the food. Their crown is therefore more edge-shaped. Humans typically have 4 incisors on each jaw with the inner pair designated as the *central incisors* and the outer ones called *lateral incisors*. The canines have a more pointy-formed crown and are used for tearing the food apart. They are located next to the lateral incisor teeth. The molar and

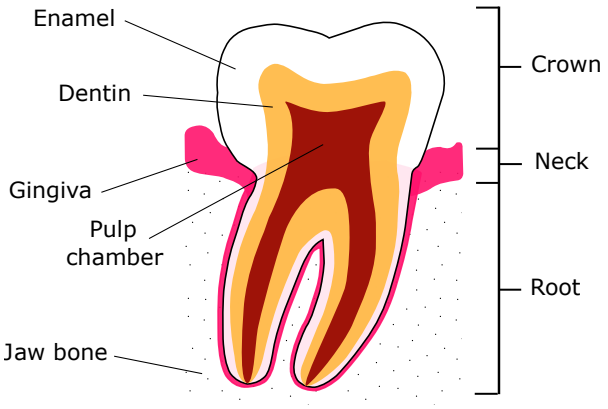


Figure 2.1.: Illustration of a human tooth showing the structure as well as the surrounding tissues. The three different parts, which tooth are generally separated into, are indicated on the right.

pre-molar teeth grind the food in preparation for digestion. They occupy the rear part of the dental arc with the pre-molars positioned before the molar teeth. Human poses 2 pre-molar and 3 molar teeth on each side of the arc. The most posterior molar (third molar) is also known as wisdom tooth. It is special in the sense that it develops later than the other molars and it is possible for adults to have fewer than four or none at all. Because of their late development, it is also possible that there is not enough space on the dental arch to fit the additional tooth. This can cause the tooth to become stuck against other teeth and be fully or partially covered by gingiva. Teeth which are only partially erupted through the gum can be more difficult to clean and promote the risk inflammation or tooth decay. Wisdom teeth are therefore often extracted.

2.2. Dental notation systems

In order to easily identify the teeth, professionals in dentistry use several different notation systems. The three most popular systems are:

(i) **the Universal Numbering System** [ADA99], which is used by most professionals in the United States;

(ii) **the Fédération Dentaire Internationale (FDI) System** [ISO16], also known as the ISO-3950 system, which is used worldwide;

(iii) **the Palmer Notation System** [Har05], which, although superseded by the FDI system, is still used in the United Kingdom;

The different systems vary from each other in the way they assign labels to refer to individual teeth. An illustration of the different numbering systems is provided in Fig. 2.3.

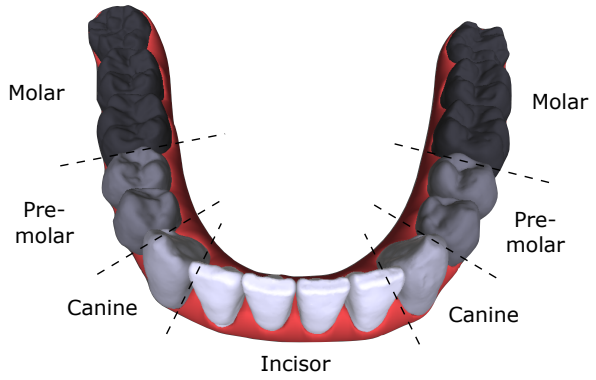


Figure 2.2.: Teeth types. Illustration of the four different teeth types (incisor, canine, pre-molar and molar) and their position on the (mandibular) dental arch.

The Universal Numbering System uses numbers in ascending order to identify teeth. The third maxillary molar on the right side (from the patient's point of view) is designated as number 1. The counting then continues along the other maxillary teeth toward the front and further to the third molar on left side which is number 16. The mandibular teeth then continue the numbering from left to right, in doing so the left third molar is assigned the number 17. The numbering finishes with the right third mandibular molar as number 32.

The ISO notation system uses a two-digit identifier. The mouth is divided into four sections called quadrants and the first digit refers to the quadrant the teeth belongs to. Starting from the maxillary right quadrant, which is assigned the number 1, the numbering continues in anti-clockwise direction with the maxillary left quadrant getting assigned the number 2, the mandibular left quadrant the number 3, and the mandibular right quadrant the number 4. The second digit then refers to the teeth within each quadrant. They are assigned a number from 1 through 8, with number 1 is being assigned to the central incisor and number 8 to the third molar (wisdom tooth).

The Palmer notation system also uses the numbers 1 through 8 in the same way as the ISO system to identify teeth within one quadrant. The individual quadrants however are referred to by symbols instead of numbers. The maxillary right quadrant is illustrated by ⌋, the maxillary left one by ⌊, mandibular left by ⌈ and the mandibular right quadrant by ⌉. These symbols provide a better visual indication of the respective quadrant. However, there are some difficulties in reproducing these symbols digitally.

It is worth noting that all notation system differentiate between primary teeth and permanent teeth. The previously explained notation for all three systems is used for permanent teeth only. To refer to primary teeth, the Universal and Palmer notation systems exchange digits by letters, while the ISO system introduces quadrants 5 to 8. All accounting for the reduced number of 20 primary teeth.

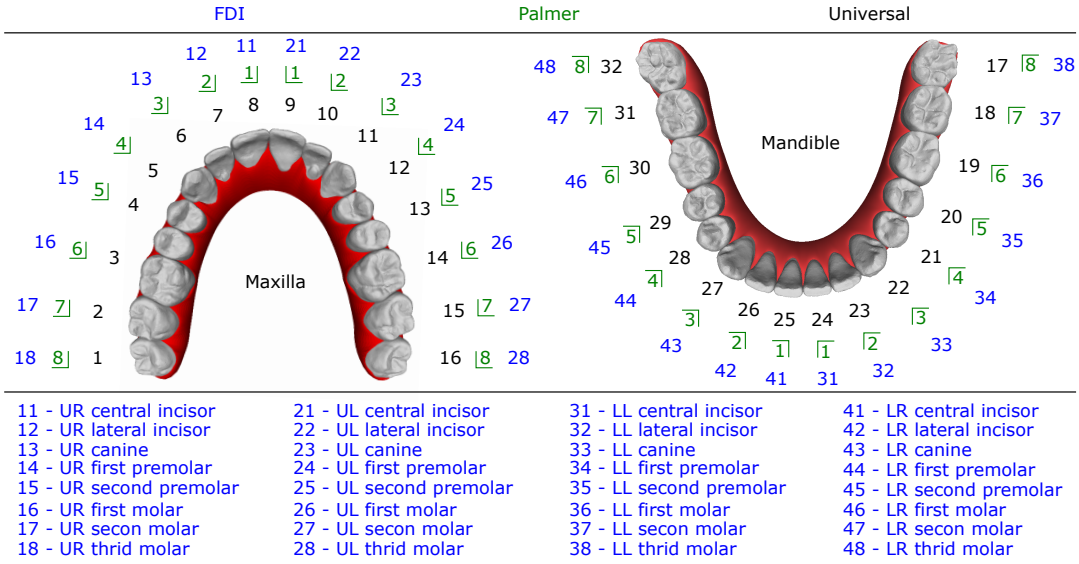


Figure 2.3.: Illustration of the most common dental notation systems: the Universal Numbering System, the Fédération Dentaire Internationale (FDI) System and the Palmer Notation System.

2.3. Dental imaging

In dentistry, the examination of a patient's set of teeth is typically done using one of several imaging techniques. Oral inspection alone does not provide sufficient information as most dental problems are not directly visible. The most commonly used image modalities to examine the teeth are various types of 2-D dental radiographs, like panoramic, bitewing or periapical (cf. Section 2.3.3). Colored photos are sometimes captured but usually only serve a documentary purpose since they do not provide any additional information compared to an oral inspection.

Dental radiographs, also referred to as dental X-rays or dental radiographic images, depict the patients hard tissues (teeth and bones) and the surrounding soft tissues (cf. Figs. 2.5 and 2.6). They are an important source of information in dental care as the entire tooth including its root, the bone composition and the surrounding soft tissue can be assessed. Identifying problems early on enables treatment of these problems before they become severe, reducing pain and discomfort for patients and saving money compared to more complex treatments later on. The relevant region of dental radiographs can be roughly divided into three main classes [SS74]:

(i) **Teeth area:** has high to medium values of gray scale (intensity) and includes the teeth of the patient,

(ii) **Dental structural area:** has medium values of gray scale and consists of gums, bone, and other periodontist structures,

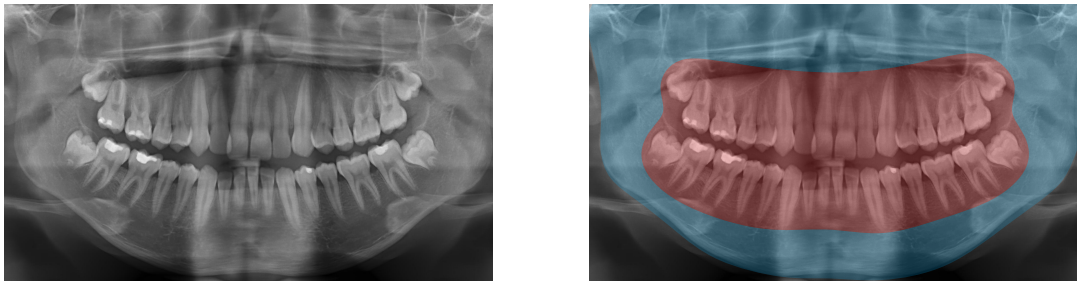


Figure 2.4.: Different areas in dental radiographs. Original image on the left, areas illustrated on the right: teeth area (red), dental structural area (blue), background area (not colorized).

(iii) **Background area:** has the lowest gray scale values among the three parts and shows the background of a teeth structure.

An illustration of these areas can be seen in Fig. 2.4. The original radiograph is depicted on the left while the different areas are indicated on the right image. The teeth area is colored in red and the dental structural area on blue. The background area remains uncolored. The boundaries between classes can be difficult to identify in certain parts. The roots of the teeth often have similar intensity values to their surrounding bone structure.

Within the teeth area, any teeth restorations like fillings, bridges or implants have even higher intensity values than the teeth due to their higher density. This makes them immediately apparent and clearly distinguishable from the natural teeth. Caries can be identified by lower intensity values in a local area with a tooth.

2.3.1. Benefits of dental radiographs

Dental radiographs enable the dentist to check the state of the teeth by visualizing hidden structures like tooth root and jaw bone. This way, the dentist can identify many dental diseases, injuries or development problems which would be hard or impossible to detect during a normal oral examination. This includes tooth decay (dental caries), especially if it is located between teeth or below restorations (fillings). Furthermore, it includes diseases in the bone or peridontal diseases – i.e. diseases affecting the soft tissue (gum) which holds the teeth in place – which in its more serious form is called peridontitis, and infections developing below the gums and certain types of tumors. Early detection of diseases, infections and injuries is important to prevent or limit further damage. Additionally, multiple radiographs captured over a period of time can be analyzed to provide information about changes in soft and hard tissues of the patient. This is useful for treatment monitoring, tracking of lesion development or following the development of teeth and jawbones in children. Dental radiographs are also utilized for treatment planing of tooth implants, braces, dentures or other dental procedures.

Digital radiographs which are state-of-the-art nowadays offer further benefits compared to the old film-based radiographs. Digital images can be enhanced and enlarged many times their actual size on

the computer screen, making it easier for the dental professional to identify and communicate dental health issues.

2.3.2. Acquisition of dental radiographs

The imaging technique used to capture radiographic images (X-ray images) is called radiography. This technique uses Röntgen radiation, also known as X-radiation or X-ray for short, to capture the internal structures of objects normally hidden from the human eye. Here, X-ray refers to electromagnetic radiation of extremely short wavelength and high frequency that can penetrate certain materials. It can pass through softer tissues like gums and cheeks and is absorbed by denser hard tissues like bones and teeth. Radiography works by measuring the energy absorption levels of objects when this type of radiation travels through the objects, i.e. it measures the difference in energy level between a source and a detector with the object in between the two.

The X-ray generator is used to create beams of X-ray with a known energy level which are directed at the target object. On the opposite side of the object, an X-ray detector 'captures' the beams of X-ray and records the remaining energy level. While the radiation travels between generator and detector, energy will be absorbed depending on the density of the material the radiation passes through. For example, harder materials like teeth and bones will absorb higher amounts of energy and therefore lower the energy levels measured at the detector, while softer tissue such as skin will absorb lower amounts of energy. From the measured absorption levels, a gray-scale (intensity) image representation can be generated where beams with high absorption appear as light-colored (brighter) areas while beams with low absorption are depicted as darker areas.

Depending on the type of radiographs, either a single or multiple X-ray images are captured and then fused the generate the final radiographs. It has to be noted that high-energy radiation can irreversibly damage tissue it passes through. The use of radiographs is therefore not without risk. The safety of dental X-ray images is discussed in Section 2.3.4.

2.3.3. Types of dental radiographs

Dental radiographs are divided into two main categories: **Intraoral**, a radiographic technique performed with the film/sensor positioned inside the mouth, and **extraoral**, in which the patient is positioned between the radiographic film/sensor and the X-ray source (the radiograph is obtained outside the patient's mouth).

Intraoral radiographs, like bitewing, periapical or occlusal X-rays, are two-dimensional radiographs that only show isolated regions of the set of teeth or individual teeth. They are the most common type of dental radiographs and provide a lot of detail and are mostly used to get more specific information about single teeth. They allow the dentist to check for cavities, the health of the root and surrounding bone or the status of developing teeth.

Figure 2.5.: Examples of intraoral radiographs: (a) bitewing (image from [WHL*16]), (b) periapical (image from [SFNA04]), (c) occlusal (image from [SKRB13]).

The various types of intraoral radiographs focus on different aspects of the teeth. Examples of these types are demonstrated in Fig. 2.5. **Bite-wing radiographs** depict both maxillary and mandibular teeth in one specific area of the mouth. They typically show the crown and neck of the teeth (cf. Section 2.1), but do not include the roots. They are useful to detect tooth decay (cavities) between teeth and gum disease related changes in the thickness of the bone. Furthermore, they support the determination of the proper fit of a crown or the integrity of fillings. **Periapical radiographs** depict the entire tooth including crown, root and surrounding area. The images focus on one area of either upper or lower jaw and typically show three to five teeth at once. This type of radiographs is used to identify any abnormalities in the root and bone structures. **Occlusal radiographs** are the largest type intraoral image and depict the complete dental arch of either maxilla or mandible from an occlusal view. They are useful to assess the placement and development of all teeth in the respective arch.

Extraoral images, like cephalometric or panoramic, are not restricted to only an isolated part of the teeth as occurs in intraoral radiographs but capture the entire teeth region as well as the surrounding areas. Fig. 2.6 depicts examples of different types of extraoral radiographs. They provide fundamental information about the teeth, jaw and skull as they also show joints between the jaws and the skull, the chin, the spine and other details originated from the bones of the nasal and face areas. Therefore, these types of radiographs are used to get an initial assessment of the overall health status of the teeth, to look for impacted teeth and to monitor growth and development of the jaws in relation to the teeth. Particularly, **panoramic radiographs**, also referred to as orthopantomography, are a common tool to complement the initial clinical examination of the teeth. Panoramic images depict the entire mouth region which includes all mandibular and maxillary teeth as well as the jaw bone and part of the nasal area. They provide an overview of the positioning of all fully-emerged teeth as well as emerging teeth which are still covered by gum. Panoramic images are also used as preoperative examination of the teeth and bones for surgeries in the temporomandibular region [AA15, WHL*16]. **Cephalometric radiographs** show a lateral projection of the entire head. They therefore include the complete skull and upper vertebrae in addition to the teeth and jaw bone. This type of radiograph is useful for the examination of the teeth in relation to the jaw bone and the relative position of upper and lower jaw. Special analysis methods have been proposed for cephalometric images. **Cone-beam computed tomography** (CBCT) images, also called digital volume tomography (DVT), is the only three-dimensional imaging

Figure 2.6.: Examples of extraoral radiographs: (a) panoramic, (b) cephalometric.

technique described here. These images visualize the entire mouth region like panoramic radiographs. They are less common due to their higher exposure to Röntgen radiation. However, they provide the most detailed information about a patient's hard and soft tissue. They are used to identify and analyze tumors or fractures and to plan surgeries and treatments related to dental implants or challenging extractions.

2.3.4. Safety of dental radiographs

Like all types of radiographic images, the image acquisition process involves the exposure of patients to x-rays (also referred to as Röntgen radiation). This form of high-energy electromagnetic (EM) radiation is a known and proven human carcinogen. Fortunately, the amount of radiation dose required to obtain dental radiographs is very low in general, but also compared to other radiographic techniques. As the head region is the only one that requires exposure during the process, the patients neck and body can be protected from the radiation by wearing a leaded collar or apron, respectively.

The effective radiation dose is measured in microSievert (μSv). According to the "European guidelines on radiation protection in dental radiology" [EUR04] the effective dose for conventional intraoral (e.g. bitewing or periapical) x-rays is between $1-8\mu Sv$, for lateral cephalographs between $2-3\mu Sv$ and for panoramic images between $4-30\mu Sv$. For comparison, the normal background radiation humans are exposed to averages to about $6.58\mu Sv$ per day. A single whole body CT scan has an effective dose of approximately $12.000\mu Sv$. Therefore, dental radiographs are considered safe. According to the America Dental Association[1] "because of the low radiation dose associated with dental radiographs, people who have received radiation treatment for head and neck cancer can undergo dental radiography safely".

3-D CBCT scans (also known as DVT) of either maxilla or mandible however have an effective dose between $100-3300\mu Sv$. These types of x-rays are typically used for implant planning as they

[1] American Dental Association, "Dental radiographs - benefits and safety", 2011, http://www.ada.org/~/media/ADA/Publications/Files/for_the_dental_patient_sept_2011.ashx (last visited on 27.04.2021).

Figure 2.7.: Exemplary set of documentary photographs to illustrate the different views: a) left view, b) anterior view, c) right view, d) maxilla view, e) mandible view. (The depicted regions have been cropped from the original photos for better visualization of the teeth.)

provide three dimensional information, especially about the roots of the teeth. This x-rays have a dose significantly higher than conventional radiographs and should only be performed if medically justified. The use of sensor-based digital radiography (like charge-coupled devices (CCD) or photo-stimulated storage phosphor (PSP)) instead of conventional film-based radiography can further reduce the amount of exposure.

2.3.5. Orthodontic records

Orthodontic records of a patient are created before any orthodontic procedure to create an accurate documentation of the condition of the dentition and soft tissue and to plan an individualized treatment for the patient. Here, orthodontic procedures refer to treatments of malpositioned teeth or misaligned teeth or jaws that employ dental braces to apply forces to the teeth in order to move them to their correct position. Orthodontic records are also used throughout the treatment to keep track of progress over time and identify deviations from the planned (ideal) treatment progression. A record can contain dental radiographs, like panoramic and cephalometric X-ray images, that provide information about the root positions and facial bone structure, and 3-D impressions (molds) of the teeth which are used to generate models of how the teeth fit together. Additionally, colored photographs of the teeth and face can also be included. While photographs do not reveal any hidden structures, they are easy to capture and can document the treatment progression without exposing the patient to unnecessary radiation. The face photos depict the patient's head from a frontal and lateral position with a smiling and non-smiling facial expression. The teeth photos are composed of a set of five colored photographs which depict the patient's teeth from five general viewing directions:

- Frontal view: an anterior view of the teeth in a closed-mouth configuration (i.e. the maxillary teeth are resting on the mandibular teeth),
- Lateral views (left and right side): views showing the teeth in a closed-mouth configuration from a left and right side, respectively,
- Occlusal views (maxilla and mandible): views depicting the teeth of the upper and lower jaw from a top-down position, typically captured using an intraoral mirror.

A set of teeth photographs is shown in Fig. 2.7. The different viewing directions are intended to maximize the amount of information about the teeth that is captured in these images while keeping the

number of images as low as possible. Since the photographs are taken manually without a fixed camera setup, the viewing direction varies slightly between images of the same category. Especially the lateral images suffer from variations in the lateral viewing angle. While this might not impact the usefulness of the photographs for documentation purposes, it makes automatic computational analysis of these photos more challenging.

3. Related Work

Medical image segmentation and image-based reconstruction are important research areas as they provide fundamental knowledge which is required for and utilized in a wide variety of applications. This chapter provides a comprehensive overview of the current state of research in the fields of dental image segmentation and teeth reconstruction. The emphasis is on methods which are linked to the technical challenges associated with the topics of this thesis, but also include some accompanying methods with similar goals.

The chapter is structured in the following way: Section 3.1 introduces the taxonomy used to categorize the related work for medical image segmentation, provides a concise overview of the approaches proposed in recent years in the dental domain and concludes with a summary and discussion of the limitations of the current state of the art. Section 3.2 briefly presents different methods for the 3-D reconstruction of teeth but focuses on image-based reconstruction. This includes a general overview of the varying concepts in the literature for image-based reconstruction in different scenarios. While the main focus is the dental domain, certain concepts that only have received limited attention in this domain will be explored through other medical fields. The conclusion summarizes the challenges involved in image-based teeth reconstruction and the limitations of existing approaches.

3.1. Dental Image Segmentation

Segmentation is a method to extract information from medical images which can then be used for diagnosis, monitoring or treatment of medical conditions. Segmentation refers to the process of partitioning a single digital image into multiple non-overlapping, connected segments which are considered homogeneous according to some characteristics [PXP00]. Typically, in the medical domain, these segments represent anatomical structures like organs, bones, teeth, blood vessels or different tissue types but can also be disease specific indicators like tumors or caries.

Image segmentation approaches can be grouped based on how they handle instances of the same object class within one image. Semantic segmentation approaches classify each pixel of the image into known classes without differentiating between object instances. In case of teeth segmentation this means all teeth will receive the same class label with no possibility to retrieve individual teeth from the segmentation result alone. Instance segmentation on the other hand combines semantic segmentation with object detection to classify, localize and segment each individual object instance of a object class.

Segmentation is considered the most important but also the most challenging task when dealing with medical images [EJ19]. The 20 year old conclusion that no single algorithm is able to robustly segment

a large variety of relevant structures over a broad range of different data sets [DA00] still hold true today, despite the advances made since then. Therefore, various approaches for each segmentation problem have been proposed which all vary in terms of complexity, achieved accuracy, robustness and speed [EGSEB11]. The challenges in medical image segmentation are the generally low image quality and resolution as well as noise or artifacts, low contrast between individual structures or large variance in terms of appearance and shape of structures belonging to the same class [Ste13]. This is also reflected in the fact that even human experts in the field show a significant inter- and intra-observer variability when annotating structures in medical images [JCCS19].

In the last decades, many different approaches to (medical) image segmentation have been presented which utilize individual or a combination of techniques to handle the associated challenges. These approaches can be grouped based on different criteria, like employed image features, required user interaction or techniques used. For example, Silva el al. [SOP18] and Son and Tuan [ST16] focus on dental image segmentation and base their taxonomy on the signal information that is used (e.g. threshold-based or cluster-based, also referred to as pixel-based, region-based or boundary-based). On the other hand, Elnakib et al. [EGSEB11] look at medical image segmentation in general and categorize the approaches into rule-based, statistical, atlas-based, and deformable model-based techniques. For this thesis, the proposed approaches for dental image segmentation have been grouped according to the amount of prior knowledge that is encoded and used for the segmentation task. Incorporating prior knowledge, like shape variation, (relative) location or appearance, into the segmentation process increases its robustness to low image quality, low contrast or artifacts. The goal is to show that incorporating such knowledge is beneficial for the tasks of teeth segmentation and teeth numbering. However, the incorporated knowledge also causes the process to be more specific to the chosen segmentation task.

Fig. 3.1 depicts the chosen taxonomy, which additionally groups the approaches into signal-based, model-based and neural network approaches. It is based on the taxonomy used by S. Steger [Ste13], but has been adapted for the dental domain and enhanced with neural network approaches. It is furthermore inspired by and built from multiple other sources [SOP18, EJ19, EGSEB11, ST16].

In the following, each category of segmentation approaches is described, references to their application in dental image segmentation are provided and the advantages and disadvantages are discussed.

3.1.1. Pixel-based

Pixel-based segmentation techniques determine the segment an individual pixel belongs to on a pixel-by-pixel basis. The only image information employed during the segmentation process is the value stored in each pixel (e.g. intensity or color). Pixel-based techniques are considered to be simple approaches and individual segments are determined based on the similarity of pixel values. Examples of such techniques are thresholding and clustering.

In thresholding, the pixels in an image are divided into two or more groups depending on their pixel value, which is either above or below a specific value (threshold) or within a certain value range. The threshold value or the lower and upper bounds of the value range can either be provided as prior

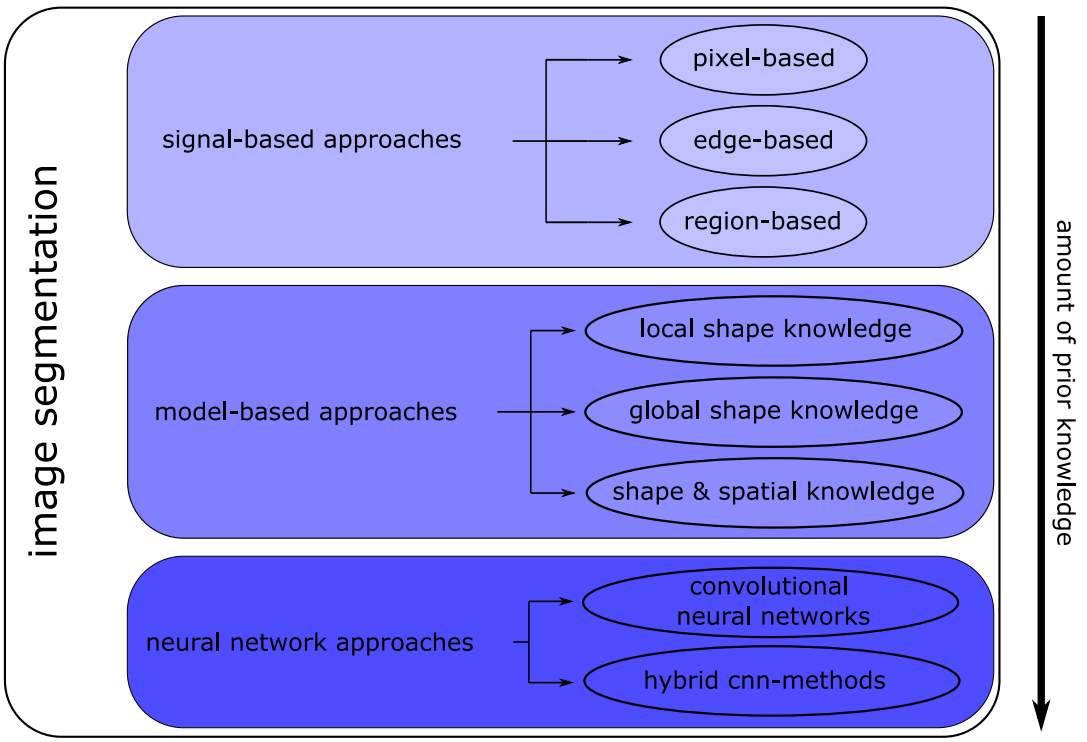


Figure 3.1.: Dental image segmentation taxonomy: The different methods are classified according to the amount of prior knowledge that is utilized/integrated. While neural networks have a larger amount of parameters and require more training data, the included knowledge cannot be modeled as specific as with model-base approaches.

knowledge or can be (semi-)automatically determined from the image. The output generated by the thresholding techniques is usually a binary image separating the pixels into foreground and background. For example, the popular Otsu thresholding method [Ots79] partitions the gray values into two groups by minimizing the variance within each group.

According to an in-depth study of the literature on segmentation methods applied in dental imaging from 2018 [SOP18], the majority of approaches (54%) at that time were threshold-based [NAM05, NAM07, NAM08a, NAM08b, LGN13, RIA*17].

Razali et al. [RAMZI14] analyzed three adaptive thresholding methods for teeth segmentation in panoramic radiographs. Adaptive methods typically produce better results for foreground-background separation than methods using a fixed threshold value. The authors compared mean-, median- and Otsu-thresholding for separating the teeth area from the background- and dental structural area. The arithmetic mean and median of the pixel values for a given image are used as threshold values, respectively. Otsu thresholding [Ots79] chooses the threshold value such that the variance within each group is minimized. The methods were compared on 20 dental radiographs and the authors concluded that median thresholding results cover more of the teeth area than the other two approaches. However, all three approaches – and specially median thresholding – also include a lot of the background area in the segmentation result.

Amer and Aqel [AA15] have proposed a threshold-based approach to extract the wisdom teeth from panoramic x-ray images. Their three-stage approach consists of a pre-processing step to improve image quality and remove unwanted regions, a region of interest (RoI) extraction step to extract image sub-regions containing the wisdom teeth, and a post-processing step to enhance the results and remove any adjacent teeth contained in the RoI. The pre-processing step uses contrast enhancement via intensity transformation techniques to better separate teeth from surrounding regions, Otsu-thresholding to remove the background area, morphological dilation to connect objects and smooth boundaries and connected component labeling to filter out unwanted regions. For RoI extraction, the image was dynamically divided into two horizontal parts in order to separate upper and lower jaw. Fixed-sized regions based on average wisdom tooth sizes in each jaw were extracted at the outer ends of the horizontal parts to extract the wisdom teeth – assuming that four wisdom teeth are present in the image. Finally, the post-processing step uses similar techniques as the pre-processing step to improve contrast (histogram equation) and remove further unwanted parts like surrounding tissue or adjacent teeth (Otsu-thresholding, morphological opening and connected component labeling). The final results are the individual region of interest images which contain the extracted wisdom teeth.

Global threshold segmentation is particularly effective, if the object of interest and the surrounding area have substantially different pixel values. However, thresholding using a single global threshold value usually fails if large variations in contrast or illumination are present in the image. This is typically the case with dental radiographic images. It is obvious from the horizontal images obtained by Amer and Aqel [AA15] that simple global thresholding (despite being combined with other pre-processing steps) is unable to accurately separate all teeth from the background, as shown in Fig. 3.2.

Figure 3.2.: Threshold segmentation example: original image (left) and extracted teeth (right). The result produced by global threshold leave room for improvement. (images from [AA15]).

A variable threshold based on the local pixel value statistics could provide better results. The approach proposed by Indraswari et al. [IANJ15] segments the teeth in panoramic radiographs by combining Multistage Adaptive Thresholding (MAT) with Decimation-Free Directional Filter Bank Thresholding (DDFBT). DDFBT is first used to decompose the input image into two directional images using horizontal and vertical directional filters. Therefore, the image is first transformed into frequency domain using Fourier transform. The domain image is then separated into vertical and horizontal directional images via Otsu-thresholding. This process enables the authors to enhance the edges in the vertical image before combining both directional images again into one image. The intention of the edge enhancement is to increase the contrast between tooth root and surrounding tissue as this areas typically suffers from low-contrast. Multistage Adaptive thresholding is then used to extract the teeth from the enhanced image. Initially, a global multilevel Otsu thresholding separates the pixels into three groups: *background*, *teeth* and *unknown*. Adaptive local thresholding is then used to match *unknown*-pixels to either *teeth* or *background* groups. The algorithm, however, only works on small sub-regions of the original panoramic radiograph which have to depict part of the teeth area.

In clustering methods, the pixels are grouped by assigning them to one of multiple clusters. In an iterative process, each pixel is assigned to the cluster it is most similar to in terms of some distance function. Many different types of clustering models exist that can be, for example, connectivity-based, centroid-based, distribution-based or density-based. The number of desired clusters usually has to be provided as input to the clustering algorithm. Clustering methods can be classified into two main groups: hard clustering methods which assign each pixel to exactly one cluster, and fuzzy clustering methods which describe the degree of cluster membership of all pixels by means of a (membership) matrix. In doing so, a pixel may belong to multiple clusters at once. In centroid-based k-means clustering data is grouped into a given number of k clusters by minimizing the within-cluster value variance in each cluster based on the squared Euclidean distance.

Son and Tuan [ST16] performed X-ray image segmentation by using a semi-supervised entropy regularized fuzzy clustering technique (eSFCM) together with Otsu thresholding and fuzzy c-means clustering. Semi-supervised (fuzzy) clustering utilizes additional user-provided information to guide and control the clustering process and enhance its performance. In semi-supervised entropy regularized fuzzy clustering [YYMS09], an entropy factor is integrated into the clustering algorithm but requires

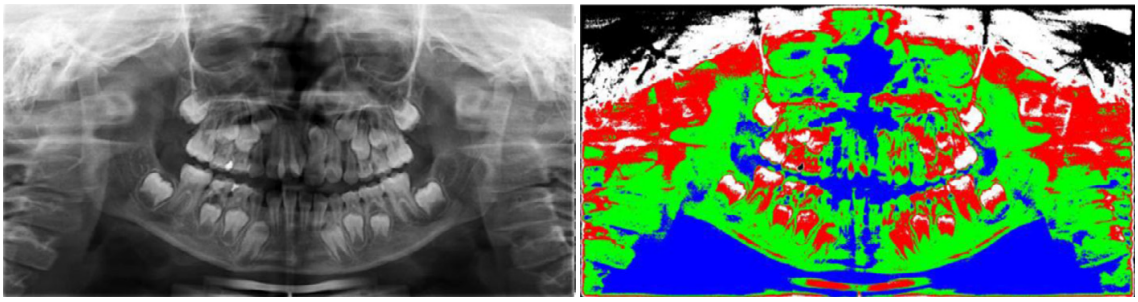


Figure 3.3.: Clustering segmentation example: original image on the left and clustering results on the right. No single cluster contains only teeth, instead all teeth are spread over multiple clusters which also contain parts of the dental structural and background area. (images from Son and Tuan [ST16]).

an initial membership matrix as additional input. The membership matrix already provides an initial guess for the cluster-affiliation of each pixel. Their motivation is to integrate clustering methods with other types of algorithms in order to improve the overall performance. In their approach, the authors start by removing the background area of the dental radiograph with Otsu thresholding and compute the membership matrix by separating the teeth area from the dental structural area via fuzzy c-means. The final segmentation result is then generated by improving the initial membership guess by using the entropy regularized fuzzy clustering technique. The authors refer to their approach as eSFCM-Otsu and tested the segmentation performance on 56 images with several cluster-focus metrics. They conclude that their approach performs better than other clustering techniques, but it could also be improved by employing spatial components in the objective function. An example of their clustering results is depicted in Fig. 3.3.

Alsmadi [Als18] proposed a hybrid Fuzzy C-Means (FCM) clustering approach for jaw lesions segmentation in dental panoramic x-ray images. The approach combines Neutrosophic sets with Fuzzy C-Means clustering. Fuzzy logic typically has problems handling indeterminate conditions well, and Neutrosophy offers a method to present the indeterminacy of pixel values by defining a neutrosophic domain. The Neutrosophic set then consists of three subsets: true, false, and indeterminate and the membership degree of each pixel for the three categories is determined. The membership degree to the Neutrosophic set is then utilized during Fuzzy C-Means clustering. The approach uses a preprocessing step to remove speckle noise via a median filter that is applied to the panoramic image. The segmentation is then performed by the proposed Neutrosophic sets with Fuzzy C-Mean clustering (NFCM) approach which performs fuzzy c-means clustering on an image transformed into the neutrosophic domain, where every pixel is represented by three properties: the probability of belonging to the foreground, the background or neither of the two. The author highlights the ability to segment lesions even in low-contrast conditions and the possibility to greatly reduce noise without blurring object boundaries. A similar approach for panoramic radiographs segmentation was proposed by Ali et al. [ASKT18] which utilized the neutrosophic orthogonal matrix with fuzzy clustering.

The simplicity of purely pixel-based methods leads to unsatisfactory segmentation results. Even when being combined with other pre- and/or post-processing techniques, these methods can only poorly handle the challenges caused by the low contrast between object classes or the intensity and illumination variation in dental radiographs. They also do not provide any kind of teeth identification or labeling as the resulting segments are not tooth specific. Nonetheless, pixel-based techniques are still used in many segmentation approaches as part of the image pre-processing, e.g. to roughly separate foreground and background.

3.1.2. Edge-based

Edge-based techniques also only rely on image information for the segmentation process but search for discontinuities (edges) instead of similarities. For this, these techniques also take the neighborhood around a pixel into consideration during the segmentation process. Typically, edge detection is performed by searching for local maxima of gradient direction with high gradient magnitudes or zero-crossings in second-order derivatives. As noise in the image can easily affect the edge detection results, a smoothing filter is usually applied as a pre-processing step. A popular method is the Canny-Edge-Detector [Can86] which utilizes the image gradient in four directions, non-maximum suppression and the removal of weak edges. For segmentation purposes, post-processing of the edge detection result is normally required to extract individual objects.

Razali et al. [RAH*14] evaluate edge-detection techniques in the context of dental age assessment. The authors propose an automatic tooth segmentation by means of extracting the tooth boundaries. Therefore, they compare the performance of the Sobel- and Canny-Edge detection methods. All input images are pre-processed with an Gaussian filter to reduce the amount of noise present in the images and improve the subsequent edge detection. Afterwards, the Sobel and Canny algorithms are applied to the pre-processed image to generate the two output images. The authors do not provide any quantitative evaluation but conclude that Sobel-Edge detection performs better compared to Canny-Edge detection. Based on the depicted results, it seems evident that edge detection alone is insufficient for separating teeth from the surrounding tissue. Especially the root area is critical as no detected edges are present in the result images for these regions.

A new approach for edge detection in orthopantanograms based on statistical measures was propsed by Gráfová et al. [GKK*13]. The fully automatic approach dynamically determines the optimal set of parameters for a user-provided conventional edge detector. Predefined parameter configurations are applied for a given input image to generate the corresponding edge detection results. Then, correspondence maps with correspondence level $1 \leq i \leq N$ are calculated – a correspondence map includes the pixel locations which where classified as an edge i or more times. The overall confidence for each map is then estimated via statistical measures (true-positive-, false-positive-, true-negative- and false-negative-probabilities) by combining the computed measures of the correspondence maps to all original edge detection results. The map with the highest confidence is selected and again compared to all original edge detection results to determine the best match, which gives the final detection result. The approach was evaluated for the most common tasks in dentistry, one of which was teeth extraction.

The authors conclude that the proposed methods demonstrated a clear advantage over conventional edge detection methods (with fixed parameters), especially in the root region. The presented results however still contain many disconnected edges at object boundaries and edges of non-teeth structures.

Similar to pixel-based methods, edge-based methods also provide undesirable segmentation results. Low contrast between structures and blurry edge in dental radiographs hamper the detection of features in certain regions. Additional information is required to facilitate a proper contour extraction in these cases.

3.1.3. Region-based

Segmentation techniques which are region-based again consider the similarity between individual pixels while also taking into account neighboring pixels. Commonly, pixels are grouped with their neighbors if a criterion for similarity between the two is fulfilled. Seeded region-growing methods iteratively extend segments from a set of provided seed-points. This guarantees a connected segment but the segmentation result highly depends on the position and number of the provided seed points, especially in the presence of noise.

Modi and Desai [MD11] identify regions of interest for tooth isolation in dental x-ray images (bitewing). They apply region growing to extract the regions containing the gap valley (area between upper and lower jaws) and the area between adjacent teeth. Afterwards, they use Canny-Edge-Detection to extract the edges of the binary image and compute the integrated edge intensity curves to identify the region of interest for tooth isolation. The results are multiple rectangular regions that each encapsulate a single tooth. No further processing is performed to extract the actual tooth contour or identify the individual teeth.

Li et al. [LSSL12] propose a watershed-based segmentation algorithm for teeth separation in dental bitewing images. The authors first improve the contrast of the original image using top-hat and bottom-hat transformations before applying thresholding to generate a binary image. Morphological operations, like erosion and hole-filling, are then used to improve the quality of the binary image. Finally, a distance transform is applied to the binary image such that each pixel contains the distance to the closest background (black) pixel and watershed segmentation [BM93] is used to retrieve the individual tooth segments.

Region-based methods rely on similarity criteria between neighboring pixels which is problematic in regions with low contrast. The presented approaches in this category only utilize bitewing radiographs with do not include the typically low contrast root region. Moreover, only teeth separation has been performed instead of segmentation. From the presented results, region-based methods seems unable to distinguish between teeth and neighboring soft tissue with makes them unsuitable for segmentation.

3.1.4. Local shape knowledge

The previously discussed segmentation techniques were purely signal-based, meaning that they only relied on the similarity or discontinuity between pixel values. Segmentation techniques using local shape information go beyond that and include additional top-down information about the expected shape of the object of interest. In order to integrate this information, the segmentation problem is typically reformulated as an energy minimization problem. The overall energy $E(S)$ of the segmentation S is given by the energy term

$$E(S) = (1-\lambda) \cdot E_{\text{image}}(I,S) + \lambda \cdot E_{\text{shape}}(S) \tag{3.1}$$

and the optimal segmentation is the one that minimizes $E(S)$. Here, $E_{\text{image}}(I,S)$ describes how good the segmentation S fits the image I based on the desired image features. For example, in the case of gradient features, how well the current segmentation matches the local image gradients computed from the image. The term $E_{\text{shape}}(S)$ on the other hand describes how well the segmentation matches the expected local shape, i.e. it can penalize jagged or improbable contours. λ is used to balance the influence of both terms. Typically, approaches using local shape information are dependent on a good initial estimate of the segmentation as the local optimization techniques used to find an optimal segmentation can get stuck in local minima. The segmentation S can be parameterized in different ways, examples are active contour or snake approaches [KWT88] or the level set method [MSV95].

A dental classification and numbering system for dental bitewing radiographs was proposed by Lin et al. [LLH10]. The approach is composed of multiple steps which include image enhancement, teeth separation, contour extraction, and tooth classification and numbering. The image enhancement is designed to reduce an uneven exposure via homomorphic filtering and to improve the contrast in between structures by active contrast stretching based on the homogeneity of image intensity and texture information. Additionally, adaptive morphological transformation is used to accentuate certain textures. Afterwards, iterative thresholding and integral projection in horizontal and vertical direction are applied to separate both jaws as well as isolate regions of interest around each individual tooth. The contour extraction is then preformed on each region of interest by applying an edge operator, sampling an equal number of contours points and applying B-spline fitting. Having extracted the teeth contours, each tooth is classified as either molar or premolar via a support vector machine (SVM) utilizing the relative width/length ratio of the tooth, the tooth pulp and the tooth crown. The final numbering is then performed by comparing formatted sequences (based on the teeth classes) to standard pattern for best matches. The presented approach provides adequate segmentation results for the task of classification and numbering, however, tooth root regions are almost never included in the processed bitewing images. The numbering system can sometimes handle missing teeth if the gap caused by the missing tooth is large enough to be manifest in the integral projection.

Rad et al. [RRN13] propose an approach for teeth segmentation and feature extraction based on the level set method. A pre-processing step is applied to the dental radiographs to reduce the noise present in the image and increase the contrast of structures of interest. Afterwards, the standard level set method is applied to extract the separation line between teeth and surrounding tissue. Even though the authors state that level set on dental radiographs is sensitive to the placement and size of the initial curves, it

Figure 3.4.: Approach by Rad et al. [RRN14]. (from left to right) Vertical integral projection is used to compute the tooth separating lines. The level set methods is then applied to extract the tooth segmentation.

is not mentioned how this problem was solved in this approach. Having extracted the semantic teeth segmentation, a gray-level co-occurrence matrix is utilized to extract additional texture features of the teeth like contrast, homogeneity or entropy. The authors conclude that level set is a promising method for teeth segmentation but requires further improvements.

In a subsequent publication, Rad et al. [RRN14] use Otsu thresholding, integral projection and the level set method to segment individual teeth in periapical dental radiographs. The processing pipeline starts with Otsu thresholding to separate the background region from the dental structural- and teeth region (i.e. separate low intensity regions from mid and high intensity regions). Then, morphological operations are applied to the thresholded image to remove unwanted objects based on geometric properties. Next, integral projection in horizontal and vertical direction is used to separate upper and lower jaw as well as individual teeth. Integral projection accumulates all pixel values in a row or column of the pixel grid. Separating lines can be found by searching for rows or columns with a local minimum, see Fig. 3.4. The level set method is then applied to each of the extracted tooth regions to get the final segmentation of each tooth. Since each region only contains a single tooth, the initial contour is given by a predefined rectangular box. The proposed method was evaluated on 32 digital periapical dental radiographs. While the authors state accurate segmentation results, the level set method fails to reliably detect the root boundary of the tooth.

Hasan et al. [HIHY16] propose to use the gradient information to segment the jaw in panoramic radiographs. The gradient information is exploited using gradient vector flow (GVF) snakes, which are an extension to traditional snakes or active contours. Snakes or active contours operate by minimizing an energy functional which depends on the curvature of the contour, the intensity values of the image and a high-level constraint energy. In case of GVF snakes, this constraint energy is given by a gradient vector flow field. The proposed algorithm is composed of four steps. First, k-means clustering with 9 clusters and subsequent thresholding based on the cluster size is used to create an initial separation of jaw and background. Next, key points around the initial separation are detected in order to initialize the gradient vector flow snake. The GVF snake is then applied to the original image to segment the jaw region based on gradient information. Since the GVF snakes often fail to include the wisdom teeth, the authors perform a post-processing step by placing a larger ellipse around the original segmentation.

The size of the ellipse depends on the width and height of the original segmentation. The authors evaluated their approach on 284 images, of which 92 % were segmented successfully.

In contrast to signal-based techniques, the usage of local shape information enables a successful segmentation even in local regions with low contrast as neighboring regions with sufficient contrast (and therefore strong image features) help to restrict the possible contour shape on low contrast regions based on the shape energy. However, an explicit modeling of the possible contour shapes also results in approaches which are more specialized to a specific problem.

3.1.5. Global shape knowledge

The next level in integrating prior knowledge into the segmentation is the utilization of a reference shape or a space of possible shapes. This way, possible segmentations can be restricted to valid shapes, i.e. shapes similar to the reference shapes. To achieve this, the shape energy $E_{\text{shape}}(S)$ (cf. Eq. (3.1)) is modified to measure the similarity between the current segmentation and the reference shapes. Approaches with global shape knowledge are useful if a lot of prior knowledge about the shape of the object of interest exists and the variance in between possible shapes is not too high. They are especially useful if the object is not clearly distinguishable from the background. However, since explicit knowledge about a single object is utilized in the segmentation approach, it is highly tuned to the specific application. Examples of such methods are Deformable Models [TF88], Active Shape Models (ASMs) [CTCG92, CTCG95] and Active Appearance Models (AAMs) [CET01].

In case of Active Shape Models (ASMs) [CTCG95], the shape c of an object is represented as a linear combination

$$c = \bar{c} + \sum_{i=1}^{m} (\vec{v}_i e_i) + \vec{r}, \tag{3.2}$$

where $\bar{c}$ is the mean shape computed from all n training instances c_j, $\vec{v}_i$ are the eigenvectors of the covariance matrix

$$A = \frac{1}{n-1} \sum_{i=1}^{n} (c_i - \bar{c})(c_i - \bar{c})^T, \tag{3.3}$$

and e_i are the shape coefficients representing the current shape. Vector $\vec{r}$ is the residual. Typically, Principal Component Analysis (PCA) is applied to reduce the dimension of vector $\vec{e} = \{e_1, e_2, \ldots, e_m\}$. The similarity between an arbitrary shape and the shape space formed by the training shapes c_j can then be measured by $|\vec{e}|$ and $|\vec{r}|$, the smaller the vectors norms are, the closer the shape is to the shape space.

Lira et al. [LGN09] have used shape models, quadtree decomposition and morphological operators to extract features from dental images. Afterwards, snake models are used to determine the individual teeth contours. Initially, low pass filtering is applied to remove noise from the original images and quadtree decomposition as well as global thresholding are each employed to generate separate masks. Overlapping regions are then removed from the thresholding mask using morphological operations and the aforementioned quadtree mask. Teeth candidates are selected based on the size of the resulting

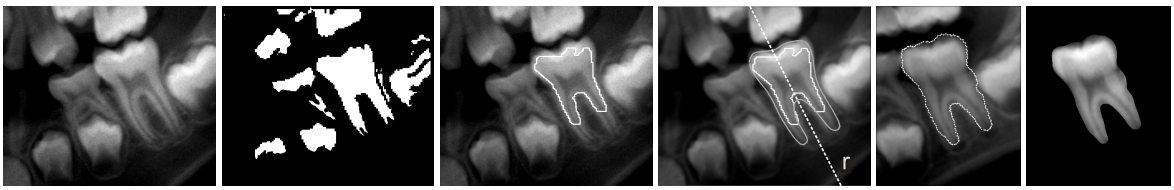


Figure 3.5.: Illustration of the approach from Lira et al. [LGN09]: a) original image, b) mask images after overlap removal, c) tooth candidate boundary, d) optimally matched mean shape model, e) tooth contour detected by snake model, f) final segmentation result.

regions, resulting in isolated but very rough tooth shapes. Trained statistical shape models based on PCA are then utilized to compute an optimal match between the resampled tooth contour and a mean shape model. The contour of the matching mean model serves as the initialization for a snake model which extracts the final tooth contour from the original intensity image. The approach is illustrated in Fig. 3.5.

Despite not being used much in the dental segmentation, active shape and appearance models have been applied successfully to radiographic images in other medical areas. Wu et al. [WM16] used an active shape model-based segmentation method for robust and accurate segmentation of the distal femur and proximal tibia in knee radiographs, while other authors used them for liver segmentation in low-contrast CT volumes [EF17] or prostate ultrasound image segmentation [BJT*20]. Bones share some of the properties of teeth, like being of denser non-deformable material and possessing a well known shape which makes teeth good candidates for statistical shape modeling. Furthermore, ASM offer potential to handle some of the challenges involved in teeth segmentation like low contrast, noise and the generally low image quality. However, the initialization dependency might be critical as automatic initialization of 32 individual teeth shape model seems challenging.

3.1.6. Spatial relations

While global shape knowledge can offer significant assistance when segmenting isolated objects, the incorporated shape knowledge might not be sufficient to segment objects if multiple similar objects are present in one image. In order to handle such situations, additional prior knowledge in form of spatial information can be utilized. In doing so, the global position of the object or the relative location of the object to other objects can be encoded.

Coupled models [TWT*03, TWT*04] combine multiple shape-based deformable active contour models using signed distance functions in order to segment multiple objects while considering their spatial relation. Atlas-based segmentation methods [CE97] use reference images with known object segmentations referred to as atlases. The shape and spatial relation of the included objects is implicitly encoded within the atlas. These reference images are registered to the input image using image-based registra-

tion techniques and the best matching atlas is selected. Having determined the transformation between best matching atlas and input image, the original segmentation can be transferred to the new image.

These models have proven useful in a variety of scenarios, like the segmentation of 22 risk structures in the head & neck area in CT images for radiation therapy planning [JSK*14], automatic localization of the sentinel lymph node and additional active nodes on SPECT-CT data [JMDW17], and automated bone segmentation and anatomical landmark localization in Cone-Beam CT [BIV*19].

While coupled models eliminate the need of initializing 32 individual shape models and offer invaluable information for the numbering of teeth due to the inherently encoded spatial relation, no approach for teeth segmentation (or numbering) has utilized this technique so far.

3.1.7. Neural networks

Convolutional Neural Networks (CNNs) have become incredibly popular over the last decade and are currently considered state-of-the-art in many (medical) image segmentation tasks [SSD20, MBP*21]. While these networks have existed for a long time, the lack of large amounts of training data and size constraints imposed by limited computing power have hindered the success of convolutional neural networks in the past. Convolutional neural networks are techniques inspired by the human visual system and consist of multiple layers of artificial neurons (feature maps). The size and depth of each feature map depends on the chosen network architecture and task to be solved. In fully connected layers, each output of one feature map has weighted connections to all input (activation units) of the succeeding feature map. This can result in millions or billions of parameters which need to be optimized during the training process. Neural networks are therefore typically dependent on a large amount of training data. Because of the large amount of trainable parameters, this taxonomy considers neural networks to include more prior knowledge than the previous categories. However, neural networks are considered black-boxes and the learned knowledge is not explicitly modeled as shape or spatial relation.

Techniques to reduce the amount of required training data include data augmentation and transfer learning. With data augmentation, the training data set is artificially enlarged by applying small deformations (augmentations), like rotation, scale, shift, reflection, etc., to the original training data in order to increase the invariance and robustness properties of the network. Transfer-learning implies that instead of a random initialization of the network parameters, parameter sets from networks already trained for similar applications are used instead to provide a better starting configuration for optimizing the network parameters.

Ronneberger et al. [WHL*16] employed a convolutional neural network that uses an encoder-decoder architecture for fully automated segmentation of dental X-ray images (bitewing radiographs). Their network is based upon the fully convolutional network (FCN) idea [LSD15] which extends the typical contracting network design by adding additional up-sampling layers to increase the resolution of the output. The general idea behind such an architecture is that the contracting path captures context information and high-resolution features while the expanding path enables a dense pixel-wise prediction by combining the information for the contracting path with the up-sampled output. Ronneberger et al.

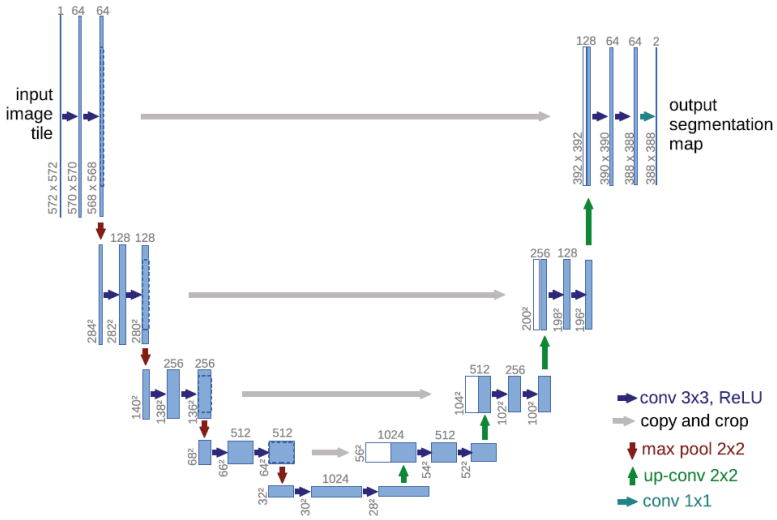


Figure 3.6.: U-Net architecture. The u-shape network structure of encoding path (left side) and decoding path (right side). Blue boxes represent multi-channel feature maps with size and channel number denoted on each box. This architecture is used frequently in medical image segmentation due to is good performance on semantic segmentation. (image from Ronnerberger et al. [RFB15]).

created a u-shaped network architecture with an expansive path that is more or less symmetric to the contracting path. Because of this, they named their network architecture 'U-Net' [RFB15]. An illustration of the network architecture is provided in Fig. 3.6. The contracting part of their network employs the typical CNN design of repeated application of 3x3 convolutions, a rectified linear unit (ReLU) and 2x2 max-pooling operations for downsampling. The number of feature channels is doubled at each down-sampling step. The expanding part utilizes repeated 3x3 convolutions, followed by a rectified linear unit (ReLU) and 4x4 upsampling convolutions which halves the number of features channels. It is followed by a concatenation with the corresponding feature map of the contracting path (skip connection) to integrate knowledge captured in these earlier layers (as well as tackle the vanishing gradient problem of deep neural networks). At the final layer, repeated convolutions (3x3, 1x1) are used for the pixel-wise prediction of the class labels. In order to cope with the limited amount of training data, the authors make excessive use of data augmentation in form of random image deformations. The U-Net achieves an average f-score of 0.567 for segmenting seven tooth structures (e.g. enamel, dentin, crown, caries) in bitewing radiographs. The segmentation results on X-ray images look promising as important structures like enamel, dentin and pulp have f-score values > 0.7 despite the low amount of available training data, which encourages the usage for other types of (dental) radiographs. However, the network only provides semantic segmentation results without an explicit separation of object instances and no numbering of individual teeth.

Silva et al. [SOP18] performed semantic teeth segmentation by applying a deep network based on mask region-based convolutional neural network (Mask R-CNN). The Mask R-CNN architecture [HGDG17] has three output branches that predict the classification (class score), localization (regressing bounding box) and mask (per-pixel segmentation) for each object candidate. It is an extension of the Faster R-CNN architecture [RHGS17] which did not include the mask prediction. In general, R-CNN networks work by first performing a feature extraction using a backbone network. Based on these features, a feature pyramid network (FPN) is constructed and regions of interest (RoI) are predicted using a region proposal network (RPN). Finally, these regions of interest are aligned to have equal size and predictions for each output branch are computed by different networks. The authors do not state any modification of the original network's structure. The network was trained on 753 annotated panoramic radiographs with an additional 452 images used for validation. Since the goal was to perform semantic segmentation, the ground truth annotation is not separated into individual teeth but is instead comprised of a single binary mask covering all teeth. The authors report an F-score of 0.79, but also state that the network failed to correctly determine the boundary of the teeth on all test images. This is explained as the result of using initial weights transferred from imagenet which was trained for person segmentation. Due to the low amount of training images, the network was unable to learn the proper features of teeth edges and borders.

Jader et al. [JFR*18] extend the work of Silva et al. [SOP18] and propose Mask R-CNN for object instance segmentation of teeth in dental panoramic X-ray images. The authors, again, use the original Mask R-CNN network architecture and do not state any modifications. In this work, the network was trained on only 193 annotated panoramic radiographs. 83 images were used for validation. However, to facilitate the new task of instance segmentation (compared to the previous task of semantic segmentation), the authors modified their image annotations such that all teeth are separate from each other. The network was trained using transfer learning with pre-trained weights taken from the MSCOCO data set [LMB*14]. Due to the very large amount of free parameters in Mask R-CNN, a training from scratch was impossible with this amount of training data. The hyperparameters (e.g., learning rate, number of epochs) were defined empirically on training experiments with validation data. The authors evaluated their approach on 1224 images, which were separated into 10 categories of varying difficulty. The authors conclude that their approach demonstrated promising results with an F1-score of 0.88 and provides highly superior results compared to unsupervised methods. Exemplary results are depicted in Fig. 3.7. While this approach provides object-level segmentations of the teeth, no labeling or numbering is performed to specifically identify individual teeth.

Koch et al. [KPIB19] use an FCN based on the U-Net architecture [RFB15] for semantic segmentation of panoramic radiographs. Instead of resizing the original input data (1127x1991 pixels) to a size the network can handle – which would cause a loss of information – the authors propose to apply the network to a grid of fixed-sized sub-regions (patches) of the original image. Therefore, they propose a different training approach by replacing the dropout applied before every max pooling operation during training with batch normalization [IS15]. Additionally, they used the Nadam optimizer and all weights of the network were initialized using the uniform Xavier initialization. The network was trained using 1201 panoramic radiographic images with ground truth segmentations from the Silva

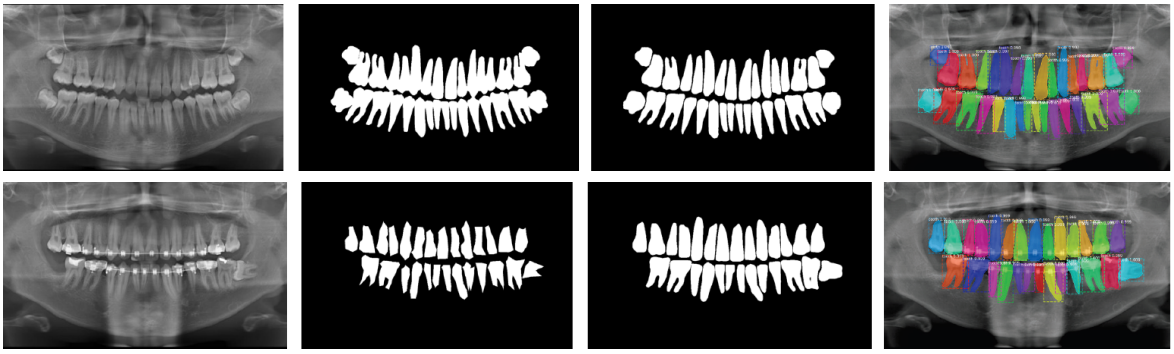


Figure 3.7.: Semantic teeth segmentation with neural networks. The first two images show the original image and the ground truth segmentation, respectively. The semantic segmentation mask (third image) is created by combining all instance segmentations of individual teeth. Mask R-CNN is used to predict the individual teeth instances and their segmentation (right picture). No teeth numbering is performed here. (images from Jader et al. [JFR*18]).

et al. [SOP18] data set, the remaining 299 images were used for testing. For the training process, the patches (512x512 pixels) were randomly selected (as a means of data augmentation) in order to increase variability in the patch standardization and batch normalization, and to reduce overfitting. The authors trained six different configurations of their U-Net variant to evaluate the effects of a varying number of classes, bootstrapping, and data augmentation by additional horizontal reflections. The achieved semantic segmentation results are between 0.9277 and 0.9342 dice score for the worst and best configuration, respectively. The best configuration uses 3 classes (background, roi and teeth), horizontal reflections for augmentation and bootstrapping. In addition to the individual network configurations, the authors also evaluated the effects of using a model ensemble. The prediction results of selected network configurations were averaged before computing the final segmentation results. The ensemble of the four configurations with the best individual segmentation results in a slight overall improvement of the semantic segmentation performance (0.9363 dice score). Again, no object-level segmentation or numbering of single teeth is performed.

Nishitani et al. [NNH*21] propose a U-Net teeth segmentation using a loss function that is weighted on the tooth edges. The approach first utilizes a region prediction network (Faster R-CNN [RHGS17]) to remove some background area from the panoramic radiographs in order to normalize the teeth size before applying the U-Net architecture. The segmentation step uses the original encoder-decoder architecture of U-Net and modifies the loss function used for training. The typical cross entropy (CE) loss used in combination with U-Net computes the entropy over the entire image, which prevents the network from learning the specific features of the tooth boundary in low contrast regions. The authors approach this problem by adding a second term to cross entropy loss function. This term specifically focuses on the tooth boundary by not computing the cross entropy over the entire image but instead only over the boundary pixels of all teeth. Therefore, an edge region with a specific width was defined

on the tooth contour line extracted from the annotated label images. The relative weight of both terms of the loss function was determined during testing. The best results were achieved with a weight of 0.6 for the CE of the entire image and 0.4 for the tooth edge. The semantic segmentation accuracy improved by 0.037 to reach a Dice of 0.927 when using the new loss function over the standard one. This is mostly due to an improved accuracy for the tooth root region. It is notable here that the segmentation accuracy also improved slightly when only the tooth boundary CE was used for learning which indicates the importance of the tooth boundary relative to the entire tooth.

Recently, Chen et al. [CZL*21] proposed a multi-scale location perception network (MSLPNet) for dental panoramic image segmentation. The authors specifically focus on enhancing fuzzy segmentation boundaries, especially in the root area. The MSLPNet extends the idea of multi-scale spatial pyramid pooling (SPP) and uses aggregation modules on each scale (branch), which enables high-level semantic features to be used as additional information on lower scales. A location perception modul (LPM) applied on each branch after the aggregation module is designed to prevent misclassification of individual pixels. After upsampling of the individual branch, all feature maps are concatenated before predicting the final result. The network is trained using a multi-scale structural similarity loss (MS-SSIM) as a patch level loss. This loss considers structural similarity information in a local neighborhood which puts a higher emphasis on the object boundary in different scales. It is combined with Bce and Dice loss to optimize the performance of the network. The network was trained on 1200 annotated images which were rescaled to 512x1024 pixel, with an additional 150 images used for validation. The authors report semantic segmentation results with a Dice score (f-score) of 0.9391 on 150 test images and improved boundary accuracy based on Pratt's figure of merit (PFOM) metric.

There are several other approaches that apply convolution neural networks with minimal modification for semantic segmentation in dental radiographs, like Sivagami et al. [SCKM20] (U-Net). More information can be found in a recent review of deep learning in the field of dentistry conducted by Hwang et al. [HJCH19].

Convolutional neural networks have contributed a significant improvement to segmentation accuracy in dental radiographic images. With sufficient training data, highly tuned networks are able to cope well with the challenges of these images like poor image quality and low contrast as well as missing teeth and artifacts. While some networks only focus on semantic segmentation (U-Net), other architectures like R-CNN also support instance segmentation which is required for any medical imaging application that intends to provide assistance to medical personal.

None of the purely CNN-based methods includes a teeth numbering to uniquely identify individual teeth. This might be due to the fact that identifying teeth based on their shape and appearance alone is very difficult and it is challenging to incorporate the explicit knowledge about the spatial relations into the network.

3.1.8. Hybrid CNN methods

This category lists approaches that combine neural networks with additional post-processing in order to enhance the segmentation result or extract additional information.

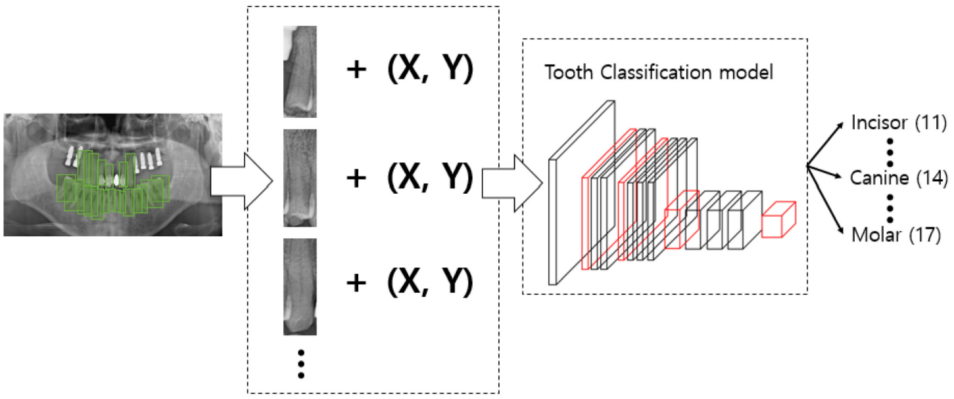


Figure 3.8.: Tooth numbering approach based on spatial relation (positional values). Teeth locations based on detected teeth instances (via R-CNN) are used to classify and number the teeth according to the FDI notation system. (image from Kim et al. [KKJ*20]).

Chen et al. [CZL*19] perform teeth detection and numbering in periapical radiographs. They employ a Faster R-CNN region proposal network and propose multiple post-processing techniques. These techniques utilize prior domain knowledge to enhance the outputs generated by the network and include a filter algorithm to eliminate overlapping regions generated by the network, a secondary CNN to identify missing teeth, and a module which imposes certain constraints on the detected tooth numberings. The Faster R-CNN network was trained on 800 images to detect all 32 individual teeth as separate classes. The validation loss was computed as the mean average precision (mAP) which incorporates the average precision for each objects class. The precision of each object class is given by the number of correctly detected bounding boxes (TP), incorrectly detected boxes (FP) and not detected boxes (FN). An intersection over union value of 0.5 was used as a threshold for correct detection. The annotation results of the network are then refined using non-maximum suppression in order to remove overlapping boxes with the same class label. Overlapping boxes with different labels are removed in a separate step. Since precision and recall numbers were still unsatisfactory after the aforementioned steps, arrangement rules were applied to the detected objects to identify missclassifications. The rules exploit the domain knowledge about the arrangement (spatial relation) of teeth in the jaw to compute a comparison score and modify any class label that is determined to be incorrect. Since missing teeth are not detected by the R-CNN network, these could impact the results of applying the arrangement rules. The authors therefore train an additional CNN which detects missing teeth based on the distance between the filtered object centers. The system achieves good results for bounding box detection and tooth numbering. The most common missclassifications were caused by confusing teeth label within the same type of teeth (e.g. incisors) or incorrectly assigning the label of the opposite side (i.e. left or right side).

The approach proposed by Kim et al. [KKJ*20] combines R-CNN and Heuristic algorithms to perform a tooth detection and numbering in panoramic X-ray images. The approach explicitly includes

tooth implants to be detected and numbered as well. The detection is performed by placing a bounding box around the desired target object – an explicit extraction of the objects contour or boundary is not part of the reported results, even-though it appears to be used for tooth numbering. The overall approach is separated into two parts: Dental object detection and tooth numbering. The Faster R-CNN architecture [RHGS17] is employed in this approach to determine the regions of interest for the three object instances of the classes in panoramic radiographs: (natural) teeth, crown implants (artificial crowns, bridges) and implant fixtures. The authors utilized a Faster R-CNN inception model which reduces the required number of computations while increasing speed and accuracy. The network was trained on 253 images which included 6446 teeth, 402 implants with fixtures and 205 crown implants. After having obtained the dental object instances for all three classes as well as their locations through the R-CNN, the tooth classification and numbering is performed. Therefore, an additional R-CNN/CNN was trained which takes the combined positional values of all teeth as input to determine the individual tooth number via classification. In doing so, the authors exploit the relative location of individual teeth, which is encoded in the combined positional values, to determine the final tooth number. Tooth shape was only used to deduce the tooth class (incisor, canine or molar). Fig. 3.8 depicts the tooth numbering approach based on positional values. The reported results include detection accuracies of 0.26 and 0.41 for implant fixtures and crowns, respectively, for an intersection over union threshold of 0.7 for positive detection. Tooth detection achieved an accuracy of 0.75 for the same criteria, while tooth numbering accuracy was 0.77

The approaches in this category all aim to perform a teeth numbering while also employing the convincing segmentation performance of neural networks. It is important to note that all approaches exploit the spatial relation between the teeth in order to achieve a teeth numbering. While Chen et al. [CZL*19] try to predict the numbering directly using the neural network, and refine incorrect prediction afterwards using arrangement rules build upon the spatial relations, Kim et al. [KKJ*20] utilize a secondary network to analyze the combined positional values of the previous teeth segmentations. Again, these combined positional values inherently include the spatial relation between individual teeth.

3.1.9. Discussion & Conclusion

This section introduced the current state of the art in dental image segmentation and tooth numbering. While many segmentation approaches have been proposed, very few include a numbering of the generated segmentation of the teeth. The fraction of those that work on panoramic X-ray images is even lower.

The focus of the first application area of this thesis is the instance segmentation and numbering of teeth in panoramic radiographs. Extraoral panoramic radiographs are generally considered the most challenging dental image type for teeth segmentation. A study of the literature on segmentation methods applied in dental imaging from 2018 [SOP18] revealed that only 20% of the proposed approaches in dental image segmentation focus on extraoral images. The majority used intraoral images (bitewing, periapical) which only depict a few teeth and do not contain much of the surrounding tissues (cf.Section 2.3.3). In addition to challenges like bad image quality and low contrast between the teeth

area and dental structural area, panoramic radiographs – unlike intraoral radiographs – capture the entire teeth area and contain all teeth but also include large parts of the surrounding structures.

As a result, purely image-based (or signal-based) methods, like pixel-, edge- or region-based methods, struggle to accurately separate foreground and background. While useful results might be achieved by signal-based methods in intraoral images because of the very restricted region that is captured, this is not the case for extraoral radiographs. Extraction of teeth boundaries between the teeth- and dental structural area is especially inaccurate due to the low contrast in these regions. To better cope with these challenges, segmentation methods have benefited from the inclusion of prior domain knowledge into the segmentation process – either as local or global shape knowledge. This enables a more specialized processing of the teeth area in panoramic radiographs compared to the background and dental structural area.

Another challenge is that while many methods for segmentation of dental radiographs have been proposed, a large part focuses on a semantic segmentation of the teeth area. In these methods, all tooth instances are treated as a single segment, i.e. only a pixel-wise classification as teeth or non-teeth is performed, which does not enable a direct numbering of the individual teeth as a delineation of the boundary between neighboring teeth is not performed. Numbering of the teeth requires extraction of the individual tooth instances and therefore relies on instance segmentation.

Moreover, all proposed methods which perform a tooth numbering also exploit the domain knowledge about the spatial arrangement of teeth. The fact that all 32 teeth only belong to 4 different teeth classes (incisor, canine, pre-molar and molar), and that teeth of the same class are also located adjacent to each other, makes it nearly impossible to assign labels based on local image or shape information alone. In order to perform teeth numbering, the spatial relation to other teeth needs to be known. Consequently, the combination of domain knowledge about the shape of individual teeth and their spatial relation to neighboring teeth seems to be required for a successful segmentation and numbering process. To this end, this thesis proposes a teeth segmentation and numbering method which includes domain knowledge about both shape and spatial relation directly into the segmentation process.

The second application area of the thesis includes a tooth segmentation and numbering in colored photographs. While the added color information generally improves the contrast between individual objects and therefore alleviates some of the segmentation challenge, a variability in viewing direction and potential reflections on the surface of the teeth add additional challenges. The challenges for numbering the teeth remain unchanged. Therefore, the same segmentation and numbering concept is applicable to this scenario as well.

3.2. Image-based 3-D teeth reconstruction

The goal of 3-D reconstruction is to infer the structure and geometry of a single or multiple real-world 3-D objects in order to generate a digital representation. It is a generally difficult scientific problem and technique that is used in many different fields, such as medical imaging [WCKE18, ACR*18], virtual reality [HLKH17], cultural heritage [RC19], personalized entertainment [ZLNW18], and many more.

Approaches to 3-D reconstruction can be roughly divided into range-based (active) and image-based (passive) methods.

Range-based methods [SBVS20] directly interfere with the environment in order to generate a depth map, i.e. an image that contains information about the distance of objects from the current viewpoint. By combining the depth information from multiple (known or tracked) viewpoints, the 3-D surface can be reconstructed. Methods in this category utilize structured light or time of flight (ToF) approaches to estimate the depth information. While structured light refers to the process of projecting a known light pattern onto a scene and inferring the depth information by analyzing the resulting appearance of the pattern in the scene, time of flight recovers the depth information by measuring how much time passed between projecting light onto the scene and being able to measure its reflection [SBVS20].

Image-based methods on the other hand utilize image understanding to retrieve information that is used to reconstruct the object's shape. They have to handle the inverse problem of trying to recover the dimension that is lost during the 3-D to 2-D projection process when capturing 2-D images of 3-D objects. Image-based reconstruction is considered ill-posed since many different geometries can result in the same 2-D projection. A recent survey by Han et al. [HLB21] covers different approaches to the problem and recent advances. A typical technique in image-based reconstruction is to use local feature points to establish correspondences between images in order to recover the lost depth information. Multiview stereo utilizes stereo vision, i.e. two images captured from sightly different angles with a known camera setup, to reconstruct the 3-D coordinates of matched feature points in both images using a triangulation principle. Structure-from-motion (SfM) techniques rely on matched feature points across multiple images to compute feature trajectories which are then used to estimate the camera motion and 3-D coordinates of these points. Both approaches rely on robust feature descriptors like traditional SIFT [Low04], SURF [BTVG06] or ORB [RRKB11] to identify the same feature points in multiple images and enable a reliable matching. More recently, deep learning-based descriptors, like L2-Net [TFW17], have also been used. Performance evaluations of local feature descriptors for image-based reconstruction have been conducted, e.g. Fan et al. [FKW*19]. Shape-from-shading (SfS) techniques assume that the location of an illumination source and the surface reflectivity of the object are known and estimate the depth information by analyzing the shading observed in an image. Shape-from-silhouette techniques are inspired by space carving [Lau94] and use a volume intersection approach to constrain the space occupied by the object. By intersecting the viewpoint-specific 3-D cones defined by the back-projected silhouettes, the object's shape is reconstructed.

Digital 3-D models of a patient's set of teeth are an important asset in dentistry for the planing of surgical procedures or the design of patient-specific treatment plans. These digital models are typically created from dental impressions, i.e. a negative imprint of the teeth, from which a (positive) model or cast can be created which is then digitalized using a laser scanner. Alternatively, they can also be computed from 3-D computed tomography (CT) scans of the patient. Both data sources are capable of delivering high-quality models, but they are typically expensive and time-consuming [WBG*16]. Moreover, in case of CT scans, they expose the patient to Röntgen radiation which is a known and proven human carcinogen (cf. Section 2.3.4). In recent years, intraoral scanners, such as the *Ture*

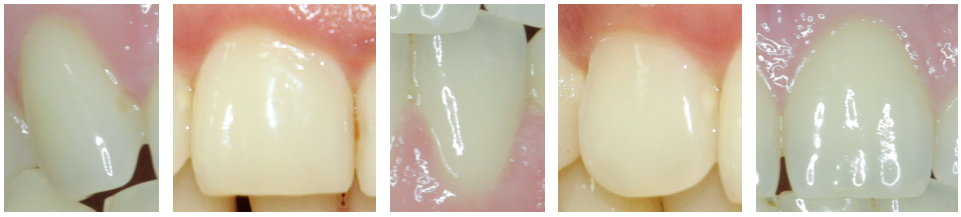


Figure 3.9.: Images for illustrating the appearance properties of teeth. The individual teeth are textureless and show no surface features while suffering from specular reflections.

Definition [1], the *iTero Element* [2] and the *TRIOS* [3], have been developed which provide another method to capture optical impressions of both soft- and hard-tissues. Most intraoral scanners utilize active reconstruction techniques based on some form of structured light scanning. Intraoral scanners can provide benefits over traditional methods like time efficiency, easier work processes for clinicians and improved patient comfort. On the other hand, these devices are highly expensive and are constantly being replaced by newer versions due to constant innovations in the field. Additionally, expert knowledge is required to efficiently operate these scanners.

Image-based reconstruction could provide a cheaper, more accessible method for generating 3-D models of the teeth. However, image-based reconstruction is very challenging, especially when trying to reconstruct teeth. Teeth possess very challenging appearance properties which makes high quality reconstructions from photographs difficult [AFSEM14, WBG*16]. Teeth are in general featureless and the material properties of teeth cause them to be highly specular (i.e. having the qualities of a mirror) while exhibiting strong subsurface scattering. Fig. 3.9 shows the appearance of several teeth in photographs, demonstrating their featureless surface and the occurring specular reflections. Additionally, almost all photographs which depict the teeth suffer from non-trivial occlusions between any number of teeth. This further limits the amount of information that can be recovered and complicates the reconstruction process.

Some approaches have been proposed over the years to tackle the aforementioned problems. In the following, the different categories of image-based reconstruction approaches for teeth are discussed and related work by other authors is reviewed. The section concludes with a discussion that summarizes the state of the art and highlights limitations of existing approaches.

3.2.1. Feature-based methods

Multiview stereo and shape-from-motion techniques which rely on corresponding feature points in multiple images have not been employed for teeth reconstruction. This can be explained because of the

[1] True Definition™ Scanner, Midmark Cooperation, https://www.midmark.com/dental/products/digital-impressions (accessed on 13.10.2021)

[2] iTero Element Scanner, Align Technologie Inc., https://emea.itero.com (accessed on 13.10.2021)

[3] TRIOS Scanner, 3Shape A/S, https://www.3shape.com/en/scanners/trios (accessed on 13.10.2021)

challenging appearance properties of teeth. Their featureless surface combined with specular reflections and subsurface scattering makes it seemingly impossible to extract sufficient robust feature points on the surface of the teeth to enable a matching between different views and therefore a reconstruction of their 3-D coordinates. This leaves only shading and silhouette information as potential sources for image-based methods to deduce the original shape.

An example of a non-image-based reconstruction method that utilizes feature points was presented by Zheng et al. [ZLS11]. They proposed a 3-D morphing technique to reconstruct the occlusal surface of a single tooth. They employ contour extraction via 3-D optimal path searching and feature point identification using a Snake model to identify matching feature points between a standard tooth model and the tooth preparation. The standard tooth surface is then deformed using a mapping function (radial basis function) to define the spatial relations between both teeth and interpolate the missing surface information.

3.2.2. Shape-from-shading

Other reconstruction techniques do not rely on the detection of specific feature points and instead use different references. Shape-from-shading techniques utilize the amount of light reflected by the surface of the object to infer their 3-D shape. Depending on the position of the light source and the orientation (and properties) of the surface, the amount of light reflected relative to the camera varies. Assuming the position of the light source and the surface reflectance model are known, the surface geometry can be estimated.

Abdelrehim et al. [AFSEM14] used a shape-from-shading (SFS) approach for 3-D teeth reconstruction. They utilized the shading from a single intraoral image with 2D-PCA shape priors (based on height maps) and an Oren-Nayar-Wolff reflectance model to reconstruct the surface of a tooth crown. Their approach reconstructs a single tooth from an image containing only that tooth and is applied only to premolar and molar teeth.

Multi-object reconstruction becomes significantly more difficult due to the presence of occlusions. In order to deal with such occlusions during the reconstruction process when reconstructing multiple teeth at once, models which incorporate prior knowledge about the tooth shape (e.g. statistical models) have also been studied. Farag et al. [FEA*13] employed statistical shape models to reconstruct the anterior part (12 teeth) of a dental arch (i.e. the arch of teeth on upper or lower jaw) from a single colored oral cavity image. They trained statistical models of texture and shape which were used in their shape-from-shading approach to model the relation between the photometric information and the underlying 3-D shape. Given an input oral cavity image and a set of anatomical landmark points, dense correspondences were established to all reference images in their aligned ensemble of oral cavity images. Based on these correspondences, the spherical harmonic projection (SHP) images which encode the illumination conditions of the input image were constructed. The inherit relation between SHP images and corresponding shape was then used to estimate the 3-D shapes of the dental arch using a partial least squares (PLS) regression model. Mostafa et al. [MEA*14] used a similar approach but constructed a combined statistical model which linked the coefficients of the texture model, the shape

model and the spherical harmonic projection images in order to estimate the 3-D shape of the dental arch.

3.2.3. Shape-from-silhouettes

Shape-from-silhouette techniques employ the silhouette, or occluding contour, of an object to gain information on the 3-D shape. The idea behind this technique is that for a given silhouette, the 3-D object is contained within the volume generated by back-projecting the silhouette according to the camera parameters of the current viewpoint. By intersecting the volumes generated from multiple viewpoints, the visual hull [Lau94] – the smallest possible volume containing the object – can be obtained. As a single silhouette defines the entire back-projected volume for that image, this type of reconstruction is heavily dependent on accurate contour segmentation in the images [HLB21]. Shape-from-silhouette approaches typically struggle with objects suffering from self-occlusions. Here, self-occlusion means that, from certain viewpoints, one part of an object is occluded by another part of the same object.

Wu et al. [WBG*16] presented a model-based teeth reconstruction approach using only extraoral photographs. They created parametric models of the teeth (one for the upper arch and one for the lower arch) in order to capture (a) the local shape variation of each tooth, (b) the pose variation of the teeth within the arch and (c) the global position and scaling of the arch. Their model was trained on 86 3-D scans obtained from high resolution plaster casts. These scans where separated into individual teeth and transformed into per-tooth meshes by fitting an artistically created tooth template mesh for each of the four teeth categories (incisor, canine, premolar and molar). The shape and pose of a single tooth τ in their parametric model Z is given by

$$Z_\tau = T\,\Phi T_\tau\,T_\tau^c\left(S_\tau^c + \sum_i^{|B_\tau|} a_\tau^i B_\tau^i\right). \tag{3.4}$$

Here, S_τ^c is the average shape of tooth τ, T_τ^c is the average pose and B_τ is the shape-subspace that encodes the shape variations. The degrees of freedom for fitting the parameterized model are the shape coefficients a_τ as well as rigid transformation parameters T_τ. Additionally, the global parameters affecting teeth are given by the rigid transformation T and anisotropic scaling along all axes Φ. The authors define three types of teeth boundaries: teeth, gums and lips, but use only teeth and gum boundaries for the reconstruction process. The boundaries are extracted from the input images using a Boosted Edge Learning (BEL) algorithm which is improved in robustness by combining it with an edge- and a smoothed orientation map. The teeth reconstruction is then formulated as a Maximum a Posteriori (MAP) estimation problem in the unknown model parameters $\mathcal{X}^*$, which consists of combined parameters of their parametric model (cf. Eq. (3.4)). The process alternates between (a) establishing correspondences between 2-D projection of the 3-D model and the extracted tooth boundaries and (b) optimizing the model parameters. Finally, they compute the color information of the teeth and incorporate the gums to produce more compelling results. Their system requires a minimum of eight extraoral photographs to perform the teeth reconstruction. In addition, a manual initialization is required by

manually identifying two teeth per dental arch (upper and lower jaw) in each input photograph in case of an uncalibrated multi-view capture. On an unknown number of test cases, their approach achieves an average Euclidean error over all non-root vertices across all teeth of 0.86 mm.

In 2019, Han et al. [HLB21] published a comprehensive survey on deep learning techniques used for image-based 3-D object reconstruction. They reviewed state-of-the-art methods and classified them into three categories: volumetric, surface-based and point-based techniques. While many different deep learning approaches exist which offer promising results for 3-D reconstruction from one or more RGB images, one main conclusion of the survey was that most techniques deal with images that contain a single object. Multi-object reconstruction in the presence of occlusion is therefore named as one of the future research directions. The authors state that approaches in this area, in general, rely on silhouette-based loss functions and, as a result, highly depend on accurate object-level segmentation.

3.2.4. Discussion & Conclusion

Image-based 3-D reconstruction of objects from individual or multiple 2-D images is an ill-posed problem that has been extensively studied. It is rarely applied in the dental domain due to the fact that teeth suffer from very challenging appearance properties which makes high quality reconstructions from photographs difficult. The material properties of teeth cause them to be specular and diffuse while exhibiting strong subsurface scattering [WBG*16]. As a result, there are almost no usable surface features available for the teeth which means that some commonly used image-based reconstruction techniques like multi-view stereo or structure-from-motion cannot be applied. This only leaves techniques utilizing the shading information or teeth outline (silhouette) as possible options. Since shape-from-shading approach rely on information about the illumination source and typically one use a single image, the most promising reconstruction technique for the use with the colored photographs of orthodontic records is shape-from-silhouette as employed by Wu et al. [WBG*16]. A limitation of the approach proposed by these authors is the dependence on user input to number the individual extracted teeth outlines in order to match the silhouettes between different images. Due to the limited amount of information available from the five photographs used in this thesis for reconstruction and the non-trivial occlusions present between teeth, a deformation-based approach that also uses a statistical model to constrain the reconstructed shapes and provide knowledge about the occluded areas was chosen. In contrast to previous approaches in this area, this thesis proposes to utilize the routinely collected set of five orthodontic documentary photos for reconstruction while other approaches instead relied on larger set of photographs.

4. Segmentation & numbering framework

This chapter introduces the general concept proposed for the segmentation and numbering task and explains its workflow in detail. After this chapter, the reader should have a clear understanding of why the framework was designed in the proposed way, what the purpose of each step in the workflow is and how each step is intended to work.

Both goals of this thesis – the teeth segmentation and numbering in panoramic radiographs and the image-based 3-D teeth reconstruction from five photographs – depend on an accurate and robust teeth *segmentation and numbering* in 2-D images. Teeth segmentation refers to the task of finding the contours of the individual teeth instances within the image while teeth numbering refers to assigning the correct number (cf. Section 2.2) to each tooth. Both tasks present several challenges which need to be solved. These include the low image quality and image characteristics of dental radiographs as well as the challenging appearance properties of teeth in photographs, the variations in the set of teeth between patients, and the difficulty of distinguishing different teeth for the purpose of numbering. As discussed during the review of state of the art methods (cf. Section 3.1), the inclusion of domain specific knowledge in the form of shape information and the spatial relation between individual teeth is required to tackle these challenges.

Because of that, this thesis proposes to use a model-based framework which combines a dynamically adapted coupled shape model (CSM) in combination with a neural network. The reasoning behind the concept is explained in the next section which also introduces the individual steps of the proposed pipeline. The subsequent sections of this chapter focus on describing the coupled shape model (Section 4.2) and explaining the individual pipeline steps in greater detail (Sections 4.3 to 4.5).

4.1. General concept

The general concept of the framework is to utilize a coupled shape model in order to encode the domain specific knowledge about the teeth shape and their relative location and combine it with the semantic segmentation performance of neural networks for handling the initialization dependency of CSMs to perform an accurate object-level teeth segmentation and numbering.

Teeth are made from the hardest material in the human body (tooth enamel) and are therefore not prone to deformations while having a well known characteristic shape, which makes them well suited for active shape models (ASM) [CTCG95,MT96] to encode the statistical information about their shape variation (global shape knowledge). The strength of an active shape model approach is that it prevents leaking of the segmentation into neighboring structures by imposing constraints on the shape of the seg-

mentation based on the statistical information extracted from the training data. This helps to handle the challenges of low image quality, low contrast and intensity variations due to artifacts, restorations and diseases (cf. Section 1.2.1). Active shape and appearance models are a well established technique to model the shape variation of structures and have been successfully applied in many different (medical) approaches like robust and accurate bone segmentation [WM16], liver segmentation in low-contrast CT volumes [EF17] or prostate ultrasound image segmentation [BJT*20]. In this work, an improved version of the ASM, the Probabilistic Active Shape Model [Kir13] is used. It increases the flexibility of the standard ASM by allowing deformations beyond the shapes observed during training, if these are sufficiently supported by image information. This should improve the segmentation performance for previously unseen shape instances.

Additionally, knowledge about the spatial configuration of the teeth is required in order to distinguish and identify individual teeth. This is mandatory for a numbering of the teeth. In order to incorporate the spatial information, the statistical knowledge about the relative location of individual teeth can be included via transformations between individual shape models, which forms a coupled shape model [TWT*03] of the teeth. The CSM then encodes both shape variation of all included structures as well as the change in relative position observed in the training data. The top-down information (domain knowledge) encoded in the coupled model can then be used in combination with the bottom-up image information (image features) to guide the segmentation process while constraining the results to valid teeth shapes and configurations.

Coupled shape models have not been applied for teeth segmentation previously, but are used in other areas for medical image segmentation. They have been applied for automatic localization of the sentinel lymph node and additional active nodes on SPECT-CT data [JMDW17], automated bone segmentation and anatomical landmark localization in Cone-Beam CT [BIV*19] or risk structure segmentation in the head & neck area in CT images for radiation therapy planning [JSK*14]. Using a coupled model has the added benefit of reducing the required initialization parameters since a single CSM needs to be initially placed instead of up to 32 individual tooth models. Moreover, since every tooth is still modeled by an individual shape model, it inherently includes a numbering of the extracted teeth contours.

A disadvantage of approaches using local search algorithms to look for relevant image features – like active shape models, probabilistic active shape models and also coupled (active) shape models – is the dependency on a good initial estimate to provide accurate segmentation results [HM09]. A local search algorithm determines the best matching image feature for every contour point separately. It is therefore limited to the vicinity around every contour point as larger search areas could move individual contour points to completely different parts of the images, which would prevent a successful contours extraction. For this reason, if the model is placed to far from the actual position of the structure(s) in the image, the local search might fail to detect the proper image features and the segmentation can fail.

A coupled model additionally suffers from the fact that it is positioned onto the image in its mean configuration. The overlap of individual model items with the relevant image regions can therefore vary and should be addressed to improve segmentation performance. The presented framework aims to handle the dependency of coupled shape models on a good initialization by exploiting the semantic

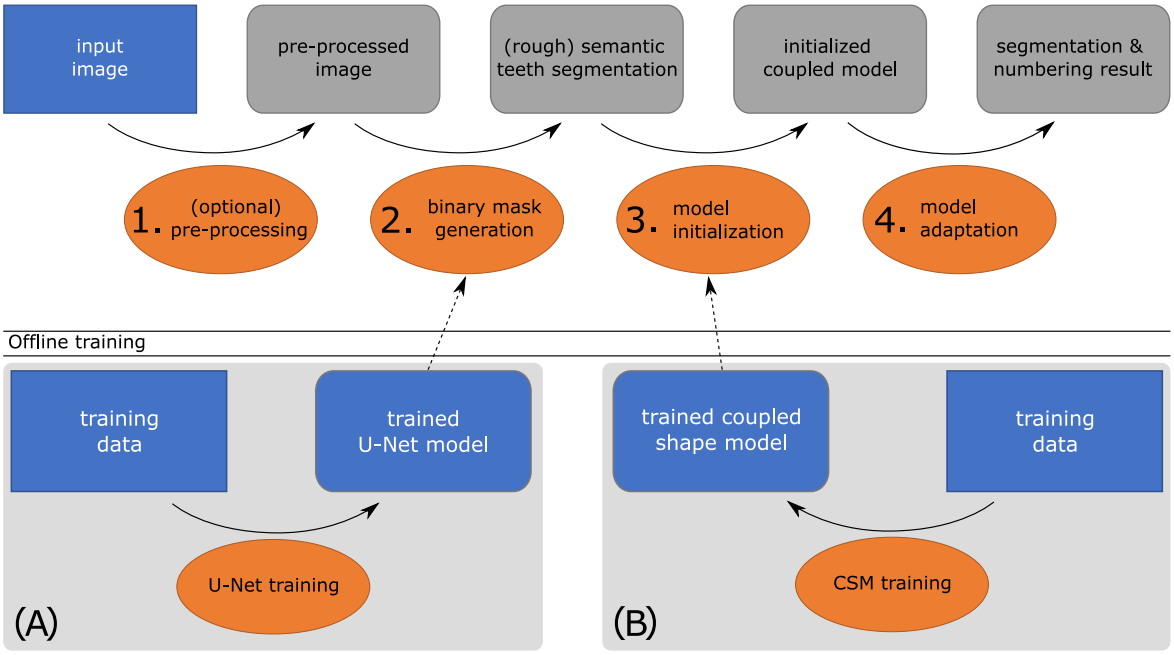


Figure 4.1.: Segmentation and numbering concept. The top part depicts the pipeline steps while the lower part depicts the offline training. Orange ellipses represent algorithmic processing, blue rectangles indicate input data and gray rectangles illustrate generated data.

segmentation performance of neural networks in combination with a dynamic adaptation to fit the shape models to the image.

The study of the related work in Chapter 3 has shown that convolution neural networks can offer good semantic segmentation performance on dental images. Other methods which only rely on image information struggle to achieve results that can be utilized for the initialization process without significant post-processing. While large amounts of training data are required for networks performing instance-based segmentation, good semantic segmentation results can already be achieved with relatively small amounts of data in combination with data augmentation [RFB15]. This framework proposes to exploit this and to utilize the semantic segmentation performance of encode-decoder networks (specifically U-Net [RFB15]) as a basis to determine a good initial estimate for the coupled shape model. The additional challenge in coupled models of sub-optimal initial placement of individual model items will be handled by dynamic adaptation strategies which aim to ensure a good fit before actively searching for image features for these items.

In order to combine all the aforementioned aspects and perform the segmentation (and numbering) of teeth according to the concept, an automatic segmentation pipeline has been devised. This pipeline consists of the following steps:

1. **Preprocessing (optional):** A preprocessing step can be included to improve the quality of the input image before applying the subsequent steps. This can include preprocessing for computing the binary mask, for the model adaptation, or both.
2. **Binary mask generation:** A trained convolutional neural network based on the U-Net architecture is used to compute a semantic segmentation (binary mask) of the teeth area. This mask is later utilized for initializing and fitting the CSM. While the architecture of the used network is designed to work for a variety of different applications, the network weights are application specific and need to be trained on training data corresponding to the application.
3. **Model initialization:** The quality of the segmentation result depends on a good initialization of the CSM onto the input image. Therefore, initialization parameters in terms of position, rotation and scale are computed from the binary mask extracted in the previous step. Depending on the configuration of the coupled shape model and the structure of the binary mask, the method for determining these parameters needs to be adjusted accordingly.
4. **Model adaptation:** The teeth segmentation and numbering is computed by adapting the coupled shape model to the image. The adaptation first uses the binary mask to refine the initialized model before performing the final adaptation in the original image. The performance of the adaptation process is enhanced by using dynamic adaptation strategies. Instead of fitting all items immediately to the image, items are dynamically added to the adaptation process to alleviate the problem of initialization dependency for local search algorithms. A free-form deformation of the model items at the end of the adaptation process (which is no longer restricted to the learned configuration space) is possible to enhance the segmentation results. The converged model then represents the final tooth segmentations with their corresponding numbering given by the coupled shape model.

Additionally, a onetime offline training step is required in order to train the coupled shape model (A) and the U-Net network (B) which are utilized during the segmentation and numbering process. The steps of the segmentation pipeline are depicted in Fig. 4.1.

The pipeline starts with an optional preprocessing step. Depending on the application, it might be advantageous to enhance the image before applying the neural network, or performing the model adaptation. This can include noise reduction, contrast enhancement, removal of imaging artifacts and many more. In this thesis, only the teeth segmentation in colored photographs uses this step. Since this step is completely application specific, it will not be discussed here further. Instead, more information can be found in the respective chapter of the thesis (cf. Chapter 6).

The second step of the pipeline is the generation of a binary mask of the relevant structures using a convolutional neural network. The U-Net architecture – inspired by encoder-decoder techniques and named after the u-shaped structure of convolutional- and de-convolutional layers – has been widely applied in medical computing since its presentation in 2015 [RFB15] due to its robust (semantic) segmentation performance in many applications. The generated mask is not only used in the next pipeline step for providing information to initialize the coupled model onto the image, but is also employed later during model adaptation. The binary mask generation pipeline step is explained in Section 4.3.

The third step in the pipeline determines the correct parameters to place the coupled model onto the input image and is explained in Section 4.4. The coupled shape model is composed of individual active shape and appearance models which rely on local search algorithms to find image features and therefore depends on a good initialization onto the input image in terms of position, rotation and scale. These parameters will be calculated based on the binary mask computed in the previous step. Depending on the structure of the coupled shape model, the shape of the binary mask and the dynamic adaptation strategies used in the next step, different methods are used to determine these parameters.

The final step of the pipeline is the adaptation of the coupled shape model to the input image in order to extract the teeth contours. Starting from the initialized model position, the model is fitted to the image by an energy minimization technique that balances between the features computed from the image and the learned shape and spatial configuration. The model is first fitted to the binary mask before switching to the preprocessed input image. During the adaptation process, dynamic adaptation strategies are used to control which model items are actively fitted to the image and which ones are only passively changed through the learned configuration space. This step is explained in Section 4.5:

The presented framework was used and improved over the course of the following scientific publications [WMW18, WWW18, WLW20, WJN*21]. Before looking at each pipeline step in greater detail, the next section will first introduce the coupled shape model.

4.2. Coupled Shape Model

The coupled shape model is a model-based technique that aims to capture the statistical variance in the relative location of individual structures as well as the shape variance of the individual structures over a set of training instances. The encoded prior information can then be utilized to reduce the complexity of the segmentation process by constraining the possible shape and location of the structures in the image. Within the coupled model, each one of the N individual structures is represented by a deformable model item and all items are coupled by their spatial relation in terms of translation, rotation and scale.

In the following, the design of model items and the modeling of their spatial relation as well as the training process of both will be explained. The CSM is partially based on the work of Steger et al. [SKW12, SJW14] and Jung et al. [JSK*14, JKW16]. Steger et al. used a 3-D articulated atlas which employed probability maps to represent an items shape, average intensity maps to model the appearance and unit transformations for the position with the atlas coordinate system. The 3-D atlas was used to segment bones in CT images of the head and neck area. Jung et al. extended the concept by adding 3-D deformable items (triangular meshes) which represent the shape as statistical shape models. They constructed a 3-D coupled shape model consisting of 22 structures and segmented bones and soft tissue structures in 3-D CT data of the head and neck area.

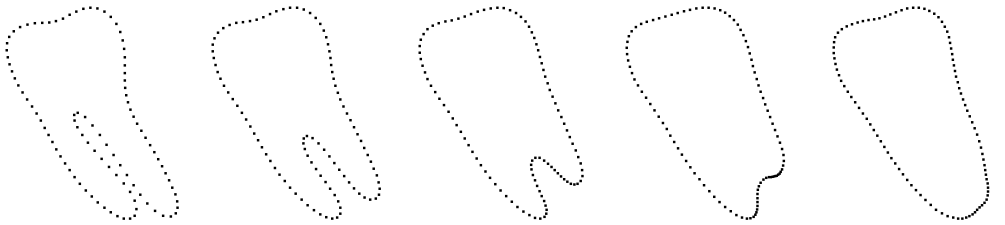


Figure 4.2.: First mode of shape variation of a deformable model item of the second molar tooth.

4.2.1. 2-D deformable model items

An individual deformable model item d_j (with $(1 \leq j \leq N)$) is designed to capture the shape variation of an individual tooth over a set of s training shapes and optionally the intensity or color appearance on each part of the tooth shape. A deformable model item is represented by a statistical shape model (SSM) with a fixed number of landmark points and is generated using a point distribution model (PDM) [CTCG92]. A landmark is a point on the contour of a tooth – mostly corresponding to a specific anatomical feature – that is consistent for every (shape) instance of the training set (cf. Fig. 4.3). It is assumed that the training shapes are provided as landmark vectors with inherent point correspondences.

The statistical information is then extracted from the s training shapes using principal component analysis (PCA) by computing eigenvectors e_m and their respective eigenvalues λ_m (with $\lambda_i > \lambda_{i+1}$) of the covariance matrix

$$S = \frac{1}{s-1} \sum_{i=1}^{s} (c_i - \bar{c})(c_i - \bar{c})^T, \tag{4.1}$$

where $\bar{c}$ is the mean shape and c_i is the vector containing the coordinates of all landmarks for training set i. i.e. $c_i = \{x_1, y_1, x_2, y_2, \ldots, x_L, y_L\}$ with L the number of landmark points. To improve the generalizability of the shape model, only the first n eigenvectors required for capturing 95% of the shape variance are kept and the remaining ones are discarded. In the context of shape modeling, these remaining eigenvectors are frequently referred to as *modes of variation* or simply *modes*. Using this approach, every valid shape $\tilde{c}$ can then be approximated by a linear combination of the n principal modes of variation:

$$\tilde{c} \approx \bar{c} + \sum_{i=1}^{n} v_i e_i, \tag{4.2}$$

where v_i are the entries of a vector $v = \begin{pmatrix} v_1 & v_2 & \ldots & v_n \end{pmatrix}$ containing the shape coefficients corresponding to the first n principal modes and e_i are the respective eigenvectors.

Shape training In order to be able to extract the statistical information about the shape variation from the training data, it is important to ensure that individual landmark points in all training instances correspond the same anatomical feature (cf. Fig. 4.3).

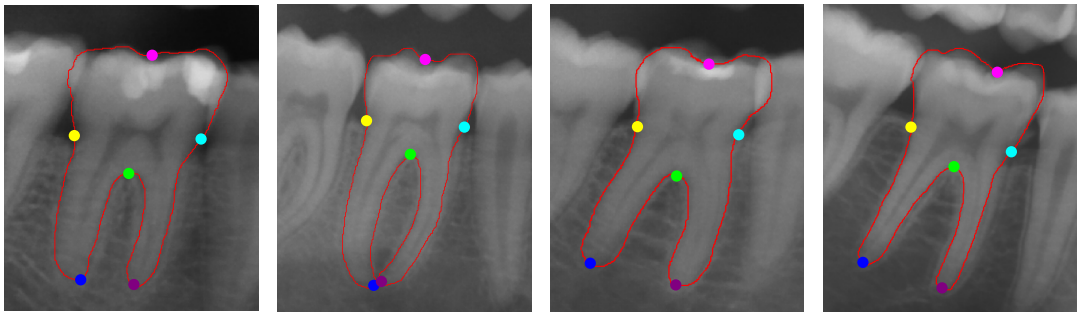


Figure 4.3.: Landmark correspondence: Landmark points on the tooth contour need to represent the same anatomical landmark in order to enable a modeling of the shape variation. Points of the same color depict the same anatomical landmark on different contours.

It is therefore necessary to explicitly establish the landmark correspondences between the individual training shapes. This process is known as *correspondence problem*. This thesis uses a pairwise method for establishing correspondences which assumes that the correspondence between two individual shapes of the training set is independent of all other training shapes. Pairwise methods make use of the transitivity of correspondences. This implies that if a landmark a of training shape s_1 corresponds to landmark b of shape s_2, and landmark b corresponds to landmark c on shape s_3, then a corresponds to c as well. It is therefore possible to register all training shapes to a single reference shape in order to establish the correspondence between the complete training set. A random training shape is chosen as the reference shape and correspondences between landmark points are established based on the curvature of the contour at the landmark position as well as the distance to the shape's centroid while preserving the order of landmark points during optimization [OT09]. The similarity ς of two landmarks a and b is given by

$$\varsigma = \sqrt{(\alpha_a - \alpha_b)^2 + (d_a - d_b)^2}, \tag{4.3}$$

where α is the normalized curvature of a landmark point and d the normalized distance to the centroid. This measure has been chosen as it provides good correspondences even in more difficult cases like molar teeth. The centroid information gives stability to the global matching of landmark points while the curvature improves the local matching. An example of the established correspondences on such a case is shown in Fig. 4.4.

Moreover, all training instances need to be properly aligned to avoid distortion of the statistical information due to external factors like scaling or rotation between individual training instances. To achieve this, the s training instances are aligned using generalized procrustes analysis [Gow75]. Generalized procrustes analysis works by superimposing a population of shapes and in doing so minimizes the squared distance between landmark points of all shapes. It is an iterative process that starts by selecting an arbitrary shape of the set of training shapes as a reference. Next, all other training shapes are superimposed onto the selected reference shape and a mean shape is computed from the set of superimposed shapes by calculating the average position of each individual landmark point. Afterwards, the mean

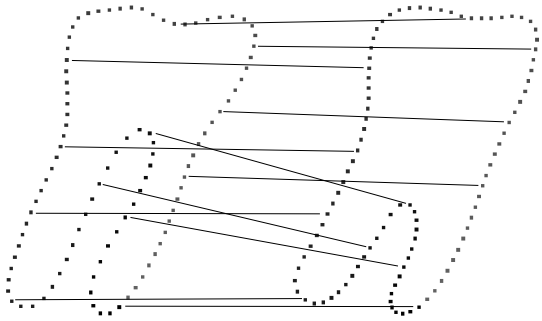


Figure 4.4.: Correspondence example. The lines between the two tooth shapes indicate selected correspondences of landmark points. Both teeth have been scaled to the same size for better visualization.

shape becomes the new reference shape and the set of training shapes is again superimposed onto the new reference. This process is repeated until the difference between the old and new reference is below a certain threshold. The final set of superimposed training shapes is considered to be properly aligned and used as input to the statistical shape model training.

4.2.2. Spatial relation

Having trained the individual deformable model items, the coupled shape model is then created by combining the relative pose of each model item in relation to the center of mass of the complete model with its shape information. Each individual model item d_j is parameterized by a vector p_j which consists of $m + n_j$ entries. Here, m are the transformation parameters and n_j the number of shape coefficients (cf. Eq. (4.2)). For a 2-D coupled model the number of transformation parameters per item is $m = 4$, 2 for translation and 1 for rotation and isotropic scaling, respectively. The transformations are expressed by linear combinations of unit transformations using a matrix exponent representation to generate a unique and linear vector (cf. Steger et al. [SKW12]) to enable a valid PCA later on. By concatenating the parameter vectors p_j of all model items for a training instance k, the configuration vector f_k is generated. Again, PCA is applied to the training configuration matrix $F = \begin{pmatrix} f_1 & f_2 & \dots & f_s \end{pmatrix}$ to describe the space of all possible configurations over all training instances, which is later used during the adaptation process. Any possible configuration of the coupled model can then be described by a vector b as

$$\tilde{f} = \bar{f} + A \cdot b + r, \tag{4.4}$$

where $\bar{f}$ is the mean configuration, r is the residual error and A is the matrix containing all eigenvectors of the covariance matrix

$$C = \frac{1}{s-1} \sum_{k=1}^{s} (f_k - \bar{f})(f_k - \bar{f})^T. \tag{4.5}$$

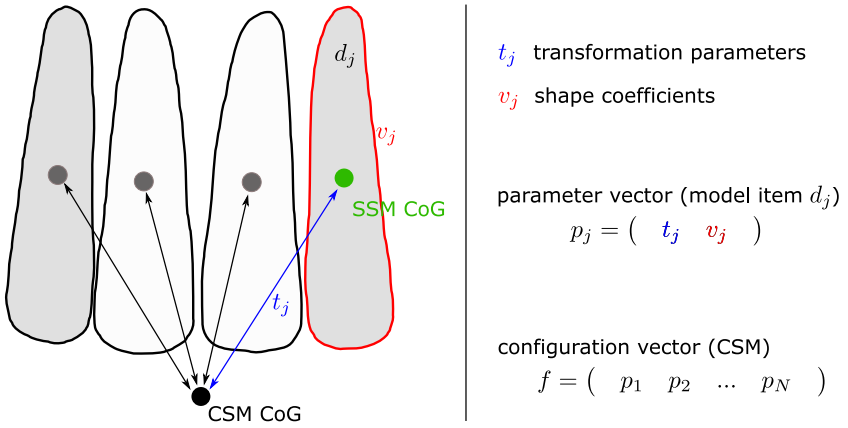


Figure 4.5.: The spatial relation of individual items of the CSM is modeled via transformations between the center of gravity (CoG) of the model item and the CoG of the CSM, as depicted on the left. The construction of the parameter vector of individual model items and the configuration vector of the CSM is summarized on the right.

The final (trained) coupled shape model can then be used for the segmentation and numbering task as is described in the following sections which outline the individual pipeline steps. In particular, the coupled model is adapted to the input image in step four (cf. Section 4.5) after being initialized onto the image.

4.3. Binary mask generation

The first obligatory step of the concept pipeline, after an optional pre-processing step, is the generation of the binary mask which is utilized for the initialization of the coupled shape model and the initial adaptation, as is described in the following sections (cf. Section 4.4 and Section 4.5). This concept proposes to use the convincing semantic segmentation performance of encoder-decoder neural networks for this task. Simpler signal-based methods, like thresholding, feature point detection or estimation of reference measures, proved ineffective in producing reliable results which could be employed for the computation of the initialization parameters (scale, rotation, and position) of the coupled model. The U-Net architecture, proposed by Ronneberger et al. [RFB15] and named after the u-shaped structure of convolutional- and de-convolutional layers, is modified and utilized here to compute the binary mask. The advantages of this kind of neural network are its good semantic segmentation performance in medical image data and that it can be trained with a limited number of training data and still produce accurate segmentation results. This is achieved by relying on data augmentation to extend the training set and a specially tuned network architecture.

The modifications made compared to the original U-Net publication to ensure the desired semantic segmentation results will be discussed in the next paragraphs and include loss function, hyperparameters and data augmentation.

Loss function An important factor for segmentation performance is the loss function used during training. The loss function measures the prediction error between the output generated by the network and the supplied training data. It is used to calculate the gradients which influence the internal weights of the network during the back-propagation process. The original U-Net uses cross entropy (CE) to measure the training loss between predictions and ground truth training samples. More recent publications in medical image segmentation suggest to use different loss functions for binary segmentation [KPIB19], especially in imbalanced data sets where the amount of background area is significantly different from the foreground area. The Tversky loss [SEG17] has shown promising results in such cases. It supports a user-defined weighting to set preference to either sensitivity (i.e. reduce the amount of teeth area which is falsely classified as background) or precision (i.e. reduce the amount of background that is incorrectly classified as teeth area). The Tversky loss is defined as

$$\tau(\alpha,\beta) = 1 - \frac{TP}{TP + \alpha * FP + \beta * FN}, \tag{4.6}$$

where TP is the number of true positives, FN the number of false negatives and FP the number of false positives when comparing ground truth and prediction in a pixel-wise fashion. A detailed explanation on TP, FP and FN can be found in Section 5.2 of the thesis. The parameters α and β are used to specify the weighting between false positives (FP) and false negatives (FN), respectively. Choosing $\alpha > \beta$ penalizes FPs more heavily than FNs, which sets the preference to increased precision – most likely at the cost of a lower sensitivity. In the case of $\alpha = \beta = 0.5$, the Tversky loss is equal to the Dice loss, which is another typical loss function used in binary segmentation for more balanced data sets. Salehi et al. [SEG17] concluded that $\alpha = 0.3$ and $\beta = 0.7$ gave the best Dice score for a highly imbalanced data set, Koch et al. [KPIB19] adopted the same setting for their U-Net-based semantic teeth segmentation in panoramic radiographs but did not perform a separate evaluation. The optimal values for the presented work will be determined during the experiments of the teeth segmentation in panoramic radiographs (Chapter 5) and colored photographs (Chapter 6).

Hyperparameters Hyperparameters also have an impact on the performance of a convolutional neural network and optimal ones can be hard to identify [MOM12]. They can be divided into two types, the ones that determine the structure of the network (model-hyperparameters) and the ones that control the training process of the network (algorithm-hyperparameters). Structure-related hyperparameters are, e.g., the size of the kernel filter, the use (and size) of padding, the number of hidden layers or the activation functions. Parameters concerting the training of the network, e.g., include the learning rate, number of training epochs or the batch size. Various different optimization methods have been proposed for determining the optimal configuration of hyperparameters. Since the U-Net architecture is utilized here, many structure-related parameters are already predetermined while others, including

some training-related ones, can be chosen based on the various related works in the field of teeth segmentation with neural networks (cf. Section 3.1.7). It is also worth noting that the main intention of the semantic segmentation result is to facilitate a robust initialization of the coupled shape model and, while the best possible semantic segmentation here might help, this step is not depending on it.

The structure-related parameters of the original U-Net are retained with some exceptions. Contrary to the original U-Net, this thesis uses padded convolutions instead of the original un-padded ones in order to retain the same image size between input and output. This is required since there might be teeth in the otherwise cropped regions and the subsequent model initialization depends on extracting all teeth, e.g. to compute an accurate scale value between input image and coupled model. The network input size will be adjusted to respect the aspect ratio of the images used in the respective application as well as any required resizing. Training parameters are chosen with a validation split of 0.25, batch-size of 8, and dropout of 0.3. The number of training epochs are set as 500 maximum with an early stopping if the validation loss did not improve within the last 50 epochs. Steps per epoch is set to 500 to achieve the desired number of training augmentations, as is explained in the next paragraph.

Data augmentation Data augmentation is a technique to (artificially) increase the amount of available data by adding slightly modified versions of already existing data. Deep neural networks with a large number of hidden layers also have a large number of trainable parameters. These networks therefore highly depend on large amounts of training data to achieve good results. The existing data can be augmented in different ways by applying random transformations to the images. The following augmentation have been utilized during the training process to increase the number of training images:

- horizontal flipping
- shearing (0.02)
- shifting (horizontal and vertical direction, 20% of image size)
- zoom (30% of image size)

During the training process, random combinations of augmentations are applies to the training data such that 4000 images are process during each training epoch.

Another option to deal with limited training data is transfer learning in which weights are copied from an already trained networks to transfer it's "knowledge" instead of starting the training process from scratch. However, this technique was not been required in this work as the achieved accuracy is sufficiently high for the purpose of model initialization.

4.4. CSM initialization

The coupled shape model uses local search methods with a limited capture range to find image features for the adaptation process. Therefore, the quality of the segmentation result and the accuracy of the numbering both depend on a robust and accurate initialization of the model onto the input im-

age. Initialization refers to ensuring a good overlap between model structures of the mean model and corresponding anatomical structures in the image.

The challenging part about coupled shape models is that it is very difficult to achieve a good overlap for all structures in the model. Moreover, any initialization method that tries to maximize the overall overlap will sacrifice overlap of some structures to compensate for others.

For this reason, this initialization method focuses on maximizing the overlap for selected structures – referred to in the following as *initialization structures* – and handling the resulting sub-optimal overlap of other structures later during the adaptation process with dynamic adaptation strategies. These are explain in the next section (cf. Section 4.5).

The goal of the CSM initialization is to compute the required initialization parameters, namely scale, rotation and translation. While translation is computed based on the initialization structures, rotation and scale will be computed based on the complete model as they provide more stable results. The computation of the parameters is based on the binary mask generated by the U-Net in the previous step. The concrete implementation to compute these parameters is application specific. On the one hand, it depends on the structure of the binary mask and, on the other hand, on the choice of initialization structures. The concrete solutions will be proposed in the application-related chapters of the thesis.

4.5. CSM adaptation

After the coupled model has been initialized onto the input image in the previous step, it will now be adapted to the image. Adaptation refers to the process of finding the optimal model configuration that best matches the teeth in the image. It is an iterative process similar to Expectation Maximization (EM) that alternates between the search of suitable image features and constraining the model to a valid configuration based on the configuration space.

The adaptation is performed by minimizing an energy functional E which depends on two sets of parameters:

- (a) the transformation parameters t describing the global position of the model in terms of translation and rotation and
- (b) the vector f describing the configuration of the coupled model.

The energy formulation is based on the Probabilistic ASM by M. Kirschner [Kir13]. The model configuration contains the encoded spatial location of the model items relative to the model center and shape coefficients of each individual item (cf. Eq. (4.4)). The functional is given by

$$E_{2D}(f,t) = E_{image}(f,t) + \alpha E_{local}(f) + \beta E_{model}(f). \tag{4.7}$$

The three energy terms E_{image}, E_{local} and E_{model} force the model to adapt to strong image features, ensure smooth contour boundaries and keep the configuration close to the learned configuration space, respectively. They will each be introduced in the subsequent paragraphs. The parameters α and β are empirically determined to balance the individual energy terms.

The model energy term E_{model} restricts the coupled shape model to stay within or close to the learned configuration space. This is achieved by modeling the probability density of the configurations by an approximate multivariate Gaussian distribution similar to the approach by Steger et al. [SJW14]. Their modeling of the probability density function is based on the approach of Moghaddam and Pentland [MP97]. The model energy is defined as

$$E_{model}(f) = \sum_{j=1}^{k} \frac{b_j^2(f)}{\lambda_j} + \frac{m-k-1}{\sum_{i=k+1}^{m-1} \lambda_j} ||r(f)||_2 \tag{4.8}$$

where λ_j with $\lambda_1 \geq ... \geq \lambda_{m-1} > 0$ are the non-zero eigenvalues of the covariance matrix C computed during model training (cf. Eq. (4.5)). The first term in the summation of Eq. (4.8) represents the *distance in feature space* (DIFS). It is calculated as the Mahalanobis distance (MD) [Mah36] in the linear subspace spanned by the k first eigenvectors. The second term in the summation denotes the *distance from feature space* (DFFS) and estimates the cost of projection to the subspace. The DFFS ensures that model configurations are not completely restricted to the linear subspace, which increases the flexibility of the model.

While the model energy E_{model} ensures that the model stays close to the learned configuration space, it only does so in terms of the magnitude of the deviation but not in terms of additional shape variation. This could potentially result in jagged and therefore implausible segmentation contours. $E_{local}(f)$ is a local shape energy that serves as a regularization force to ensure smooth boundaries. It penalizes landmarks that move too far away from their neighbors. The local shape energy is defined as

$$E_{local}(f) = \sum_{i=1}^{N} E_{local}^{(i)}(f) \tag{4.9}$$

with

$$E_{local}^{(i)}(f) = \sum_{\forall j \in \mathcal{L}(i)} \sum_{\forall k \in \mathcal{N}(j)} ||g(f,i,j) - \mu_j - g(f,i,k)||_2, \tag{4.10}$$

where $\mathcal{L}(i)$ are the landmarks of model item i, $\mathcal{N}(j)$ are the neighbors of landmark j, $g(f,i,j)$ is the position of landmark j (of model item i) in the current model configuration f and μ_j is the mean relative position of j in relation to its neighbors in the training data.

The term $E_{\text{image}}(f,t)$ is responsible for drawing the contour of model items in the direction of strong image features. An intuitive approach to model the image energy is to minimize the sum of weighted squared distances (WSD) between the current model configuration and the detected feature points in the image, resulting in

$$E_{\text{image-wsd}}(f,t) = \sum_{i=1}^{N} E_{\text{image-wsd}}^{(i)}(f,t) \tag{4.11}$$

with

$$E_{\text{image-wsd}}^{(i)}(f,t) = \sum_{\forall j \in \mathcal{L}(i)} \left\| w_j^{(i)} \left(x_j^{(i)}(f,t) - \hat{x}_j^{(i)}(f,t) \right) \right\|_2. \tag{4.12}$$

Here, $w_j^{(i)} \geq 0$ are weights determined by the appearance model describing the quality of the image feature, $x_j^{(i)}$ is the vector of the optimized landmark positions of structure i, and $\hat{x}_j^{(i)}$ are the detected image features. However, this type of image energy is unable to counter the shrinking force caused by the local shape energy $E_{local}(f)$.

Instead, the Probabilistic ASM uses a different definition of image energy. It penalizes not only the deviation of the landmark itself but also considers the neighboring landmarks. This aims to compensate a movement of the landmark due to the local shape energy by moving neighboring landmarks in the opposite direction. More information on this is available in [Kir13]. The image energy term $E_{\text{image}}(f,t)$ is therefore defined as

$$E_{\text{image}}(f,t) = \sum_{i=1}^{N} E_{\text{image}}^{(i)}(f,t) \tag{4.13}$$

with

$$E_{\text{image}}^{(i)}(f,t) = \sum_{\forall j \in \mathcal{L}(i)} \left\| w_j^{(i)} \left(x_j^{(i)}(f,t) - \hat{x}_j^{(i)}(f,t) \right) + \sum_{\forall k \in \mathcal{N}(j)} w_k^{(i)} \left(x_k^{(i)}(f,t) - \hat{x}_k^{(i)}(f,t) \right) \right\|_2 . \tag{4.14}$$

Here, the first inner term is the difference between the optimized landmarks $x_j^{(i)}$ of structure i and the detected corresponding image features $\hat{x}_j^{(i)}$, while the second inner term is designed to counteract a shrinking force introduced by the local shape energy.

The optimization process is done using a gradient descent optimizer. The transformation parameters t are optimized first, and then the configuration and transformation parameters f,t are optimized jointly. Initially, the model is adapted to the binary mask using gradient features to ensure a good placement of all teeth. The final adaption is then performed on the input image using appropriate appearance features (gradient features for dental radiographs and HSV color features for photographs) which, after convergence, provides the final teeth contours detected in the image.

Local search strategy The optimal image features $\hat{x}$ which are the driving force behind the model adaptation are detected using a local search strategy. This search is limited to the vicinity around the current shape configuration of individual model items. A linear search centered at the landmark points and oriented in normal direction is performed for each landmark in order to find the best matching feature. Here, different types of image features, like intensity values, gradient magnitude and/or direction, or appearance vectors can be used. Fig. 4.6 illustrates the search of best matching features points on a binary image using gradient magnitude as image features. The strongest image features are therefore found on the boundaries of the white area.

Dynamic adaptation strategy The iterative process of adapting the coupled shape model – and therefore the individual deformable items – to the input image is governed by many different parameters. These parameters and the individual items they apply to can be modified for different phases of the

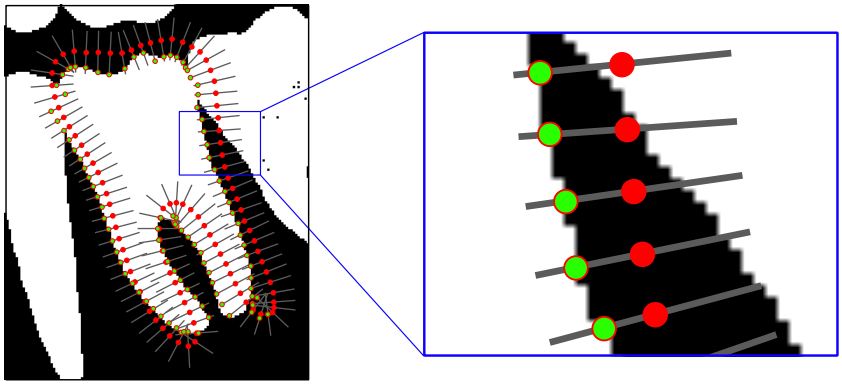


Figure 4.6.: Local feature search: A linear search in normal direction is performed at each landmark position to determine the best matching image features. In this case, gradient-magnitude and -direction are used as image features. Red dots indicate the current landmark locations of the shape model, dark gray lines depict sampled locations inside and outside the shape, and green dots depict the identified best matching image features.

adaptation process. The individual phases of the adaptation process are referred to as *adaptation steps*. The set of adaptation steps defines an *adaptation strategy*.

The motivation behind this approach is the following: As stated before, the presented coupled shape model utilizes a local search strategy to find the optimal appearance features for each model item which requires a good initialization of the coupled model. However, in contrast to single shape models where a good initialization ensures a close proximity of mean shape model and image features, the coupled model initialization needs to balance the initialization of all its included model items. Even good positioning of the mean model's center can result in large discrepancies between some model items and their corresponding image features. Fig. 4.7 depicts a good initialization of an individual molar tooth model versus the initialization of the mean coupled model. It is evident that since the mean model is initialized according to its center of mass, the teeth close to the center of the model show good overlap with the input image while the teeth farther away from the center (e.g. molars) might not match as good. The bigger the variation in relative position between individual items is, the larger the discrepancy can be when placing the mean model. This could lead to model items being adapted to the wrong structure in the input image due to incorrect appearance features being found. Instead of trying to balance the proximity of each model items to its image features, a dynamic adaptation strategy is used which initially limits the image-driven adaptation to more reliable items before including the others.

For this reason, the image energy given by Eq. (4.13) is modified to:

$$E_{image}(f,t,p) = \sum_{i=1}^{N} w_i(p) E^{(i)}_{image}(f,t), \tag{4.15}$$

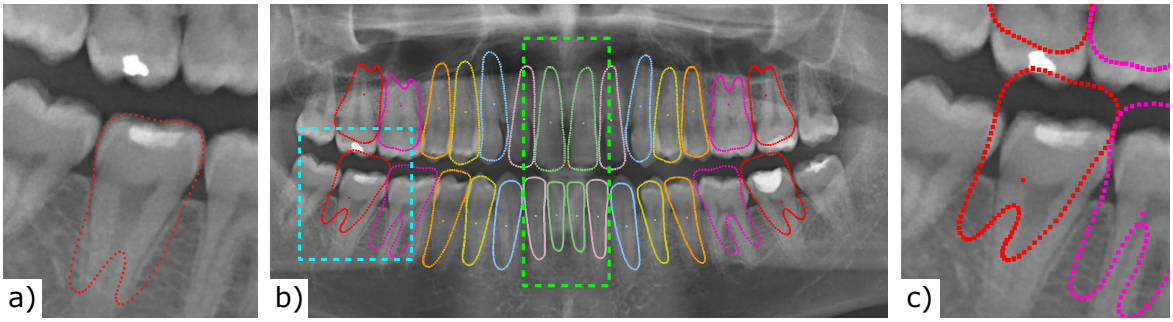


Figure 4.7.: Comparison of ASM and CSM initialization: (a) theoretically optimal initialization of single molar tooth ASM, (b) optimal initialization of CSM based on CoG (green box), (c) initial molar tooth position resulting from CSM initialization (cyan box in center picture). It is evident that the position of the molar tooth model resulting from the CSM initialization is sub-optimal compared to the single ASM.

where p is a parameter describing the progress of the adaptation process and $w_i(p) \in \{0,1\}$ are element weights depending on the adaptation progress. Initially, only the weights of the reliable model items, i.e. items which are expected to have close proximity to their image features after in initialization process, are set to one while all others are set to zero. Model items which are not adapted during an adaptation step are only changed passively through the learned (spatial) configuration. During the adaptation process, additional items will be added dynamically via individual adaptation steps until all items are actively adapted and Eq. (4.15) becomes equal to Eq. (4.13).

By adapting the initially unreliable items at a later time, they have already been (passively) moved closer to their correct position and more reliable image features will be found once they are actively adapted.

Additionally, each adaptation step allows to modify the adaptation parameters like the weights used to balance the energy terms in Eq. (4.7) and number of iterations, the parameters for the local search like search radius, or parameters related to the appearance features.

The dynamic adaptation strategy is manually designed and fine-tuned for each application scenario to ensure an optimal adaptation process.

4.6. Conclusion

This chapter introduced the general concept for the segmentation and numbering task. It utilizes a coupled shape model to encode the domain-specific knowledge about the shape variation of individual model items together with information about their spatial relation. This top-down information is exploited during the adaptation process which fits the model to the input image. By using dynamic adaptation strategies and optimizing an energy-functional that balances the encoded top-down information

and the bottom-up gradient image features retrieved by a local search algorithm, the teeth contours can be extracted from the image. The model-based approach helps to handle the associated challenges like low image quality, image characteristics of dental radiographs as well as the challenging appearance properties of teeth in photographs. It also inherently provides the tooth numbering corresponding to the extracted contours. The initialization dependency of local search-based approaches is tackled using a convolutional neural network. The network generates a binary mask of the teeth area which is then used to place the coupled model onto the input image. The binary mask is utilized during parts of the adaptation process as well to restrict the image feature search.

This concept was designed in order to provide an accurate and robust teeth segmentation and numbering in 2-D images which is required for both goals of this thesis – the teeth segmentation and numbering in panoramic radiographs, and the automated silhouette-based 3-D reconstruction of teeth from photographs. The application of the presented concept for these task is covered in the next two chapters. Chapter 5 contains the teeth segmentation and numbering in panoramic radiographs and Chapter 6 the 3-D teeth reconstruction from five photographs.

5. Teeth segmentation & numbering in panoramic radiographs

This chapter presents the application of the segmentation and numbering concept explained in the previous chapter for the first research question of this thesis – the fully-automatic segmentation and numbering of teeth in dental panoramic radiographs. It is the first approach for teeth segmentation (and numbering) to explicitly model shape knowledge as well as knowledge about the spatial relation of teeth. The goal of this chapter is to show that the integration of domain knowledge about teeth shapes and their spatial relation can handle the technical challenges, like image quality and characteristics of panoramic images or variations in the set of teeth, involved in this task. To this end, the performance of the proposed approach is evaluated on two different data sets.

The workflow of the teeth segmentation and numbering in panoramic radiographs follows the workflow of the proposed concept (cf. Chapter 4) and is depicted in Fig. 5.1. No pre-processing is performed on the input images – the zoomed-in images contained in the workflow figure are just for illustrative purposes, internally the complete image is used during processing.

In the offline training phase, a coupled shape model containing 28 teeth is created and trained for each data set (Fig. 5.1 B). Similarly, U-Net networks are trained for semantic segmentation of the same 28 teeth (Fig. 5.1 A). The four teeth which are not included are the wisdom teeth. The fact that many people are missing some or all of their wisdom teeth would significantly reduce the amount of available training data as model training requires training instances that include all structures included in the model. Moreover, the variation in the presence of the wisdom teeth would also increase the difficulty for model initialization.

The segmentation and numbering is then performed according to the proposed pipeline of the segmentation and numbering workflow. The trained U-Net network is used to extract a binary mask of the teeth area (without wisdom teeth). This mask is employed in the model initialization step to compute the initialization parameters in terms of position, rotation and scale for the coupled shape model to maximize the overlap of the initialization structures. Afterwards, the initialized coupled model is adapted to the input image using custom dynamic adaptation strategies which gradually extend the set of actively adapted structures while utilizing gradient image features. The model adaptation step utilizes both the binary mask computed by the U-Net and the input image for different adaptation steps. The final teeth contours and their numbering are then provided by the converged coupled shape model.

Throughout this chapter, the performance of the individual pipeline steps will be discussed directly in the respective section. Therefore, the data sets and evaluation metrics are introduced at the beginning of

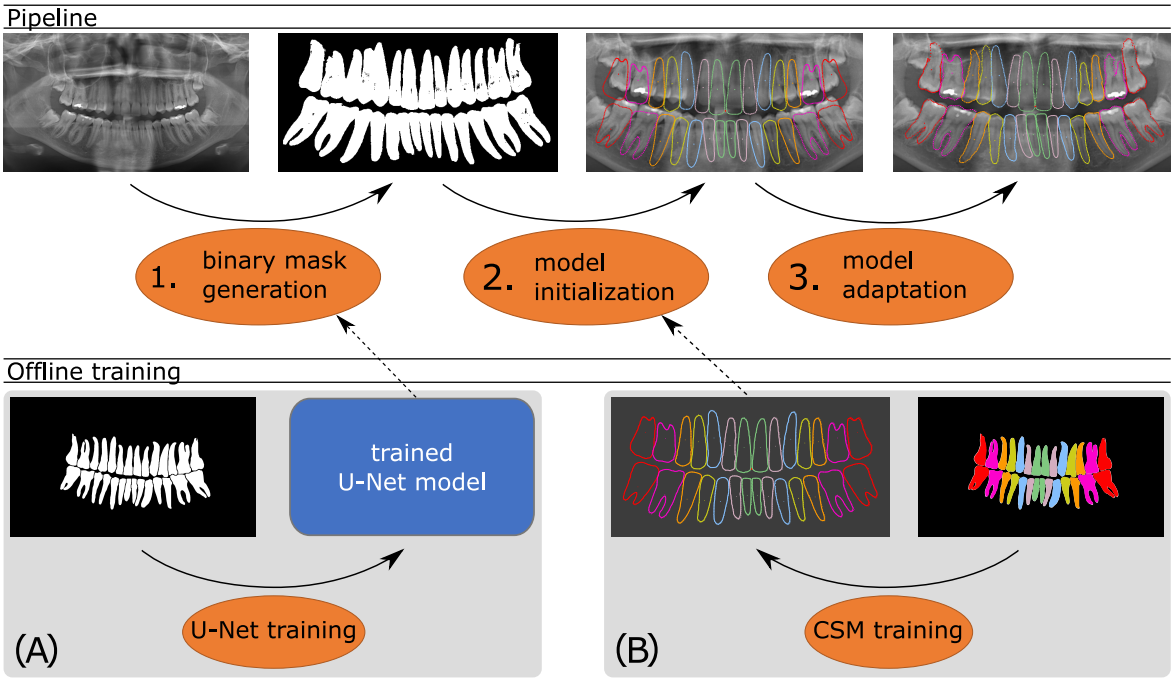


Figure 5.1.: Teeth segmentation and numbering workflow for panoramic radiographs. The required U-Net network (A) and CSM model (B) will be trained in an offline training phase based on manually annotated training data. The (online) pipeline then utilizes these to compute the teeth segmentation and numbering for an unseen input image (top half).

the chapter. The evaluation of the final segmentation and numbering will be done in a separate section followed by a separate discussion section.

The structure of this chapter is as follows. Section 5.1 introduces and discusses the data set used for the training and evaluation for the approach and recaps the challenges involved with panoramic radiographs. The metrics used for performance measurements are then explained in Section 5.2. The offline (training) phase of the workflow is discussed thereafter. Section 5.3 describes the design and training of the coupled shape model used for this approach. Section 5.4 covers the training of the U-Net network and the estimation of optimal values for selected hyperparameters. Then, the implementation of individual pipeline steps for the approach are introduced. Section 5.5 handles the first pipeline step, the binary mask generation. Section 5.5 explains the implementation of the second pipeline step, the computation of the initialization parameters for the coupled model from the binary mask. Section 5.6 presents the final pipeline step, the model adaptation including the dynamic adaptation strategies. Section 5.7 describes the experiments and results of the presented approach for the two data sets used for evaluation. The chapter concludes with a discussion of the approach and its results in Section 5.8 and a final conclusion in Section 5.9.

This chapter is partially based on the publications entitled "Automatic Teeth Segmentation in Panoramic X-Ray Images Using a Coupled Shape Model in Combination with a Neural Network" [WMW18] and "Automatic Teeth Segmentation in Cephalometric X-Ray Images Using a Coupled Shape Model" [WWW18].

5.1. Data

Throughout this chapter, the proposed segmentation and numbering approach is trained with and evaluated on two different data sets. The approach was originally designed and developed on data set 1, but has been updated for the only recently available, much larger data set 2.

Data set 1 This set is a small in-house data set consisting of only 24 panoramic radiographic images. It is used to train and evaluate the approach for the case of limited training data - which was one of the criteria for the design of the framework. The set includes images of a variety of cases with several difficulties like missing or broken teeth, artificial teeth, fillings and bridges. It also covers patients that have all, none or only some wisdom teeth. A more detailed description of the included images via a separation into eight categories is available in Table 5.1. The individual image categories will be discussed in more detail in a separate paragraph later in this section. The set was randomly divided into a training set containing 12 images and a test set containing the remaining 12 images. Gold-standard segmentations and numberings of all teeth were manually created for all images. All images have a resolution of 2440x1280 pixels.

Data set 2 The second data set is the DNS Panoramic Images data set[1] published by Silva et al. [SPOP20]. The set consists of 543 panoramic radiographs with tooth-wise ground truth segmentation and corresponding label. The authors divided the data into a training set of 432 images and a test set consisting of the remaining 111 images. The same separation is used in this thesis. All images in this set have a resolution of 1991x1127 pixels. This data set is a subset selected from the UFBA-UESC Dental Images data set published by Silva et al. [SOP18] in 2018. The original data set is comprised of 1500 panoramic radiographs which were categorized into ten groups based on certain criteria like presence of all teeth, presence of restorations, etc. This data set has been used in a number of publications already [KPIB19, ZLG*20, CZL*21]. However, the annotations included with this set were binary masks with a pixel-wise classification as teeth or non-teeth. Such annotations are only suitable for training and evaluation of semantic segmentation approaches. The new DNS (Detection, Numbering, and Segmentation) Panoramic Images data set was selected from eight of the original ten categories and enhanced by separating the binary mask into tooth-wise segmentations and adding the corresponding tooth number.

[1] https://github.com/IvisionLab/dns-panoramic-images

Table 5.1.: Image categories and average number of teeth per category for the two data sets used in this thesis. The category description is taken from the original publication (UFBA-UESC Dental Images data set [JFR*18]).

Category	**Description**	**Data set 1**		**Data set 2**		**Average number**
		train	test	train	test	**of teeth**
1	Images depicting **all the teeth**, containing teeth with restoration and with dental appliance	0	0	16	7	32
2	Images depicting **all the teeth**, containing teeth with restorations and without dental appliance	0	3	140	35	32
3	Images depicting **all the teeth**, containing teeth without restorations and with dental appliance	0	1	32	8	32
4	Images depicting **all the teeth**, containing teeth without restorations and without dental appliance	4	0	72	21	32
7	Images depicting **missing teeth**, containing teeth with restorations and with dental appliance	0	2	28	7	27
8	Images depicting **missing teeth**, containing teeth with restorations and without dental appliance	5	2	102	25	29
9	Images depicting **missing teeth**, containing teeth without restorations and with dental appliance	0	3	12	2	28
10	Images depicting **missing teeth**, containing teeth without restorations and without dental appliance	3	1	28	6	28
Σ		12	12	432	111	

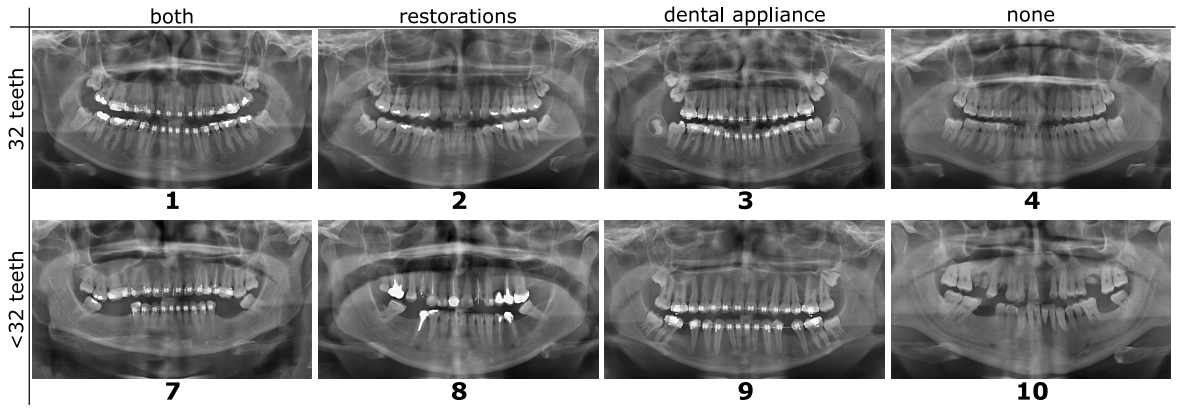


Figure 5.2.: Image categories for panoramic radiographs. The top row shows the categories containing all 32 teeth with/without restoration/dental appliance (left to right: both, restorations, appliance, none). Bottom row shows the corresponding categories with less than 32 teeth.

Image categories The image categories used in this thesis are based on the categories used by Silva et al. [SOP18] for their UFBA-UESC and DNS data sets and will be described in the following. The images of our in-house data set have also been categorized according to this scheme. An overview of the eight categories is given in Table 5.1 which also lists the amount of images in each category – separated by test and training set for each of the two data sets. The category numbers relates to the numbering used in the original publication of the UFBA-UESC data set [SOP18]. Categories 5 and 6 of this data have been omitted by Silva et al. for their DNS data set. These categories includes images with dental implants and image depicting more than 32 teeth – both cases are considered to be beyond the current scope of the tooth numbering task. The categories 1 through 4 cover images which contain all 32 teeth, images where this is not the case are included in the categories 7 through 10. Exemplary images for each category are depicted in Fig. 5.2. Each group of four categories then distinguishes between different combinations regarding the presence or absence of restorations and dental appliance. The term restorations refers in this context to fillings, inlays, bridges, or artificial teeth. Categories 2 and 8 cover such images and some restorations can be seen in the corresponding exemplary images (cf. Fig. 5.2). Dental appliance indicates the use of metal-wired braces (cf. Fig. 1.5). Images containing these are included in categories 3 and 9, exemplary images are again shown in Fig. 5.2. Of the remaining categories, 1 and 7 include images which show both restorations and dental appliance while images containing neither of both are put in categories 4 and 10.

Category 4 can be considered the easiest one as the images contains all teeth and no artifacts. However, these images still suffer from low image quality, low contrast and variations in the teeth's shape, orientation and position in between patients. Category 10 also does not include artifacts, but missing teeth already significantly increase the difficulty of matching the correct tooth model to the corresponding tooth in the image, i.e segment and number the present teeth correctly. Additionally, missing teeth increase the challenge for the automatic initialization of the coupled shape model. The difficulty in-

creases further for the other categories because of the presence of image artifacts caused by restorations and/or dental appliances. Category 7 can therefore be considered the most difficult category.

5.2. Metrics

The segmentation and numbering performance is evaluated using the metrics introduced in this section. In case of 2-D image segmentation, common evaluation metrics are based on the confusion matrix [MBP*21, WWZ20]. Any segmentation problem can also be interpreted as a (binary) classification problem of assigning the correct class label of either 'object' (or 'foreground') or 'background' to every pixel in the image. The confusion matrix is used to accumulate the results for the different prediction categories which are computed by comparing the predicted label to the true (ground truth) label. The result of this comparison can be one out of four categories:

- **True Positive (TP):** the pixel was correctly classified as 'object', i.e. predicted and true label are both 'object',
- **True Negative (TN):** the pixel was correctly classified as 'background', i.e. predicted and true label are both 'background',
- **False Positive (FP):** the pixel was falsely classified as 'object', i.e. predicted label is 'object' but true label is 'background',
- **False Negative (FN):** the pixel was falsely classified as 'background', i.e. predicted label is 'background' but true label is 'object'.

By counting all occurrences of the four categories and denoting them in matrix form, the confusion matrix is generated. Using these four categories of prediction results, several well known metrics can be defined to provide useful information about the segmentation performance. Fig. 5.3 depicts a visual representation of the confusion matrix and its connection to the metrics used for evaluation which will be explained in the following.

The Accuracy of the segmentation is described by the number of correct predictions over the number of total predictions. It is calculated as

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN}. \tag{5.1}$$

While Accuracy might be the most intuitive metric to measure the performance, it is very problematic in cases with a high class imbalance. Class imbalance refers to the problem that the number of 'object' and 'background' pixels might be vastly different. Therefore, predictions for one class are more likely to occur compared to the other which can greatly affect the reliability of the Accuracy measure. For example, imagine a small object that only accounts for approximately 5% of the pixels in an image. Classifying the complete image as 'background' will result in an accuracy of 95% even though no object is actually segmented. For this reason, multiple metrics or combinations of metrics are typically provided to enable an accurate assessment of the segmentation performance.

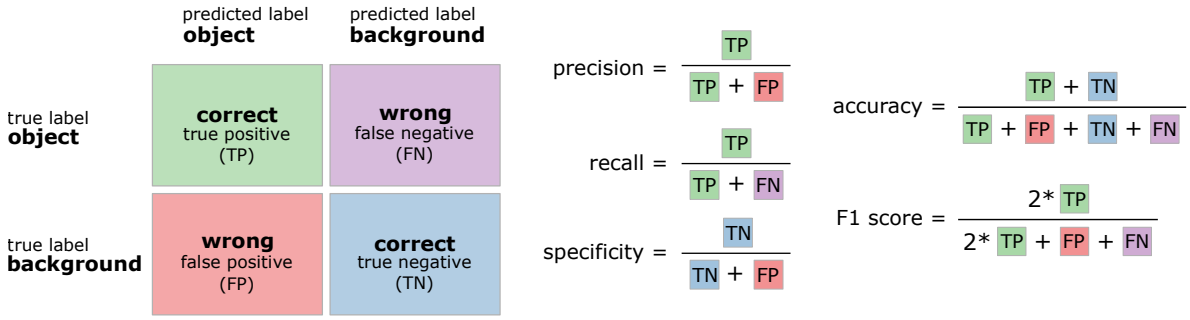


Figure 5.3.: The confusion matrix (left) and the utilization of its four classes for different performance metrics (right). The colored boxes in the metric formulas correspond to the individual classes of the confusion matrix with the same color.

Recall, also known as Sensitivity, expresses the ability of the approach to capture all occurrences of the 'object' within the image in the segmentation result. In other words, an approach with a high recall/sensitivity reduces the number of pixels which are incorrectly classified as 'background' and therefore increases the confidence of the 'background' predictions. Recall is defined as

$$recall = \frac{TP}{TP+FN}. \tag{5.2}$$

Specificity describes the same concept but focuses on the 'background' instead. Similarly, approaches with high Specificity increase the confidence of the 'object' predictions. It is defined as the portion of background predictions which are correctly labeled:

$$specificity = \frac{TN}{TN+FP}. \tag{5.3}$$

The Precision of an approach describes how reliable the predictions for the 'object' class made by the approach are. It is defined as the ratio of correct object prediction over all objects predictions:

$$precision = \frac{TP}{TP+FP}. \tag{5.4}$$

Precision and Recall are generally considered to be a trade-off. Increasing the precision value typically means only assigning the 'object' label to pixels with a high confidence. This reduces the overall number of pixels which are classified as 'object' but typically removes more false positive predictions than true positive predictions, resulting in an increased precision. On the other hand, more pixels are then falsely classified as 'background' (i.e. true positive with a low confidence will be converted to false negatives) leading to a reduced recall value. For this reason, the harmonic mean of precision and recall is therefore commonly used as a measure to assess the performance of a segmentation method. This harmonic mean of precision and recall is referred to as the F1 score and is calculated as follows,

$$F1\ score = 2 * \frac{recall * precision}{recall + precision} = \frac{2*TP}{2*TP+FP+FN}. \tag{5.5}$$

In case of binary data, the F1 score is equal to the dice coefficient which is another typical segmentation performance measure mainly used in 3-D (cf. Section 6.3.2):

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}. \tag{5.6}$$

The metrics accuracy, specificity, precision, recall and F1 score will be used in the following sections to assess the performance of the proposed segmentation and numbering approach on the two data sets by comparing the generated tooth-specific segments with their ground truth. Metrics for a single test instance are calculated by averaging the results for all teeth of that instance. Average results for multiple test instances are given by the arithmetic mean of all individual teeth results of these instances (to put the same weight on each tooth in case instances are not containing all 28 teeth). Standard deviation from the mean over the population of tooth results is provided as an additional performance information.

The numbering aspect of the proposed method can also be interpreted as an object detection problem. Object detection (in contrast to object segmentation) normally focuses on localizing the object within the image by means of a bounding box, instead of a pixel-wise classification. In order to determine the validity of a detection, the overlap between ground truth bounding box G and predicted bounding box P is typically used [PNdS20]. This overlap is measured by the Intersection over Union (IoU), also known as Jaccard Index [Jac01]:

$$IoU(G,P) = \frac{|\mathrm{P} \cap \mathrm{G}|}{|\mathrm{P} \cup \mathrm{G}|}. \tag{5.7}$$

The IoU ranges from 0.0 to 1.0, and a threshold α is used to distinguish a valid detection ($\alpha > IoU(G,P)$) from an invalid one ($\alpha \leq IoU(G,P)$). The value of the threshold is chosen based on the application, typically 0.5. A higher threshold demands a more precise localization of the object. Since segmentation and numbering are closely connected in this approach, a similar approach to Hatiharan el al. [HAGM14] is used. The computation of the IoU is done based on the pixel-wise classifications (confusion matrix) instead of using the bounding box around the segmented object. The IoU is therefore given by

$$IoU = \frac{TP}{TP + FN + FP}. \tag{5.8}$$

The ratio of correct detection will be provided for thresholds of $IoU > 0.5$, $IoU > 0.7$ and $IoU > 0.9$.

5.3. Coupled shape models

A coupled shape model – as introduced in Section 4.2 – is used to capture the shape variation and spatial relations of the individual teeth in dental panoramic radiographs. The coupled model used for the task of teeth segmentation and numbering consists of $N = 28$ deformable model items representing 28 of the total 32 teeth. The included teeth are illustrated in Fig. 5.4. In the image, the number of each tooth corresponds to the FDI notation system and each color indicates a different type of tooth (i.e. central incisor or first molar). The same color scheme is used throughout this chapter whenever the model

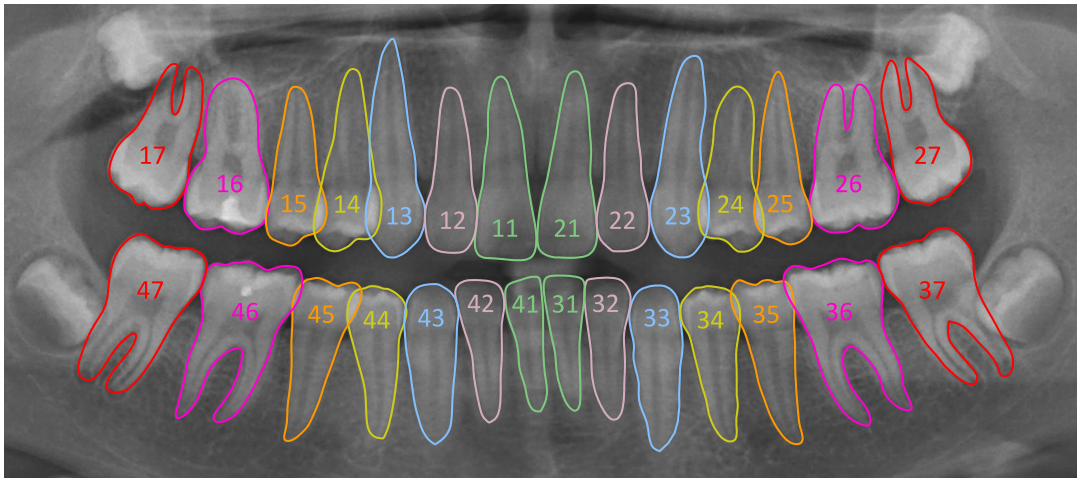


Figure 5.4.: Teeth included in the coupled shape model for panoramic radiographs. Each tooth is represented by a statistical shape model within the coupled model. The number refers to the FDI notation system and the each color indicates a different tooth type, like central incisor (green) or second molar (red).

is depicted. Wisdom teeth have been omitted from the model due to the implications their inclusion would have on model training and initialization. The training process for the coupled model requires the training data to contain all structures included in the coupled model in order to extract the statistical information about the spatial relation and generate the model configurations (cf. Section 4.2.2). Due to the fact that many people are missing some or all of their wisdom teeth, this would greatly reduce the amount of available training data. Moreover, the much larger variation in the shape and orientation of wisdom teeth compared to other teeth would demand additional training data to cover the space of possible variations.

Offline training To capture the shape variation, the contour of an individual tooth is represented by $L = 100$ landmark points and denoted as vector $c = (x_1, y_1, \ldots, x_{100}, y_{100})$. The shape training for each model item is performed as described in the previous chapter (cf. Section 4.2) by first establishing landmark correspondences between all training instances based on curvature and centroid-distance. Then, all instances are aligned using generalized procrustes analysis to remove any shape variation related to scale, rotation or translation. Finally, the statistical information is extracted using Principal Component Analysis to generate the SSM. The coupled model is then trained by capturing the spatial relations between model items, combining the spatial and shape information of all model items and performing PCA on the model configuration vectors.

Any possible model configuration of the teeth in panoramic radiographs can then be described by a vector b as

$$\tilde{f} \approx \bar{f} + A \cdot b, \tag{5.9}$$

where $\bar{f}$ is the mean configuration and A is the matrix containing all eigenvectors of the covariance matrix of all possible configurations (cf. Eq. (4.4)). For more details see Section 4.2.2.

The coupled shape model has been trained on two different data sets of largely different size. The coupled model for data set 1 has been trained on the 12 images in the training set while the model for data set 2 has been trained on the 432 training images of that set. Training images of data set 2 which did not contain all the required 28 teeth have only been used for training the shape models of the individual model items but not for spatial relations.

Fig. 5.5 depicts the first two modes of variation for both models. As a reminder, the modes of variation refer to the principal components (eigenvectors) of the underlying shape space describing shape variance in descending order (cf. Section 4.2.1). Due to the larger amount of training data in data set 2, the teeth shapes in the second model are more smooth as fine details have been averaged out and are hidden behind the much larger amount of modes of variation. The larger amount of modes also results in smaller changes for individual modes as the statistical information is more evenly distributed, i.e. the space of possible configurations is more densely populated because of the larger sample size. The difference in overlap between neighboring structures is mostly related to a different approach to labeling of the gold standard segmentation in both data sets. While the annotations in data set 2 do not overlap at all (i.e. in case of overlapping teeth contours in the image, one tooth takes preference and effectively occludes the other tooth), the annotations of data set 1 include overlaps if these are visible in the image.

5.4. U-Net & binary mask generation

The extraction of the binary mask of the teeth area (i.e. semantic teeth segmentation) for the first step of the online pipeline is performed by the convolutional neural network architecture (modified U-Net) introduced in Section 4.3. The networks used for this application (one for each data set) are trained during the offline training phase based on the respective training data. Since the networks only perform semantic segmentation, the tooth-based gold-standard annotations for each training instance are combined into a single semantic segmentation mask which classifies pixels as either tooth or non-tooth but excludes wisdom teeth. The workflow illustration (Fig. 5.1) depicts the (colorized) teeth instance-based training data on the lower right and the combined binary training data on the lower left.

Offline training The input size of the network has been chosen as half of the original image size while maintaining the original aspect ratio, resulting in an input size of 992x560 pixels. This provided the best compromise between image resolution and memory requirement during training. Input images therefore need to be down-sampled for the extraction of the binary mask, while the computed binary masks are up-sampled to the original resolution for model initialization and adaptation.

Some network and training parameters are chosen based on prior experience as mentioned in the concept. Particularly, 500 epochs with early stopping, 500 steps per epoch, a batch-size of 8, validation split of 0.25, and dropout of 0.3. Image augmentation is applied to increase the number images the

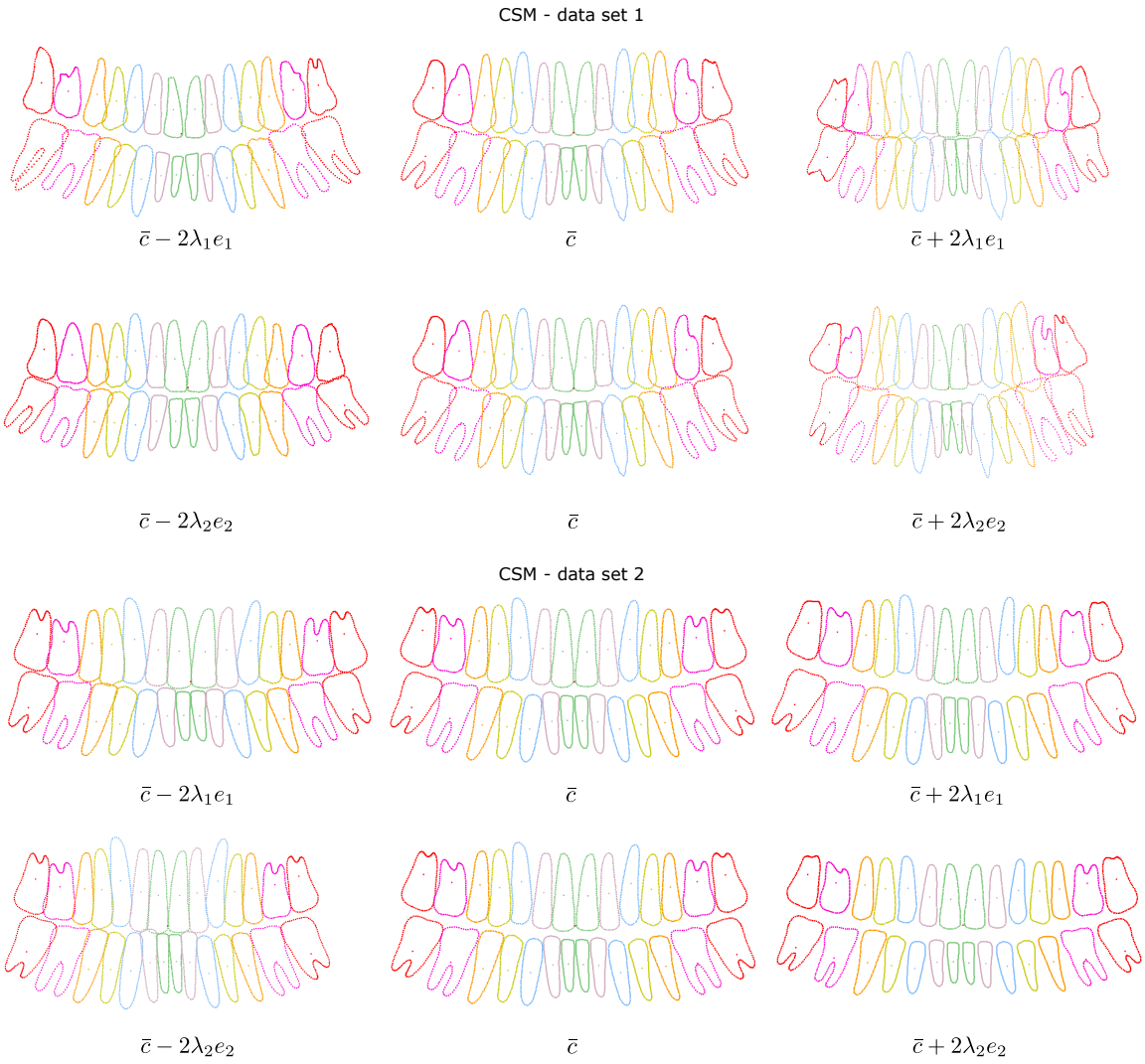


Figure 5.5.: Top: First two modes of variation for the coupled model from data set 1 (12 training images), Bottom: First two modes of variation for the coupled model from data set 2 (432 training images).

Table 5.2.: U-Net training parameter optimization. The table shows different training parameter configurations and the resulting F1 score on the test and validation data of data set 2. The configuration with the highest test score was kept. This configuration was then also used to train a second network on data set 1.

model parameter			training parameter			best	F1 score	F1 score
width	height	filters	loss: α	loss: β	learning rate	epoch	validation	test
992	560	64	0.4	0.6	0.001	48	0.954	**0.923**
992	560	32	0.4	0.6	0.001	40	0.902	0.922
992	560	32	0.3	0.7	0.001	95	0.907	0.922
992	560	64	0.3	0.7	0.001	32	0.950	0.921
992	560	32	0.3	0.7	0.010	127	0.900	0.920
992	560	32	0.7	0.3	0.001	75	0.902	0.915
992	560	64	0.5	0.5	0.001	40	**0.955**	0.913
992	560	32	0.7	0.3	0.010	53	0.902	0.913

network is trained on in each epoch to 4000. These parameters were not optimized because the given set of parameters already provided results sufficient for model initialization and initial adaptation.

Other parameters, like learning rate or loss parameters, are evaluated to determine the best configuration since they have a larger impact on segmentation performance. The evaluation is performed on the larger data set (data set 2) by repeating the training process for different parameters configurations and evaluating the performance on the separate test set. The training was conducted on a GPU-cluster with 2 x 16 core Intel Xeon Gold 6130s, 260GB of random-access memory and 8 x Nvidia GeForce RTX 2080 Ti GPUs with 11GiB graphics memory each. The results of these tests are presented in Table 5.2. A Tversky loss with parameters $\alpha = 0.4$ and $\beta = 0.6$ works best to counter the imbalance between foreground and background area and provides the best F1 score on the test set. A lower learning rate of 0.001 compared to 0.01 also slightly improves the network predictions but has less impact than the loss function. Using double the amount of filters throughout the network significantly increases the F1 score on the validation set but has only a small impact on the test score. This seems to indicate that the increase in network parameters allows the network to be more specialized towards the validation set without gaining any generalization performance. Typically, dropout during training should be increased in these scenarios to reduce the bias towards the validation set. For this scenario, however, the achieved results were considered robust and accurate enough for the intended task.

Binary mask generation The best performing parameter configuration with a test F1 score of 0.9228 was chosen and also used for the smaller data set (data set 1). Here, a test F1 score of 0.8967 was achieved with a network trained on this data set's 12 training images using the same configuration. Table 5.3 lists more detailed performance measures for each data set.

Table 5.3.: Semantic segmentation performance of the two U-Net networks trained with the optimal parameter configuration on their respective data set. The table shows the chosen performance metrics together with their standard deviation.

	Precision	Recall	Accuracy	Specificity	F1 score
U-Net 1 (data set 1)	0.891 ± 0.040	0.904 ± 0.034	0.974 ± 0.007	0.984 ± 0.008	0.897 ± 0.024
U-Net 2 (data set 2)	0.938 ± 0.041	0.912 ± 0.060	0.971 ± 0.011	0.986 ± 0.035	0.923 ± 0.035

The achieved semantic segmentation results with a F1 score of 0.8967 and 0.9228, for data set 1 and 2 respectively, are accurate and robust enough to enable a good initialization and initial adaption of the coupled shape model. Both networks also manage to correctly omit the wisdom teeth from the binary segmentation mask in the vast majority of cases. Therefore, present wisdom teeth should not greatly negatively impact the model initialization process. Fig. 5.6 shows some panoramic images and their corresponding computed binary mask from both data sets. The first row depicts the aforementioned successful exclusion of wisdom teeth from the binary mask that present in the panoramic image. The second and third row show the best and worst result (in terms of F1 score) for each data set.

Closer inspection of the worst result for data set 2 reveals that the provided gold standard annotation for this case is very inaccurate. In fact, the semantic segmentation generated by the trained U-Net delineates the actual teeth boundary better than the gold standard annotation. The inaccurate gold standard annotation in data set 2 seem to be limited to image category 7. Results for this category are therefore reported worse than they actually are.

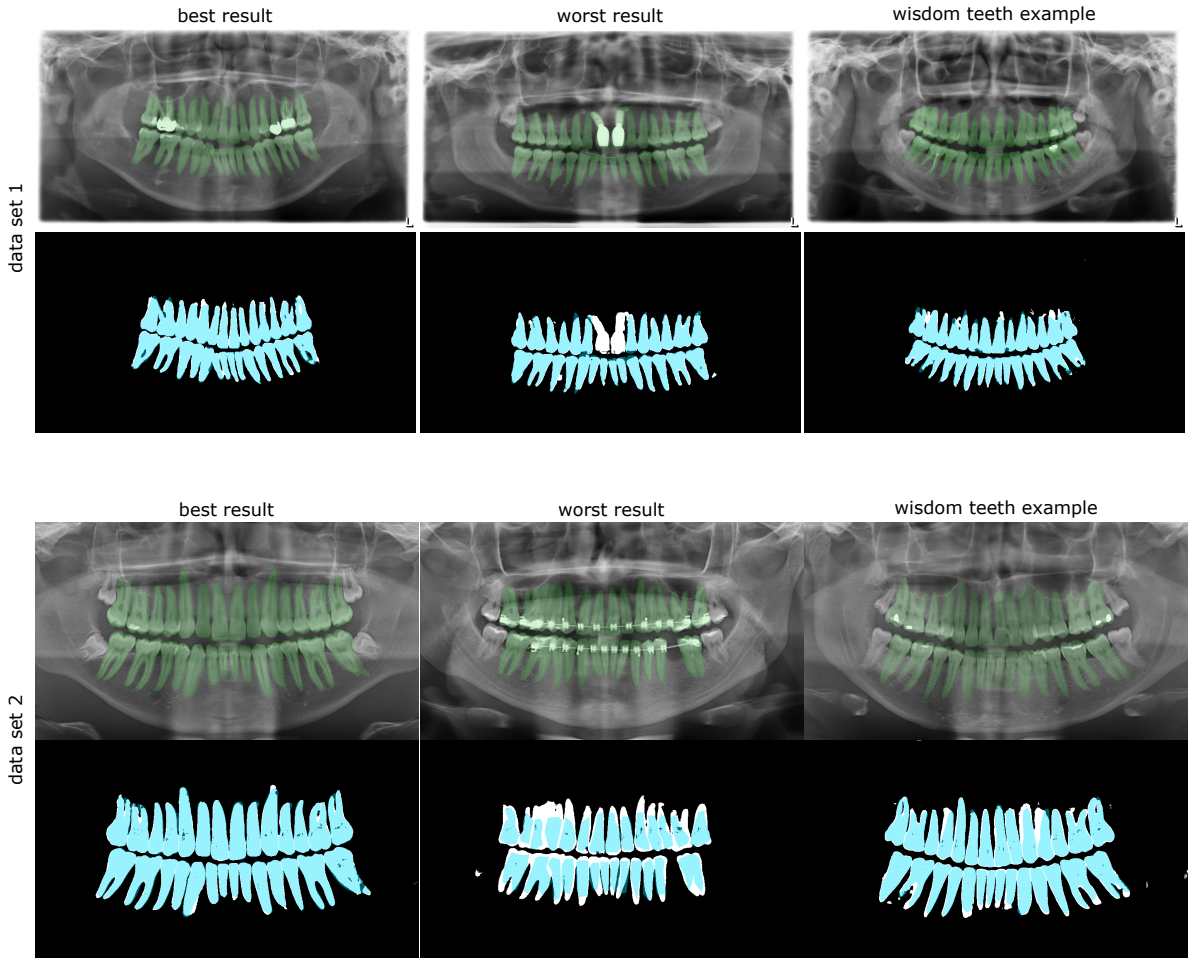


Figure 5.6.: Semantic teeth segmentation results from both data sets. The results are each depicted as a green overlay on the original image and as the computed binary mask with a blue overlay of the ground truth annotation. The first column shows the best result for each data set and the second column the worst result (based on F1 score). The third column contains examples of segmentation results with correctly omitted wisdom teeth.

5.5. Model initialization

After computing the binary mask of the teeth area (excluding wisdom teeth), the trained coupled shape model needs to be placed onto the input image. For this, the binary mask is utilized to compute the initialization parameters in terms of position, rotation and scale. This initialization process is explained in this section. As explained in the concept chapter (cf. Section 4.4), the initialization focuses on correctly placing selected structures of the coupled model – *initialization structures* – rather than trying to maximize the mean overlap of all structures. In case of panoramic radiographs, the four central incisors are the most natural choice as initialization structures. Due to the symmetric structure of the teeth area, these teeth are typically located close to the center of gravity (CoG) of the teeth area with upper and lower jaw being more or less aligned, which allows for an easier estimation of their initial position. In contrast, selecting the outer molar teeth for initialization would require the estimation of four positions as upper and lower molars are less often aligned while the two pairs are also at opposite ends of the teeth area. Variations in their positions will also make it more difficult to achieve good overlap between the coupled model in its mean configuration and all four molar teeth.

The initialization parameters for the model position therefore require the extraction of the center point between all four central incisors. The scale parameter aims ensure a similar size between the teeth area in the image and the coupled model, while the rotation parameter ensures a similar orientation.

For the reference position, instead of simply computing the center of mass of the binary teeth area, which can be easily affected by missing or broken teeth and unusual teeth configurations, the x- and y-coordinates are computed separately. The x-coordinate is calculated as follows, which is also illustrated in Fig. 5.7). First, an axis-aligned bounding box around the teeth area in the binary mask is computed – depicted by the cyan box. Next, an inverted mask is generated by flipping all pixel values withing the bounding box region. Then, the center of gravity of the foreground (white) pixels in both the normal and the inverted bounding box region is calculated. The final x-coordinate is given by the average of the x-coordinates of both CoGs. The reason behind this approach is that missing teeth can shift the CoG more towards the opposite side. By also considering the inverse mask, this shift can be offset and more robust results are achieved.

The y-coordinate is determined using a horizontal projection to identify the location of the gap valley between upper and lower jaw. A small region of interest at the center of the image is defined and shifted down by about 10% of the image height. This decision is motivated by the fixed setup that is used to capture panoramic radiographs. It results in limited variance of the vertical position of the teeth in the image. The chosen size of the region of interest was determined empirically. Within this region, the horizontal projection is applied and the minimum in the resulting histogram defines the location of the separation line between upper and lower jaw (cf. Fig. 5.7), which is used as the y-coordinate.

The scale value is initially approximated by the ratio between the size of the bounding box of the binary mask and the bounding box of the mean model. This value is then refined by placing the model according to the previously computed position and maximizing the overlap between mask and model. The rotation of the coupled shape model was left unchanged as the symmetric structure of the dental

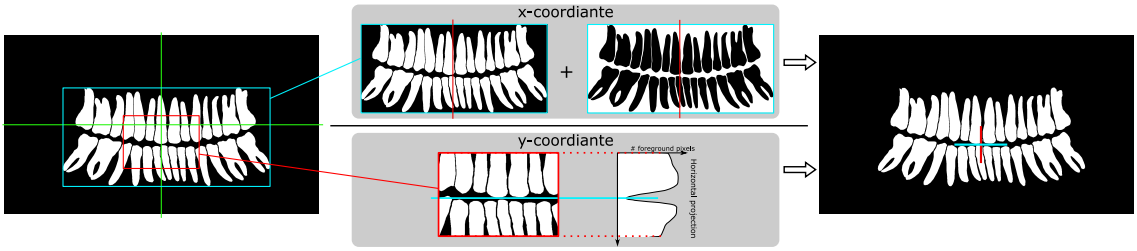


Figure 5.7.: Calculation of reference position for model initialization. First, the bounding box (cyan) of the teeth area is computed and the region of interest (red) for the horizontal projection is positioned relative to the image center (indicated by the intersection of the green lines). Afterwards, x- and y-coordinate are calculated. The final reference position is shown in the right picture.

arches did not require any adjustment. Having computed all initialization values, the coupled model can be placed onto the input.

5.5.1. Evaluation & discussion

The performance of the automatic model initialization was evaluated to determine if the model was positioned and scaled correctly. The difficulty for such an evaluation is that there is no ground truth of the "optimal" initialization for a given image available. Instead, a subspace of parameter combinations will result in a valid adaptation. While some combinations will result in better adaptation results, these results are both affected by the initialization as well as the adaptation process, i.e. the design of the adaptation strategy and the parameters of the individual adaptation steps. A more precise initialization allows for a more fine-tuned adaptation process and therefore better adaptation results. However, achieving higher precision requires more specialized methods which increases the risk of a failed initialization on more difficult cases. The goal is to find an approach that generalizes well while still being as precise as possible. Generally speaking, model initialization can be considered successful, if the incisor model items are adapted to the corresponding image structure during the first model initialization step. Consequently, it was considered unsuccessful if the incisor model item were adapted to incorrect structures. The scale value was considered correct, if the distance between left and right molar teeth is approximately the same between the mean CSM and the teeth in the image.

Fig. 5.8 depicts examples for three correctly and three incorrectly initialized coupled shape models. Each image shows the position and scale of the mean CSM after the initialization process. The green landmark points (tooth outlines) correspond to the incisor teeth (cf. Fig. 5.4) and the red overlay displays the ground truth segmentations of the incisor teeth in the image. For a better visualization, the failed cases also include a picture of the CSM drawn onto the binary mask. These cases will be discussed in more detail in the following paragraphs.

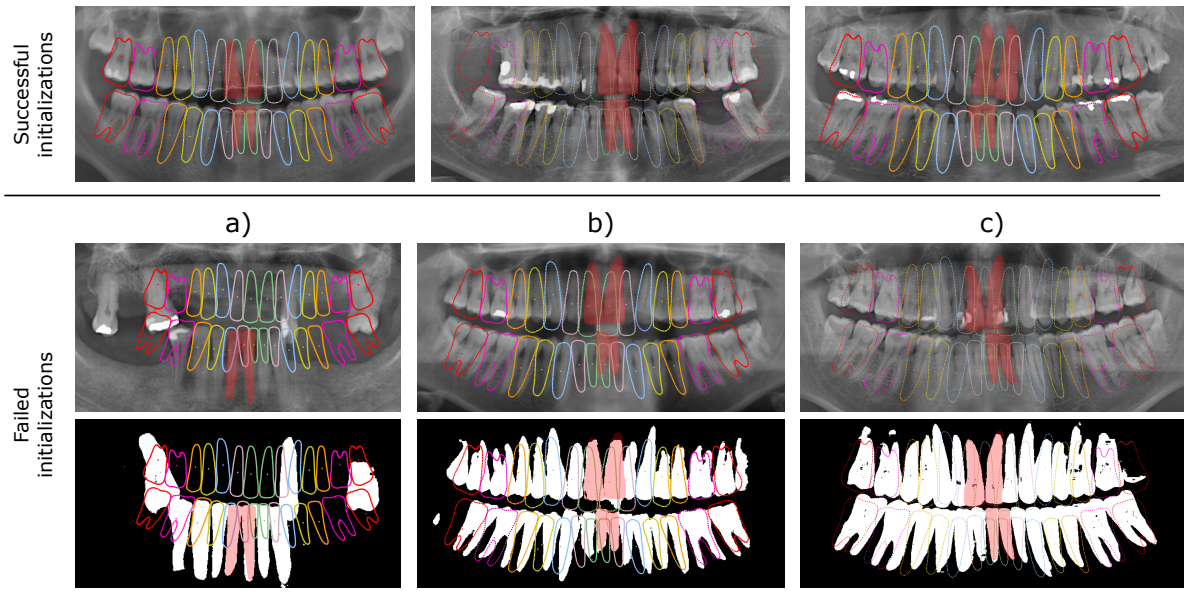


Figure 5.8.: Model initialization results. The top row depicts the coupled shape model after successful initializations. The bottom cases depict failed initializations illustrated on both the original image and the binary mask (semantic teeth segmentation). The incisor teeth are highlighted in all images as red overlays since initializing the model aims to maximize the overlap between these teeth and the corresponding model items (green outlines).

Overall, visually inspecting all test instances manually revealed that initialization was successful for 11 of 12 instances (92%) in data set 1 and 103 of 111 (92%) in data set 2. The different parameters of the initialization will be discussed in detail in the following.

Horizontal position A correct initialization of the coupled model in horizontal direction requires that neither maxillary nor mandibular incisors are offset by more than roughly half an incisors teeth width in either direction. A greater offset will cause them to be adapted to the neighboring tooth instead, leading to a failed adaptation. Since the mandibular incisors have a smaller width compared to the maxillary ones, they are typically defining the possible tolerance in the horizontal offset. Fig. 5.9 depicts two (manual) initializations on the same image with slightly different horizontal offsets to the left (from an ideal position). While the mandibular incisors in the left example (green box) will be adapted to the correct structures in the image (thin red outline), the incisors in right example (red box) are offset too much to the left and will result in an incorrect adaptation. The results after adaptation are displayed next to the initialization examples, indicated by the green (successful) and red (failed) arrows, respectively. Because the width of the mandibular incisors is rather small, the initialization is relatively dependent on the good estimate of the x-coordinate.

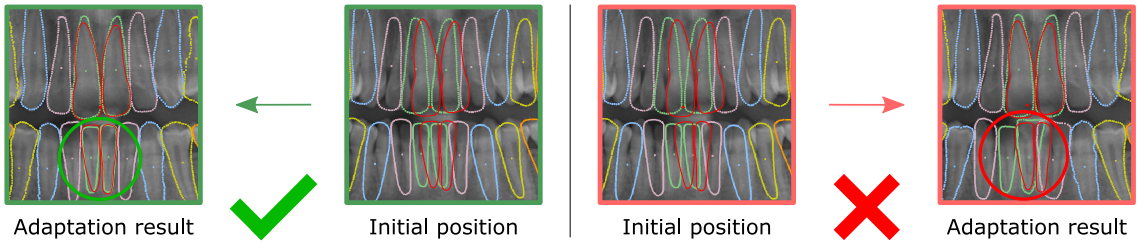


Figure 5.9.: The figure illustrates the amount of deviation in horizontal direction that the coupled model is able to handle after model initialization. The two pictures in the middle show two manual initializations with different offsets to the left (from an ideal position). While the left one results in a successful adaptation of the incisor teeth (model items in green, ground truth in red), the right one fails.

A limitation are cases where upper and lower incisor are not aligned, i.e. the lower incisor are significantly offset (more than a teeth width) in horizontal direction compared to the upper incisors. As the CSM is positioned in its mean configuration (with upper and lower incisors aligned), it will be impossible to find a position in which all incisor are correctly adapted to their corresponding image structure. This scenario occurred in three out of the 111 test cases – even trying to manually initialize the CSM resulted in a failed adaptation for these cases. Fig. 5.8c) depicts such a case where the red overlays depict the ground truth segmentations of the incisor teeth. A successful initialization would only be possible, if the presence and direction of the offset would be known beforehand and the configuration of the CSM could be adjusted to account for that. Other options would be to use two separate models for upper and lower jaw, which would require a different initialization approach.

Missing teeth can also impact the accuracy of the initialization. While a small number of missing teeth can be handled by the approach (cf. Fig. 5.8, top row, second image), a large number of missing teeth is beyond the scope of this approach (cf. Fig. 5.8a)).

Vertical position The initial model position in vertical direction aims to position the CSM in such a way that each jaw overlaps with the respective image structures. For this process, the acceptable tolerance is significantly higher compared to the horizontal position. On one hand, this is because the incisor teeth have a larger extent in vertical direction than in horizontal directions, on the other hand, the vertical initialization is supported by the subsequent adaptation process. The binary mask is used to mask out the non-teeth region in the input image. This results in no image features outside the teeth region, while the detectable image features inside the teeth region help to pull the model in the correct direction. Fig. 5.10 depicts the acceptable tolerance in vertical direction (left and right image), while the center image shows the adaptation result for both initializations after the first adaptation step. The y-coordinate computed by detecting the separation between maxilla and mandible was always well within the acceptable tolerance. Even for cases with multiple missing teeth, the gap valley was detected correctly. As long as the selected region of interest that is used for its detection (cf. Fig. 5.7,

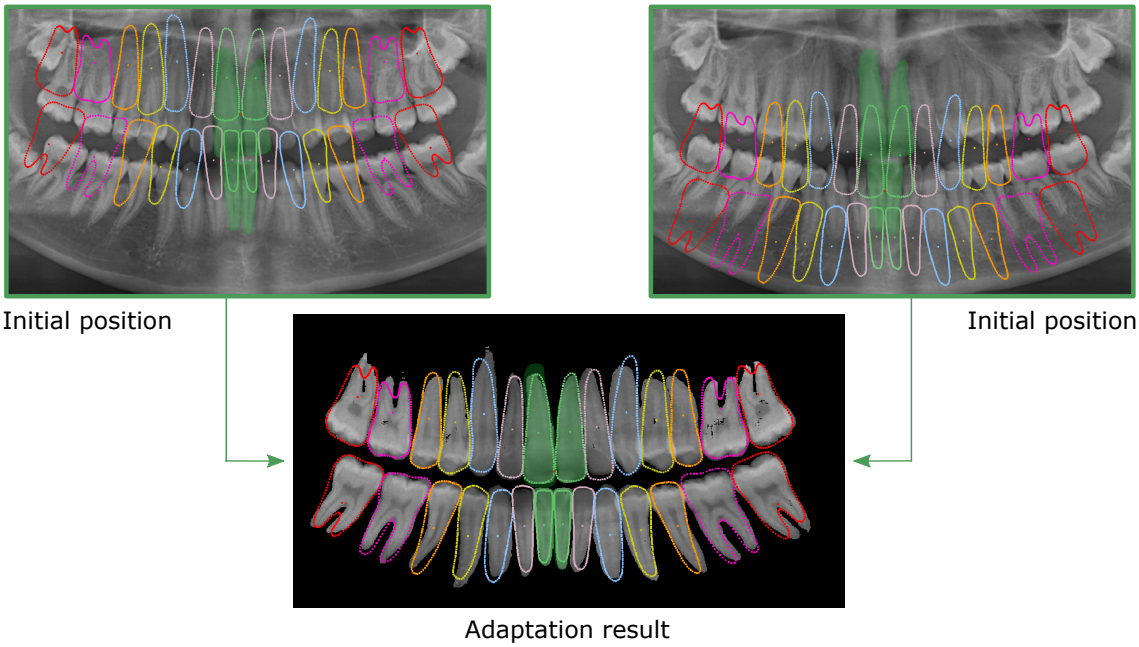


Figure 5.10.: Robustness of the adaptation to variations of the model position in vertical direction. The depicted variation (left and right image) is the maximum amount the model can handle. Since the model adaptation uses a masked image (input image masked by binary teeth mask; middle image) during the first part of the adaptation, it can recover from large deviations as image features will only be found in the masked region.

red rectangle) contains at least one tooth of both upper and lower jaw, the valley can be detected. This was true for all test cases. The vertical position is considered correctly detected for all test instances of both data sets.

Scale The scale value is intended to match the size of the mean coupled model to the teeth in the image. Because of the chosen dynamic adaptation strategy, scaling the mean model too small is less harmful compared to overestimating the scale factor. The reason for this is that the adaptation starts in the center (with the incisor teeth) and gradually increases the set of active structures outwards. In doing so, model items are more likely to be push outwards, compensating a too small scale factor, rather than being pulled inwards to counteract a too large scale factor. On an individual tooth level, underestimating scale is also preferable. There are (normally) almost no image features in the inside area of a tooth. Consequently, a too small tooth model is unlikely to shrink further (assuming its initial position is correct) but instead will grow to match the contours. Contrary, a too large tooth model is easier matched to image features in the surrounding of the actual tooth.

Another important aspect is that the scale estimation depends on the binary mask not containing any wisdom teeth. While the neural network was trained on semantic segmentations which excluded wisdom teeth, the generated binary masks can rarely contain wisdom teeth or parts of them nonetheless. This will result in a scale factor that is too large since the bounding box of the teeth region is used as reference, potentially causing teeth to be adapted to incorrect image features. Fig. 5.8b) depicts a failed scale estimation caused by the maxillary wisdom teeth being (partly) included in the binary mask as can be seen in the rightmost image. Missing teeth can have a similar effect. Depending on the location and amount of missing teeth, the bounding box will be smaller, resulting the in CSM being scaled to small. Fig. 5.8a) shows an incorrect scale estimation caused by too many missing teeth.

5.6. Model adaptation

The final step of the pipeline is the adaptation of the coupled model to the input image using gradient image features in order to extract the teeth contours and numbers. As explained in detail in Section 4.5, the adaptation is done by minimizing an energy functional E which depends on two parameters: (a) the transformation t describing the global position of the model in terms of translation and rotation and (b) the configuration vector f describing the configuration of the coupled model. The functional is given by

$$E_{2D}(f,t) = E_{image}(f,t) + \alpha E_{local}(f) + \beta E_{model}(f). \tag{5.10}$$

Here, the term E_{image} is responsible for ensuring that the contour of model items moves in the direction of strong image features (gradient features), the model energy term E_{model} restricts the model to stay within or close to the learned configuration space and the local energy E_{local} enforces a smooth contour of model items.

As introduced in Section 4.5, the coupled shape model uses dynamic adaptation strategies to ensure an optimal progress of fitting the model items to the respective structures in the image. Therefore, a specialized adaptation strategy for the segmentation of teeth in panoramic radiographs has been designed. The *adaptation strategy* is internally separated into multiple *adaptation steps*. Each step controls which structures are actively adapted to the images, i.e. it defines the element weights $w_i(p)$ for each model item i in the model energy function (cf. Eq. (4.15)):

$$E_{image}(f,t,p) = \sum_{i=1}^{N} w_i(p) E_{image}^{(i)}(f,t). \tag{5.11}$$

An adaptation step signifies the progress p in the adaptation process. Model items that are not adapted during an adaptation step (i.e. $w_i(p) = 0$ in Eq. (5.11)) are only changed passively through the learned (spatial) configuration (i.e. E_{model} in Eq. (5.10)). The intention behind adapting a different set of items at a different stages in the adaptation process is to ensure a close proximity between model items and corresponding image features. The adaptation of selected items will affect the position of other model items due to the learned spatial configuration. Adaptation steps are also used to define certain

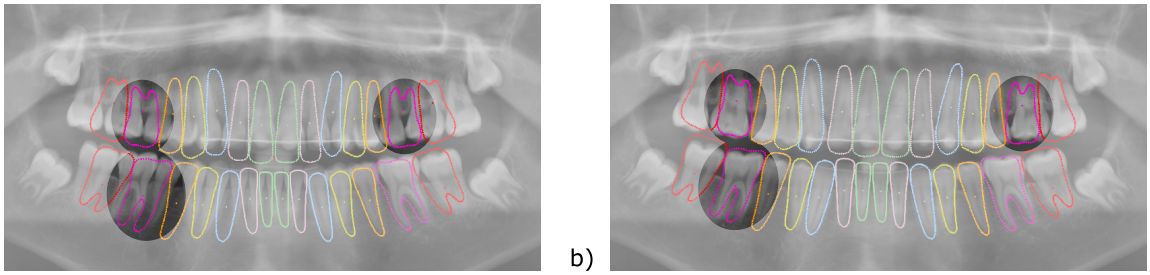


Figure 5.11.: Comparison of molar teeth positions: (a) positions after CSM initialization; (b) positions at the beginning of adaptation step six, during which the molar models are actively adapted for the first time. It is evident that the positions of the molar tooth models are significantly closer to the corresponding structure in the image compared to the position in (a).

adaptation parameters like search radius of the local search algorithm, number of iterations with each step, or disabling scale or model deformation for individual model items.

During model initialization, the CSM is positioned onto the image in such a way that the central incisor teeth of the CSM overlap well with the incisors in the image. Consequently, starting from the initialization structures – the 4 central incisors – the set of model items (teeth) which are actively adapted to the image is gradually extended. The gradual extension of the set of adapted items is done because the mean model is initialized the maximize the overlap of initialization structures with their corresponding structures in the image. Therefore, the central incisors (initialization structures) show good overlap with the input image while the teeth farther away from the center (e.g. molars) might not match as good – depending on the location of the teeth in the input image. The left image in Fig. 5.11 illustrates the discrepancy between the position of the first molar teeth in the image and their corresponding model items after model initialization. The discrepancy can cause the local search algorithm to incorrectly detect features of neighboring teeth as good candidates for the molar tooth model, leading to an incorrect segmentation. By adapting the outer teeth at a later time, they have already been (passively) moved closer to their correct position and more reliable image features will be found once they are adapted. The position of the molar tooth model at the beginning of adaptation step six, when image feature points for the molar contours are first searched, is shown in the right image of Fig. 5.11. It is evident that the position is much closer to the corresponding teeth in the image.

The order in which model items are adapted during the model adaptation is illustrated in Fig. 5.12. Starting from the central incisors, more teeth will be added based on the natural order of teeth from anterior to posterior teeth. A higher step also includes all the teeth already actively adapted in the lower steps. From step 7 onward, all teeth are adapted to the image.

To further increase the robustness of the adaptation, the initial seven adaptation steps fit the coupled model to a masked version of the input image instead of utilizing the complete input image right away. The masked version is created by using the binary mask generated by the neural network. All intensity

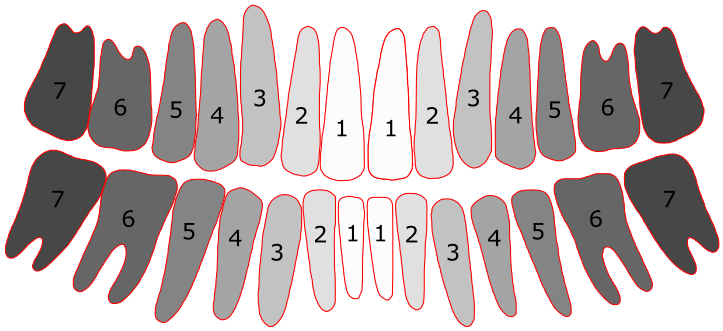


Figure 5.12.: Dynamic adaptation strategy: The numbers in the illustration represent the order in which the groups of model items are adaptation to the image. A higher number includes all items marked with that number as well as all items with lower number - i.e. the third adaptation step includes the canine (3) teeth as well as the central (1) and lateral (2) incisors.

information in the input image which was classified as background in the binary mask is set to zero and only the information classified as teeth is retained. This masked version is then employed to compute the gradient image features used during adaptation. This results in strong gradient features at the edges of the masked region, while the included intensity information also provides features inside the masked region. This approach ensures a good placement of all teeth before using the intensity image. The final adaption steps are then performed on gradient features computed from the complete (un-masked) intensity image while adapting the entire model. The adaptation concludes with a free-form deformation where each model item is no longer constrained by the CSM but instead is adapted independently of all other items. The items are then only constrained by their individual statistical shape models, but to a much lesser extent compared to the CSM. This step is performed to ensure an optimal adaptation of each tooth model to its respective image features. Constraints cannot be completely removed because of the challenges introduced in the beginning, like low contrast, noise and artifacts (dental restorations).

Fig. 5.13 illustrates the adaptation progress on an exemplary image. Starting form the automatic initialization (first image), each subsequent image depicts the configuration of the coupled shape model after the next step in the adaptation strategy, i.e the second image shows the CSM after adaptation step one, the third image after step two, and so on. The last image contains the final configuration of the coupled model and therefore the segmentation and numbering result. The segmentation result of each individual tooth is then stored as a binary image.

After explaining the process of adapting the initialized coupled teeth model to the panoramic radiographs in this section, the adaptation parameter estimation and the results of this final adaptation step – and therefore the final teeth segmentation and numbering results of the presented approach – are discussed on the next section.

5.7. Experiments and results

The parameter optimization and performance evaluation of the final pipeline step, the model adaptation, was done on the two data sets introduced in Section 5.1. The optimization of the adaptation parameters was performed on the respective training data subset while the evaluation of the segmentation pipeline including the final step, and therefore the final segmentation and numbering performance, was performed on the unseen test subset.

Adaptation parameters The training data of both sets was utilized to determine the best parameters for the adaptation strategy for each data set. The main parameters are the included model items, the search radius for the local search strategy (cf. Section 4.5), the maximum number of iterations and the weights of the individual terms of the energy functional that is minimized during model adaptation (cf. Eq. (4.7)). More parameters are used internally, but the listed 'main' parameters have the largest impact on segmentation performance. Each step of the adaptation strategy has its own set of parameters, but the optimal values for the parameters of each step are closely related. The two major differences between the adaptation steps that impact the parameter values are the set of included model items (cf. Fig. 5.12) and the image type that is utilized for the step (either the masked or non-masked input image, cf. Fig. 5.13). To ensure that failed automatic initializations (on the training data) do not negatively impact the estimation of the adaptation parameter values, the CSM was manually placed onto each training image for the optimization process. The value for each parameter of each adaptation step was determined empirically by trying to maximize the segmentation performance (F1 score) on the individual training set.

Results After having determined all parameters, the two trained instances of the fully automatic teeth segmentation and numbering approach (one for each data set) were evaluated on their corresponding set of 12 and 111 test images. The performance was assessed by comparing the computed tooth-specific (numbered) segmentations to manually generated gold standard annotations in a pixel-wise fashion using the metrics described in Section 5.2. These metrics are Precision, Recall, Accuracy, Specificity and F1 score. The overall measures are calculated by averaging the tooth-specific metrics of all test cases. Standard deviation from the mean is provided as an additional measure for all metrics to indicate the amount of variation over all test cases.

Table 5.4 summarizes the performance of the proposed approach for both test sets. The entries referring to *all cases* indicate the overall performance of the entire pipeline on the respective test set and include the segmentation results of cases with failed initializations. The proposed approach achieves a F1 score of 0.833 ± 0.108 on the small data set and 0.823 ± 0.189 on the large data set. Since the performance of local search-based algorithms are known to be initialization dependent, additional entries are provided showing the performance only for cases with successful and failed initializations, respectively. The eight test instances with failed initializations in data set 2 show significantly reduced segmentation performance with a F1 score of 0.372 ± 0.264, the single failed case on data set 1 evalu-

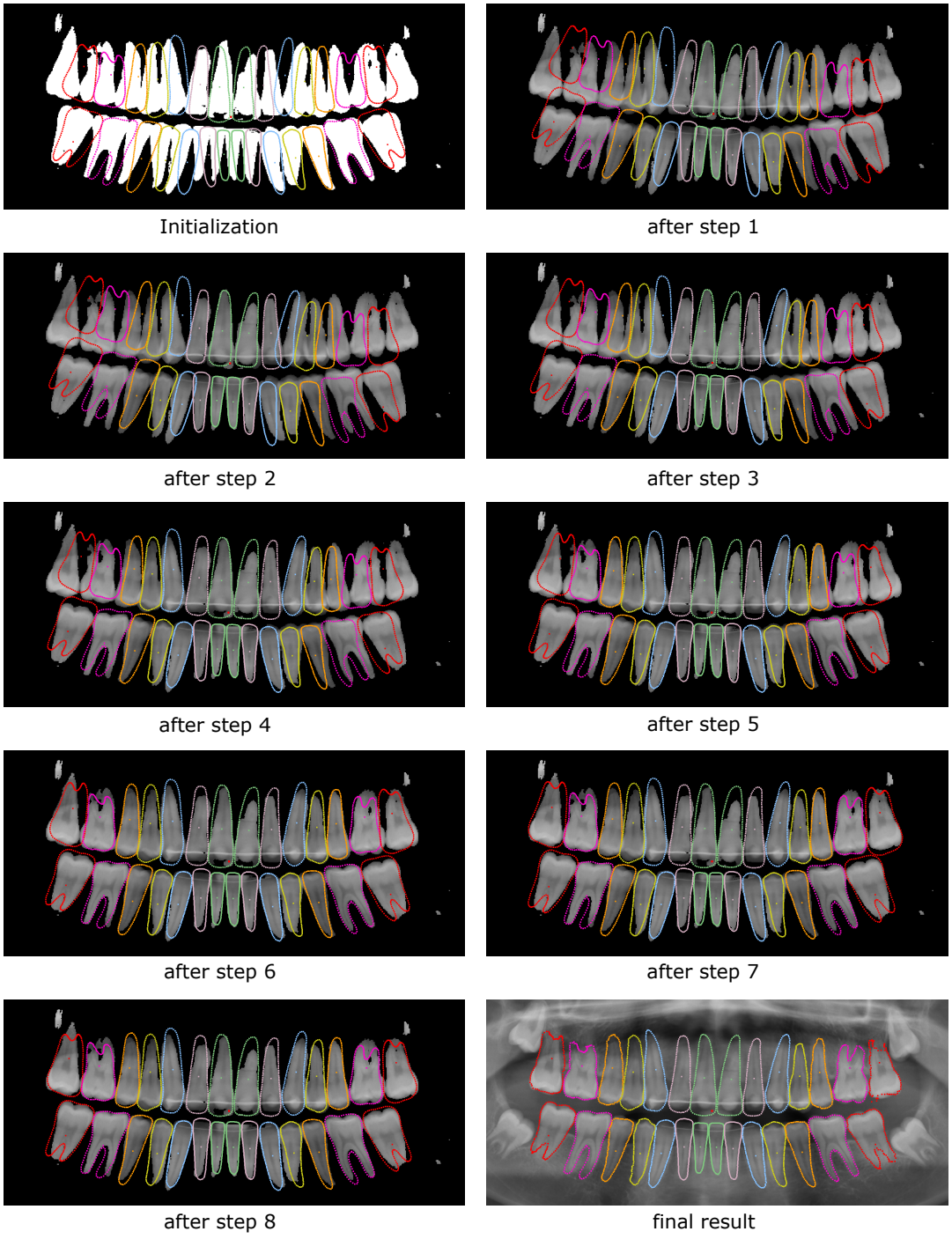


Figure 5.13.: Coupled shape model configuration after each step in the adaptation strategy. Top left: After model initialization, bottom right: final result. (Original images haven been cropped for better visualization.)

Table 5.4.: Evaluation of the teeth segmentation and numbering approach on two different data sets. Separate U-Net networks and CSMs were trained for each data set, adaptation strategy parameters were individually optimized as well. The table contains the overall performance on all cases of the test set as well as separate entries for the average performance in case of successful and failed automatic initializations.

	images	Precision	Recall	Accuracy	Specificity	F1 score
data set 1 (all cases)	12	0.847 ± 0.107	0.827 ± 0.110	0.999 ± 0.001	0.999 ± 0.001	0.833 ± 0.108
data set 1 (successful init.)	11	0.875 ± 0.054	0.854 ± 0.069	0.999 ± 0.001	0.999 ± 0.000	0.861 ± 0.060
data set 1 (failed init.)	1	0.536 ± 0.000	0.535 ± 0.000	0.996 ± 0.000	0.998 ± 0.000	0.527 ± 0.000
data set 2 (all cases)	111	0.808 ± 0.194	0.851 ± 0.202	0.998 ± 0.003	0.999 ± 0.001	0.823 ± 0.189
data set 2 (successful init.)	103	0.834 ± 0.066	0.876 ± 0.080	0.998 ± 0.001	0.999 ± 0.001	0.849 ± 0.066
data set 2 (failed init.)	8	0.355 ± 0.258	0.399 ± 0.271	0.993 ± 0.003	0.996 ± 0.001	0.372 ± 0.264

ates to a F1 score of 0.527. Consequently, the performance improves if the failed cases are omitted and only successfully initialized cases are considered for the evaluation.

The overall segmentation and numbering performance per image category on each data set is shown in Table 5.5. The number of images belonging to a specific category within each data set is denoted as well to illustrate the significance of the individual category and the impact on the overall result. These measures include all results regardless of the initialization outcome. For data set 1, categories 1 and 4 have no measures as there are no images in the test set that belong to these categories.

Visual representations of exemplary segmentation results are depicted in Fig. 5.14. It depicts the best and worst test cases in terms of average F1 score for each data set. Each depicted result includes the final positions of the landmark points of each model item colorized by tooth type on the intensity image (left image) and a comparison of the gold standard annotations (filled colorized shapes) and the corresponding extracted teeth contour (colorized contour line) shown in the right image. The color of each tooth corresponds to the CSM design (cf. Fig. 5.4) which also indicates the respective tooth number. Further results depicted in Fig. 5.15. Additional examples of the segmentation behavior in case of missing teeth are illustrated in Figs. 5.16 and 5.17.

The numbering performance is assessed using the pixel-wise IoU (cf. Section 5.2) between tooth segmentation and ground truth. Detections are validated based on five different IoU thresholds, with higher thresholds putting more emphasis on accurate pixel-wise classification. Table 5.6 shows the overall results as well as separate entries for cases with successful and failed initialization. Numbering performance in terms of correct detections is 0.937 ± 0.145 on data set 1 and 0.906 ± 0.196 on data set 2 for an IoU threshold of 0.5. The ratios of correct detections drop to 0.807 ± 0.193 and 0.760 ± 0.245, respectively, for an IoU threshold of 0.7. Again, performance varies significantly between successful and failed initializations. The presented results will be discussed in the following section.

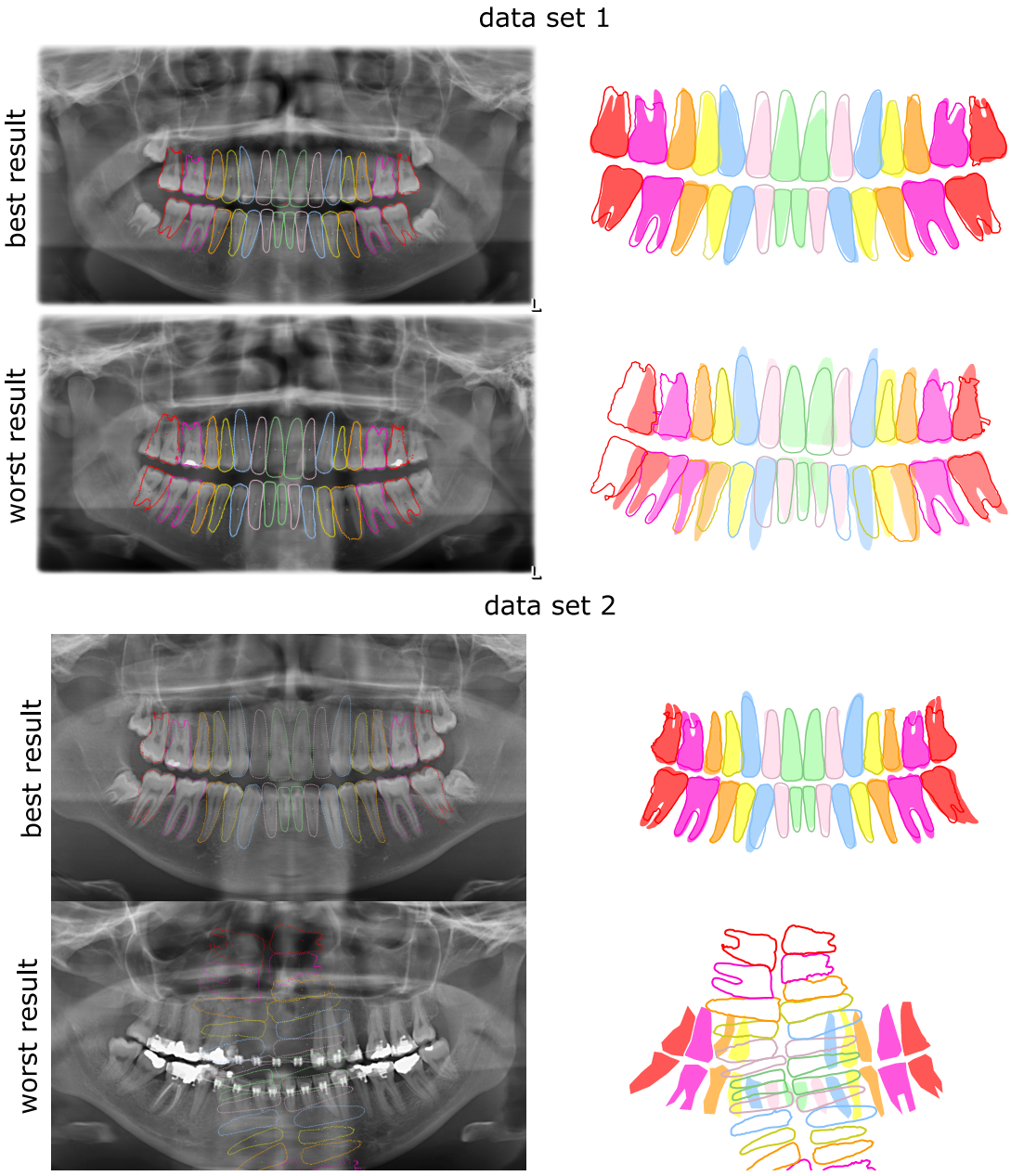


Figure 5.14.: Teeth segmentation (and numbering) results. For each result, the detected tooth boundaries are illustrated as landmarks points on the intensity image (left images) and as contours lines drawn over the colorized gold standard annotations (right images). Contours and annotations of the same tooth are colorized with the same color for each quadrant of the set of teeth (which also corresponds to the introduced model design).

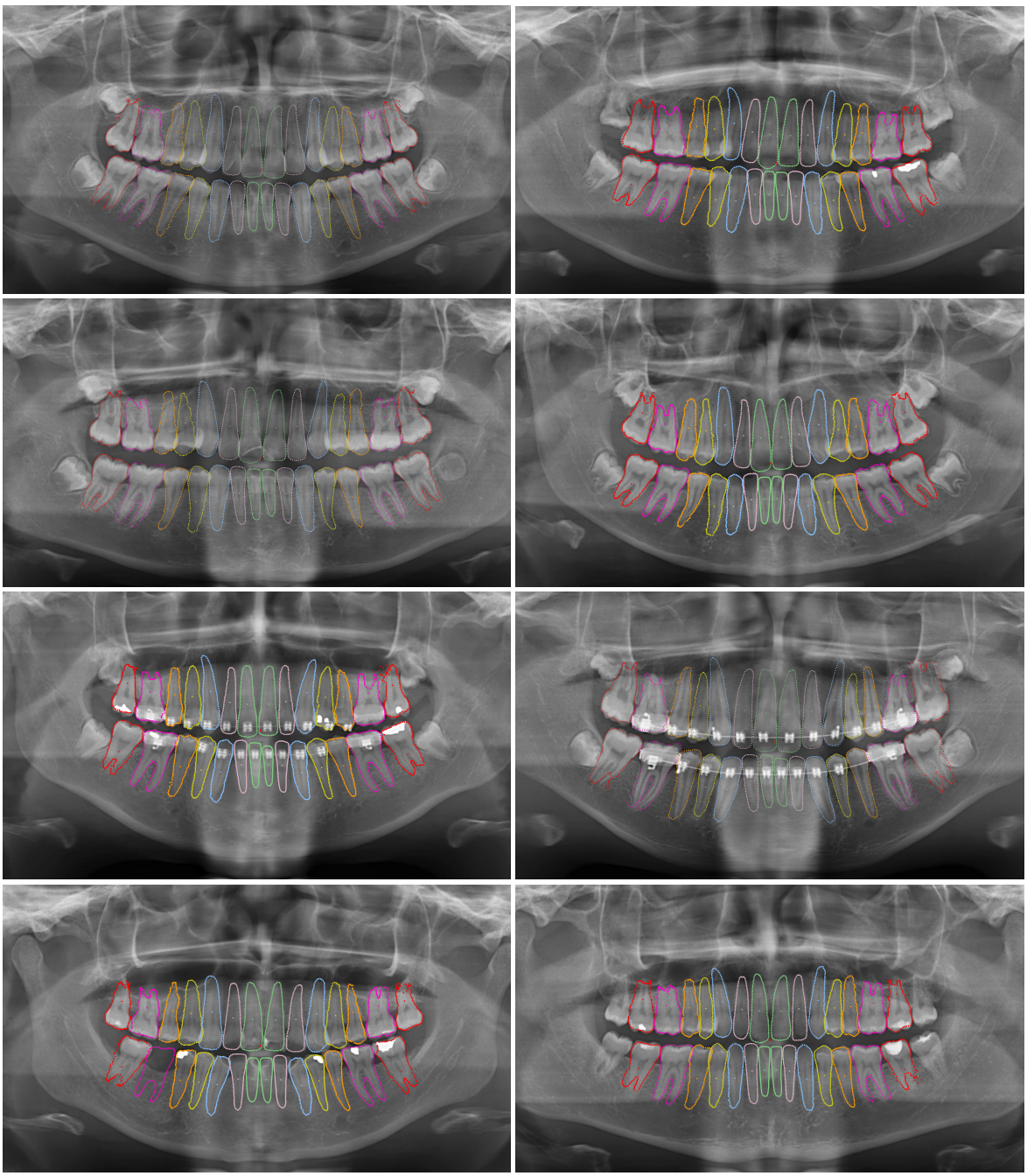


Figure 5.15.: Collection of teeth segmentation (and numbering) results. The detected contours are illustrated as landmarks points on the intensity image. The contour color indicates different tooth types and numberings for each quadrant (cf. Section 5.3).

Table 5.5.: Evaluation of teeth segmentation and numbering performance for each image category from the two test sets. For details about each category please refer to Section 5.1.

	data set 1		data set 2	
category	images	F1 score	images	F1 score
1	0	-	7	0.874 ± 0.046
2	3	0.751 ± 0.159	35	0.861 ± 0.077
3	1	0.905 ± 0.000	8	0.824 ± 0.113
4	0	-	21	0.872 ± 0.060
7	2	0.887 ± 0.005	7	0.635 ± 0.126
8	2	0.873 ± 0.009	25	0.728 ± 0.162
9	3	0.881 ± 0.013	2	0.880 ± 0.047
10	1	0.679 ± 0.000	6	0.810 ± 0.106

Table 5.6.: Evaluation of the teeth numbering on the two data sets using different pixel-wise IoU thresholds for correct detections. The table contains the overall performance on all cases as well as separate entries for the performance in case of successful and failed initializations.

	images	IoU > 0.5	IoU > 0.6	IoU > 0.7	IoU > 0.8	IoU > 0.9
data set 1 (all cases)	12	0.937 ± 0.145	0.916 ± 0.169	0.807 ± 0.193	0.456 ± 0.151	0.016 ± 0.024
data set 1 (successful init.)	11	0.977 ± 0.064	0.954 ± 0.107	0.844 ± 0.155	0.479 ± 0.138	0.017 ± 0.026
data set 1 (failed init.)	1	0.500 ± 0.000	0.500 ± 0.000	0.393 ± 0.000	0.214 ± 0.000	0.000 ± 0.000
data set 2 (all cases)	111	0.906 ± 0.196	0.867 ± 0.214	0.760 ± 0.245	0.431 ± 0.224	0.024 ± 0.038
data set 2 (successful init.)	103	0.949 ± 0.093	0.907 ± 0.133	0.799 ± 0.188	0.454 ± 0.210	0.026 ± 0.039
data set 2 (failed init.)	8	0.359 ± 0.311	0.335 ± 0.322	0.256 ± 0.318	0.120 ± 0.169	0.005 ± 0.012

5.8. Discussion

The presented fully automatic segmentation and numbering approach for teeth in panoramic radiographs has been trained and evaluated on the two data sets introduced in Section 5.1. The first data set (data set 1) contains 24 images split into 12 train- and 12 test-instances. Data set 2 includes 432 training- and 111 test images. For each data set, a separate U-Net network and coupled shape model were trained on the training images as described in Sections 5.3 and 5.4, respectively. The training data together with the trained models were also used to optimize the model initialization process as well as the parameters of the model adaptation for each data set (cf. Sections 5.5 and 5.6).

The experimental results on the two test set indicate good tooth-specific segmentation and numbering performance with an average F1 score of 0.833 ± 0.108 for the small data set 1 and 0.823 ± 0.189 for

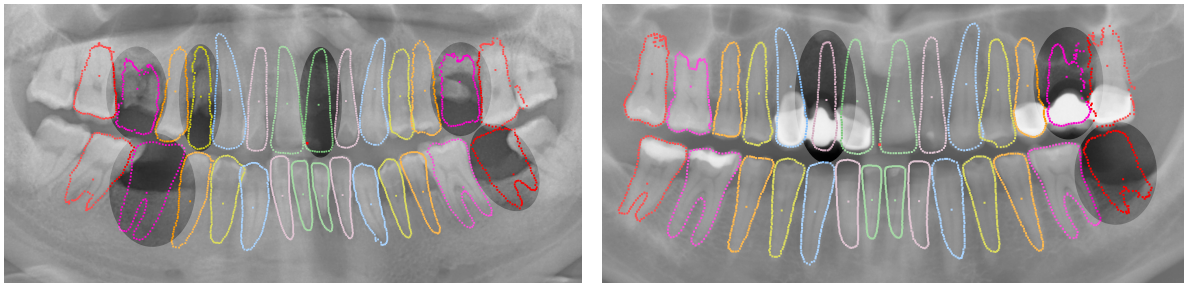


Figure 5.16.: Exemplary cases of successfully handled missing teeth. If the gap left by the absent teeth is still present, the coupled model places the shape model corresponding to the original in the gap (as indicated in the pictures) and adapts the remaining teeth successfully. (Original images haven been cropped for better visualization.)

the second data set. Given a successful initialization, the proposed combination of dynamic adaptation strategy and masked input image (utilizing the binary mask) allows the approach to correctly detect the individual tooth structures in the image that correspond to the model items.

The approach is able to handle the technical challenges outlined in the Introduction chapter for this kind of task. It is able to handle the challenging *image quality and characteristics* of dental panoramic radiographs, like noise or low contrast, as well as the difficult *tooth numbering* by utilizing the shape and spatial information encoded in the coupled shape model in addition to the features extracted from the images. It is able to deal with the inter-patient *variations in the set of teeth* like variations in the teeth's shape, size, position and rotation in the majority of cases by modeling the statistical variation in the input data within the coupled model and considering the deviation from expected configurations during model adaptation. Dental appliance and restoration can sometimes negatively affect the detected tooth contour since their denser material appears brighter in X-Ray images and results in stronger gradient features. Examples will be provided later in this section (cf. paragraph 'General observations').

In general, the detection of fine contour details is affected by the chosen balance between model energy and image energy during model adaptation (cf. Eq. (5.10)). A stronger influence of the image energy (which allows the segmentation to deviate further from statistically valid shapes and spatial relations) supports a better and more accurate adaptation, if the detected image features are artifact-free. However, it can also lead to unrealistic or incorrect segmentations in the presence of noise, low contrast or artifacts (e.g. dental appliance) as the detected image features are less reliant and more prior knowledge is required. Determining the correct balance between reliance on image features and restriction based on the learned statistical information is key for achieving an overall optimal segmentation performance.

The following paragraphs discuss selected topics, like impact of missing teeth, dependence on initialization and general observations made on the final segmentation results.

Missing teeth An extracted or otherwise removed tooth leaves a gap in its original place. The coupled shape model is able to handle missing teeth if the space previously occupied by that tooth is still vacant and big enough to fit the original tooth – and therefore also the corresponding shape model. The model item corresponding to the missing tooth can be placed in the existing gap and even though no valid image features are detectable for such a model item, all subsequent items will still overlap with the corresponding tooth in the image. The validity of an image feature can be assessed via the gradient magnitude and the gradient direction. The ratio of landmark points with valid image features can also be used to identify missing teeth (in cases where the gap is still present). Fig. 5.16 shows examples of successfully handled missing teeth and highlights the model items placed in the vacant space. However, gaps caused by missing teeth can shrink over time because neighboring teeth naturally move into the open space. More likely, gaps are closed by orthodontic procedures using dental braces which apply forces to the teeth to move them to fill the gap. Partly or completely closed gaps will cause the adaptation to fail as the coupled model still attempts to fit the model item of the original tooth to the one that has been moved to its place. This is illustrated in Fig. 5.17. This behavior is to be expected as the coupled shape model encodes the spatial relation of a complete set of teeth. Accounting for changes in the spatial relation due to missing teeth would require changes to the coupled model as well as a way to detect missing teeth in cases where a gap is not longer present. Missing teeth without the respective gap still present are therefore beyond the scope of this approach.

Another aspect is the total number of missing teeth. While isolated missing teeth even in multiple locations can be handled (cf. Fig. 5.16, left image), multiple consecutive missing teeth can cause either the initialization or the adaptation to fail. Fig. 5.14 (bottom case) shows a failed adaptation caused by all upper incisor teeth missing. Despite a correct initialization, the initial model adaptation step was unable to stabilize the model in a horizontal position because no image features for the upper incisors could be found. This imbalance caused a rotation of the complete model from which the algorithm was unable to recover. Even though such a large rotation is unlikely to occur naturally in any panoramic image, the global transformation that defines the position of the coupled model in image space is not constrained by any prior knowledge as it is considered an external factor (as is typically the case with any active shape model-based approach). Future restrictions on the global transformation could reduce the likelihood of such failed adaptation. In its current state, the model adaptation struggles with multiple missing incisor teeth. Similarly, a large number of missing teeth (10+) will negatively affect the performance.

Initialization dependence The performance of the presented approach is highly dependent on a good initial placement of the coupled shape model. It is evident from Table 5.4 that an incorrect placement (i.e. the defined initialization structures do not sufficiently overlap with their corresponding structures in the image) will result in low segmentation and numbering performance. The F1 score for the single failed case of data set 1 (which is mainly caused by a missing lower central incisor where the gap was no longer present) was 0.5375. The average F1 score for the eight failed cases of data set 2 dropped to 0.3721 ± 0.2636.

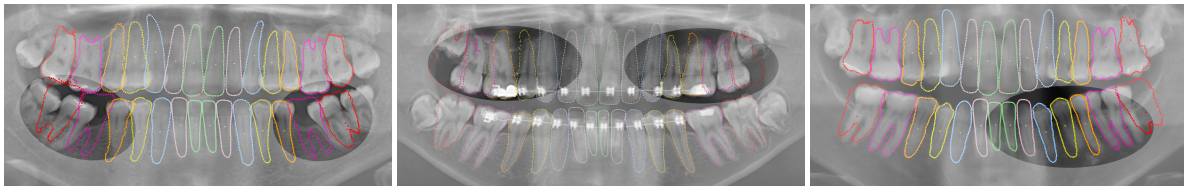


Figure 5.17.: Exemplary cases of failed adaptations caused by missing teeth without a gap present. The areas highlighted in red illustrate that the tooth model of the missing tooth is instead matched to the tooth which filled the gap. This results in incorrect adaptation of all subsequent teeth. (Original images haven been cropped for better visualization.)

The used local search method is only able to find strong image features close to the initial position. Since incorrect initialization can put image features of neighboring teeth closer a model item than the true features, the model is unable to recover from this incorrect initial placement during model adaptation. This behavior of algorithms which used local search methods to find image features is well known [HM09]. The coupled shape model consists of many individual SSM which all use local feature search and therefore require good initialization. However, trying to find a position that allows for an immediate adaptation of all model items is almost impossible and would potentially result in many failed or sub-optimal segmentations and numberings. Instead, dynamic adaptation is used in this thesis which gradually increases the set of actively adapted model items and therefore only requires an good initialization (i.e. good overlap between model items and corresponding structures in the image) of a small subset of model items. This approach simplifies the initialization task and enables a successful adaptation in the large majority of cases. A detailed discussion of the challenges faced during initialization as well as the initialization performance was done in Section 5.5.1. In summary, the initialization was successful for $\sim$92 %of the test cases for both data sets – 10 out of 11 for data set one and 103 out of 111 for set two. The main reasons for failed initialization are an erroneous placement in horizontal direction resulting in the initialization structure being adapted to incorrect image structures, and an erroneous scale estimation which can lead to incorrect segmentation of the posterior teeth. Especially the small width of the mandibular incisor teeth leaves very little room for error for the horizontal placement of the model.

General observations Careful observation of the segmentation performance revealed areas which offer potential for future improvements. Malpositioned teeth that are too far away from their expected position in the image might not be segmented correctly. Fig. 5.18 a) and b) show examples of such cases. The red overlay in the images illustrates the ground truth annotation and the (colored) dotted line represents the final segmentation result. The distance between the actual position of the tooth and the expected position (based on statistical knowledge) means that the local feature search is unable to detect the relevant image features.

Dental appliance can cause inaccurate segmentations as the denser material appears very bright in the image and therefore offers strong gradient features at its boundary. The adaptation process is

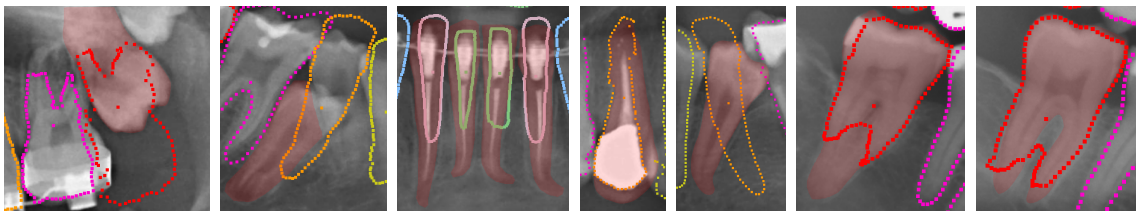


Figure 5.18.: Potential challenges which contribute to inaccurate segmentations: a) and d) - malpositioned teeth, b) and c) - dental appliance, e) - low overlap, f) and g) - molar root shape variance.

likely to be attacked to those features if the boundary of the dental appliance is close to the actual tooth boundary. Such cases are depicted in Fig. 5.18 c) and d) where the (dotted) colored contour representing the detected tooth boundary is incorrectly fitted to the dental appliance resulting in a low segmentation performance. (The ground truth is again shown as a red overlay).

In rare cases the adaptation of the model item to the image feature can fail because not enough valid feature points were detected. Feature points are considered valid if the gradient magnitude is above a certain threshold to filter out features detected in the dental structural area and to reduce the influence of bad image quality and noise. This also prevents the adaptation of model items to random image features in case the corresponding tooth is missing in the image. Unfortunately, low overlap between the model item and the corresponding structure in the image can also result in invalid features which then prevents a successful adaptation. An example of such a case is illustrated in Fig. 5.18 e).

The segmentation accuracy of the molar teeth sometimes suffers from the fact that, depending on the tooth's orientation, either one or two roots are visible in the image. Because the 2-D shape model encodes the visible outline of the tooth, there is a large shape variation in the root part of the molar tooth models (cf. Fig. 4.2). The large variation of possible shapes limits the ability of the SSM to guide the adaptation process based on prior knowledge. Additionally, since the possible shapes are modeled as a linear sub-space, any shape configuration 'between' the one- and two root outlines is also considered a valid shape. As such, the model adaptation is sometimes unable to correctly fit the tooth model to the structure in the image. Fig. 5.18 f) and g) show examples of molar teeth with one and two visible roots, respectively, where the model adaptation failed to correctly segmented the tooth boundary in the root area.

State of the art A comparison to other approaches is difficult as only Silva et al. [SPOP20] have also performed a simultaneous tooth segmentation and numbering. Other approaches only focused on semantic teeth segmentation [KPIB19, NNH*21, SCKM20], instance-based segmentation without numbering [SOP18, JFR*18], or teeth detection and numbering without a contour extraction (segmentation) [TTB*19, CLP*21]. Silva et al. [SPOP20] trained four established CNN architectures on their DNS data set (cf. Section 5.1) to perform teeth instance segmentation and numbering. However, they

only report classification metrics in terms of average precision with cut-offs (AP50 and AP75), and mean average precision for the Mask R-CNN [HGDG17], PANet [LQQ*18], HTC [CPW*19] and ResNeSt [ZWZ*20] architectures. The metrics used in this thesis are only reported for the semantic segmentation performance of the analyzed networks, but this evaluation was instead performed on a much larger data set of 778 test images. It is worth noting that the CNN-based approaches are not dependent on a separate initialization method and completely failed segmentations are unlikely assuming sufficient training data. However, as they are considered as black-box approaches, it is also more difficult to judge why incorrect segmentations or numberings occur.

5.9. Conclusion

In this chapter, an automatic approach for teeth segmentation and numbering of 28 teeth in panoramic radiographic images was presented. It is based on the segmentation and numbering concept introduced in the previous chapter. The proposed approach is the first one for teeth segmentation (and numbering) to explicitly model tooth shape knowledge as well as knowledge about the spatial relation of teeth. Most state of the art methods focus only on semantic teeth segmentation and do not provide instance-based tooth segmentation or tooth numbering because of the increased difficulty. The human set of 32 teeth is composed of only 4 teeth types, which requires the inclusion of knowledge about the geometric configuration of teeth (i.e. the spatial relation between individual teeth) to perform a tooth numbering. The presented approach utilizes a coupled shape model of the teeth to encode this information. The 2-D coupled model is composed of a statistical shape model (SSM) for each tooth which is coupled with all other individual models using their spatial relation. This not only provides the information required for tooth numbering but also enables a more robust segmentation process using gradient image features (bottom-up) in combination with a priori statistical knowledge about the teeth in order to guide the segmentation process (top-down) [MT96]. The coupled model is used in conjunction with a neural network and employs dynamic adaption strategies to robustly segment (and number) teeth in panoramic radiographs. The neural network provides a binary mask of the teeth area for calculating the necessary information for the initialization of the coupled model. Additionally, the mask is useful for the early adaptation steps and is used together with the dynamic adaptation strategy to ensure a good placement of individual teeth after the first 8 adaptation steps. This way, the parameters of the final adaptation on the intensity image can be tailored more towards fine tuning the tooth contour rather than localizing the structure in the image.

The proposed approach was evaluated on two data sets and achieved a good tooth-specific segmentation and numbering performance in terms of F1 score of $\sim$0.83 and $\sim$0.82 on the two sets, respectively. The approach was able to handle the challenging *image quality and characteristics* of dental panoramic radiographs, like noise or low contrast as well as the tooth numbering by utilizing the shape and spatial information encoded in the coupled shape model. It is able to deal with the inter-patient *variations in the set of teeth* like variations in the teeth's shape, size, position and rotation in the majority of cases. This includes missing teeth, given that the space originally occupied by the missing tooth is still present such that the mean shape model of that tooth can be placed into the gap and subsequent teeth can be po-

sitioned correctly. Dental appliance and restoration can sometimes negatively affect the detected tooth contour since their denser material appears brighter in X-Ray images and results in stronger gradient features. Limitations of the presented approach – which will result in incorrect segmentation of some teeth – are missing teeth where the gap has been closed over time or by orthodontic procedures (since the encoded spatial relation is no longer accurate in these cases). Additionally, like any algorithm that utilizes a local search method to detect feature points, the coupled shape model is dependent on a good initialization to produce accurate segmentation and numbering results. Wisdom teeth were not included due to the impact on model training and especially model initialization.

6. Image-based 3-D teeth reconstruction

This chapter covers the second research question of this thesis, the image-based 3-D teeth reconstruction from five colored photographs, to support computed-aided systems for an automatic initial assessment of the severity of teeth malposition and a more accurate monitoring during orthodontic treatments. It proposes the first fully-automatic teeth reconstruction approach based on (colored) photographs which aims to reconstruct the complete set of teeth. It is also the first reconstruction approach to utilize the documentary photographs typically taken during orthodontic procedures and, as a result, will perform the reconstruction only based on the limited information available from five photographs. The colorized photos are captured as part of the orthodontic treatment record and depict the patient's teeth from a frontal view, lateral views from the left and right as well as top-down views of maxilla and mandible (cf. Section 2.3.5). The predefined general viewing direction of each photo (referred to as *view* within this chapter) intends to maximize the amount of information about the teeth that is contained in these images for documentation purposes. This aspect enables an image-based reconstruction despite the low amount of images.

Nonetheless, the overall amount of information about the teeth which can be extracted from just five photographs is limited (cf. Section 1.2.2). Some posterior teeth, namely second molars and wisdom teeth, are at most visible in a single view and, therefore, these are not included in the reconstruction as there is simply too few information available for a meaningful reconstruction. Moreover, most other teeth are only visible in two or three of those pictures and all teeth suffer from heavy occlusions. For those reasons, a model-based approach that incorporates prior knowledge about the shape and location of the teeth is best suited for the reconstruction task, as it provides vital top-down information to complement the available image information, especially about the expected shape of occluded regions. Additionally, the prior knowledge can be used to constrain the reconstructed teeth to valid shapes.

The reconstruction process is complicated further by the fact that the teeth possess very difficult appearance properties (cf. Section 1.2.2). This makes their (2-D) silhouette (i.e. the boundary between tooth and surrounding tissue) the only feature that can be reliably extracted from the photographs. Consequently, a silhouette-based loss function is used during the reconstruction process. It is employed to minimize the difference between 2-D projections of the reconstructed 3-D teeth and their corresponding 2-D contours in the images. Deviations between projections and contours are utilized in a deformation-based reconstruction method, in which a 3-D model providing an initial guess of the teeth is deformed such that it optimally matches the contours extracted from the photographs. For a fully-automatic reconstruction process, a numbering of the extracted 2-D teeth contours is required in order to automatically establish the correspondences between the 2-D silhouettes and the 3-D model teeth.

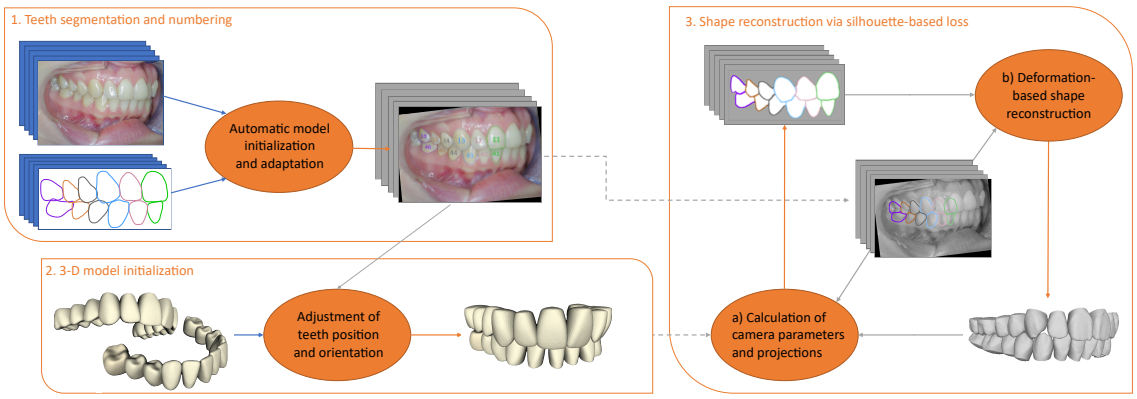


Figure 6.1.: Overview of the general workflow of the reconstruction. Input data is depicted in blue, generated data in gray and algorithmic progressing in orange.

The overall workflow of the model-based 3-D teeth reconstruction, which is also illustrated in Fig. 6.1, can be roughly divided into three main steps:

1. **Teeth segmentation:** the segmentation and numbering of the 2-D teeth contours in the colored photographs,
2. **3-D model initialization:** the initial adjustment of the teeth positions in the 3-D models,
3. **3-D shape reconstruction:** the iterative deformation-based shape reconstruction of the teeth while minimizing the silhouette-based loss.

The first step of the reconstruction workflow (Fig. 6.1 (1.)) is concerned with retrieving the relevant information from the provided images. This is done by extracting the teeth contours (silhouettes) and their numbering from each of the five colored photographs. As concluded in Section 1.2.3, the teeth contour extraction and numbering in (colored) photographs is similar to the research question presented in the previous chapter – the teeth segmentation and numbering in panoramic radiographs. Therefore, the same concept of coupled shape model, neural network and dynamic adaptation introduced in Chapter 4 will be applied for this task. Fig. 6.2 illustrates the segmentation and numbering workflow that will be applied separately to each view. This means that for each viewing direction, a view-specific 2-D coupled shape model (CSM) (Fig. 6.2B) and a view-specific neural network (Fig. 6.2A) are trained in the offline training phase. The online processing of unseen photograph starts with a preprocessing step (removal of specular reflections). Then, the U-Net network computes a binary mask of the teeth area which is utilized for the initial placement of the coupled model. The teeth contours (and numbers) are then extracted by adapting the positioned CSM to the corresponding photograph. The adaptation employs dynamic adaptation strategies to improve the robustness by gradually increasing the set of adapted teeth.

After extracting the contour and numbering information from all photos, the model-based reconstruction process starts with initializing the employed 3-D models (Fig. 6.1 (2.)). These 3-D coupled

shape models, one for upper and one for lower dental arch, were trained in an offline training phase (cf. Fig. 6.1) to capture the shape and spatial variance of the 3-D teeth. The two models are adjusted in this second step to assure a good initial starting point for the deformation-based reconstruction. The adjustments include the scale and relative position of both arches but also the local teeth positions within each arch. The correspondences between 3-D model teeth and extracted 2-D contours are automatically established using the included numbering information from the first step.

In the third step, the adjusted 3-D models are employed to perform the iterative teeth reconstruction. A silhouette-based loss between 2-D projections of the 3-D models and the 2-D teeth outlines of the input images is minimized to determine the optimal 3-D model configurations that best explain the input. This expectation maximization-like process alternates between estimating the projection parameters for each view based on the current 3-D model configurations and deforming the 3-D models such that the 2-D projections match the extracted contours from the photographs. The encoded statistical information of the 3-D models provides additional information about the expected shape of the teeth and constrains the reconstruction to statistically valid shapes. The iterative process repeats until the silhouette-based loss no longer changes or until a certain number of iterations have been reached.

The performance of the presented model-based approach for teeth reconstruction is evaluated on 89 data sets that are randomly separated into 50 sets for training and 39 sets for testing. The evaluation includes an assessment of the contour extraction (and numbering) accuracy from the five types of photographs by comparing the generated segmentations to manually annotated gold standard segmentations. The final 3-D reconstruction results are compared against ground truth teeth meshes retrieved by laser scanning plaster imprints.

The chapter is structured as follows: Section 6.1 covers the first step of the workflow and explains the application of the segmentation and numbering concept to extract information from each of the five photographs. This includes the design of the coupled shape model for each view, the pre-processing of the photographs, the binary mask generation using the neural network, and the CSM initialization and adaptation. The experiments and results cover the offline training phase as well as the performance evaluation on a set of test data. The steps of the deformation-based reconstruction process which tries to recover the 3-D teeth shapes and spatial relation are then presented in Section 6.2. The two 3-D models for lower and upper dental arch are initialized in terms of relative position, scale, and local teeth positions (second step) and then iteratively deformed to minimize the silhouette-based loss between segmented teeth contours and projected teeth silhouettes of the 3-D models. The iterative process includes a camera parameter estimation for each view, the generation of 2-D projections to calculate the silhouette-based loss, and the teeth deformation to minimize said loss. Experiments and results are shown in Section 6.3, followed by a discussion of the results in Section 6.4. The chapter ends with a conclusion about the presented teeth reconstruction approach in Section 6.5.

This chapter is partially based on the publication "Automatic model-based 3-D reconstruction of the teeth from five photographs with predefined viewing directions" [WJN*21].

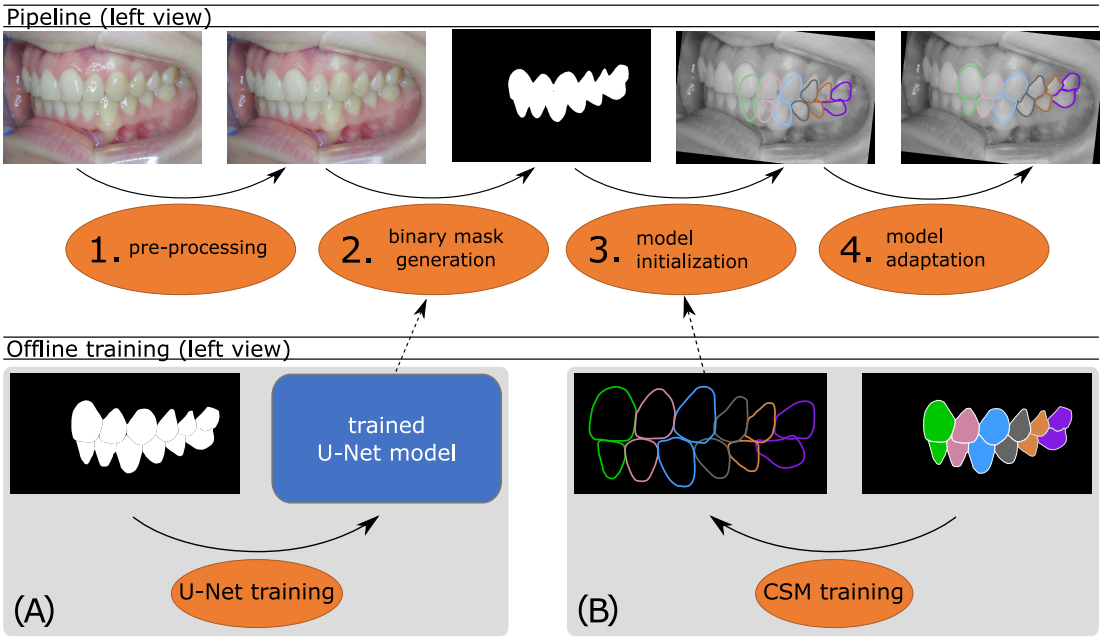


Figure 6.2.: Teeth segmentation and numbering workflow for each colored photograph (shown using the left view as an example). The required U-Net network (A) and CSM model (B) will be trained in an offline training phase based on manually annotated training data. The (online) pipeline then utilizes these to compute the teeth segmentation and numbering for an unseen input image (top half).

6.1. Teeth segmentation & numbering

The first step of the reconstruction workflow is the extraction of the numbered teeth outlines from each of the five photographs. The outlines are used during the reconstruction process as a reference to deform the 3-D teeth models and recover the true 3-D shapes via a silhouette-based loss. This process mandates that the correspondences between 2-D outlines and 3-D teeth model are known, which is achieved through the inherent numbering assigned by the chosen approach. This workflow step presents the second application of the segmentation and numbering concept introduced in Chapter 4. Since the five photographs depict the teeth from five different viewing directions, the teeth that are visible in each view are different. Moreover, the expected 2-D outlines of teeth are directly related to the viewing direction. Therefore, the segmentation and numbering concept is applied separately for each type of photograph (view). The workflow for the processing of each view follows the workflow of the proposed concept and includes an optional pre-processing step (cf. Section 4.1). The workflow for one selected view (left view) is depicted in Fig. 6.2, but is applied in the same way for the other four views as well.

A 2-D coupled shape model (Fig. 6.2B), to encode the statistical shape variation of the teeth silhouettes as well as knowledge about the spatial relation, and a U-Net network (Fig. 6.2A), for semantic teeth segmentation, are trained during an offline training phase for each view. Additionally, an appearance model is trained for each contour point of the 2-D shape models based on the color information of the photographs (HSV color space). The online extraction of the numbered teeth outlines on an unseen photograph then includes a *pre-processing step* to remove specular reflections, the *generation of a binary teeth mask* via the convolutional neural network (U-Net), a *model initialization step* to position the corresponding coupled shape model correctly onto the respective image and a *model adaptation step* to fit the CSM to the image and extract the final teeth contours.

In contrast to the concept's application for the teeth segmentation and numbering in panoramic radiographs (as presented in the previous chapter), colored photographs are used here instead of gray-scale radiographs. The viewing direction of these photographs varies significantly more as the photos are captured manually with a handheld camera compared to the fixed generator-detector setup used for dental radiographs. The added color information, on the other hand, provides additional information to help distinguish teeth from the surrounding tissue. The amount of feature channels for the U-Net is therefore increased to account for the additional color information. Moreover, the model adaptation will also incorporate the color appearance of contour points to determine suitable image features instead of relying solely on gradient features.

The structure of this section is as follows. The first two subsections cover the design of the 2-D coupled shape model for each view, i.e. the teeth included in each model (Section 6.1.1), and the general design of U-Net networks (Section 6.1.2). Afterwards, the individual steps of the segmentation and numbering workflow are presented. This includes pre-processing and mask generation (Section 6.1.3), automatic CSM initialization (Section 6.1.4), and CSM adaptation (Section 6.1.5). Experiments and results of the teeth segmentation and numbering are discussed at the end of this section (Section 6.1.6). Section 6.1.6 also covers the offline training of the coupled shape models and neural networks.

6.1.1. 2-D CSM designs

The view-specific 2-D coupled shape models are designed to encode the shape variation and spatial relations (cf. Section 4.2) of the most prominent teeth outlines in each view. This information is used to support the extraction of the teeth contours in the photographs and to provide the corresponding numbering. The numbering is required during the reconstruction process to establish the correspondence between extracted teeth contours and the teeth in the 3-D models (cf. Section 6.2).

The view-specific CSMs each contain the teeth listed below (the number in parenthesis after each tooth refers to the FDI notation system):

- **Maxilla view** ($N = 12$ items): maxillary right first molar (16) to maxillary left first molar (26),
- **Mandible view** ($N = 12$ items): mandibular right first molar (46) to mandibular left first molar (36),

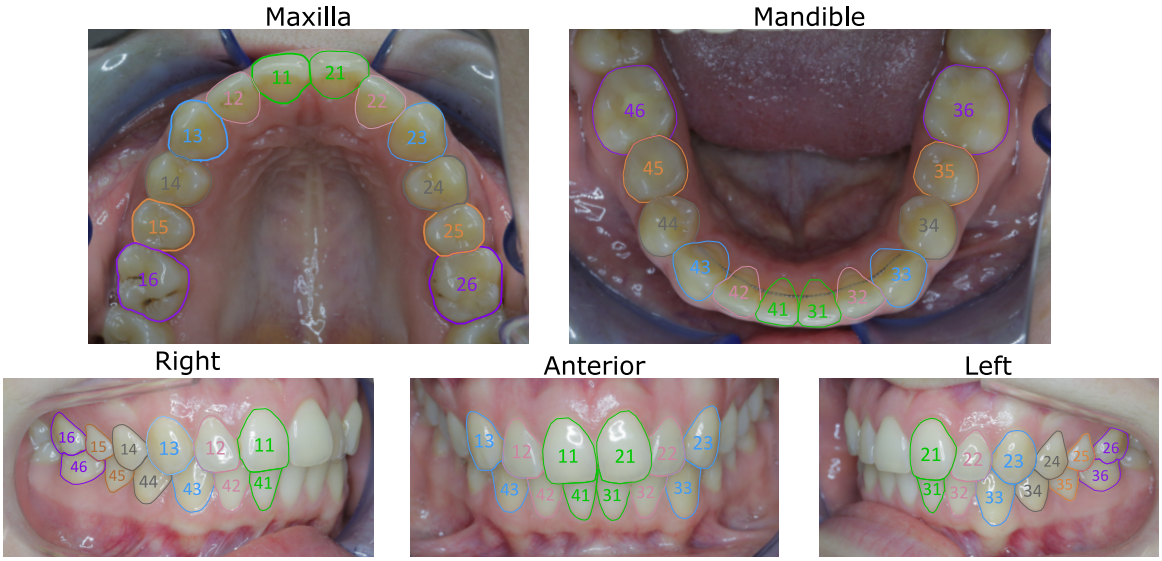


Figure 6.3.: CSM design for each of the five views included in orthodontic records. Each colored tooth outline corresponds to a shape model in the CSM. The colors indicate different tooth types, the numbers refer to the FDI notation system. The color scheme is used consistently throughout this chapter to identify the individual model items.

- **Right view** ($N = 12$ items): maxillary right first molar (16) to maxillary right central incisor (11) and mandibular right first molar (46) to mandibular right central incisor (41),
- **Anterior view** ($N = 12$ items): maxillary right canine (13) to maxillary left canine (23) and mandibular right canine (43) to mandibular left canine (33),
- **Left view** ($N = 12$ items): maxillary left central incisor (21) to maxillary left first molar (26) and mandibular left central incisor (31) to mandibular left first molar (36).

A visual representation of the teeth included in each model is also provided in Fig. 6.3. The number assigned to each tooth again refers to the corresponding number in the FDI notation system, while the color indicates the different teeth types (i.e. canine or first molar). The color-code is consistently used throughout this chapter to depict the outline of the respective teeth.

The reason for including the listed teeth in the respective model is that these teeth are the most prominent ones in the respective view. This means that these teeth are visible in the majority of photographs of that view. For example, anterior teeth (incisors and canines) are easily visible in the anterior view (cf. Fig. 6.3, 'Anterior' image), while posterior teeth are only visible to a small extend and can sometimes be occluded completely. This adds additional challenges to the posterior ones and makes them harder to segment accurately. They are therefore not included in the coupled shape model because information about their outline cannot be robustly extracted from all photographs. An incorrect segmentation or numbering (in case a missing tooth is not detected) will negatively affect the reconstruction, while

Figure 6.4.: Illustration of possible tooth landmark appearance categories: tooth to gum (blue), tooth to background (turquoise) and tooth to tooth (pink). Exemplary appearance profiles (centered on the landmark location) for each class are depicted on the right.

the information gained from a successful segmentation is limited. Given this trade-off, these teeth have been excluded from the coupled shape model. The anterior teeth not included in the left and right views suffer from large variations in their shape caused by variations in the camera angle. Other teeth, like second molars, are almost never (completely) visible in any of the photographs. An additional aspect is that the photographs are captured at close distance to the teeth. This often results in the lateral teeth in the image being out of focus and appearing blurred. CSMs which included these teeth resulted in inferior segmentation performance during early testing.

All coupled shape models are trained in an offline training phase as described in Section 4.2. The statistical information about the shape and spatial relation of teeth outlines in each view is extracted from manually annotated training data by establishing landmark correspondences, and applying generalized procrustes analysis and principal component analysis (PCA). The shape models of the individual teeth (deformable model items) are constructed using $L = 100$ landmark points. The same number of landmark points has already proved reliable for teeth segmentation and numbering in panoramic radiographs (cf. previous chapter). From the modes of variation (eigenvectors) computed during PCA, only the eigenvectors required for capturing 95 % of the shape variance are kept and the remaining ones are discarded in order to improve the generalization ability of the model. The models generated during the offline training process as well as the training data used are discussed in the experiments and results part of the teeth segmentation and numbering section (cf. Section 6.1.6).

In order to make use of the color information included in the photographs, an appropriate appearance model is added to the design of the CSMs.

Appearance model The appearance of individual landmark points of a model item was captured using the color distribution around that contour point. This appearance profile was extracted from all training instances by sampling the HSV color information (using linear interpolation) in normal direc-

tion of the contour at the landmark location. On either side of a landmark point, m samples are extracted using a fixed step size which results in appearance profiles of length $2m+1$. A profile p is then given by a vector $p = (hsv_1, \ldots, hsv_{2m+1})$ where each entry hsv_i contains the color information with respect to hue, saturation and value. Instead of computing the mean appearance (and covariance) of a landmark point from all these profiles as proposed by Cootes and Taylor [CT*04], a clustering-based method is used to retrieve the most discriminative appearances. This is done because the method of Cootes and Taylor [CT*04] assumes a normal distribution of profiles, but the color appearance varies significantly depending on the neighboring tissue resulting in a non-normal distribution. Depending on the shape and position of the tooth and neighboring teeth, a contour point can be located at the border of tooth to gum or lip, tooth to tooth, or tooth to background (i.e. a black background caused by (small) gaps between teeth or missing teeth which could provide a view inside the oral cavity but appear black due to the missing light). A similar separation was used by Wu et al. [WBG*16] who classified the tooth boundaries based on the occluding object. Fig. 6.4 illustrates the different categories of landmark appearances by colorizing a tooth contour according to the corresponding category. For each category, an exemplary appearance profile is provided as well. Using only a single mean appearance per landmark point would not only fail to adequately capture the distinct appearances of the different categories but could also create unrealistic appearance profiles by averaging appearance values over different categories. K-Means clustering [JMF99] with $k = 6$ clusters is used to group the appearance profiles – the median appearance of each cluster is used as respective center during th clustering process. A total number of 6 clusters is used as this should also separate large appearance discrepancies within categories into different clusters and offers better encapsulation of unexpected appearances. The median appearance in each cluster (instead of the mean) is used to not generate unrealistic average values.

In order to compare two profiles, the distance between two appearance profiles p^i and p^j is calculated as the average Euclidean distance of the profile entries:

$$dist(p^i, p^j) = \frac{1}{2m+1} \sum_{e=1}^{2m+1} \left\| p_e^i - p_e^j \right\|, \tag{6.1}$$

where $2m+1$ is the length (i.e. number of entries) in each appearance profile. The individual H, S and V values are normalized to a range between 0 and 1 to weight all components equally. The final centers (median appearances) are later used during the adaptation process to determine suitable image feature points. The appearance model is trained together with the coupled shape models during offline training (cf. Section 6.1.6.3).

Having described the coupled shape model design and the motivation behind it, as well as the appearance model in this section, the next section will cover the neural networks for the extraction of the binary mask (semantic teeth segmentation).

6.1.2. U-Net networks

Individual U-Net networks are trained for each view to perform a semantic teeth segmentation in the corresponding photographs. This semantic segmentation classifies each pixel in the image content as

either tooth or not-tooth, resulting in a binary mask of the teeth area. As explained in the concept Section 4.3, the goal of this binary mask is to facilitate a robust initialization of the coupled shape model onto the images and support the initial model adaptation. The semantic segmentation via CNNs is designed to only cover the teeth included in the CSM design of the respective view (cf. previous Section 6.1.1). Excluding all other teeth (that are present in the photograph) supports a more robust model initialization because the CNNS should provide more consistent results between images.

A similar approach was already successfully used in the previous chapter to perform a semantic teeth segmentation in (gray-scale) radiographic images while excluding wisdom teeth during the segmentation process. The photographs introduce the challenge that viewing angle, scale and location of the teeth in the image vary significantly between photos as they are captured by hand. This can also results in blurred teeth and boundaries, if the teeth area was out of focus when the photo was taken. Fortunately, the added color information provides additional clues to separate the teeth from the surrounding area.

The five networks, one for each view, are trained during the offline training phase. The original U-Net architecture was modified slightly for the use with color images. Compared to the original U-Net design, the input layer was changed to accept 3 channel images to account for the additional color information. Other parameters like input size in terms of width and height, number of filters on the input layer, the color space used for the input image, or loss parameters α and β are evaluated during the experiments to determine the optimal values. For this, the same training data that was used to train the coupled shape models is utilized. The tooth-specific gold-standard annotations belonging to a training instance are merged into a single (semantic) segmentation mask for each photograph to serve as training data for the network. A binary semantic segmentation mask is depicted in Fig. 5.1A, compared to the colorized teeth-specific segmentation in Fig. 5.1B. The experiments to determine the optimal parameter configurations are discussed at the end of this section, cf. Section 6.1.6.

After having trained five coupled shape models and five neural networks (one of each for every view) in the offline training phase, the following sections cover the online part of the teeth segmentation and numbering for each view – starting with the preprocessing and the mask generation.

6.1.3. Pre-processing and mask generation

The online processing of an unseen set of photographs begins with a pre-processing step (cf. Fig. 6.2). Because of the teeth's thin outer translucent enamel coating, the photographs contain a lot of specular reflections on the surface of the teeth. This can negatively affect the search for image feature points during the model adaptation step because the borders of reflection areas result in strong image gradients (when using gradient features) while the presence of reflections results in unusual color variance (when using color appearance features). Therefore, all five input images are processed to remove specular reflections. Dynamic global thresholding on the value channel of the HSV color space is used to identify regions in the image showing these very bright reflections. The HSV color space is used as it describes colors as a combination of hue (color value), saturation (departure of a hue from white), and value (departure of a hue from black) [Smi78]. Specular reflections are the only image content with

Figure 6.5.: Image pre-processing to remove specular reflections. Original image regions are shown on the left; middle images are the binary masks of the detected reflection computed via global thresholding in the value channel of HSV color space; the final result after in-painting the detected reflections using the approach of Telea [Tel04] are depicted on the right.

close to maximum value, which enables an easy separation of these regions from other image content in the value channel. Almost identical results can be achieved using the brightness (L) channel of the HSL [Smi78] color space. An image in-painting technique [Tel04] based on the fast marching method for level set applications is then utilized to reconstruct the image information in the detected regions, thus removing the specular reflections. The missing image information is reconstructed based on the known color information around the region to be in-painted. This technique is iteratively applied to advance the region boundary further inside until the whole region has been in-painted. Examples of the image pre-processing step are depicted in Fig. 6.5. It shows a sub-region of two input photographs containing specular reflections (left images), the binary mask of detected reflection areas computed via global thresholding (middle image), and the image sub-region after in-painting (right image). These reflection-free images are subsequently used in future pipeline steps.

The next step, after pre-processing all five photographs, is the extraction of the binary teeth masks via the offline trained U-Net networks. These networks are designed to only segment the teeth that are part of the coupled shape model of the respective view in the binary mask. This will simplify subsequent processing as the masks are intended for a robust initialization of the coupled shape models onto the photographs and to support the initial model adaptation. In order to apply the corresponding neural network, each photograph is down-sampled to a resolution of 336x224 using linear interpolation. An input size of 336x224 provided the best semantic segmentation results on an unseen test set during

network training (cf. Section 6.1.6.4). The binary masks generated by the different networks are then up-sampled again to the original image resolution. The up-sampling uses nearest neighbor interpolation to maintain the binary nature of the mask. The step following the pre-processing and teeth mask generation for all input images is the initialization of the individual coupled shape models onto the input images.

6.1.4. Automatic 2-D CSM initialization

The individual coupled shape models need to be placed onto the respective input images in terms of position, rotation and scale in order to enable a robust segmentation of the teeth contours. All of these values are computed using the binary mask of the teeth generated by the neural network in the previous step. Slightly different approaches are used to determine these values for the individual views because of the different characteristics the binary teeth mask has for different image groups.

The maxilla and mandible views depict the dental arches of the respective jaw from a 'top-down' perspective. These views therefore do not suffer from large occlusions and have relatively stable viewing angles because of the way these images are captured. The calculation of the initialization parameters for these views relies on detecting parts of the U-shaped teeth mask. From the center of the image, a ray-based search is performed to find the left and right inner and outer boundaries of the teeth area. Fig. 6.6a shows the results of this search, with detected boundary points illustrated in red. Based on the detected boundary points, the center line (blue) is established on both sides and the average x-coordinate of the center line is used as reference (vertical yellow lines on both sides) to estimate both the horizontal position and scale reference required for initialization. For establishing the vertical position of the model, a similar technique is used to determine the position of the anterior teeth in the mask (cf. Fig. 6.6b). A ray-based search to the top (maxilla) or bottom (mandible) is performed and the average y-coordinate of the center line is used as reference. Fig. 6.6c shows an exemplary initialization based on these references. The initializations for the remaining three views are more challenging to handle robustly.

The anterior view is also relatively consistent in terms of viewing direction because the photographs are taken directly from the front (of the patient). The difficult part here is that the photos are taken in a 'closed-mouth' position. As a result, the anterior maxillary teeth can partly or completely occlude the anterior mandibular teeth – which is also referred to as overbite in medical terms. This requires additional processing to correctly position the model in vertical direction. The optimal model position in horizontal direction (x-coordinate) is determined by identifying the teeth separations of the maxillary teeth in the binary mask (cf. Fig. 6.7). Analyzing the upper contour of the teeth area for local minima (in terms of y-coordinate) provides candidates for such separations. These candidates are then filtered based on expected (relative) teeth width and expected number of teeth in order to eliminate false positive detections. Assuming that the central incisor teeth are the widest teeth in this view, the candidate that maximizes the margin to its neighbors is chosen as the separation between the central incisors. The location of this separation also provides the x-coordinate for the model position. This separation is typically close to the center of teeth area, due to the symmetry of the teeth included in the anterior

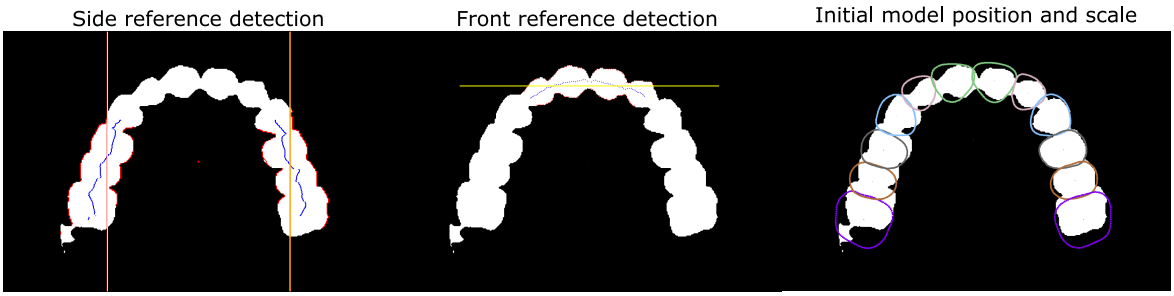


Figure 6.6.: Computation of references for the model initialization (maxilla and mandible views). A ray-based search from the image center is used to establish the center points (blue) for the sides and front of the U-shaped teeth mask. Average coordinates of those points are then used as references (yellow lines) for the initialization.

CSM. The separations are illustrated Fig. 6.7a. The separation between the central incisors is colored in red while the remaining ones are colored in green. The teeth separations are also used to calculate the initial scale for the CSM. Based on the separation between the central incisors, one can estimate the horizontal distance between the two central incisors and the two lateral incisors. Both distances are indicated in Fig. 6.7a in blue and turquoise, respectively. The (average) ratio between the distances computed on the binary mask and the equivalent distances in the coupled shape model is then used as initial scale value. Finally, the y-coordinate of the model position is determined by searching for the separation between maxillary and mandibular teeth. Therefore, the vertical image gradients of the input image within the masked region are analyzed to find the local maxima in terms of gradient magnitude. Fig. 6.7b shows the detected maxima in red (blue and green dots indicate the begin and end of the vertical search area – these are computed based on the binary mask boundary with some margin to the inside). After removing outliers, the average y-coordinate of these maxima is used for model initialization. An exemplary model initialization based on these parameters is depicted in Fig. 6.7c.

The left and right views provide the biggest challenge for model initialization. All images are captured with a freely movable handheld camera, but, in contrast to the anterior, maxilla and mandible views, left and right views are not taken from a frontal position but at an angle from the left and right side (of the patient). This means that different interpretations of a left and right view result in large variations of camera position. These variation (and the resulting changes to the teeth outlines) will naturally be encoded in the statistical models as variations in global teeth location and teeth shape (cf. Section 6.1.6.3). This means that the differences in teeth configuration are larger along the modes of variation compared to other models. This makes an optimal initial placement and a subsequent adaptation more difficult since the coupled shape models are initialized in their mean configuration. To support a more robust adaptation, the configuration of the CSM for these views will be optimized during initialization to account for these variations – relying solely on the model adaptation proved ineffective to handle this challenge. The parameters for initial position (i.e. the center between canine

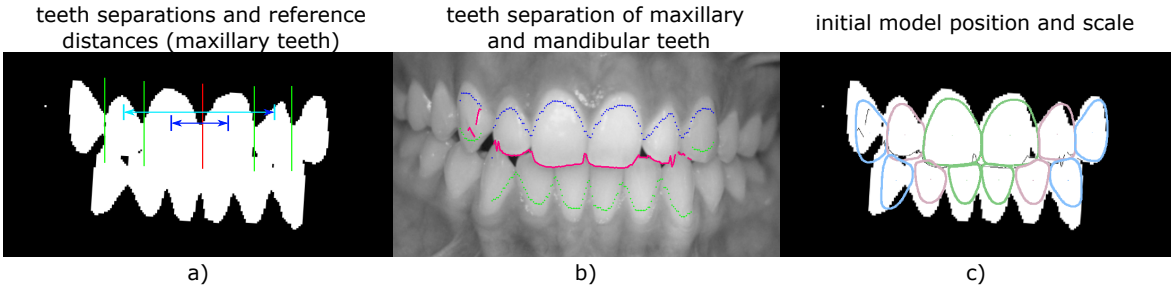


Figure 6.7.: Model initialization for anterior view. a) The binary mask is used to compute the maxillary teeth separations and identify the reference separation between the central incisor teeth. b) Gradient magnitude in vertical direction is used to determine the reference separation between maxillary and mandibular teeth. c) Resulting initial model position based on computed references.

and first pre-molar teeth) and scale are first computed by applying a similar process as for the anterior view. Next, the detected vertical teeth separations are used to calculate the width of the maxillary teeth in the image. The first mode of variations of the CSM (which corresponds to the largest variation in the coupled model and represents the shifting camera position – cf. Section 6.1.6.3) is then optimized to find the best match between the teeth width in the CSM and the calculated reference values.

After initialization, the next step is the adaptation of the coupled shape models to the input images to detect the teeth outlines in these images.

6.1.5. 2-D CSM adaptation

The last pipeline step of the teeth segmentation and numbering part is the adaptation of the initialized models to their corresponding input images. The adaptation is performed according to the concept (cf. Section 4.5) by minimizing an energy functional E which depends the global position of the coupled model (in the image) and the configuration of the CSM (i.e. teeth shapes and relative locations). Internally, the energy function aims to balance the energies of moving the landmark points in the direction of strong image features while restricting the model configuration to be close to the learned configuration space. The search for suitable image features is performed based on gradient image features during the first part of the adaptation and an appearance model during the second part – both parts will be explained later in this section. The appearance model contains the most discriminative color appearance profiles for each landmark point (cf. Section 6.1.1) and is trained together with the coupled shape model in the offline training phase.

The adaptation process for each view uses a specific dynamic adaptation strategy to ensure that model items have good overlap with their corresponding image structure before starting the search for image features. This is done because the mean model, which represents the average spatial configuration of its model items, is initialized only according to its center of mass and therefore cannot guarantee

a good overlap for all individual items after initialization. Instead, the items close to the center will have good overlap with the corresponding tooth in the image while the items farther away from the center might not match as good, which could lead to an incorrect adaptation of these items. Active shape and appearance models are known to be highly dependent on their initial position [HM09]. The dynamic adaptation strategies are therefore used to ensure that only items with (expected) good overlap are actively adapted while other items are passively moved closer to their correct position through the learned statistical information. The dynamic adaptation process is explained in more detail in Section 4.5 and consists of view-specific *adaptation strategies* which are separated into multiple *adaptation steps*. The strategy defines which model items are actively adapted during each adaptation step.

For the anterior CSM, the four central incisor model items have the biggest overlap with the corresponding image structures. The adaptation process starts to adapt these items first – the optimal model configuration that matches the detected image features for these four items will also affect the shape and location of the other model items through the learned configuration space. This should increase their overlap with the teeth in the image, enabling a more robust adaptation. The adaptation then gradually extends the set of active items, adding the four lateral incisor items next and finally the four canine items. Since the anterior model is relatively compact, the benefit of a dynamic adaptation process is limited. However, it should still provide superior results as it is able to compensate sub-optimal initializations.

Maxilla and mandible models will benefit significantly from dynamic adaptation strategies due to their U-shape structure. Both models are initially placed according to the central incisors. An incorrect scale or rotation of the model compared to the teeth in the image will result in low or no overlap of the items at the ends of the U-shape (i.e. molar model items). The gradual extension of the set of adapted items starting from the central incisors will allow the adaptation to follow the dental arch of teeth in the image. The adaptation strategy starts with the central incisors items and iteratively adds items farther away from the center.

The left and right CSMs behave similar to the anterior model as the included teeth have a similar overall spatial configuration. The items closest to the model center are the canine and first pre-molar items, which are adapted first. The subsequent steps continue to gradually include the other items.

Fig. 6.8 illustrates the order in which the model items are adapted for the different models. It displays the mean configuration of the anterior, maxilla and left models together with the order in which the individual items are adapted. Each number corresponds to the adaptation step during which the item is actively adapted for the first time – a steps i also includes the items of all previous steps j, with $j < i$. The two model that are not displayed have similar layouts to other model: the mandible model has the same adaptation order as the maxilla model (only for mandibular teeth) and the right model is identical to the left model.

Similar to the first application of the segmentation and numbering concept in the previous chapter (cf. Chapter 5), the robustness of the adaptation can be increased further by utilizing a masked version of the input images during the first part of the adaptation process. This masked input image is created

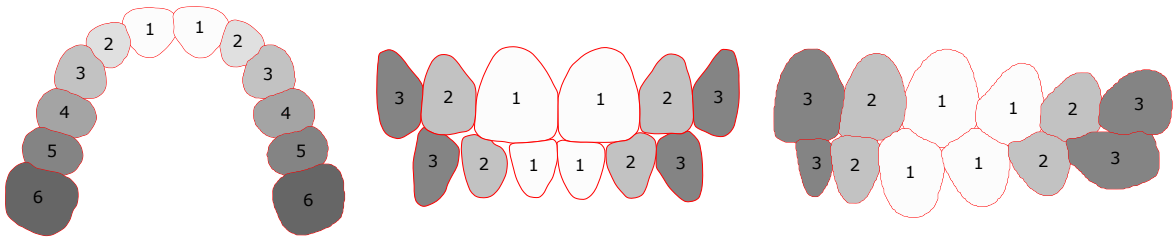


Figure 6.8.: Dynamic adaptation strategies: The numbered model items indicate the order in which they will be fitted to image features during model adaptation. The mandible model follows the same pattern as the maxilla model and the right model behaves similar to the left model.

by converting the photograph to a gray-scale version which is then masked by the binary teeth mask generated by the neural network. All pixels that are classified as teeth (white in the binary mask) keep their intensity (gray-scale) value while all other pixels are set to black (zero intensity). Gradient image features (based on gradient magnitude and direction) are then employed to fit the CSM to the masked input image. Similar to the previous application, this results in strong gradient features at the edges of the masked region, while the included intensity information also provides features inside the masked region (compared to just using the binary mask). This approach results in superior segmentation performance as it supports the dynamic adaptation strategies in ensuring an optimal overlap between mode items and corresponding image structures before switching to appearance features for the second part of the adaptation. The second part is then performed on the color image using the HSV appearance profiles trained during the offline training phase as the appearance model. After convergence, the adapted model provides the final teeth contours detected in the image together with the numbering encoded in the model.

6.1.6. Experiments and results

This section summarizes the experiments for the teeth segmentation and numbering in the five photographs of orthodontic records, as well as the achieved results. It starts by describing the data used for training and testing, and the performance metrics used for evaluation. Next, it covers the offline training of the coupled shape models and the optimization of the network parameters for the U-Net neural networks. The final sub-sections then present the results of the mask generation, the model initialization and the model adaptation steps.

6.1.6.1. Data

The experiments have been performed on 89 data sets. Each data set consists of one photograph for each the five views and the corresponding numbered gold standard annotations of the teeth. The gold

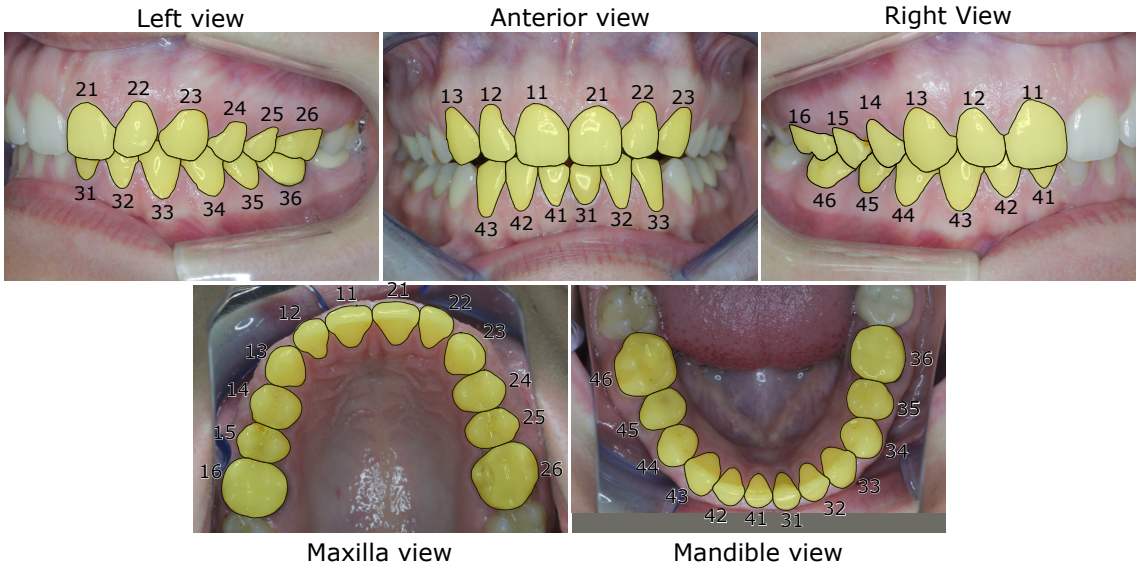


Figure 6.9.: Example of the gold standard annotations for one data set. The visible part of the teeth has been manually annotated and labeled with the corresponding number of the FDI notation system.

standard teeth contours and numberings have been obtained by manual annotation of the 12 relevant teeth contours (that are included in the design of the respective CSM) in each of the 445 photos (i.e. five photos for each of the 89 data sets). Fig. 6.9 illustrates exemplary manual annotations and numberings for one data set. The data sets also include the laser-scanned ground truth 3-D surface meshes of the individual teeth. These are relevant for the 3-D reconstruction part and will be introduced in the respective section (cf. Section 6.3.1). The colorized photographs each have a resolution of 5480x3662 or 5760x3840 pixels. Different camera were used to capture the individual data sets – however, all photographs belonging to one data set are always captured with the same camera. The 89 data sets were randomly separated into distinct training and test sets for the experiments.

A total of 50 data sets were assigned to the training set. In order to increase the number of training data, all training images have been mirrored to increase the total number of training samples to 100. This is made possible by the symmetric structure of the set of teeth (cf. Section 2.1). For each view, the training images and corresponding gold standard annotations were mirrored horizontally and the numbering of the annotations was updated accordingly. During this process, left and right views became the respective other view. The CSMs of both view are therefore effectively trained on the same data, with one being the mirrored version of the other. Nonetheless, they are treated as separate models due to the different numberings assigned to the model items (teeth). The remaining 39 data sets are considered the test set and used for performance evaluation.

6.1.6.2. Metrics

The performance of the teeth segmentation and numbering is evaluated individually for each view by comparing the numbered segmentations generated by respective CSM to the ground truth segmentations with the corresponding number. The comparison is performed using the same metrics that were used for the performance evaluation of the same task introduced in the previous chapter (cf. Section 5.2). These metrics are *precision*, *recall*, *accuracy*, *specificity* and *F1 score*. All metrics are based on the confusion matrix which summarizes the pixel-wise performance of the segmentation by accumulating the occurrences of four prediction categories. These categories are defined by the relation between predicted label and true label.

The accuracy measures the ratio of correct predictions over the total number of predictions, but is misleading when object and background cover different amounts of the image. Recall captures the ability of the approach to include the complete target object in the segmentation, while precision describes how much of the segmentation is actually correct. Specificity expresses the performance for correctly covering the complete background. The F1 score is the harmonic mean of precision and recall and is the most meaningful of the five metrics for judging the performance of the segmentation approach.

6.1.6.3. Offline training: 2-D coupled shape models

The five different coupled shape models are trained during the offline training phase as described in Section 4.2. The statistical information about the shape and spatial relation of the teeth outlines in each view is extracted from the training data (cf. Section 6.1.6.1) by establishing landmark correspondences between different training instances, applying generalized procrustes analysis to align the shapes, and utilizing principal component analysis (PCA) to find the modes of variation for both shape and spatial relation. For each statistical shape model (deformable model item) utilized in the CSMs, the contour is represented using $L = 100$ landmark points and 95 % of the shape variance are retained during training.

Fig. 6.10 illustrates the results from the offline training. It depicts the first mode of variation of the coupled shape models of the anterior view, left view and maxilla view. The models for right and mandible views behave very similar to the left and maxilla models, respectively. The individual model items are represented as colored closed contours with the color corresponding to the presented CSM design (cf. Fig. 6.3). The center of gravity of the different items is also shown since it is used to model the spatial relations.

The first mode represents the principal component with the largest variance in the learned shape space. For the anterior view, this seems to be the progressive occlusion of the mandibular anterior teeth by the maxillary ones (Fig. 6.10, left image of Anterior View CSM), and the opposite movement of creating a gap between lower and upper teeth (right image). The medical term for the first case (occlusion of anterior mandibular teeth) is overbite [Dof00], while a gap between the anterior teeth of both dental arches in a closed mouth position is referred to as open bite [Dof00]. This phenomenon is specifically important for the automatic initialization of the coupled shape model onto the anterior

image. In case of the left (and right) view models, the first mode represents a shift in camera position. The right image (Fig. 6.10, Left View CSM) corresponds to an almost lateral view of the teeth from the left side (of the patient). The viewing direction then shifts more and more towards a frontal view when one compares the right image to the middle and left image. This illustrates the large variance in the angle from which the left and right photographs are captured and presents a challenge for the model initialization process. The change of viewing direction results in large changes in the relative distance (and width) of teeth which makes it impossible to place the mean model onto the image and ensure an adequate overlap between all model items and their corresponding teeth in the image. This needs to be considered for the model initialization and adaptation process (dynamic adaptation process). The maxilla (and mandible) models show a smaller amount of variation in the first mode. The most prominent change is the shape of the arch which results in an increased distance between central incisors (Fig. 6.10, maxilla view, green contour) and first molars (pink).

6.1.6.4. Offline training: U-Net networks

The five U-Net networks for the semantic teeth segmentation (mask generation) are also trained in the offline training phase. Each view-specific network was trained on the respective photographs from the training set. As a reminder, the tooth-specific annotations of an image are merged into a single mask to serve as training data for the network. However, only the 50 non-mirrored versions (cf. Section 6.1.6.1) were used to enable an unbiased validation split. Horizontal mirroring is still performed on the two splits as part to the data augmentation process for anterior, maxilla and mandible view. Left and right view do not use mirroring as part of the data augmentation because the teeth structure in these images is not symmetric. Instead, more of the remaining augmentations are performed to generate a same amount of augmented images as for the other views.

The training was performed for 500 epochs with early stopping (i.e. the training stopped if no reduction in minimal loss was observed for at least 50 epochs), a batch size of 16, 250 steps per epoch, a validation split of 0.3 and a dropout rate of 0.3 in each pooling layer. The number of training images used in each epoch was increased to 4000 via the data augmentations (cf. Section 4.3). The training was conducted on a GPU-cluster with 2 x 16 core Intel Xeon Gold 6130s, 260GB of random-access memory and 8 x Nvidia GeForce RTX 2080 Ti GPUs with 11GiB graphics memory each.

Table 6.1 lists the performance of selected parameter combinations tested during the training process. The presented results are from the U-Net model for the maxilla view and are ordered by the achieved F1 score on the test set, from highest to lowest. Analyzing at the results provides clues on how individual parameters affect the performance. Using a lower input size for the network seems to improve the segmentation performance even though it also results in rougher boundaries when up-sampling the output back to its original size. Since the up-sampled output is used in subsequent pipeline steps, the lowest tested input size was 336x224 pixel. It performed slightly better then a resolution of 672x448, and significantly better than a resolution of 1344x896. Converting the image to HSV color space does not improve the results compared to the traditional RGB color space. While the HSV color space more closely models the way the human vision perceives color, this representation seems sub-optimal for

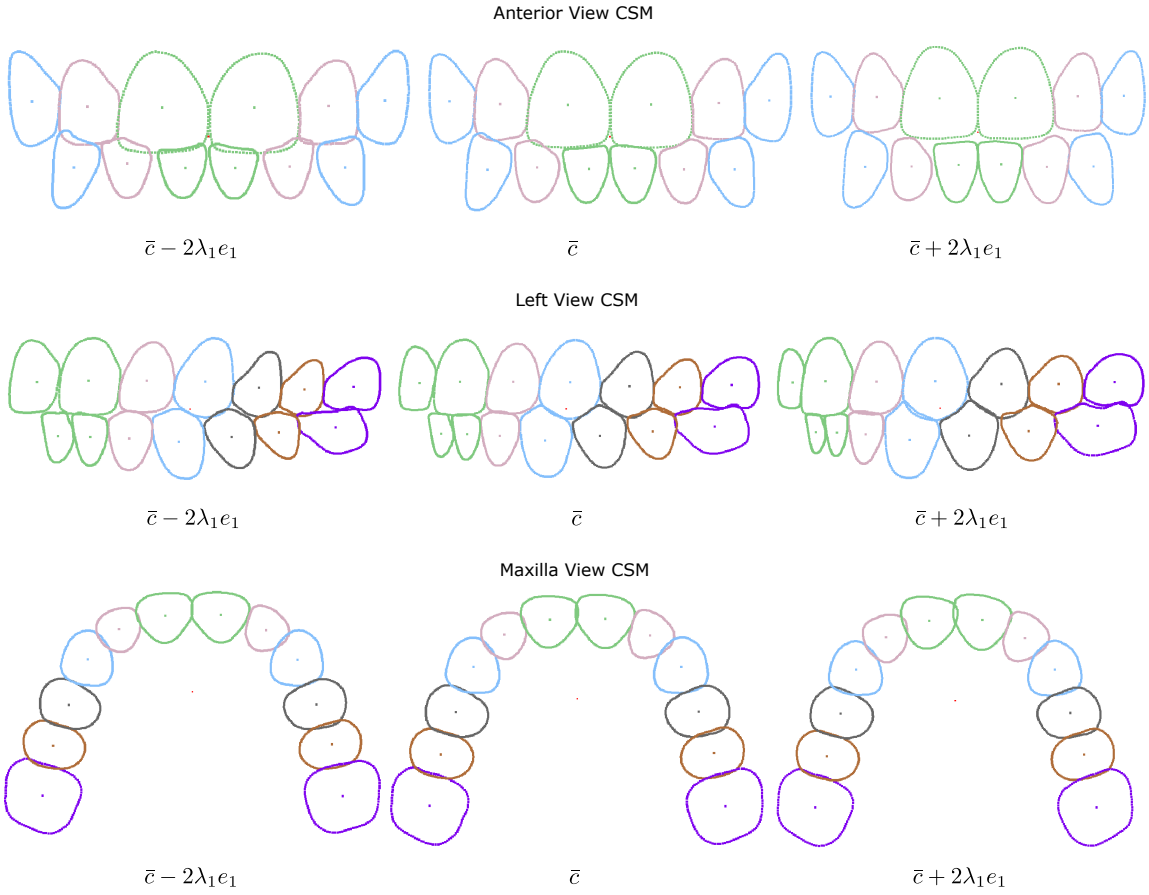


Figure 6.10.: First mode of variation of the coupled shape models for anterior, left and maxilla views which have been trained on 100 training samples. The colors of the individual model items correspond to the presented CSM design (cf. Fig. 6.3).

training neural networks. For the Tversky loss, $\alpha = 0.4$ and $\beta = 0.6$ generally outperformed other combinations. The best combination was found as input size of 336x224x3 with 64 filters, an initial learning rate of 0.001, and a Tversky loss with $\alpha = 0.4$ and $\beta = 0.6$.

Because all five photographs are relatively similar in terms of content, the time-consuming parameter optimization procedure was not repeated individually for the other four views. Instead, the parameter combination with the overall best performance for the maxilla view was used to train the other four networks as well. The trained networks are then used during the online processing to extract the binary teeth mask from the respective images.

Table 6.1.: U-Net training for semantic teeth segmentation (maxilla view). The table shows selected parameter configurations and the resulting F1 score on the validation and test data.

model parameter			color	training parameter			best	validation	test
width	height	filters	space	loss: α	loss: β	learning rate	epoch	F1 score	F1 score
336	224	64	RGB	0.4	0.6	0.001	35	**0.934**	**0.963**
336	224	64	RGB	0.3	0.7	0.001	60	0.934	0.962
672	448	64	RGB	0.4	0.6	0.001	48	0.917	0.958
672	448	64	RGB	0.4	0.6	0.010	120	0.932	0.955
336	224	64	HSV	0.4	0.6	0.001	17	0.929	0.954
672	448	64	RGB	0.3	0.7	0.001	71	0.901	0.953
672	448	64	RGB	0.2	0.8	0.001	70	0.921	0.949
672	448	64	HSV	0.3	0.7	0.001	46	0.910	0.947
672	448	32	RGB	0.3	0.7	0.001	42	0.891	0.944
672	448	32	RGB	0.4	0.6	0.001	18	0.909	0.944
672	448	64	RGB	0.3	0.7	0.010	29	0.909	0.936
1344	896	64	RGB	0.4	0.6	0.001	20	0.801	0.856

6.1.6.5. Mask generation

The mask generation is the second step of the online pipeline and utilizes the trained U-Net networks to extract a binary mask of the prominent teeth (i.e. only the teeth included in the design of the coupled shape models) from the individual photographs. The evaluation of this step was performed on the 39 test data sets. The tooth-specific gold-standard annotations of an image are again merged into a single mask for comparison. Table 6.2 lists the quantitative evaluation results when comparing the generated masks to the gold-standard masks.

All five networks achieve very good overall results. The highest F1 score is achieved by the anterior view U-Net with 0.9674 and standard deviation of 0.0261. The U-Net for the mandible view has the lowest F1 score of all five networks, with a value of 0.9329 and a standard deviation of 0.0740. Semantic segmentation performances of this level should generate binary masks that are usable for the automatic model initialization in the next pipeline step. The achieved results outperform the results of the semantic segmentation of teeth in panoramic radiographs performed in the previous chapter - despite a significantly lower amount of training data (50 vs. 432). The added color information of the photographs and the resulting higher contrast between teeth and surrounding tissue seems to be the most likely explanation for the increase in performance.

A visual representation of the results is provided in Fig. 6.11. It shows the best and worst segmentation results for each view. Each depicted result consists of two images: the left image includes a green overlay of the computed binary mask onto the photograph, the right image shows the binary mask with the gold-standard annotation overlayed in turquoise. The best results show almost perfect

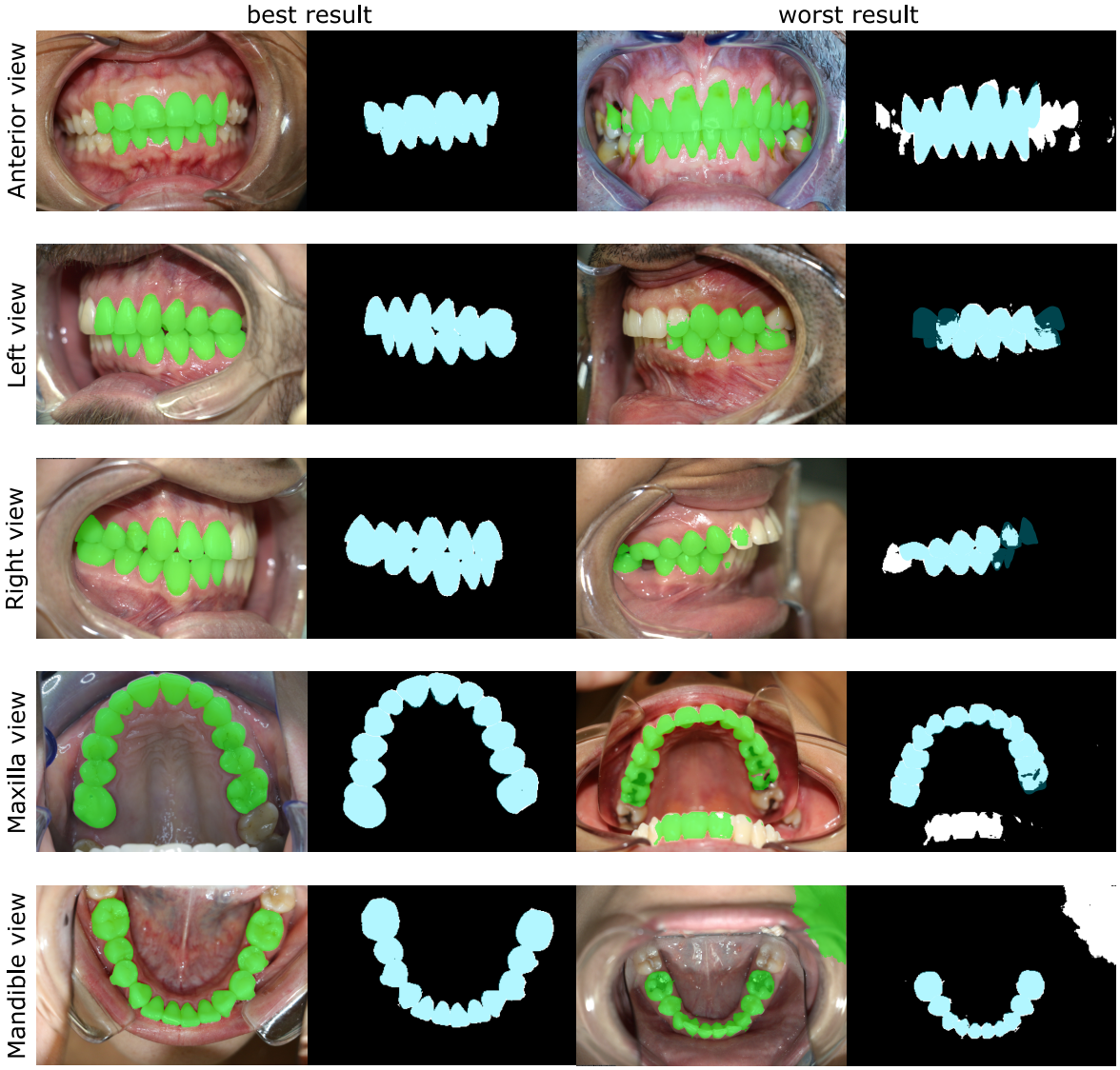


Figure 6.11.: Semantic teeth segmentation results (binary mask) for all views. The results are each depicted as a green overlay on the original image and as the computed binary mask with a turquoise overlay of the ground truth annotation. Each row shows the best and worst segmentation result on the test set for each view (based on F1 score).

Table 6.2.: Semantic segmentation performance of the five view-specific U-Net networks trained with the optimal parameter configuration from the maxilla view network. The table shows the chosen performance metrics together with their standard deviation.

	Precision	Recall	Accuracy	Specificity	F1 score
U-Net (anterior view)	0.980 ± 0.015	0.956 ± 0.047	0.992 ± 0.009	0.998 ± 0.002	0.967 ± 0.026
U-Net (left view)	0.934 ± 0.068	0.954 ± 0.043	0.986 ± 0.011	0.991 ± 0.009	0.942 ± 0.046
U-Net (right view)	0.969 ± 0.041	0.951 ± 0.055	0.990 ± 0.011	0.996 ± 0.005	0.959 ± 0.041
U-Net (maxilla view)	0.962 ± 0.013	0.966 ± 0.052	0.989 ± 0.007	0.993 ± 0.003	0.963 ± 0.030
U-Net (mandible view)	0.980 ± 0.034	0.899 ± 0.117	0.982 ± 0.020	0.997 ± 0.007	0.933 ± 0.074

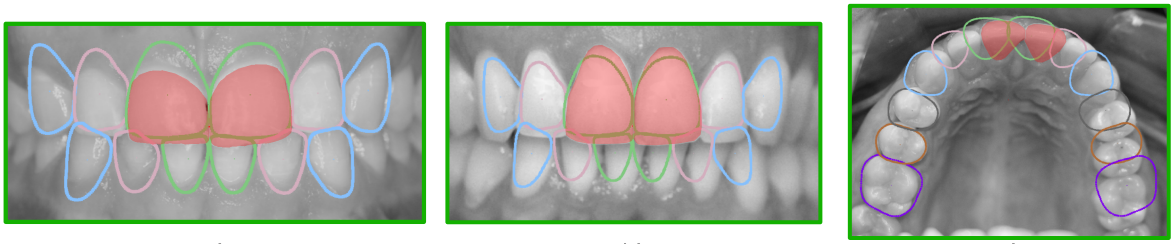


Figure 6.12.: Automatic model initialization results for anterior and maxilla view. The images depict the CSM position and scale on the photograph (images are cropped for better visualization). A good initialization shows sufficient overlap between central incisor model items (green outlines) and central incisor teeth (red overlays).

overlap between the generated output and the gold standard annotation for all five views. The worst results include incorrect segmentations in other image regions (maxilla and mandible views), and over- or under-segmentation in the teeth area (anterior, left and right view). The automatic initialization performance on these binary masks will be assessed in the next section.

6.1.6.6. Model initialization

The performance of the automatic computation of the initialization parameters is assessed by visually inspecting the position and scale of the coupled shape models on the image after the initialization as well as the result after the first adaptation step. This method of evaluation is used as there exists not only a single set of correct parameters but a subspace of (valid) parameter combinations that will result in a valid adaptation. Whether or not a given parameter combination leads to a valid adaptation is also highly dependent on the chosen adaptation parameters like search radius or used image features. For

this reason, simply calculating the overlap between the model items and corresponding teeth in the image also does not provide sufficient information to distinguish a successful from a failed initialization. Assessing if the first adaptation step manages to adapt the included model items to their correct image structures provides the most reliable method of judging the initialization performance on the test set.

Anterior view The chosen initialization method for the anterior view was able to correctly initialize the CSM onto the photograph on all 39 test instances. This means that the maxillary central incisor model items are close enough to the maxillary central incisors in the photographs and are adapted successfully during the first adaptation step. Additionally, the scale of the entire coupled shape model matches the respective teeth in the photograph. Consequently, the horizontal teeth separations used as reference for the initialization were detected successfully on the binary mask and the separation between the two central incisors was identified correctly. Examples of some initializations are depicted in Fig. 6.12a) and b). Central maxillary incisors are highlighted in red for better reference.

Maxilla & mandible view The automatic initialization for maxilla and mandible photographs aims to position the models according to the central incisors as well. A ray-based search is performed to find the U-shaped teeth arch in the image and position and scale the model accordingly. Due to the stable viewing direction observed on the training data, no rotation correction is performed. Using this method, the maxilla CSM was initialized correctly on all 39 instances of the test set. One such initialization is shown in Fig. 6.12c. The mandible model was initialized correctly in 37 of the 39 test cases (i.e. 95 % of initializations are correct). Some initialization results are depicted in Fig. 6.13. The central incisors are highlighted in all images with a red overlay, as these should overlap with the central incisor model items (green outline). Successful initializations are indicated by a green border, failed ones with a red border. One of the failed cases was caused by the photo being taken from a different viewing direction which resulted in a rotation of the U-shaped teeth arch. This rotation meant that the incisor model items were not close enough to the correct teeth in the image, which lead to an incorrect adaptation (cf. Fig. 6.13c). The second failed case was due to a missing lateral incisor with the gap originally left by the missing tooth already closed. This again caused the incisor model items to not line up with the correct teeth in the image (cf. Fig. 6.13b).

Left & right view Left and right views are the most challenging ones for the automatic initialization because of the large variation in viewing direction. For these views, the initialization process aims to position the models according to the canine and first pre-molar teeth/model items. The process tries to detect the teeth separations in horizontal direction based on the binary mask and place the model by identifying the separation between canine and first molar. This process is challenging as (relative) teeth width changes with different viewing directions. Contrary to the other views, the CSMs are not initialized in their mean configurations but the model configuration is optimized to better match the viewing direction. However, this optimization is not considered when judging a successful initialization because a successful adaptation of the items further away from the model center is only evident during

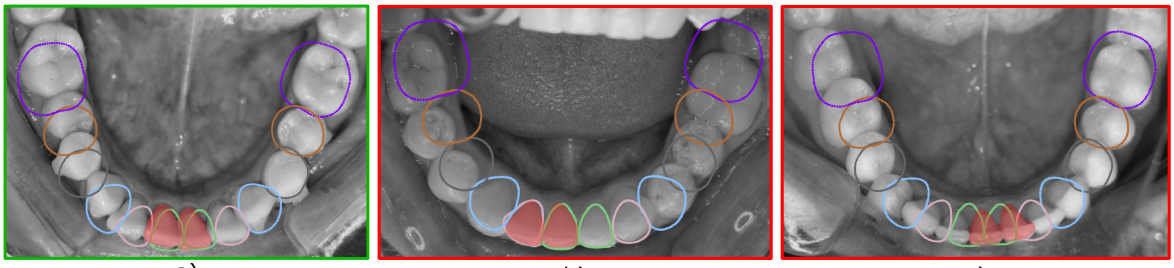


Figure 6.13.: Automatic model initialization results for the mandible view. Shown are the CSM position and scale on the photograph (images are cropped for better visualization). A good initialization shows sufficient overlap between central incisor model items (green outlines) and central incisor teeth (red overlays). Valid initializations are indicated by a green border, invalid ones with a red one.

later steps of the adaptation due to the dynamic adaptation process. The optimization quality will be reflected in the quantitative evaluation of the final segmentation performance.

For the left view, 33 out of 39 models (85 %) were correctly initialized, for the right view, 36 of 39 (92 %). Fig. 6.14 depicts examples of successful and failed initializations of both views. The canine and first pre-molar teeth are again highlighted in red, the corresponding model items are depicted with a blue (canine) and gray (first pre-molar) outline. A green border indicates a successful initialization, failed initializations have a red border.

6.1.6.7. Model adaptation

The final step of the teeth segmentation and numbering pipeline is the model adaptation. During this step, the view-specific CSMs are fitted to the respective photographs as described in Section 6.1.5. This process is defined by adaptation strategies that control which model items are actively adapted during the individual adaptation steps (dynamic adaptation). Additionally, each adaptation step includes several parameters like, e.g., local search radius, number of iterations, and weights for the energy functional. These parameters were optimized empirically during the experiments by maximizing the segmentation performance (F1 score) on the 50 training data sets. The CSMs were manually placed onto each training image for the optimization process to ensure that failed automatic initializations (on the training data) did not negatively impact the estimation of the adaptation parameter values. The parameter optimization is performed individually for each CSM with the exception of the left and rights views. Here, the same parameters are used for both models while accounting for the fact that one is the mirrored version of the other. The determined optimal parameters for each view are then used to evaluate the performance for the model adaptation on the test data. The performance is assessed by comparing the tooth segmentations given by the final model configurations to the gold standard annotations using the metrics described in Section 6.1.6.2 (precision, recall, accuracy, specificity, F1 score).

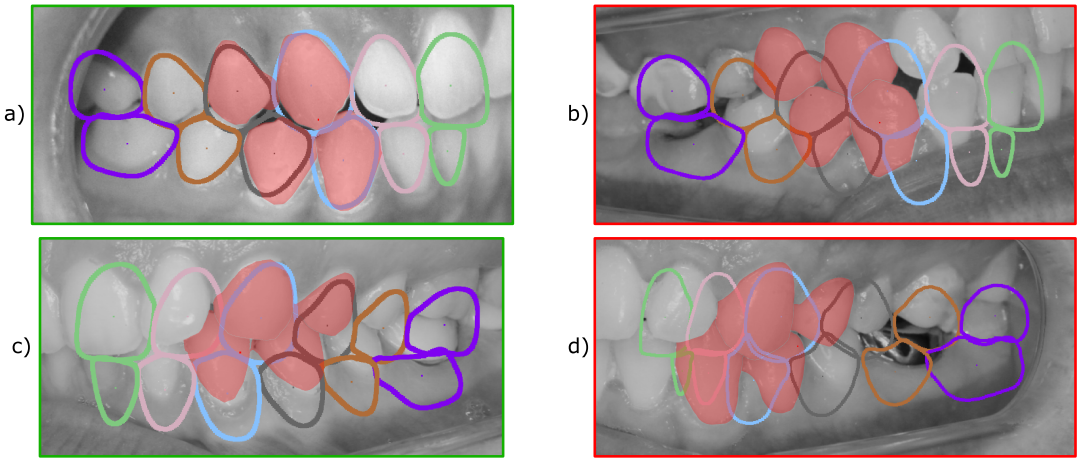


Figure 6.14.: Automatic model initialization results for lateral views. The relevant teeth for the initialization (canine and first pre-molar) are highlighted in red. A good initialization result should show good overlap with the corresponding model items (blue and gray outlines). Successful initializations are depicted with a green border, failed ones with a red border.

The overall measures are calculated by averaging the tooth-specific metrics of all test cases for the specific view. Standard deviation from the mean is provided as an additional measure to indicate the amount of variation.

Table 6.3 summarizes the performance of the proposed approach to extract the prominent teeth for each of the five photographs. The highest segmentation performance is achieved on the maxilla view with an F1 score of 0.931 ± 0.107. The lowest F1 score was 0.631 ± 0.389 for the left view. It is noteworthy that the best performing views had no failed initializations (maxilla and anterior), while the lowest performing view suffered from six failed initializations (left view). A failed initialization generally results in a very poor segmentation performance (i.e. F1 score < 0.2). This is also reflected in large standard deviations for views with failed initializations. A collection of segmentation results is depicted in Fig. 6.15.

The numbering performance is assessed using the pixel-wise IoU (cf. Section 5.2) between individual tooth segmentation and corresponding manual annotations. A numbering is considered correct, if the IoU is above a predefined threshold. Five different threshold values are evaluated, with higher thresholds putting more emphasis on accurate pixel-wise classification. Table 6.4 lists the individual results for the five views. The best results are achieved for anterior and maxilla views with detection rates of 0.950 and 0.982 for an IoU threshold of 0.5. The mandible view reaches a ratio of 0.894, while right and left views perform the worst with 0.757 and 0.627, respectively. Higher thresholds naturally result in lower performance scores, since more accurate segmentations are required. While successful detection rates slowly decline for thresholds of 0.6 and 0.7, a threshold of 0.8 results in a large perfor-

Table 6.3.: Evaluation of the teeth segmentation and numbering for the five different views. The table contains the overall performance (mean and standard deviation) for different metrics computed by comparing the generated segmentations to the gold standard annotations.

	Precision	Recall	Accuracy	Specificity	F1 score
Anterior view	0.891 ± 0.160	0.884 ± 0.155	0.998 ± 0.002	0.999 ± 0.001	0.882 ± 0.161
Left view	0.633 ± 0.397	0.653 ± 0.390	0.994 ± 0.008	0.996 ± 0.005	0.631 ± 0.389
Right view	0.756 ± 0.326	0.763 ± 0.314	0.996 ± 0.006	0.998 ± 0.003	0.751 ± 0.317
Maxilla view	0.940 ± 0.103	0.924 ± 0.115	0.998 ± 0.004	0.999 ± 0.002	0.931 ± 0.107
Mandible view	0.852 ± 0.262	0.832 ± 0.259	0.996 ± 0.007	0.998 ± 0.003	0.840 ± 0.258

mance drop. Segmentation and numbering performance are closely related since both are based on the numbered teeth segmentation provided by the adapted coupled shape model.

6.1.7. Discussion

The presented fully automatic segmentation and numbering approach for teeth in five colored photographs is used to extract the teeth outlines which will be used in the next part of the workflow to reconstruct the 3-D shape and spatial configuration of the teeth. The contour extraction presents the second application of the segmentation and numbering concept introduced in Chapter 4. It was trained on 50 training data sets (which was increased to 100 by mirroring the images horizontally) and tested on 39 data sets. After removing specular reflections from the photographs, trained U-Net networks are used to generate a binary mask of the teeth area for each image which is then used to initialize a trained view-specific coupled shape model onto the respective image. These models are fitted to the images using dynamic adaptation strategies by employing gradient image features of the masked input image during the first half of the adaptation and color appearance profiles on the full input image during the second half. The final fitted models then provide the teeth segmentations and numberings for each of the five photographs. Fig. 6.16 illustrates the results of selected steps from the teeth extraction process for one set of photographs. For each individual view, the results of the binary mask computation, the model initialization and the model adaptation are visualized.

The experimental results on the test data revealed varying performance between different views. While the segmentation and numbering performance was good for anterior, maxilla and mandible views with F1 scores of 0.882, 0.931, and 0.840, respectively, the results for right and left views were lower with 0.751 and 0.631, respectively. The different factors impacting the performance will be discussed in the following paragraphs.

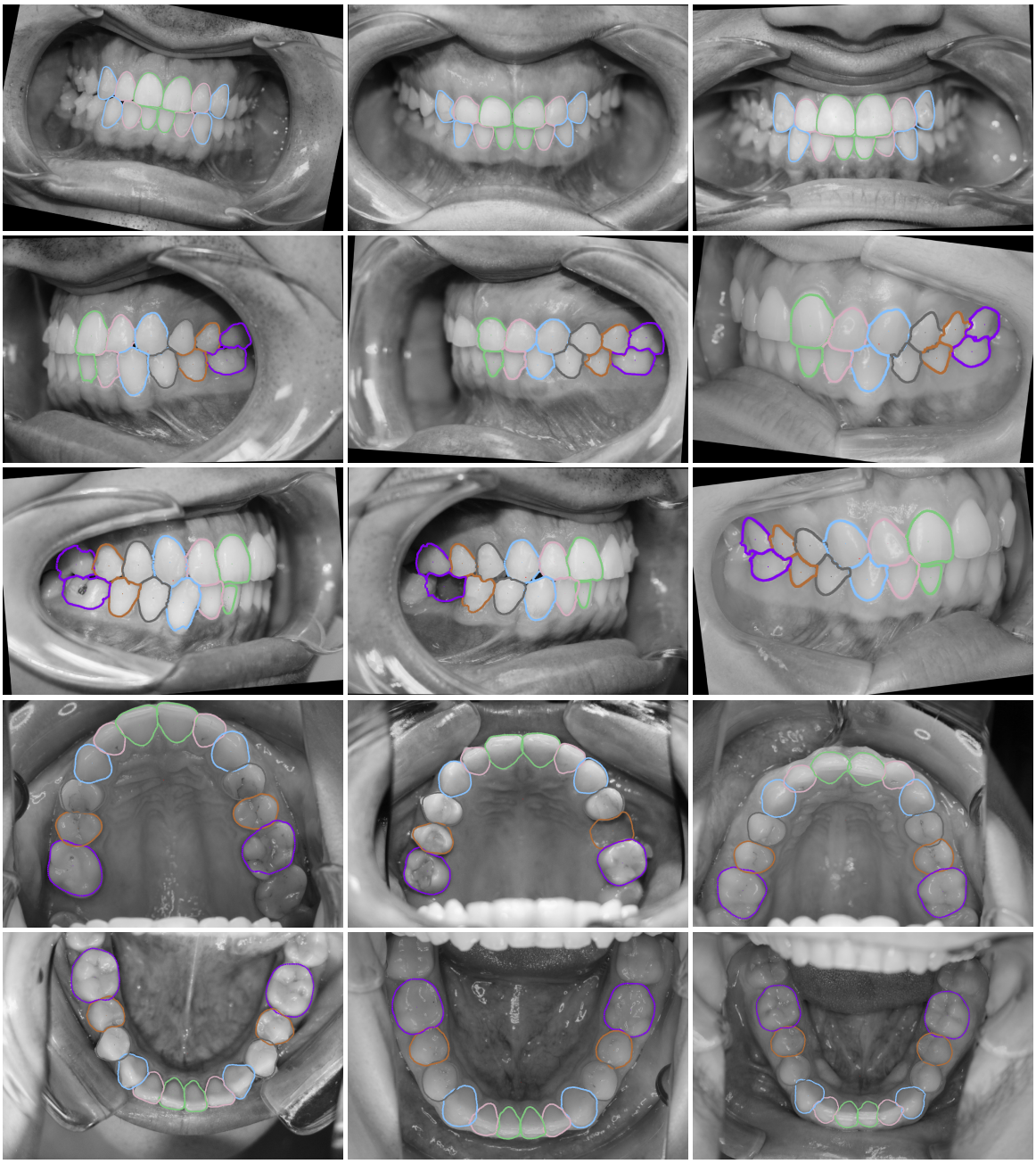


Figure 6.15.: Collection of teeth segmentation (and numbering) results. The detected contours are illustrated as landmarks points on the intensity image. The contour color indicates different tooth types as described in the CSM design (cf. Section 6.1.1). The results in one column do not belong to the same patient.

Table 6.4.: Evaluation of the numbering performance for the five views using different pixel-wise IoU thresholds for correct detections. The table contains the overall performance on all cases of the test set as well as separate entries for the average performance in case of successful and failed automatic initializations.

	IoU > 0.5	IoU > 0.6	IoU > 0.7	IoU > 0.8	IoU > 0.9
Anterior view	0.950 ± 0.212	0.939 ± 0.233	0.909 ± 0.279	0.760 ± 0.414	0.226 ± 0.403
Left view	0.627 ± 0.473	0.593 ± 0.481	0.555 ± 0.486	0.483 ± 0.487	0.281 ± 0.437
Right view	0.757 ± 0.419	0.718 ± 0.439	0.675 ± 0.457	0.593 ± 0.479	0.353 ± 0.465
Maxilla view	0.982 ± 0.130	0.975 ± 0.152	0.970 ± 0.165	0.940 ± 0.230	0.594 ± 0.476
Mandible view	0.894 ± 0.304	0.883 ± 0.318	0.864 ± 0.339	0.777 ± 0.409	0.331 ± 0.457

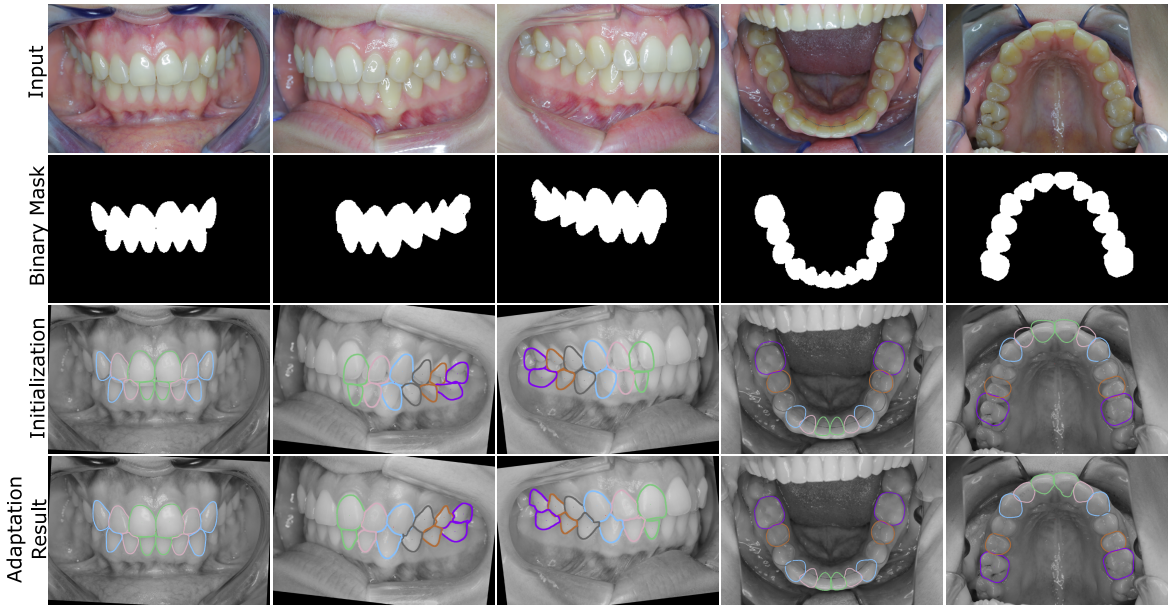


Figure 6.16.: Steps of the teeth contours extraction. First row: input photographs for all views; Second row: binary mask of the teeth area; Third row: placement of the mean CSM after the initialization step; Forth row: final adaptation (and numbering) result.

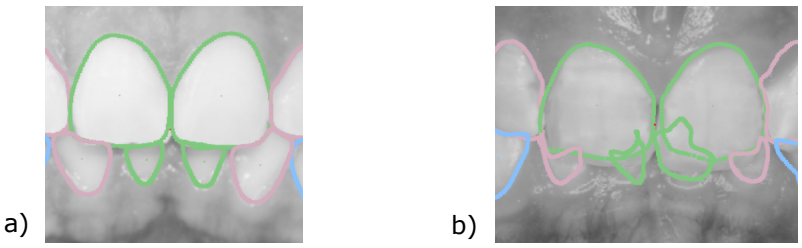


Figure 6.17.: Handling of partially and completely occluded mandibular teeth in anterior view photographs. Partly occluded teeth are correctly segmented using the shape variance encoded in the CSM (left image). Complete occluded teeth result in incorrect segmentations as the CSM adapts to wrong image features (right image).

Because the coupled shape models consist of multiple active shape models with local search algorithms, a good initial placement is paramount for a robust adaption. If the model items are placed too far from their corresponding image structures, the local search will not be able to fit the items correctly. Instead, model items might be adapted to neighboring structures or random features, resulting in a failed adaptation and very low performance scores. The utilized concept makes use of binary masks to tackle this problem. These semantic segmentation masks are generated by U-Net networks which achieve a F1 score between 0.933 and 0.967 for the different views. The distinct color of teeth compared to their surrounding likely contributes to these high performance values. Moreover, the trained networks also correctly omit the teeth that are not part of the coupled shape model and therefore where omitted during network training. The high performance supports the initialization processes for the individual views. Overall, only 11 out of 195 initializations were considered incorrect, 9 of those are from the left or right view. The failed initializations result in very low segmentation performance for the affected images, which is reflected in the large standard deviations for the affected views (mandible, left, and right views). Initialization for the lateral views is especially challenging because of the large variation in camera positions these images are captured from. The failed initializations for these views are mostly related to an incorrect placement in horizontal direction or an incorrect rotation correction (cf. Fig. 6.14).

Apart from the initialization dependence, some views also suffered from other challenges. The mandibular incisors in anterior view photographs can be partly or completely occluded by the maxillary incisors (overbite condition). While the model is able to handle partly occluded teeth via the encoded statistical variance, completely occluded teeth are problematic because the CSM still tries to fit the corresponding model item to the image. Since no 'true' image features of these teeth exist, the model item is inevitably fitted to incorrect features. This can also negatively affect the adaptation of other items and accounts for some of the performance difference between maxilla and anterior views despite both having no failed initializations. Fig. 6.17a shows correctly handled partly occluded teeth, Fig. 6.17b depicts a case with completely occluded mandibular central incisors and how an incorrect

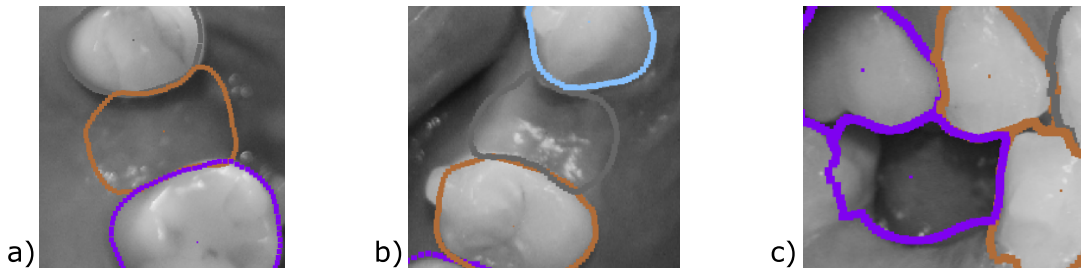


Figure 6.18.: Model adaptation results in cases with missing teeth: a, b) the model item corresponding to the missing tooth was correctly placed in the resulting gap. c) Incorrect adaptation result if the gap does not match the original tooth size (due to missing teeth revealing previously occluded teeth).

adaptation of the corresponding model items (small green contours) affects other items (i.e. incorrect lower boundary of maxillary central incisors).

The adaptation results of left and right views are affected by large variations in viewing direction. Despite an initial optimization during the initialization process to better align the model configuration to the viewing direction of the photograph and the dynamic adaptation process, model items still do not overlap well at the beginning of their adaptation. While this can be compensated by using a larger local search radius when looking for image features, this also increases the risk of fitting model items to incorrect features – which results in lower segmentation performance. Another potential problem is the large teeth shape variance introduced by the changing viewing direction. This reduces the models ability to constrain the shapes since more variations are considered valid. In general, variations in camera position need to be reduced for left and right views (e.g. by a more precise guideline on how to take the photographs) to enable a more robust adaptation. Views with more consisted camera positions, like anterior view, perform significantly better.

Missing teeth generally provide another challenge that affects the performance of the presented approach. Similar to the concept application for panoramic radiographs (cf. previous chapter), the approach can handle missing teeth if the resulting gap is still present since the gap is visually different compared to the tooth. The corresponding model item can then occupy this space and other model items can still be adapted correctly (cf. Fig. 6.18a,b). Since no valid appearance features are detected for all (or the majority of) landmark points of this model item, this clue can be used to flag the tooth as missing and exclude it from the subsequent 3-D reconstruction.

On the other hand, if the gap is no longer present, the CSM is unable to detect that a specific tooth is missing. Left and right models sometimes struggle to deal with missing teeth. Because of the sideways perspective, a missing tooth with gap reveals the previously occluded part of the neighboring tooth which results in a gap that is smaller than the expected size of the tooth outline. This can negatively affect the model adaptation as the model energy constrains the item to valid shapes (according to the training data) and will prevent it from shrinking correctly (cf. Fig. 6.18c).

Overall, the presented teeth segmentation and numbering provides a fully-automatic approach to extract the tooth outlines and the corresponding numbering from the five photographs. The relatively stable viewing directions of anterior, maxilla and mandible views enable a robust modeling of the statistical variance in tooth shape and relative location in form of a view-specific coupled shape model. The CSM then supports an accurate tooth segmentation via gradient features and color appearance profiles and inherently provides the numbering information as each tooth is modeled via a specific model item. The left and right views offer a bigger challenge due to the large variability in viewing direction. This leads to more variance in the statistical information encoded in the respective view-specific CSMs and a less robust segmentation performance. The initialization dependency of such types of statistical models is successfully handled by utilizing a semantic teeth segmentation mask. These masks are computed by offline trained neural networks (U-Net) and achieve high accuracy due to the available color information. The presented application of the teeth segmentation and numbering concept offers the information that is required to enable a fully-automatic reconstruction of the 3-D teeth shapes and spatial configuration via a silhouette-based loss. This reconstruction process is described in the next section.

6.2. Deformation-based reconstruction

The extracted numbered teeth contours of all input images are now used to recover the 3-D shapes and relative locations of the teeth. The presented model-based reconstruction approach combines the extracted teeth contours with prior knowledge about the 3-D shape and location of the teeth to handle the limited amount of information that is available from the five photographs. Two 3-D coupled shape models, one for the upper and one for the lower dental arch, encode the prior knowledge and are deformed using a silhouette-based loss. In this context, the silhouette-based loss aims to minimize the difference between 2-D projections of the reconstructed 3-D teeth and their corresponding 2-D contours in order to determine the two 3-D CSM configurations that best explain the numbered contours extracted from the photographs. This fully-automatic process requires that the correspondences between 2-D outlines and 3-D teeth model are known, which is achieved through the inherent numbering provided by the chosen contour extraction approach (cf. previous section).

The reconstruction process is separated into two parts and corresponds to the second and third steps of the reconstruction workflow introduced at the beginning of the chapter (cf. Fig. 6.1). First, the two 3-D CSMs need to be initialized in terms of overall scale, teeth positions and relative position of both models to ensure a good initial starting point for the deformation-based reconstruction. The contours extracted from maxilla and mandible views are used here to optimize the scale and spatial configuration of the teeth in the respective model as these contours suffer the least from occlusions and the top-down viewing direction offers a good representation of the teeth relative location. The relative position of the coupled shape models to each other is then estimated using the contours from anterior, left and right views. This initialization adjustment of both models should provide a good starting point for generating the 2-D projections that are used for computing the silhouette-based loss.

The final part of the reconstruction workflow is the deformation-based reconstruction of the teeth shapes that aims to find the model configurations such that the projected 2-D silhouettes of the 3-D model items match the 2-D teeth outlines extracted from the input images (silhouette-based loss). This expectation maximization-like process estimates the camera parameters that correspond to each view based on the current model configurations, uses these camera parameters to calculate 2-D projections of the 3-D model items (teeth), and deforms the 3-D model items in order to minimize the silhouette-based loss. Because the silhouettes and contours only provide information about the visible part of a tooth, the prior statistical knowledge encoded in the CSMs complements this with information about the (expected) shape of the occluded part and helps to constrain the reconstruction to valid teeth shapes. The iterative process repeats until a predefined number of iterations has been reached or the silhouette-based loss drops below a certain threshold.

Before these two reconstruction steps are explained, the design of the coupled shape models, the silhouette-based loss function and the camera parameter estimation are described as all are utilized in these reconstruction steps. Therefore, the structure of this section is as follows. Section 6.2.1 covers the design of the 3-D CSMs. The silhouette-based loss is explained in Section 6.2.2, followed by the projection parameter estimation in Section 6.2.3. The initialization of the 3-D CMS is then covered in Section 6.2.4 and finally, in Section 6.2.5, the deformation-based shape reconstruction is introduced. Experiments and results are then presented in the following section.

6.2.1. 3-D CSM designs

The employed 3-D coupled shape models represent the maxillary dental arch and mandibular dental arch, respectively, and encode both the statistical shape variance of each tooth as well as the statistical variance in their relative position. The construction and training of both models is similar to the 2-D coupled shape model introduced in Section 4.2, but uses the implementation of Jung et al. [JSK*14]. Deformable model items are represented as triangle meshes and again modeled using a point distribution model. Landmark correspondences between triangular meshes of the same tooth are established using a pairwise registration method [Kir13] and are aligned using generalized procrustes analysis [Gow75]. For modeling the spatial relation, 3-D transformations with $m = 7$ parameters are used which includes 3 parameters for translation, 3 for rotation and 1 for scaling. The statistical knowledge contained in the CSM is utilized to support the reconstruction process as the information obtained from the photographs is limited and the depicted teeth suffer from heavy occlusions. The two dental arches are represented by individual models to enable a separate optimization of the teeth positions and shapes in both models as well as the relative location of both models.

The two models contain the same teeth that are included in the five 2-D coupled shape models for contour extraction:

- **Maxilla model** ($N = 12$ items): maxillary right first molar (16) to maxillary left first molar (26),
- **Mandible model** ($N = 12$ items): mandibular right first molar (46) to mandibular left first molar (36).

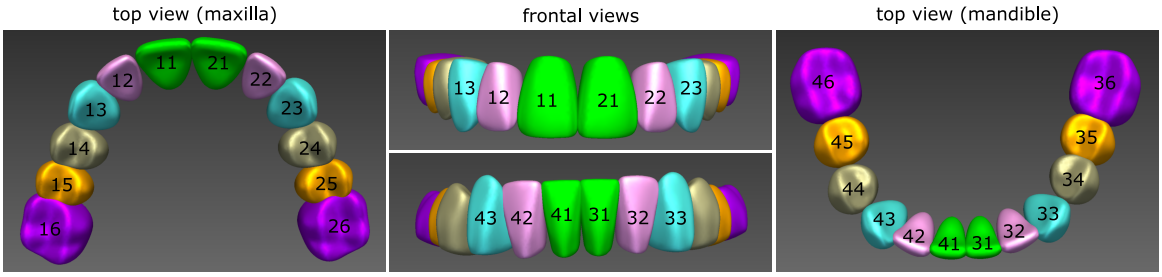


Figure 6.19.: CSM design for the two 3-D coupled shape models used for teeth reconstruction. The colors indicate different tooth types, the numbers refer to the FDI notation system. The color scheme corresponds to the 2-D CSM design to allow easy identification of matching items.

Second molars and wisdom teeth are not included since there is no contour information available to perform any kind of reconstruction – no model items for contour extraction were included in the 2-D CSMs because these teeth were rarely visible in the photographs. While it would be possible to include these teeth in the model and recover their most likely shape and position based on the statistical relation to the other (reconstructed) teeth, this was not the scope of the presented approach. Fig. 6.19 depicts the teeth included in both models together with the corresponding numbering from the FDI notation system. The teeth are depicted from a top-down view (left and right images) and a frontal view (middle images). The individual meshes are colored in the same way as the contours of 2-D CSMs to allow for an easy visual identification of matching items. The models are trained in an offline training phase which is explained in Section 6.3.3.

6.2.2. Silhouette-based loss

The teeth reconstruction and camera parameter estimation is based on a silhouette-based loss between projected 2-D silhouettes (of the reconstructed 3-D teeth) and the extracted teeth outlines. The general idea behind such a loss function, which is inspired by space carving [KS00], is that the 2-D silhouettes should (given the correct intrinsic and extrinsic camera parameters) match the 2-D contours extracted from the input image. Any discrepancy between the two implies that the 3-D shape needs to be deformed to generate a better estimate of the true 3-D shape. The difference $\mathcal{E}_{SL}$ between projected points and extracted contour points for a model item i in view v is given by

$$\mathcal{E}_{SL}(i,v,f,t) = dist\left(P\left(f,t,i;\theta^{(v)}\right),\mathcal{C}^{(v)}(i)\right), \tag{6.2}$$

where $dist(\cdot,\cdot)$ is an appropriate distance metric between both points sets (e.g. L_2, binary cross-entropy loss or negative intersection-over-union) [HLB21] and $P(\cdot)$ is a 3-D to 2-D (perspective) projection function which takes occlusions into account. The parameters f and t represent the current model configurations and global transformations, $\theta^{(v)}$ are the camera parameters corresponding to view v

and $\mathcal{C}^{(v)}(i)$ is the 2-D contour extracted from view v that corresponds to model item i. The overall discrepancy of the reconstructed model is given by the combined loss of all model items in all views:

$$\mathcal{E}_{CSL}(f,t) = \sum_{\forall v \in \mathcal{V}} \sum_{\forall i \in \mathcal{N}(v)} \mathcal{E}_{SL}(i,v,f,t). \tag{6.3}$$

Here, $\mathcal{V}$ is the set of all available views which are anterior, left, right, maxilla and mandible views, and $\mathcal{N}(v)$ are the teeth included in the 2-D CSM of view v.

In this approach, the negative intersection-over-union (IoU) is used as the distance metric $dist(\cdot,\cdot)$ to measure the dissimilarity between the areas enclosed by the projected and extracted contours. Similar to the IoU defined as an evaluation metric for the numbering accuracy (cf. Section 5.2), the metric is again computed in a pixel-wise fashion using the confusion matrix to get more accurate results (compared to the version based on bounding boxes). The negative IoU is therefore given by

$$L_{IoU} = 1 - \frac{TP}{TP+FN+FP}. \tag{6.4}$$

6.2.3. Estimation of projection parameters

The projection of a 3-D model tooth onto a 2-D image plane to determine its 2-D silhouette is an integral part of the reconstruction process. On the one hand, the projected silhouettes are utilized to calculate the loss compared to the extracted outlines in order to measure the quality of the reconstructed 3-D shapes. On the other hand, the projection is used to identify the 3-D vertices of the model tooth that define the silhouette. These vertices will be updated directly during the optimization process to deform the 3-D model such that it better matches the visual hull [Lau94]. To be able to get an accurate representation of 2-D silhouettes for each view, the projection (or camera) parameters need to resemble the original camera parameters used to capture the photo as close as possible.

Because all photographs are captured using a freely movable camera (i.e. an uncalibrated camera setup), an estimation of the camera parameters $\theta^{(v)}$ for each view v is required. These parameters are used as input to the perspective projection function $P\left(f,t,i;\theta^{(v)}\right)$ (cf. Eq. (6.2)) which then generates the silhouettes for the respective view. It is important to note that the projection respects occlusions between teeth and only outputs the silhouette of the visible part of the tooth i. The camera parameters to be estimated are position, up-vector and focal point – all other camera parameters are assumed to be fixed. The optimization is performed independently for each view which can be considered a special case of rigid structure-from-motion [WA82] and allows for an accurate and robust calibration of projection parameters while optimizing the model configuration in the same process [HLB21].

Starting from predefined default parameters for each view (according to the expected viewing directions), the individual optimal parameters $\theta^{(v)}_{opt}$ are determined by minimizing the silhouette-based loss

for this view only while using the corresponding items of both 3-D models:

$$\begin{aligned}\theta_{opt}^{(v)} &= \underset{\theta^{(v)}}{\operatorname{argmin}} \sum_{\forall i \in \mathcal{N}(v)} \mathcal{E}_{SL}(i,v) \\ &= \underset{\theta^{(v)}}{\operatorname{argmin}} \sum_{\forall i \in \mathcal{N}(v)} dist\left(P\left(f,t,i;\theta^{(v)}\right), \mathcal{C}^{(v)}(i)\right).\end{aligned} \tag{6.5}$$

Again, $v \in \mathcal{V}$ is one of the available views, $\mathcal{N}(v)$ are the teeth included in the 2-D CSM of view v, $dist(\cdot,\cdot)$ is the negative IoU, $P(\cdot)$ is the projection function and $\mathcal{C}^{(v)}(i)$ is the tooth outline exacted from the photograph.

Since the optimal camera parameters depend on the current configuration of the 3-D coupled models f and the model positions t, they need to be recomputed when the configuration or position of the 3-D models changes which implies an iterative approach for the reconstruction.

6.2.4. Automatic 3-D CSM initialization

After explaining the silhouetted-based loss and the estimation of the camera parameters in the previous two subsections, this subsection continues with the reconstruction workflow by covering the initial alignment of the 3-D coupled shape models to the image data. This alignment determines the initial scale, teeth positions and relative position of both models to ensure a good starting configuration for the deformation-based reconstruction. For the individual models, these parameters are best deduced from the respective maxilla and mandible view. Because of the top-down view characteristics of these two views, the teeth outlines are only slightly affected by occlusions between neighboring teeth. This enables a matching between the centers of gravity (CoG) of the extracted 2-D contours and 3-D model items of the corresponding CSM to determine a global scale factor and the relative teeth positions. It is performed by adapting the CSM to the CoGs of the contours by minimizing an energy functional $E_{3D-Init}(f,t)$ that aims to find the optimal model configuration f and global transformation t. The energy functional is defined as

$$E_{3D-Init}(f,t) = E_{CoG}(f,t) + \alpha E_{model}(f), \tag{6.6}$$

where the model energy $E_{model}(f)$ is identical to Eq. (4.8), again ensuring the model stays close to the learned configuration space, $E_{CoG}(f,t)$ aims to match the centers of gravity of the model items and their corresponding contour, and α is an empirically determined weighting parameter used to balance both terms. The term $E_{CoG}(f,t)$ is given by the root mean square Euclidean distance between CoGs:

$$E_{CoG}(f,t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\| cog(f,t)^{(i)} - cog(\mathcal{C})^{(i)} \right\|_2^2}, \tag{6.7}$$

with $cog(f,t)^{(i)}$ being the center of gravity of model item i and $cog(\mathcal{C})^{(i)}$ the CoG of the contour corresponding to item i. Since the centers of the contours can only be computed in the two-dimensional

image plane, the aforementioned energy ignores the elevation (z-coordinate) of the model items when calculating the Euclidean distance. This adaptation is performed independently for the maxilla and mandible models. Both model configurations then represent the scale and teeth locations that best match the extracted contours without applying any kind of shape deformations (i.e. tooth shapes are only changed through the updated model configurations).

The next step is to establish the relative location of both models. For simplification, the position of the maxilla models is fixed and only the position of the mandible model is optimized. The contours extracted from the left, right and anterior views are utilized for this step as they show both maxillary and mandibular teeth. Since the teeth outlines suffer from occlusions in all three of these views, the only reliable way to match models and contours is via the silhouette-based loss. Therefore, the camera parameters of the three views are estimated as described in Section 6.2.3, except that only the maxillary teeth are used (since the position of the mandibular teeth is still to be determined). Using the maxilla model as reference here has the added benefit that the corresponding contours suffer less from occlusions (in comparison to the mandibular contours) which allows a more accurate parameter estimation. Having computed the individual camera parameters, the position of the mandible model relative to the maxilla model is optimized (in terms of translation) by minimizing the silhouette-based loss of the mandible model for these three views:

$$t^*_{mand} = \underset{t_x,t_y,t_z}{\operatorname{argmin}} \mathcal{E}_{CSL}(f,t), \tag{6.8}$$

where t_x, t_y, t_z are the translation parameters of the global model transformation t to be optimized, and $\mathcal{E}_{CSL}(f,t)$ is the silhouette-based loss defined in Eq. (6.3). After the initialization step, the two models are aligned and consist of teeth which have been adjusted in terms of their (relative) position, but not in terms of any shape deformations. The remaining final reconstruction step now aims to recover the 3-D shapes based on the extracted numbered contours.

6.2.5. Teeth reconstruction

The deformation-based shape reconstruction utilizes the numbered teeth contours of all five input images and the initialized 3-D coupled shape models of the teeth in order to recover the 3-D shape and relative location of 24 teeth. This iterative process aims to minimize the silhouette-based loss between 2-D projections of the 3-D models and the extracted teeth outlines while employing the statistical information encoded in the coupled shape models to fill in the missing information about the occluded parts. Therefore, the model configurations $\hat{f}$ (describing the position and shape of the individual teeth) and global transformations $\hat{t}$ (describing the relative position of the two arches) that best match the contours minimize the functional

$$E_{reco}(\hat{f},\hat{t}) = E_{CSL}(\hat{f},\hat{t}) + \alpha E_{model}(\hat{f}). \tag{6.9}$$

Here, $E_{CSL}(\hat{f},\hat{t})$ is the silhouette-based loss defined in Section 6.2.2 and $E_{model}(\hat{f})$ ensures the models stay close to their learned configuration space (cf. Eq. (4.8)).

The silhouette loss depends on the the camera parameters θ^v for each view v as these are required to compute the 2-D projections of the current model configuration. Therefore, each optimization step *it* starts with the re-computation of the camera parameters θ^v for all views based on the model configurations of the previous iteration $\hat{f}^{(it-1)}, \hat{t}^{(it-1)}$ as described in Section 6.2.3. Having determined the camera parameters, the projections for all views that will be used for the deformation-based shape reconstruction can be computed. Additionally, the individual 3-D landmark points for each model items which define the 2-D silhouette in each view can be identified. These landmark points will be moved in the deformation step in order to determine the model configurations that minimize the difference between silhouettes and observed contours – these configurations then define the optimal 3-D teeth shapes via encoded statistical shape knowledge. Similar to the local search for feature points in the 2-D CSM adaptation, the landmark points will be moved in normal direction relative to the surface at the given landmark location. The amount of movement is calculated based on the mismatch between silhouette and contour.

For each projected silhouette point, the corresponding point on the extracted contour needs to be identified. This is achieved by matching each silhouette point to its closest contour point based on the L_2 norm, i.e. for each silhouette point s_i the corresponding contour point c_i^* is given by

$$c_i^* = \underset{c_j}{\operatorname{argmin}} ||s_i - c_j||_2. \tag{6.10}$$

This process is repeated for each pair of silhouette and corresponding contour for all five views.

Knowing the correspondences, the 2-D displacement vectors $d_i = \overrightarrow{s_i c_i^*}$ between the projected silhouette points and their corresponding contour points are computed. These displacements are then utilized in order to adjust the position of the 3-D vertices that define the silhouette along their normal direction. Additionally, the displacements are propagated to the neighboring vertices and scaled according to a Gaussian distribution. Again, this process is repeated for all model items and views to determine the optimal displacements that should minimize the discrepancy between 3-D models and extracted contours. Since individual vertices can have multiple displacements (via propagation from neighbors or from different views), the displacements are stored for each vertex and only the largest one is kept as the optimal displacement. While applying the displacements results in deformed models that better explain the input (and therefore minimize $E_{CSL}(\hat{f}, \hat{t})$), the limited amounts of views and heavy occlusions also result in unrealistic teeth shapes. Therefore, the energy functional (cf. Eq. (6.9)) also includes the model energy $E_{model}(\hat{f}$ that constrains the model configurations by the learned configuration space in order to determine the closest 'valid' model configuration. The shape reconstruction process is repeated until convergence or until a predefined number of iterations has been completed.

The reconstruction process will result in model configurations (i.e. teeth shapes and positions) that best fit the input while still being constrained by the learned model configuration space. This could yield a sub-optimal result if the theoretically best model configuration is outside of the learned configuration space. Therefore, the final step is a free-form deformation of the teeth shapes that is no longer restricted by the configuration space.

6.3. Experiments and Results

The proposed fully-automatic reconstruction approach of 24 teeth from five colored photographs (of the orthodontic record) has been tested to assess its performance on real data sets. The deformation-based reconstruction process utilizes the teeth outlines from the five photographs as the reference to minimize the silhouette-based loss and recover the best matching model configurations of the 3-D CSMs which then define the final teeth shapes and locations. The utilized teeth outlines and the required corresponding numbering have been extracted from the individual photographs by the approach covered in Section 6.1. The performance of this prerequisite step was already analyzed in Section 6.1.6. Therefore, this section focuses on the experiments and results of recovering the best matching model configurations that determine the shapes and locations of the reconstructed teeth. It starts by describing the data used for the experiments, the different metrics for comparing the reconstructed teeth to the ground truth scans, and the offline training of the 3-D coupled shape model for the two dental arches. The achieved reconstruction results on the test sets, which utilize the teeth outlines and numbering extracted from the photographs, are then presented at the end of this section. A discussion of the presented results will be done in the next section.

6.3.1. Data

Different data sets are used during the experiments to train the 3-D coupled shape models in the offline training phase, and to evaluate the reconstruction performance of the proposed approach. Both data sets are described in the following paragraphs.

Training data (3-D CSMs) The training data for the 3-D coupled shape model consist of triangular meshes representing the shape of individual teeth. The meshes have been obtained from dental impressions (plaster imprints) which were digitized using a laser scanner and subsequently separated into individual meshes for each tooth as well as the gums. During this process, the teeth meshes were slightly extended in root direction and then closed to generate a valid surface. Each tooth mesh therefore contains an accurate representation of the structure of tooth surface and size (in mm). Each mesh is also labeled with the corresponding tooth number to easily identify and match meshes between different scans, gum meshes were not used for model training. Fig. 6.20 shows an exemplary scan with mandible (a) and maxilla (b) depicted from a top and frontal direction. It also contains the aligned scans showing the relation of both arches both with and without the gum included (c). The complete training set for the 3-D coupled shape models consists of 1000 scans.

Test data The reconstruction performance has been evaluated on the same 39 test data sets that were used for the evaluation of the teeth segmentation and numbering part of this approach (cf. Section 6.1.6.1). The teeth outlines and numbers extracted from the corresponding photographs are used for the teeth reconstruction of the respective data set. The segmentation accuracy for those outlines was discussed in Section 6.1.6.

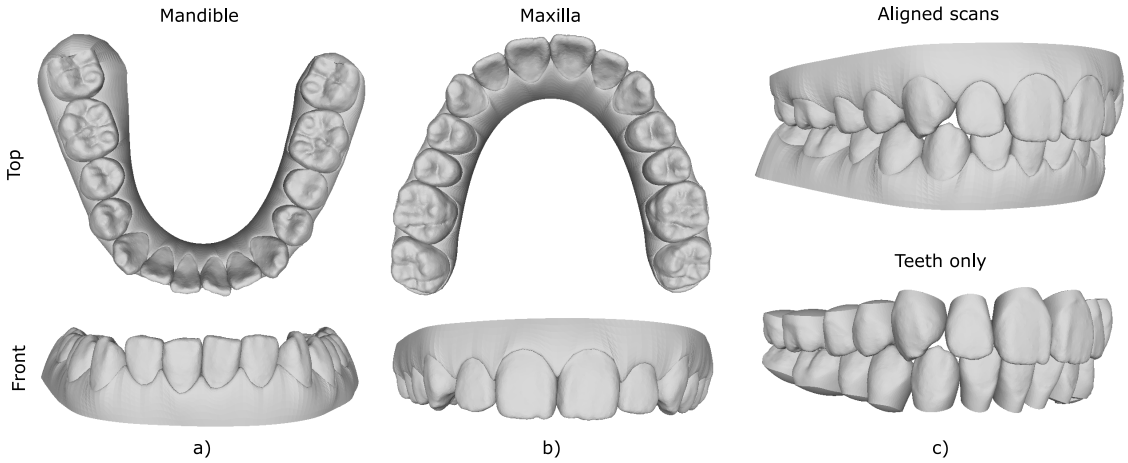


Figure 6.20.: Example of a 3-D plaster scan used for model training. Each training sample consists of individual meshes (c) for each tooth and the gums that were extracted from the original scan (a,b).

Each test data set also includes ground truth meshes of all teeth representing the shape and relative location of the teeth. The meshes were generated in the same way as the training data for the 3-D coupled shape models, i.e. laser scanned plaster imprints were separated into individual meshes and labeled accordingly (cf. Fig. 6.20). The gum meshes were not used for the evaluation process.

6.3.2. Metrics

The performance of the proposed reconstruction approach is evaluated using several different metrics. The most common metrics used to compare the generated results to the ground truth references measure the similarity (or dissimilarity) based on the volumetric overlap or the surface distance. Individual variants of metrics for both categories differ in the mathematical definitions of how overlap and distances are measured [HvGS*09]. Volumetric and surface-based metrics provide information about different aspects of the generated results. Thus, metrics from both categories will be employed here. Since the reconstructed teeth are in a surface-based representation (triangular meshes), they are converted to a voxel-based representation for calculating the volumetric metrics.

Volumetric overlap error (VOE) The volumetric overlap error is used to measure the overlap between a ground truth volume (set of voxels) G and the reconstruction result (set of voxels) R. It is given in percent and defined as

$$VOE(R,G) = 100\left(1 - \frac{|\mathrm{R} \cap \mathrm{G}|}{|\mathrm{R} \cup \mathrm{G}|}\right). \tag{6.11}$$

The ration between the intersection and the union of both volumes ($|\mathrm{R} \cap \mathrm{G}|/|\mathrm{R} \cup \mathrm{G}|$) is also known as the Jaccard index [Jac01]. The VOE is a popular metric to measure the segmentation accuracy. 0 % VOE indicates a perfect match between ground truth and results, while 100 % error means both volumes do not overlap at all.

Dice similarity coefficient (DSC) The dice similarity coefficient [Dic45, Sor48], also known as Sorensen-Dice coefficient, is used to measure the similarity between the ground truth G and the reconstruction result R. It ranges between 0 and 1, with larger values indicating a more accurate result. It is defined as

$$DSC(R,G) = \frac{2\,|\mathrm{R} \cap \mathrm{G}|}{|\mathrm{R}| + |\mathrm{G}|}. \tag{6.12}$$

Compared to the Jaccard index used in the volumetric overlap error, the DSC puts more emphasis on overlapping voxels (i.e. voxels inside the intersection of both volumes).

Relative volume difference (RVD) The relative volume difference is given in percent and used to assess if the generated reconstructions are too big or too small in relation to the ground truth. It is defined as

$$RVD(R,G) = 100 \left(\frac{|\mathrm{R}| - |\mathrm{G}|}{|\mathrm{G}|} \right), \tag{6.13}$$

with G and R the set of voxels of the ground truth volume and the reconstruction result, respectively. Technically, the RVD is not a metric because the computed measure is not symmetric. Because of this reason, the value of the RVD is typically provided as a signed number to indicate if the result is bigger (positive sign) or smaller (negative sign) than the ground truth.

Average symmetric surface distance (ASD) ASD is the average distance (in millimeters) between the reconstructed surface S_R and the ground truth surface S_G. For each surface point (landmark) r of S_R, the distance $dist(r, S_G)$ to the surface of S_G is calculated. This distance is defined as the shortest Euclidean distance ($\|\cdot\|_2$) of an arbitrary surface point $g \in S_G$ to point r:

$$dist(r, S_G) = \min_{g \in S_G} \|r - g\|_2 . \tag{6.14}$$

The same process is repeated to determine the distances $dist(g, S_R)$ of every surface point g of the ground truth surface S_G to the reconstructed surface S_R. The average symmetric surface distance is then defined as the average of all those distances:

$$ASD(S_R, S_G) = \frac{1}{|S_R| + |S_G|} \left(\sum_{r \in S_R} dist(r, S_G) + \sum_{g \in S_G} dist(g, S_R) \right). \tag{6.15}$$

Root mean square symmetric surface distance (RMSD) The root mean square symmetric surface distance is similar to the average symmetric surface distance and is also provided in millimeters. However, the shortest Euclidean distance of an arbitrary (surface) point a to surface S_B is squared and RMSD is given by the root of the average symmetric distances. Consequently, it is defined as

$$RMSD(S_R, S_G) = \sqrt{\frac{1}{|S_R| + |S_G|} \left(\sum_{r \in S_R} dist(r, S_G)^2 + \sum_{g \in S_G} dist(g, S_R)^2 \right)}. \tag{6.16}$$

The RMSD has the advantage compared to the ASD that large deviations from the ground truth contour are punished stronger.

Hausdorff distance (HD) The Hausdorff distance [HKR93] or maximum symmetric surface distance is also related to the previous two metrics. Instead of averaging the minimal distances between both surfaces, the HD is given by the maximum of all those distances:

$$HD(S_R, S_G) = \max \left(\max_{r \in S_R} dist(r, S_G), \max_{g \in S_G} dist(g, S_R) \right). \tag{6.17}$$

As a result, it provides a measure for the true maximum error and is sensitive to outliers.

6.3.3. Offline training: 3-D coupled shape models

The 3-D coupled shape models are designed to capture the statistical variance of the teeth shapes and locations in the respective dental arch and are trained on an offline training phase. The training process is similar to the 2-D CSMs, but uses triangular meshes to model the 3-D statistical shape models of the teeth and a pairwise registration method [Kir13] to establish landmark correspondences between individual training samples. Knowing the correspondences, the samples are aligned using Generalized Procrustes Analysis (GPA) and the extraction of the statistical information from the set of aligned training shapes is again performed using principal component analysis (cf. Section 4.2) while retaining 95 % of the shape variance. The creation of the 3-D coupled model then follows the same methods as the 2-D CSM, but utilizes 3-D affine transformations with seven parameters to describe the relative location of deformable model items. Each deformable model item is represented by a triangular mesh consisting of 1500 vertices, $\mathcal{L} = (x_0, y_0, z_0, \ldots, x_{1499}, y_{1499}, z_{1499})$.

The two coupled shape models have been trained on the 1000 training samples described in Section 6.3.1. Fig. 6.21 illustrates the resulting coupled shape models for maxilla and mandible. For each model, the first mode of variation is depicted at three states while showing the model from two different viewing angles (top and front). The individual model items are colorized according to the different teeth types. The colors correspond to the coloring of the 2-D CSMs for easy visual matching.

Due to the large amount of training data covering many different teeth configurations, the models contain over 200 modes of variation and the observable changes in a single mode are rather small. The

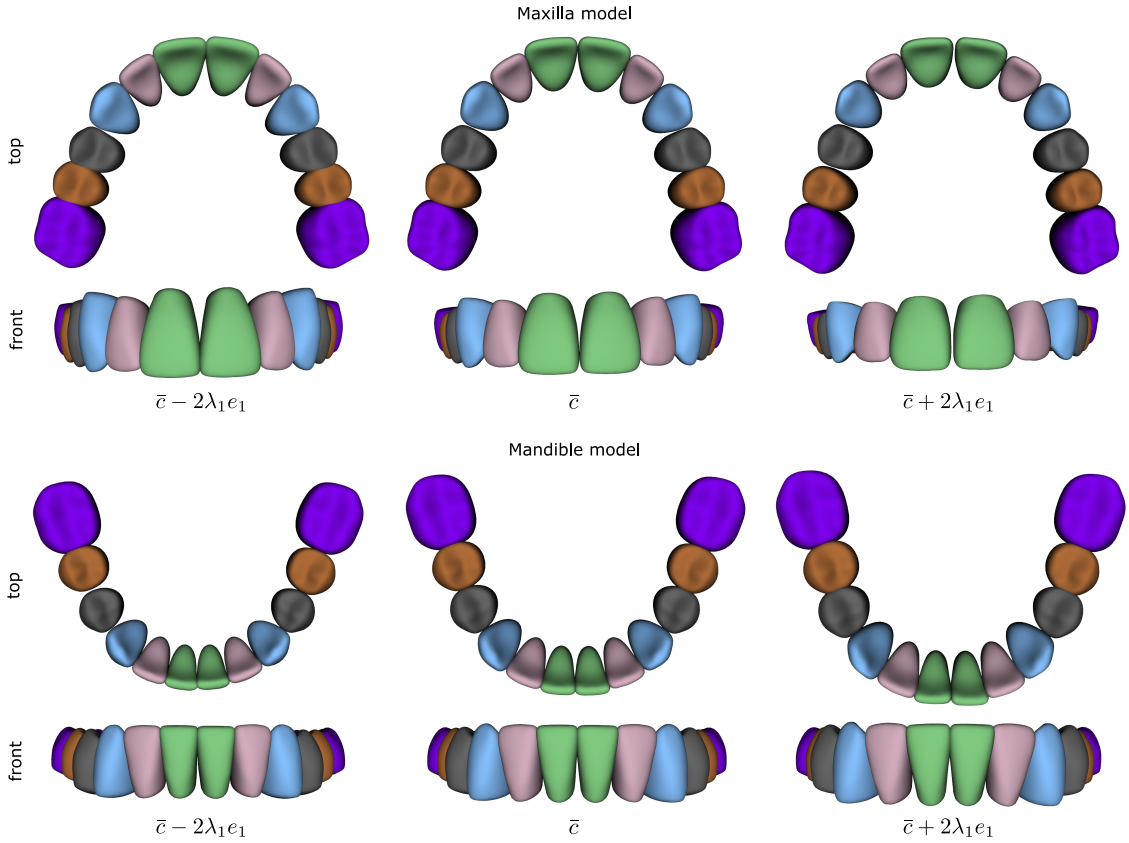


Figure 6.21.: First mode of variation of the 3-D coupled shape models for maxilla and mandible teeth. Both models were trained on 1000 training samples. The colors of the individual model items illustrate the different tooth types and correspond to the colors chosen for the 2-D CSMs (cf. Fig. 6.3).

most prominent changes in the first mode is a variation in the U-shaped arch that affects how much of the posterior teeth is visible in the frontal view (i.e. the second pre-molar teeth (colored in orange) are almost not visible in the right pictures and more revealed in the left ones). Another change is the tilting of the anterior teeth (colored in green) which is observable in the top view.

The two trained models are used in their mean configuration during the reconstruction process as a starting point for the iterative deformation-based approach. The encoded statistical information is used to complement the extracted teeth outline with shape information about the occluded parts and to constrain the reconstruction to valid teeth shapes and relative locations. The results of this approach are presented in the next subsection.

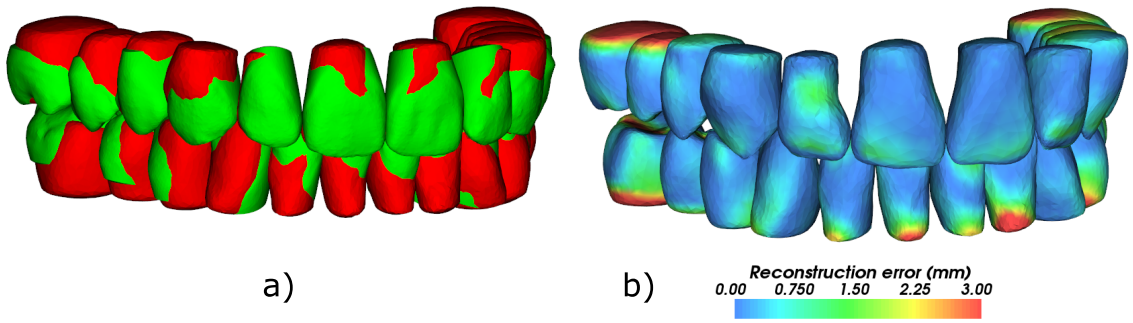


Figure 6.22.: Exemplary reconstruction result. a) reconstructed teeth shapes and location in red, ground truth in green; b) surface distance between reconstruction and ground truth color-coded on the reconstruction result.

6.3.4. Evaluation

A quantitative evaluation has been performed to assess the accuracy of the proposed fully-automatic reconstruction approach. The extracted numbered teeth outlines of the 39 test sets from the first part of the reconstruction workflow (cf. Section 6.1) are used to perform a deformation-based reconstruction of the 3-D teeth shapes and relative locations. This consists of the automatic model initialization for the two 3-D CSMs and the iterative optimization to determine the best matching model configurations which provide the final shapes and (relative) locations. Since the initialization process only uses the teeth outlines from the previous step and no other references are computed, no separate evaluation is performed (as was the case for the teeth segmentation and numbering). Instead, the final reconstructed teeth meshes of the 39 test sets have been compared to the corresponding ground truth meshes in order to determine the accuracy of the reconstruction.

Since each reconstruction has an arbitrary scale and position, each combined set of reconstructed meshes (i.e. model items of maxilla and mandible model are considered as one set) has been registered to the corresponding set of ground truth meshes by an individual global similarity transformation in order to retain teeth shapes and relative locations and sizes. These global transformations were obtained via an iterative closest point (ICP) method [BM92] based on the centers of gravity (CoGs) of the reconstructed and ground truth tooth meshes. The performance is then assessed by comparing the individual reconstructed teeth meshes to their corresponding ground truth using the metrics described in Section 6.3.2. The overall measures are computed by averaging the tooth-specific results of all test cases. Standard deviation from the mean is provided as an additional measure to quantify the variation in the results. The evaluation results are summarized in Table 6.5 which lists the overall performance of the approach in terms of volumetric overlap error (VOE), dice coefficient (DC), relative volume difference (RVD), average symmetric surface distance (ASD), root mean squared surface distance (RMS), and Hausdorff distance (HD). Minimum and maximum values entries for each metric show the highest and lowest value that was achieved for a single tooth, but are susceptible to outliers. It is important to

Table 6.5.: Quantitative evaluation results on the test set for different metrics: average symmetric surface distance (ASD), volumetric overlap error (VOE), dice coefficient (DSC), Hausdorff distance (HD), relative volume difference (RVD), and root mean squared surface distance (RMSD), .

Metric	Minimum	Maximum	Mean $\pm$ Std. dev.
ASD (mm)	0.246	3.582	0.807 ± 0.379
VOE (%)	20.854	98.991	47.209 ± 13.886
DSC	0.020	0.884	0.679 ± 0.133
HD (mm)	1.030	8.264	2.475 ± 0.888
RVD (%)	-46.826	47.638	-1.979 ± 14.826
RMSD (mm)	0.321	4.211	0.982 ± 0.443

note that all metrics except dice coefficient (DSC) measure the reconstruction error, due to this lower numbers are better. For the DC, a higher value is better as it measures the overlap between both volumes. The presented approach achieves an average symmetric surface distance of 0.807 ± 0.379 mm, volumetric overlap error of 47.209 ± 13.886 % and dice coefficient of 0.679 ± 0.133.

Table 6.6 contains the average results for each individual tooth and includes volumetric overlap error (VOE), dice coefficient (DSC) and average symmetric surface distance (ASD). Performance varies between individual teeth with posterior teeth (i.e. the 4th, 5th, and 6th tooth of every quadrant) generally being reconstructed a little better in terms of volumetric overlap error than anterior teeth (1st, 2nd, and 3rd tooth). The highest overlap error can be observed for the four mandibular incisor teeth. Dice coefficient indicates similar results. Average symmetric surface distance is more inconsistent when comparing individual result.

An example of a reconstruction is depicted in Fig. 6.22. The left image shows the reconstruction result in red and the ground truth in green, on the right, the reconstruction results are colored to illustrate the surface distance to the ground truth at each vertex.

Table 6.6.: Quantitative evaluation results on the test set for each individual tooth for selected metrics: volumetric overlap error (VOE), dice coefficient (DSC), and average symmetric surface distance (ASD). The teeth are numbered according to the FDI notation system.

Tooth		VOE (%)	DC	ASD (mm)
FDI index	Type	(avg. ± std-dev.)	(avg. ± std-dev.)	(avg. ± std-dev.)
16	molar	40.569 ± 14.601	0.734 ± 0.136	0.865 ± 0.495
15	pre-molar	47.661 ± 14.156	0.674 ± 0.145	0.846 ± 0.415
14	pre-molar	43.999 ± 13.325	0.708 ± 0.126	0.745 ± 0.328
13	canine	42.375 ± 14.197	0.721 ± 0.121	0.739 ± 0.343
12	incisor	51.167 ± 11.933	0.647 ± 0.116	0.751 ± 0.282
11	incisor	49.210 ± 11.011	0.666 ± 0.107	0.834 ± 0.315
21	incisor	48.162 ± 12.470	0.673 ± 0.119	0.824 ± 0.343
22	incisor	51.845 ± 14.846	0.635 ± 0.154	0.793 ± 0.433
23	canine	41.433 ± 13.686	0.727 ± 0.138	0.743 ± 0.497
24	pre-molar	44.202 ± 15.559	0.701 ± 0.160	0.801 ± 0.564
25	pre-molar	44.208 ± 14.865	0.703 ± 0.149	0.806 ± 0.572
26	molar	40.577 ± 13.210	0.736 ± 0.114	0.857 ± 0.392
36	molar	45.573 ± 11.558	0.698 ± 0.101	0.994 ± 0.356
35	pre-molar	43.095 ± 12.336	0.717 ± 0.108	0.737 ± 0.306
34	pre-molar	40.792 ± 17.027	0.727 ± 0.165	0.708 ± 0.492
33	canine	49.902 ± 13.475	0.656 ± 0.132	0.843 ± 0.339
32	incisor	56.896 ± 9.050	0.597 ± 0.092	0.819 ± 0.204
31	incisor	57.069 ± 12.420	0.589 ± 0.140	0.827 ± 0.376
41	incisor	58.309 ± 11.284	0.579 ± 0.130	0.849 ± 0.345
42	incisor	57.731 ± 11.840	0.584 ± 0.129	0.849 ± 0.282
43	canine	46.598 ± 11.427	0.689 ± 0.099	0.755 ± 0.256
44	pre-molar	39.960 ± 11.392	0.743 ± 0.102	0.648 ± 0.273
45	pre-molar	43.821 ± 9.623	0.715 ± 0.081	0.731 ± 0.218
46	molar	47.736 ± 8.936	0.682 ± 0.077	1.031 ± 0.257

6.4. Discussion

The presented fully-automatic teeth reconstruction from only five colored photographs utilizes the numbered teeth contours extracted from the photographs to reconstruct the 3-D shape and relative location of the teeth by minimizing a silhouette-based loss while constraining the shapes using coupled shape models. The performance of the approach has been evaluated on 39 test data sets that were distinct from the data used for training.

The teeth segmentation and numbering presented in Section 6.1 achieves varying performance on the different photographs ranging from an average dice overlap of 0.931 for the maxilla view to 0.631 for the left view. Because the reconstruction only relies on the teeth outlines as a reference for the reconstruction, an accurate object level segmentation is required for good reconstruction performance. Because of this, the sub-optimal results for left and right views will have a negative impact on the reconstruction accuracy that can be achieved by this approach. Nonetheless, the teeth segmentation and numbering provides an automatic way to extract both teeth outlines and corresponding numbers that enable a fully-automatic reconstruction pipeline – this approach is the first to provide such an automatic pipeline for image-based reconstruction of multiple teeth. The model-based reconstruction process uses these contours to deform two 3-D coupled shape models (one for upper and one for lower dental arch) by maximizing the overlap between 2-D projections of the model items and extracted contours, while constraining the deformations to valid teeth shapes.

On the 39 test data sets, the approach achieves promising results despite the limited amount of information that is available for recovering the 3-D information. The evaluation reveals an average symmetric surface distance of 0.807 ± 0.379 mm and dice coefficient of 0.679 ± 0.133.

While minimizing the silhouette-based loss leads to good matches between contours and projections in maxilla, mandible and anterior views, the projection generated for the lateral views (left, right) do not overlap as well with their corresponding contours. This might be related to the sub-optimal segmentation performance for these views, but can also be attributed to the larger occlusions occurring in these views.

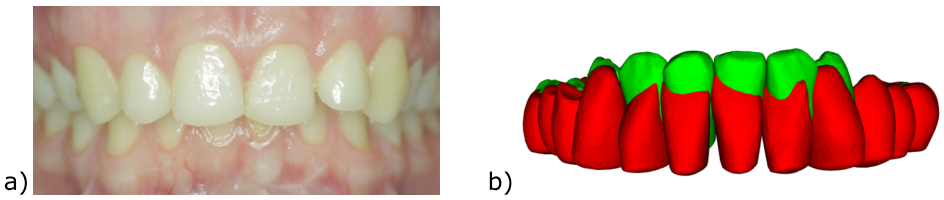


Figure 6.23.: Reconstruction result in case of overbite: a) anterior view photo with occluded lower frontal teeth; b) comparison of the resulting reconstruction of the lower dental arch (red) and the corresponding ground truth (green). The missing information about the lower teeth due to the occlusion leads to incorrect reconstructions.

Analyzing the tooth specific results in Table 6.6 shows that the majority of teeth reaches an average dice coefficient of about 0.70. However, it is evident that the reconstructions of the anterior teeth in the lower jaw perform worse than the other teeth with a DSC of slightly below 0.6. This phenomenon can be explained by the fact that the mandibular incisors are often occluded by the maxillary teeth. Because the images are captured in a 'closed-mouth' position, the maxillary teeth will naturally overlap with the mandibular ones to a small extend. This affects the estimation of the teeth height as the incisal edge (top end) of the anterior mandibular teeth is not visible. Largely or completely occluded lower teeth are considered an overbite condition and will drastically reduce the amount of information that can be extracted from the image. This affects both the segmentation accuracy of the teeth contours in the photographs as well as the reconstruction accuracy. In extreme cases, only the mandible view remains to provide silhouette information about the anterior mandibular teeth. Such a case is illustrated in Fig. 6.23, where image a) depicts an overbite case and image b) contains the reconstruction result of the lower dental arch. The missing information for the anterior and left or right views leads to a mismatch between reconstructed teeth and ground truth.

Another aspect that impacts the reconstruction accuracy of the anterior teeth is the missing information about their rotation and tilt. This type of information could be deduced from the maxilla and mandible views, and (to some extent) the left and right views as well. However, simply relying on the contour to reconstruct this information is insufficient when only using five images as is evident from the results. More robust methods to extract, e.g., the incisal edge of the incisors are needed to tackle this problem.

6.5. Conclusion

In this chapter, a fully-automatic image-based approach for the 3-D reconstruction of 24 teeth from only five photographs was presented. It applies the segmentation and numbering concept introduced in Chapter 4 to extract the teeth outlines from the colored photographs and label them accordingly. The outlines are the only usable image feature that can be employed for reconstruction because of the difficult appearance properties of the teeth which results in almost no detectable surface features. The numbering is required to group teeth outlines extracted from different views and enable a fully-automatic reconstruction. The contour extraction is limited to 24 teeth since the other 8 teeth (second and third molars) are almost never contained in any of the photographs. The segmentation and numbering utilizes view-specific U-Net networks to perform a semantic teeth segmentation and 2-D coupled shape models (CSM) to encode the statistical information about the expected shape and relative location of the outlines. The semantic segmentation masks are used to initially place the CSMs onto the images and dynamic adaptation strategies are employed to fit the models to the teeth contours in the image. While anterior, maxilla, and mandible views achieve satisfactory performance results on 39 test sets, left and rights views suffer from large variations in camera perspective which negatively affects the accuracy of the segmentation results. Stricter guidelines for capturing the photographs are required to provide more consistent teeth contours extraction. Automatic model initialization works correctly in 95 % of cases but could be made more robust as a failed initialization results in very poor segmentation

performance. Nonetheless, the extracted contours and numbering enable a 3-D reconstruction of the teeth.

The image-based reconstruction uses a model-based approach since five photos only provide limited information and because the teeth suffer from heavy occlusions which make reconstruction more difficult. 3-D coupled shape models for maxilla and mandible teeth encode the statistical knowledge about their shape and spatial relation and support the reconstruction by providing information about the occluded parts. By minimizing a silhouette-based loss between the extracted contours and 2-D projections of the CSMs, the original 3-D shape and configuration of the teeth is recovered. The numbered contours extracted from the 39 test data sets were used to evaluate the performance of the proposed approach. Overall, the achieved average symmetric surface distance was 0.807 ± 0.379 mm with an volumetric overlap error of 47.209 ± 13.886 % an a dice coefficient of 0.679 ± 0.133. Given the limited amount of information, these results are promising. Reconstructions of the mandibular anterior teeth suffer from the fact that these are often (partly) occluded by the maxillary teeth which further reduces the available information for an accurate reconstruction. The tilt and rotation of the anterior teeth is also difficult to estimate based on so few images. Nevertheless, the presented approach offers the first fully-automatic image-based 3-D teeth reconstruction of the . More importantly, it is also the first approach to utilize the five photographs taken as part of the orthodontic record and performs the image-based reconstruction solely based on these images. This increases the difficulty because the five photographs only provide limited information about the teeth.

7. Cephalometric landmark localization

This chapter presents the third application of the model-based segmentation and numbering concept. In contrast to the previous chapters, it does not focus on extracting object-level segmentations and numbering of relevant structures but instead uses the concept to predict the position of several landmarks by exploiting their spatial relation to each other and to other structures in the image. These predictions are then refined in a second step to determine the exact locations of those landmarks. The landmarks in question are cephalometric landmarks which refer to predefined anatomical features in cephalometric radiographs and are utilized in the analysis of those images. This *cephalometric analysis* is an important tool in orthodontics used for clinical diagnosis and the planing of treatment and surgeries. It aims to extract fundamental information about the soft tissue-, dental- and bony structures of a patient from lateral cephalometric radiographs. For this purpose, several diagnostic measurements like angles and distances are computed based on the predefined cephalometric landmarks. For each of these measurements, a clinical acceptable range has been defined and measures which fall outside of this range indicate a potential medical condition. The type and number of measurements - and thus the set of utilized landmarks - varies between different methods of analysis like Steiner analysis [Ste53], Downs analysis [Dow56] or Ricketts analysis [RRC*82]. Fig. 7.1 shows a cephalometric radiograph and the defined landmarks used for the Steiner analysis method.

In clinical practice, the identification of these landmarks is done manually or semi-automatically via (digital) manual tracing [LWH*16]. This is a time consuming and subjective process because the lateral projection of the cranio-facial area (i.e. the parts of the head enclosing the brain and the face [Hux76]) onto a 2-D image leads to superimposed anatomical structures and because the morphology of the facial area varies highly in between patients [KSP*20]. Additionally, cephalometric radiographs suffer from bad image quality like intensity variation, noise and low contrast. This can result in a high inter- and intra-observer variability on the traced landmark locations and, as a result, the design of the patient's treatment plan heavily relies on the dentist's experience and visual perception [WHL*16]. An automated cephalometric landmark localization could mitigate these problems and provide a substantial benefit to the dentist but requires a robust and accurate localization of the individual landmarks in cephalometric radiographs.

In this context, the next section will summarize the challenges for this task. Section 7.2 covers the related work conducted by other authors and summarizes the conclusions of the literature study. The method proposed for the landmark localization task is introduced in Section 7.3. It includes the utilization of the concept for predicting landmark locations and landmark-specific Hough Forests to refine the prediction. Experiments and results are presented in Section 7.4. The chapter concludes with a discussion of the results (Section 7.5) and a final conclusion (Section 7.6).

This chapter is partially based on the publication "Automated Cephalometric Landmark Localization using a Coupled Shape Model" [WLW20].

7.1. Challenges

An automated landmark detection in cephalometric images is difficult to achieve with high precision and success rate due to the following challenges:

- **Image quality:** The image quality of cephalometric radiographs suffers from the same problems as the panoramic radiographs that are used for teeth segmentation (cf. Section 1.2.1). The noise present in these images as well as the low contrast between different regions are challenging for automatic image processing techniques. Additionally, cephalometric radiographs suffer from superimposition of several structures as these images are captured from a lateral position.
- **Ill-defined landmarks:** The local intensity appearance of a landmark can pose a challenge for manual or automatic localization. Some landmark locations do not coincide with any clearly identifiable image features like structure, texture or intensity gradients while others share similarity with different landmarks. Moreover, some landmarks suffer from high variance in their local appearance which is either due to variations in the surrounding anatomical structure or related to the previous challenge of superimposition due to the lateral projection. This is also reflected in a high inter- and intra-observer variability on manually traced landmark locations by clinical experts [WHL*16]. Consequently, the exact location of certain landmarks is difficult to determine (automatically) in the 2-D projection image.
- **Clinical accuracy range:** For use in clinical practice, all detected landmark positions need to be within an accuracy range of 2.0 mm around the actual landmark location [LWH*16]. Depending on the physical resolution of the image, this requires a detection of all landmark locations within a small number of pixels around the true locations. This alone is difficult to achieve and even more demanding in combination with the previously listed challenges.

As a result, most approaches in the literature rely not only on the local appearance of a landmark around the their true location. Instead, they incorporate the spatial relation between landmarks to predict or constrain the possible locations and support the detection process. Even some of the presented deep learning-based methods employ this idea to more accurately identify the landmark locations. The related work in the area of cephalometric landmark localization is reviewed in the next section and includes a summary of the different techniques.

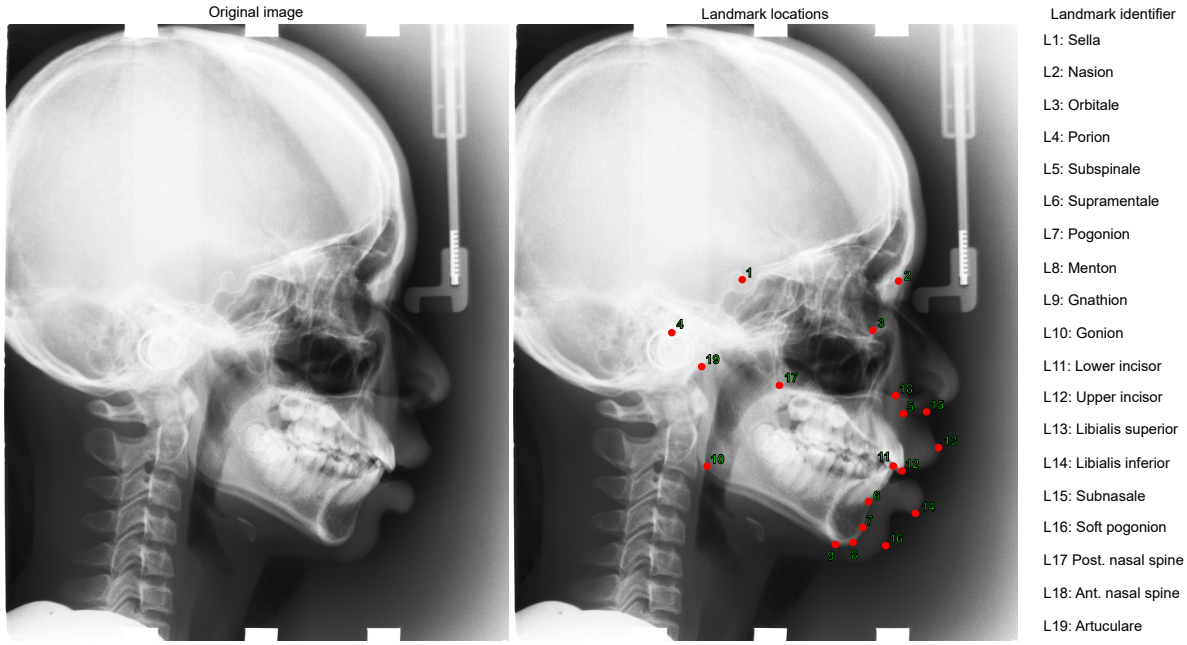


Figure 7.1.: Cephalometric landmarks used for the Steiner analysis method. The left image shows the original image and the right image depicts the locations of the individual landmarks. The identifier for each landmark which refers to the anatomical structure is also provided.

7.2. Related work

Several methods for the automatic detection of cephalometric landmarks have been proposed in the past. In 2015, Wang et al. [WHL*16] organized a Grand Challenge at the International Symposium on Biomedical Imaging (ISBI) on the topic of cephalometric landmark localization. They provided a publicly available data set which consists of 400 images, 150 for training and 250 for testing, and corresponding gold standard annotations for 19 landmarks used in the Steiner analysis method [Ste53] from two clinical experts. Performance was measured in terms of success detection rate (SDR) which indicates the percentage of correctly detected landmarks for a given accuracy threshold, i.e. predicted locations are considered correct if the distance to the gold standard is below the threshold. As one of two teams who reached the final round of this challenge, Ibragimov et al. [ILPV15, ILPV14] modeled the appearance of the landmarks using a Random Forest-based classifier. Haar-like features extracted at different image scales were utilized to capture both the global and local appearance. To better constrain the individual locations within the image, the spatial relations to neighboring landmarks were encoded in a shape model and both shape and appearance were used to determine the optimal landmark locations. The approach achieved a SDR in the 2.0 mm range of 68.13 %. The other team participating in the final round, Lindner and Cootes [LC15], applied Random Forest regression-voting in a Constrained Local Model framework (RFRV-CLM) to automatically detect the 19 landmarks. For each

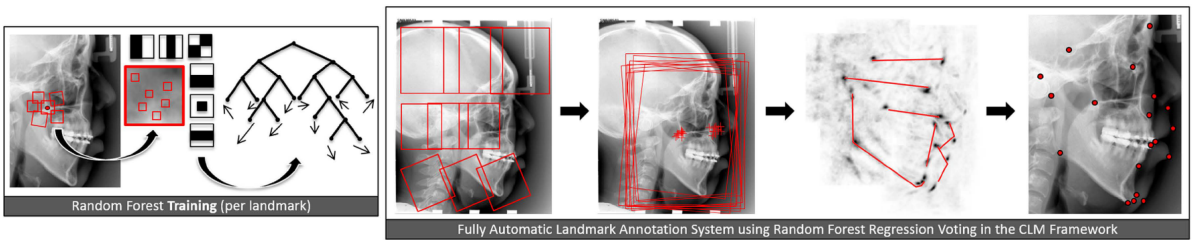


Figure 7.2.: The Random Forst Regression-Voting within a Constrained Local Model (RFRV-CLM) proposed by Lindner et al. [LWH*16]. Landmark-specific random forests encode the local appearance and vote for landmark locations which are constrained by a shape model to determine the final locations. (images from Lindner et al. [LWH*16])

landmark, a separate Random Forest was trained which predicted the position of the landmark relative to a sub-region of the original image. A statistical shape model over all landmark positions was used to exploit the spatial relations and restrict the space of possible landmark locations. Fig. 7.2 depicts the Random Forest training (left image) and the regression voting in the CLM framework (right image). Their approach resulted in a SDR of 70.65 %.

Lindner et al. [LWH*16] subsequently optimized their approach by performing extensive experimental evaluation and analyzed the impact of several parameters of the RFRV-CLM framework on the achieved results. A larger patch size of 30x30 pixels was eventually used for the Random Forest training – larger patches allowed for more of the surrounding appearance to be encoded. The sampling range around the original landmark locations for extracting training samples was increased and finally, the search radius for sampling candidate locations when searching for landmark locations in an unknown image was also increased. With these changes, they could improve the SDR in the 2.0 mm range to 84.70 %.

In 2017, Arik et al. [AIX17] employed a deep learning-based method for the detection of the landmarks. They trained individual convolution neural networks for each landmark using patches of 81x81 pixels to capture the intensity distribution around the original landmark location in the training images. The networks were trained to output a probabilistic estimation for the landmark being located at the center pixel of the input patch. These networks were then used to compute the pixel-wise probabilities for all landmarks in an unseen image. Again, the results could be improved by refining the landmark locations using a probabilistic shape-based model which constrained the predictions using the spatial relations between landmarks. The accuracy was 72.29 % on the ISBI 2015 test set.

Another CNN-based framework for anatomical landmark detection was proposed by Payer et al. [PŠBU19]. Their Spatial Configuration-Net (SCN) utilized a two-stage approach for the localization of landmarks: ambiguous but accurate candidate prediction and reducing ambiguities by incorporating spatial relations. The first part had a network structure similar to fully convolutional networks and used the local appearance of the landmarks to generate potentially ambiguous heatmaps (for each individual landmark). The second part received the (ambiguous) heatmaps as input and implicitly incorporated a

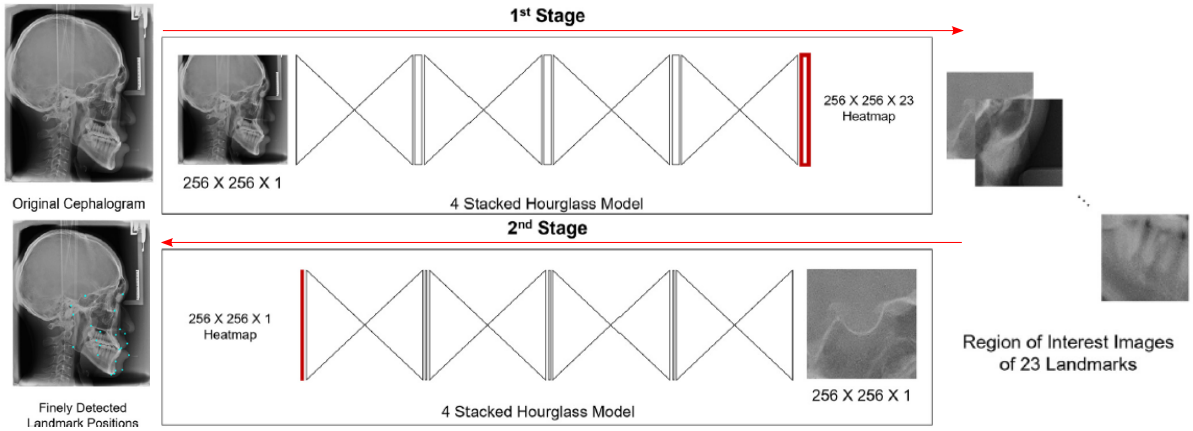


Figure 7.3.: Two-stage method for cephalometric landmark detection proposed by Kim et al [KSP*20]. A Stacked Hourglass (SHG) network is used for detecting initial landmark locations which are then refined using a second SGH. (image modified from [KSP*20])

geometric model of the spatial configuration of landmarks to predict the final location of each landmark. The success detection rate of the SCN on the ISBI 2015 data set was 73.33 %.

Kim et al. [KSP*20] used a stacked hourglass (SHG) network to localize 23 landmarks in a cephalometric radiograph. The SHG architecture was first introduced by Newell at al. [NYD16] and repeatedly applies ('stacks') so called hourglass modules to generate heatmaps of probable landmark locations. A single hourglass module is similar to a fully convolutional network and related to encoder-decoder architectures in the sense that it aims to capture information at every scale. Convolutions and max-pooling layers are used to transfer the information to lower resolutions while nearest neighbor interpolation and skip connections are utilized to combine features across scales. The stacked hourglass design allows supervision via the intermediate heatmaps generated after every module which improves the training performance. Kim et al. [KSP*20] utilized the network architecture in a two-stage process. First, a network of four stacked hourglass modules is used to predict the landmarks locations on a cropped and resized (256x256 pixels) version of the original image. Then, a second network of the same architecture is utilized to refine the 23 locations by using regions of interest (256x256 pixels) around the predicted locations taken from the original cephalometric image. The two-stage approach is illustrated in Fig. 7.3. The approach was trained on 2075 cephalograms and achieved an accuracy of 1.16 ± 0.06 mm for the 19 landmarks on the ISBI 2015 data set with an SDR of 84.53 %.

Hwang et al. [HMK*21] applied the YOLO v3 network architecture [RF18] for predicting 19 landmark locations. YOLO (short for 'You Only Look Once') is a fully connected network which is designed for real-time object detection. It consists of 106 fully convolutional layers and performs object detection at three different scales to cover small, medium and large objects. The authors used 1983 cephalometric images which were manually annotated by a clinical expert to train the network. The authors do not state any modification of the original YOLO v3 network structure or parameters. The

trained network achieved an average SDR of 75.45 % on the ISBI 2015 test set in the 2.0 mm accuracy range.

In summary, many different approaches for automatic landmark localization in cephalometric radiographs have been proposed. The majority try to encode the appearance of each landmark in a local region around the original location. This learned appearance is then used to identify potential landmark locations in previously unseen images. However, these local appearances alone do not allow for a robust localization of all landmarks. The varying image quality (low contrast, noise) as well as variations in the appearance due to the superimposition of anatomical structures caused by the lateral projection make the detection more difficult. Moreover, some landmarks are too similar in their appearance to reliably prevent incorrect detection. As a result, many approaches also utilize the spatial relations between landmarks. They either eliminate false positive detections and constrain the locations after an initial detection, or predict probable candidate locations and refine those in a second step. Either way, incorporating spatial relations between landmarks helps to improve the robustness and overall accuracy of the predicted landmark locations.

7.3. Methods

In order to handle the challenges of an automatic localization of cephalometric landmarks, previous approaches benefited from the inclusion of the spatial relation between landmarks to increase the accuracy of the detected locations. Since the landmarks represent (anatomical) features in the cranio-facial area which are tied to the morphology of the skull, the inclusion of references to the shape of the skull and head in the radiographic image could offer additional information. The outline of these structures is much easier to detect in the cephalometric images (compared to the actual landmark locations) and could be used to more accurately predict the landmark positions. For this reason, the presented approach for cephalometric landmark localization utilizes the segmentation and numbering concept to not only encode the spatial relations between landmarks but also include their relative position to the skull and head (skin) outlines in the coupled shape model. The coupled shape model (CSM) is extended to enable the inclusion of single point items that are not modeled using a statistical shape model but simply contain a point coordinate.

The workflow of the proposed approach can be roughly divided into two steps. First, the landmark locations are predicted based on the shape and location of skull and head outlines in the image. The search space for each landmark can therefore be narrowed down to a small region of interest around the predicted position. Then, the optimal location of each landmark is refined based on the local intensity appearance by a landmark-specific Hough Forest. Fig. 7.4 illustrates the workflow steps and the required offline training step. A U-Net network for segmenting the head area in the cephalometric images (A), a 2-D CSM which encodes the statistical shape variation of skull and head outlines as well as knowledge about the spatial relation between those structures and the landmark location (B), and the landmark-specific Hough Forests that contain the local appearance information of each landmark (C) are trained in the offline training phase. The landmark localization in an unseen image is then performed

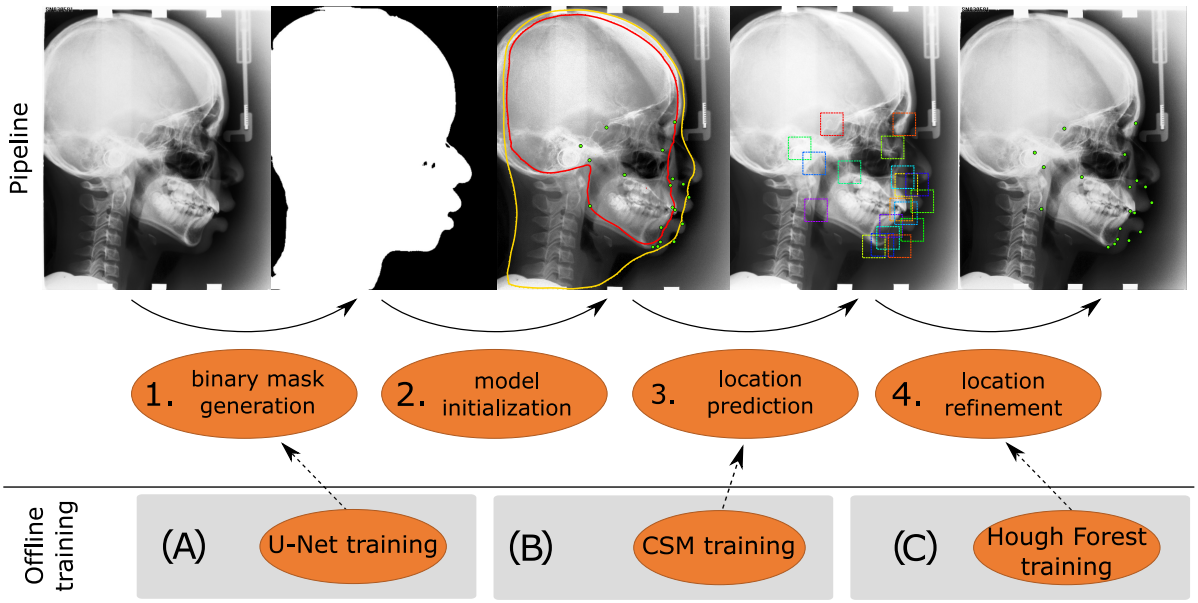


Figure 7.4.: Workflow of the automatic landmark localization in cephalometric radiographs. The *segmentation and numbering* concept is used to predict the landmark locations (1-3), which are then refined using landmark-specific Hough Forests (4). The training of the required U-Net, CSM and random forests is performed in an offline training phase.

by generating the binary mask of the head area (1), placing the CSM according to the binary mask (2), adapting the CSM to the skull and head contours to predict the landmark locations (3) and finally, utilizing the landmark-specific Hough Forests to detect the landmark positions within the predicted regions of interest (4).

The following subsections cover the individual workflow steps in more detail, starting with the design of the coupled shape model. The offline training step (which requires a set of training data and corresponding annotations) is explained in Section 7.4.

7.3.1. CSM design

The coupled shape model is used in this approach to predict the locations of the cephalometric landmarks in relation to the shape of the skull and head of the patient. In order to include the landmarks (which are represented by simple 2-D coordinates) into the CSM, the model was extended to support 'single point items' that do not include a statistical shape model like the deformable model items. These items therefore do not have any shape coefficients v and a single point item i is simply parameterized by a vector p_i that contains their transformation parameters t_i (cf. Section 4.2.2). These transformation parameters t_i describing their relative location to the center of the CSM have the scale and rotation parameter set to 1 and 0, respectively, as these do not apply to these items. The training process of the

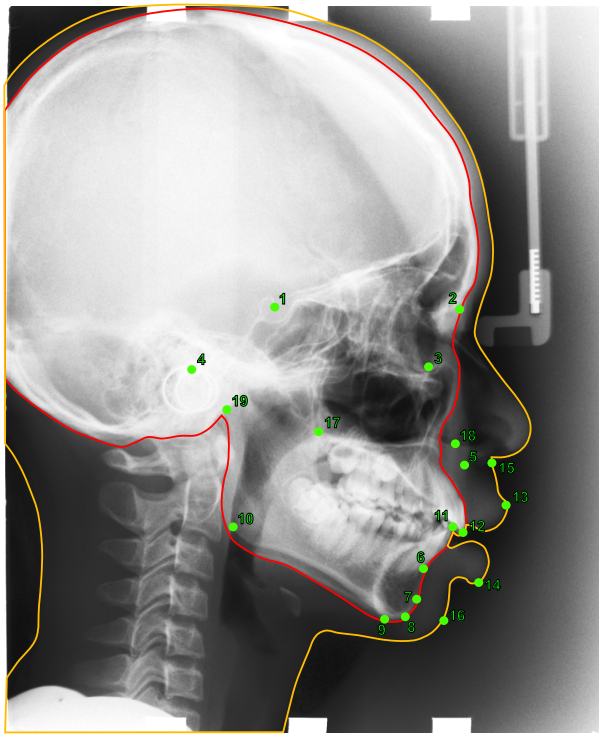


Figure 7.5.: CSM design for cephalometric landmark prediction. Skull (red) and head (yellow) contours are modeled as statistical shape models (deformable model items) while the individual landmarks are represented by their relative location (single point items).

coupled shape model remains unchanged. The resulting CSM then consists of deformable model items for the skull and head outline and a single point item for each landmark. Fig. 7.5 illustrates the CSM with the skull contour depicted in red, the head contour in yellow and the different landmarks in green.

7.3.2. CSM initialization

The CSM needs to be optimally placed onto the cephalometric image to enable a successful adaptation of the deformable model items (skull and head outlines) to the corresponding image structures. The initialization process is again based on a binary mask computed by a U-Net neural network. Opposite to the previous concept applications, where the binary mask contained several model items of the respective CSM, the initialization process is simplified here by the fact that the U-Net can be trained to segment the head area in the cephalometric image which corresponds exactly to the deformable model item representing the head outline. Because of this, the optimal initial CSM placement can be determined by matching the model item to the binary mask and no additional processing is required to identify more references (like in the previous applications). The mean configuration of the CSM

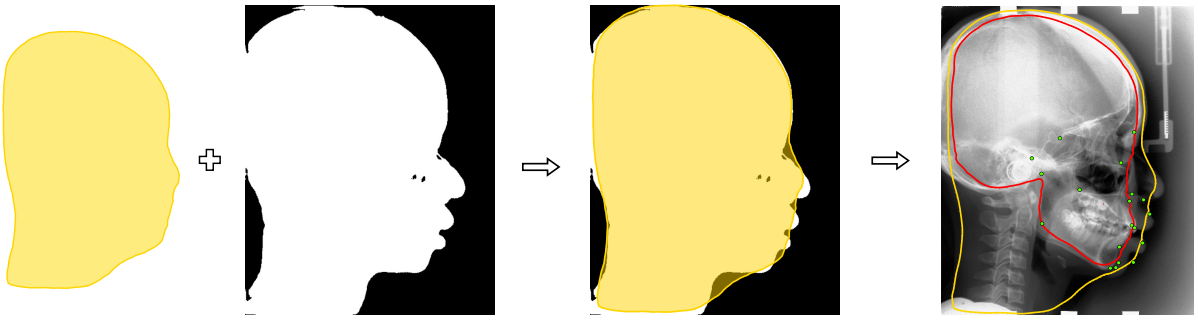


Figure 7.6.: Model initialization process. Since the binary mask corresponds to a single model item of the coupled shape model, the initialization simply maximizes the overlap between the mean shape of this item and the binary mask to determine the required parameters (translation and scale).

is utilized to maximize the overlap between the area enclosed by the mean shape of the head model item and the computed binary mask. The maximum overlap is determined by minimizing the pixel-wise negative IoU between mean shape and mask while optimizing the translation and scale of the mean CSM. The optimal parameters resulting from this can then be used to for placing and scaling the coupled shape model onto the input image. Fig. 7.6 illustrates the initialization process on an exemplary binary mask and shows the resulting initial model position.

7.3.3. Location prediction

The prediction of the landmark locations is performed by adapting the initialized CSM to the skull and head outlines in the image. The resulting model configuration that minimizes the energy functional E (cf. Section 4.5) then provides the landmark positions that best match the input image. During the adaptation process, only the structures with statistical shape information are fitted to the image data using gradient image features (bottom-up information) which applies to the model items of the skull and the head. The single point items that represent the landmark locations are only changed passively by the encoded statistical knowledge about their spatial relation to the other structures (top-down information), but not by any image features. The predicted landmark locations are then refined in the next step to determine the final landmark positions.

The dynamic adaptation strategies introduced in the concept chapter are not utilized for this application. The reason for this is that the CSM only includes 2 model items which are actively adapted to image features and both items are close to their corresponding structure in the image after initialization. Adapting the items at different stages throughout the adaptation process is therefore not required.

7.3.4. Location refinement

The landmark locations given by the final configuration of the adapted coupled shape model are refined using landmark-specific Hough Forests [GYR*11]. A Hough Forest is an application of the Random Forest (RF) framework [Bre01] and consists of multiple decision trees. Here, each set of leaf nodes from an individual tree can be regarded as a discriminative codebook of the landmark's appearance [GL13]. Essentially, a random tree in the Hough Forest learns a mapping between the local appearance of an image patch (implicit shape model) and its probabilistic vote about the object's class and location. By combining the independent predictions from the included random trees, the forest is able to form a more accurate and robust estimate about the object's position within the provided region of interest.

The Hough Forests utilized for location refinement are constructed according to the approach by Gall and Lempitsky [GL13]. Sets of training samples (image patches) $P = \{\mathcal{P}_i\}$ are extracted from the training data, where each sample $\mathcal{P}_i = (\mathcal{I}_i, c_i, d_i)$ consists of the local appearance $\mathcal{I}_i$, the class label c_i and the offset d_i. The class label indicates whether the sample originates from a positive ($c_i = 1$) or negative ($c_i = 0$) example, where a positive example is a patch containing the actual landmark location while a negative (or counter-) example does not contain the actual location. The offset d_i describes the 2-D offset vector between the center of the training sample and the landmark location – in case of a negative example the offset is undefined. The appearance of a sample $\mathcal{I}_i$ is defined by a vector of extracted image features. For the cephalometric landmarks, a combination of raw intensity values, gradient orientations (histogram of oriented gradients) and gradient magnitudes is used. The individual trees of the forest are then trained independently on a random subset of the training samples following the standard framework [Bre01]. Starting from the root, nodes are inserted into the tree and an optimal binary test is assigned (from a pool of randomly generated tests) that splits the data while reducing the uncertainty in both class label and offset. A leaf-node is created if a data set arriving at the node is small enough or a predefined depth has been reached. The information about the training samples reaching the leaf-node L is then stored (proportion C_L of positive examples and list of offsets $D_L = \{d_i\}$). At runtime, the stored information is utilized to cast the probabilistic Hough votes about the object's existence at different locations. The votes of the individual trees are combined by accumulating them in a 2-D Hough image $V(x,y)$ in an additive way. Each pixel location (x,y) in the Hough image then represents the sum of votes for this location. The maxima within the Hough image $V(x,y)$ are then used as detection hypotheses for a location (x,y) while the value at $V(x,y)$ serve as the confidence measure.

The location refinement is performed using region of interest of 300x300 pixels around the predicted landmark location. Final locations are determined based on the maximum of the resulting probability map (Hough image). In case multiple maxima with the same value exist, the location closer to the predicted one is chosen. Fig. 7.7 illustrates the steps of the refinement process, from region of interest (left) over probability map (middle) to final landmark location (right).

The landmark-specific forests utilized here consist of 20 random trees and are used to encode the local appearance of a landmark within a 200x200 pixel region around the original landmark location.

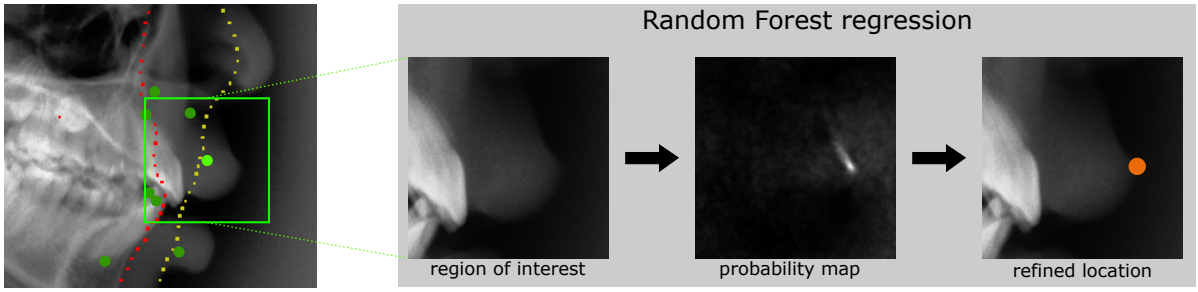


Figure 7.7.: Location refinement using landmark-specific Hough Forests. The predicted location (left) is used to define a region of interest (RoI) and the appearance information encoded in the Hough Forests is used to determine the most probable landmark location within the RoI.

7.4. Experiments & results

This section presents the experiments that were conducted to assess the performance of the presented approach as well as the achieved results. It starts by covering the data used for training and evaluation (Section 7.4.1). It explains the metrics used for measuring the quantitative results (Section 7.4.2) and the offline training of the neural network, coupled shape model, and Hough Forest (Section 7.4.3). Finally, it presents the experimental results (Section 7.4.4) which will be discussed in the next major section (Section 7.5).

7.4.1. Data

The approach was trained and tested on the ISBI 2015 Challenge[1] data set [WHL*16]. It includes lateral cephalograms from 400 patients who were between 6 and 60 years old. All radiographs were captured using the same device and stored in TIFF format with an image resolution of 1935x2400 pixels. The physical resolution is 0.1 mm per pixel along both directions. All images were manually annotated by two clinical orthodontists to generate gold standard positions for 19 cephalometric landmarks. These landmarks correspond to common anatomical structures that are used in different methods for the cephalometric analysis, like Steiner analysis (cf. Fig. 7.1). The 400 annotated cephalometric images were separated into 150 images that are designated for training, while the remaining 250 images are to be used for testing. The big age range of the subjects results in a large variety of different head sizes, morphologies and postures in the images and as a result a large variance in the overall position of the landmarks.

For this approach, additional annotations for the skull and skin (head) outline have been manually created for the first 50 training images. These annotations are required for training the U-Net and coupled shape model. Fig. 7.8 illustrates the manually annotated outlines on some exemplary cephalograms. The outline of the skull is depicted in red and the skin (head) outline in yellow.

[1] http://www-o.ntust.edu.tw/~cweiwang/ISBI2015/challenge1/

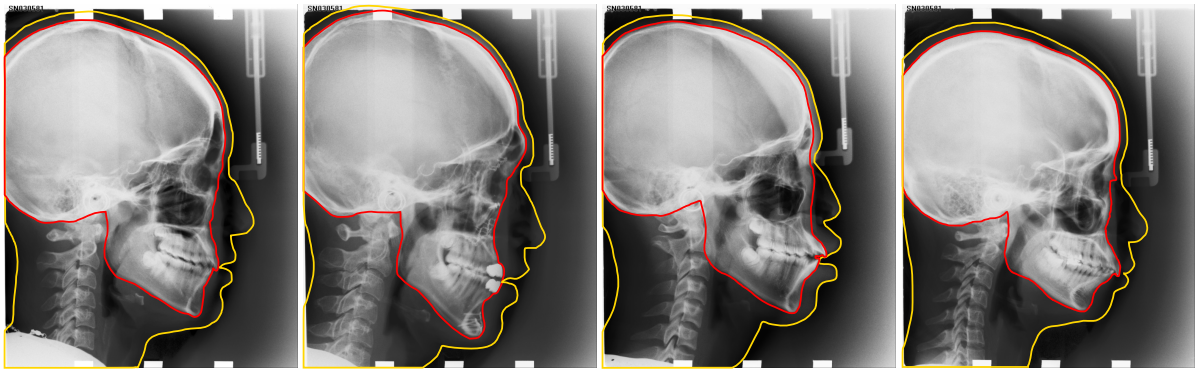


Figure 7.8.: Examples for the manually created annotations of the skull and head outlines. The skull outline is depicted in red and the head outline in yellow.

7.4.2. Metrics

The performance evaluation of the landmark localization uses the same metrics as in the original publication [WHL*16] of the ISBI Challenge. Results are reported as point-to-point errors and are measured using the absolute distance (in mm) between the gold standard annotations and the detected landmark locations. Additionally, the ratio of correctly detected locations is computed and reported as success detection rate (SDR). Here, a detection is considered correct if the point-to-point error is below a predefined threshold.

The point-to-point error is defined as the radial error $RE(d,r)$ between the detected landmark position d and the corresponding gold standard position r. It is computed by

$$RE(d,r) = \|d-r\|_2 = \sqrt{(d_x - r_x)^2 + (d_y - r_y)^2}, \tag{7.1}$$

where x and y indicate the point coordinates in x- and y-direction. The landmark-specific mean radial error MRE_l for a single landmark l is then defined as

$$MRE_l = \frac{1}{N}\sum_{i=1}^{N} RE(d^{(l,i)}, r^{(l,i)}), \tag{7.2}$$

where N is the number of images and $d^{(l,i)}$,$r^{(l,i)}$ are the detected and gold standard positions of landmark l in image i. The standard deviation of the landmark-specific mean radial error is given by

$$SD_l = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(RE(d^{(l,i)}, r^{(l,i)}) - MRE_l\right)^2}. \tag{7.3}$$

The landmark-specific success detection rate (SDR) is then defined as the fraction of correct detections over the total amount of detections (i.e. number of images). A landmark location is considered to

be correctly detected, if the radial error $RE(d^{(l,i)}, r^{(l,i)})$ between the predicted landmark position d and the gold standard position r is not greater than a threshold T:

$$SDR_l = \frac{\left|\left\{i \,|\, 1 \leq i \leq N, RE(d^{(l,i)}, r^{(l,i)}) \leq T\right\}\right|}{N} \times 100. \tag{7.4}$$

The SDR is reported for the four different thresholds that are used in the challenge, i.e. $T \in \{2.0\,\text{mm}, 2.5\,\text{mm}, 3.0\,\text{mm}, 4.0\,\text{mm}\}$.

The overall performance measures in terms of mean radial error and success detection rate are computed by averaging the individual landmark-specific results.

7.4.3. Offline training

The U-Net network for extracting the binary mask of the head area, the 2-D coupled shape model which encodes the shape variation of skull and head as well as the spatial relations between skull, head and landmarks, and the Hough Forest for refining the landmark locations have to be trained once before they can be utilized for the approach (offline training).

U-Net The U-Net for generating the binary mask of the head area was trained on the 50 training images with corresponding skin (head) outline annotations. The 50 images were randomly split into 30 training, 10 validation and 10 test images. The training was again performed as described in Section 4.3. However, no horizontal mirroring was performed during data augmentation as the cephalometric images are not symmetric and are captured from a single predefined orientation.

The network was trained for 500 epochs with early stopping, a batch size of 32, 250 steps per epoch and a dropout rate of 0.3 in each pooling layer. The images and masks were resized to a resolution of 496x609 pixels and 64 filters were used on the input layer. The Tversky parameters were chosen as $\alpha = \beta = 0.5$ since the head area covers a large amount of the image. The starting learning rate was 0.001. The training was performed on a GPU-cluster with 2x 16 core Intel Xeon Gold 6130s, 260GB of RAM and 8x Nvidia GeForce RTX 2080 Ti GPUs with 11GiB graphics memory each.

Using these parameters, the network achieved an F1 score of 0.9950 on the 10 test images. The performance was considered accurate enough so that no further parameter optimization was performed. Since there are no gold standard annotations of the head area available for the 250 test images of the ISBI data set, a visual inspection of the 250 masks generated by the network for these images was performed. It revealed only small inconsistencies which should not negatively impact the model initialization process.

Coupled shape model The 2-D coupled shape model consists of deformable model items for skull and head and single point items for the 19 landmarks. It was also trained on the 50 training images with additional annotations for skull and head outlines. During the training process (cf. Section 4.2), the statistical information about the shape variation and spatial relation is extracted from the training

Figure 7.9.: Exemplary results of the binary masks of the head area generated by the U-Net on the test data of the ISBI 2015 data set.

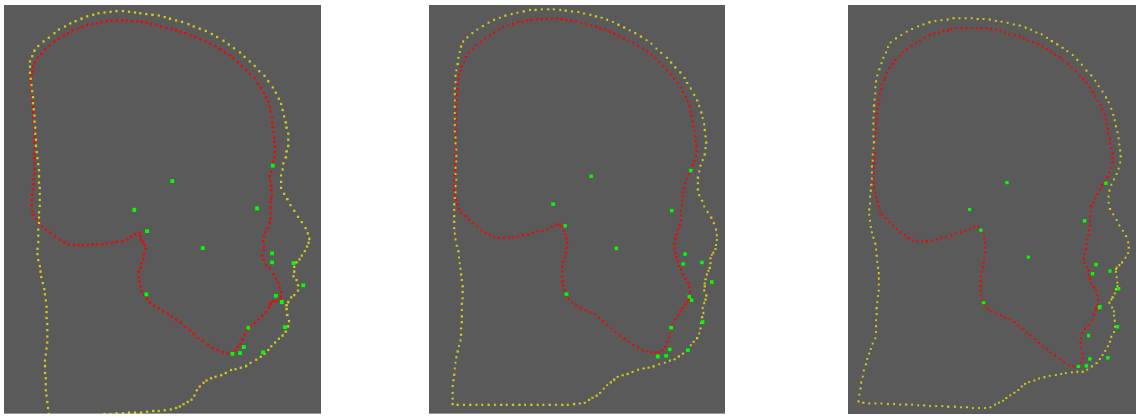


Figure 7.10.: First mode of variation of the coupled shape model containing the cephalometric landmark locations relative to the skull and head shape. The head item is colored in yellow, the skull in red and the single point landmark items are depicted in green.

data and encoded in the principal modes of variation of the model configuration vector. The contour of skull and head is represented using $L = 200$ points.

Fig. 7.10 illustrates the trained coupled shape model. It depicts the model changes related to the first mode of variation. The skull outline (red) and head outline (yellow) are depicted via their individual contour points and all landmark locations are shown in green. Because of the modeled spatial relation, the landmark locations change accordingly when the shape of skull or head is modified. This can easily be seen when looking at the chin area of the skull (lower right area of the red outline). Despite a large change in shape from the left to the right images, the three (green) landmarks remain close to the chin contour. While the true landmark locations (for these three landmarks) are actually located on the skull contour, the small mismatch in the model results from the fact that the spatial relations are modeled relative to the center of gravity of the skull and not specific contour points. Nonetheless, the locations are close enough to enable a prediction of the true landmark location.

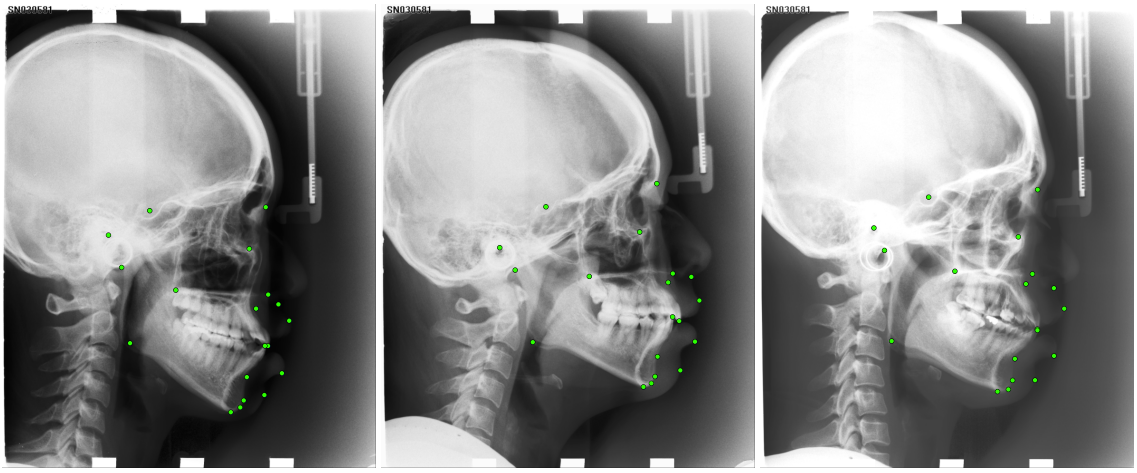


Figure 7.11.: Exemplary results of the fully automatic landmark localization in cephalometric radiographs.

Hough Forest The landmark-specific Hough Forests that encode the local appearance of the landmarks are each trained on 150 positive and 70 negative examples. The positive examples correspond to the 150 training images in the data set. In a first iteration, the forests are trained using only the positive examples and are then applied to the same training data to predict the landmark locations. Using those predictions, 70 negative examples were chosen for each landmark that incorrectly received high probabilities. The final forests are then trained using both the positive and negative examples. Each landmark-specific Hough Forest consists of a set of 20 random trees [Bre01] and is trained on patches of 200x200 pixels. The individual random tree are constructed according to the standard random forest framework [GYR*11, Bre01]

7.4.4. Landmark localization results

The performance of the presented landmark localization approach is evaluated on the separate set of 250 test images. For each test image, the landmark locations are determined by applying the workflow of location prediction and location refinement. The binary mask of the head area is computed by the U-Net network and used to place the CSM onto the image. The CSM is adapted to the skull and head outlines in the image using gradient image features and the final model configuration is used to predict the landmark locations via the encoded spatial relations. Regions of interest are then extracted around the predicted locations and the final landmark positions (within these regions) are determined by the landmark-specific Hough Forests.

As discussed during the offline training of the U-Net (cf.Section 7.4.3), no gold standard annotations of the skull and head outlines are available for the test set to quantify the accuracy of the binary masks or model adaptation. A visual inspection of the generated binary masks revealed only small inconsis-

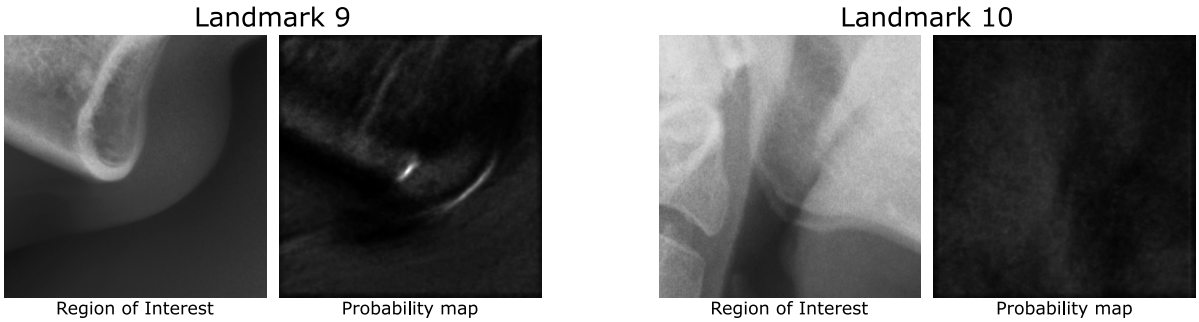


Figure 7.12.: Exemplary results of the fully automatic landmark localization in cephalometric radiographs.

tencies, and the model initialization was considered successful for all test images. The quality of the model adaptation to the skull and head contours is evaluated by the accuracy of the related landmark prediction.

To quantify the accuracy of the landmark prediction step, the landmark predictions provided by the coupled shape model are compared to the gold standard landmark locations using the mean radial error (MRE) and success detection rate (SDR), but using different thresholds. The MRE over all predicted landmark locations in the 250 images is 7.65 ± 4.21 mm. 92.94 % of predicted locations are within 10.0 mm around the true location, and 98.7 % are within 15 mm. Given the physical resolution of 0.1 mm per pixel, almost all predicted locations are therefore less then 150 pixels away from their true position. The chosen region for location refinement of 300 x 300 pixel around the predicted location should therefore contain the true position and enable the Hough Forest to detect it based on the learned local appearance.

Finally, the localization performance is assessed by comparing the detected locations to the gold standard annotations. Table 7.1 lists the localization results for each landmark in terms of mean radial errors (MRE) with standard deviation (SD) and success detection rate (SDR) (cf. Section 7.4.2). The SDR is given for four different accuracy threshold, where 2.0 mm is the clinically relevant threshold. The average detection performance over all landmarks and test images is given by a MRE of 1.953 ± 2.977 mm and a SDR of 76.04 % for the 2.0 mm accuracy threshold. The achieved localization results vary significantly between different landmarks. The landmarks 1 and 8 are the only ones with a success detection rate higher than $\sim$90 % for the 2.0 mm threshold. Other landmarks, like 5 and 13, are only detected correctly in $\sim$63 % and $\sim$66 %, respectively, for the same threshold. The worst results are achieved for landmark 10 which is within 2.0 mm of the true position in only $\sim$36 % of the images. Fig. 7.11 depicts exemplary results of the detected landmark locations.

Table 7.1.: Experimental results on the ISBI 2015 Challenge dataset [WHL*16]. Mean Radial Error (MRE) with Standard Deviation (SD) and Successful Detection Rate (SDR) in four categories are shown for each landmark as well as the total average (AVG) over all landmarks.

ID	MRE ± SD	SDR (in %)			
	(in mm)	2.0 mm	2.5 mm	3.0 mm	4.0 mm
1	1.215 ± 2.830	91.20	92.40	92.80	94.80
2	1.364 ± 2.582	80.00	84.40	86.40	91.60
3	1.453 ± 2.962	78.00	84.40	89.20	94.80
4	2.863 ± 4.916	72.40	76.00	79.20	81.60
5	2.015 ± 1.684	62.80	71.60	76.40	87.20
6	1.841 ± 2.045	69.20	76.00	82.80	90.80
7	1.628 ± 3.486	82.40	89.60	92.40	95.60
8	0.749 ± 0.490	92.00	95.20	97.20	99.60
9	1.056 ± 1.729	88.00	95.60	97.20	98.40
10	6.727 ± 7.579	36.00	41.60	46.00	55.60
11	1.512 ± 2.979	79.60	83.20	85.20	87.60
12	1.351 ± 3.402	84.80	85.20	87.60	90.80
13	2.607 ± 3.653	66.00	84.00	90.80	94.00
14	2.268 ± 4.382	80.40	86.80	92.40	94.80
15	1.517 ± 2.111	80.40	85.20	89.60	93.20
16	1.872 ± 2.772	72.00	82.00	87.20	91.60
17	1.253 ± 1.618	87.60	92.00	95.20	96.40
18	1.694 ± 2.020	70.80	79.60	84.00	90.80
19	2.126 ± 3.326	71.20	77.60	84.40	88.40
AVG	1.953 ± 2.977	76.04	82.23	86.11	90.40

7.5. Discussion

While the landmark prediction based on the adapted coupled shape model puts 98.7 % of the landmarks close enough to their true location so that the landmark-specific Hough Forest could detect it (based on the selected region of interest around the predicted position), only 76.04 % of the refined landmarks are placed within 2.0 mm (20 pixels) of their true location. The reasons for this are likely related to the challenges introduced at the beginning of the chapter. Certain landmark locations are ill-defined in the sense that they do not correspond to any clearly identifiable image features or have large appearance variations due to superimposition.

Analyzing the resulting probability maps for the regions of interest reveals the varying performance of the location refinement. While some landmark positions are reliably identified by the corresponding Hough Forest with high probabilities, probability maps for other landmarks are more uniformly distributed with no clear indication of the most likely position. For example, Fig. 7.12 depicts one region of interest after landmark prediction and corresponding probability map for landmarks 9 and 10. While the position of landmark 9 is identified with high probability (white area in the probability map), this is not the case for landmark 10.

Landmark 10 greatly suffers from the fact that the superimposition of the left and right parts of the jaw bone generates largely varying local appearances. While the spatial relation helps to predict the landmark location – the mean radial error after the prediction step is among the lowest of all landmarks – no robust Hough Forest for the localization based on the local intensity appearance could be trained. Similar issues are responsible for the lower performance of landmarks 5, 6 and 18. While landmarks 5 and 6 are located on the skull contour, their location is not well defined in vertical direction resulting in inaccurate localizations.

A more robust refinement of the landmark locations is need in order to increase the overall performance of the approach. While a more precise location prediction will also increase the localization performance because larger constrains can be imposed on the possible positions, the location refinement does not suffer from multiple possible positions with high probability within the predicted region, but rather no highly probable position at all. Despite the fact that random forest-based classifiers were the preferred choice in previous works in this area, the trained Hough Forest for this approach seem unable to replicate this performance. Lindner et al. [LWH*16] were able to increase the performance of their Random Forest Regression-Voting (RFRV) based approach by 10 % (in terms of average SDR) by fine tuning the RF parameters in their approach, which might also be an option here.

A comparison to the performance of other approaches on the 2015 Challenge dataset is shown in Table 7.2. The presented approach performs better than some recent approaches like Hwang et al. [HMK*21] or Payer et al. [PŠBU19] in the clinically relevant 2.0 mm accuracy range, but is also significantly behind the best approach from Kim et al. [KSP*20] and Lindner et al. [LWH*16]. Fortunately, the two-step approach offers many possibilities to implement changes and improve the local refinement of the locations – either by increasing the localization performance of the random forest-based approach or utilizing different methods like neural networks for the location refinement.

Table 7.2.: Comparison of different methods for cephalometric landmark localization on the ISBI 2015 Challenge data set. Successful Detection Rate (SDR) in four categories is listed for each individual approach.

Method	SDR %			
	2.0 mm	2.5 mm	3.0 mm	4.0 mm
Ibragimov et al.	68.13	74.63	79.77	86.87
Lindner and Cootes	70.65	76.93	82.17	89.85
Lindner et al.	**84.70**	89.38	92.62	96.30
Arik et al.	72.29	78.21	82.27	86.80
Payer et al.	73.33	78.76	83.24	89.75
Kim et al.	84.53	**90.11**	**93.21**	**96.79**
Hwang et al.	75.45	83.66	88.92	94.24
Proposed method	76.04	82.23	86.11	90.40

7.6. Conclusion

The automatic landmark localization in cephalometric images combines the segmentation and numbering concept for location prediction with a random forest-based regression to detect the final landmark positions. The extended coupled shape model includes the landmark locations as single point items in order to encode their spatial relation relative to the shape of skull and head (skin). The trained U-Net network was able to accurately segment the head in the cephalometric images and the resulting binary mask enabled a successful initial placement of the CSM. While the utilization of the spatial relation to the skull and head outline via the coupled shape model successfully constrains the search space of possible landmark locations to a small region of interest, the identification of the true position within this region still leaves room for improvement. The trained landmark-specific Hough Forests were unable to robustly detect the position of ill-defined landmarks based on their local appearance. On a test set of 250 images the approach achieved a success detection rate of 76.04 % in the clinically relevant 2.0 mm accuracy range. While it was able to outperform some recently presented methods, it is roughly 8 % behind the best performing approaches in this area. However, the chosen two-step approach enables the possibility to combine the location prediction with a different method for location refinement and potentially improve the overall performance.

8. Conclusions and Future Work

Oral health is an important part of personal healthcare and oral diseases like tooth decay in primary or permanent teeth, or abnormal position of teeth and misalignment of teeth or jaws affect billions of people worldwide. Regular examination of the oral health status by a medical professional is required to detect oral health conditions in their early stages and treat them before they become too severe. Digital imaging is typically employed for this task to perform image-based diagnosis and treatment planing. In dentistry, the most commonly used image modality is X-ray as it provides information about the complete tooth which is partly hidden below soft tissue and embedded in the jaw bone. In addition, the dental branch of orthodontics also utilizes photographs and 3-D models of the teeth (generated from imprints) to assess and treat malposition and malocclusion of teeth. Medical imaging applications that automatically extract clinically relevant information from these medical images would be of great assistance since they could significantly reduce the time required for analysis, enhance the clinical workflow, and remove any inter- or intraobserver variability.

In this context, this thesis focused on extracting model-based object-level segmentations of the teeth together with their corresponding numbering from medical images. The main research questions focused on i) a **teeth segmentation and numbering in dental panoramic radiographs** and ii) an **image-based 3-D teeth reconstruction from five photographs** which relies on numbered 2-D teeth outlines extracted from the different photographs. Both topics utilized the same general concept which encodes top-down information about the statistical information about the teeth shapes and spatial relations in a coupled shape model and uses dynamic adaptation to fit the model to bottom-up image features. The same concept has also been applied in a different scenario to support the **localization of clinically relevant landmarks in dental cephalometric radiographs**. This chapter provides a summary of the contributions and research results presented in the thesis in Section 8.1 and outlines open issues and ideas for future research related to this work in Section 8.2.

8.1. Research results

This thesis has presented a model-based concept for simultaneous teeth segmentation and numbering. It includes tooth-specific statistical shape models to capture the shape variance of each tooth in order to provide prior information to guide the segmentation process while restricting it to valid shapes. In addition, the individual shape models are coupled by their statistical spatial relations to form a coupled shape model. The crucial task of initially positioning the coupled shape model onto the image employs the semantic segmentation performance of convolutional neural networks to enable a fully-

automatic processing of unseen data. The adaptation is controlled by specialized, dynamic strategies which determine the order in which model items are fitted to image features in order to increase the robustness of the model adaptation.

8.1.1. Teeth segmentation & numbering in panoramic radiographs

This thesis presented the first approach that explicitly combines spatial and shape domain knowledge for simultaneous teeth segmentation and numbering via the model-based segmentation and numbering concept. Object-level tooth segmentation and numbering is challenging due to the image quality and characteristics of panoramic images, patient-specific variations in the set of teeth, and the fact that the 32 teeth are only composed of 4 different types which makes them hard to distinguish based on shape alone. The proposed concept successfully handles those challenges by combining bottom-up image information with top-down knowledge about the teeth shape and spatial relation. The initial placement of the initialization-dependent coupled shape model was solved using references computed based on a binary mask of teeth area (CNN output). The CSM was then fitted to the panoramic radiograph using specialized multi-step adaptation strategies which control the order of adapting individual model items in order to alleviate the limitations of local search algorithms and ensure an optimal segmentation process.

The proposed approach was trained and evaluated on two different sets of data. On a small in-house data set it achieved an overall teeth-specific segmentation performance in terms of F1 score of 0.833 ± 0.108. Similar results were measured for a larger data set with a F1 score of 0.823 ± 0.189. The numbering performance in terms of correct detections was 0.937 ± 0.145 on the small data set and 0.906 ± 0.196 on the larger one for an IoU threshold of 0.5, and 0.807 ± 0.193 and 0.760 ± 0.245, respectively, for an IoU threshold of 0.7. The approach is able to handle missing teeth if the space previously occupied by that tooth is still vacant and big enough to fit the original tooth (and therefore the corresponding model item of the CSM). Otherwise, contour extraction and numbering will be incorrect for subsequent teeth (following the order of adaptation in the dynamic strategy). The contour extraction is also sometimes negatively affected by malpositioned teeth, dental appliance and a varying number of visible roots (molar teeth). The known initialization dependence of local search-based methods also had a large impact on performance. While initialization was successful in 92 % of cases, a failed initialization resulted in very poor segmentation and numbering performance. In conclusion, the chosen approach was able to handle the challenges associated with the task of object-level tooth segmentation and numbering and provided promising results on two separate data sets.

8.1.2. Image-based 3-D teeth reconstruction

In orthodontics, 3-D models of the teeth extracted from DVT images or via 3-D scanners are used for treatment planning in case of misaligned teeth while colored photographs are only used for documentation purposes. This thesis presented the first fully-automatic image-based 3-D teeth reconstruction approach which aims to reconstruct the complete set of teeth. Moreover, it is the first approach to uti-

lize the five documentary photographs typically taken for orthodontic procedures. The challenges for such an approach are the minimal amount of images which significantly limits the information available for the reconstruction, the very difficult appearance properties of the teeth which results in very few usable features, the simultaneous reconstruction of multiple objects which suffers from non-trivial occlusions, and general variations in the set of teeth between patients. The proposed approach utilized a deformation-based reconstruction method. A 3-D coupled shape model of the teeth was used to encode prior information and was then deformed using a silhouette-based loss function to find the optimal match between model parameters and photographs. The required individual teeth contours were extracted from the photographs using the segmentation and numbering concept which enabled a fully-automatic reconstruction process.

On 39 test data sets, the approach achieved promising results despite the limited amount of information that is available for recovering the 3-D information. The average symmetric surface distance was 0.807 ± 0.379 mm and the dice coefficient was 0.679 ± 0.133. The reconstruction performance was most frequently affected by inaccurate contour extraction in the left and right view photographs. While maxilla and anterior views showed a teeth segmentation performance of 0.931 ± 0.107 and 0.882 ± 0.161 in terms of F1 score, respectively, the left view only reached 0.631 ± 0.389 (0.751 ± 0.317 for right view). The minimization of the silhouette-based loss therefore induces a reconstruction error and stricter shape constrains need to be enforced via the coupled shape model. Other reconstruction inaccuracies result from (completely) occluded teeth where the shape is only reconstructed based on the encoded statistical knowledge. In conclusion, the presented fully-automatic reconstruction approach achieves good results given the very limited information offered by the five photographs included in orthodontic records.

8.1.3. Cephalometric landmark localization

The analysis of dental cephalometric radiographs aims to extract fundamental information about the soft tissue-, dental- and bony structures of a patient by computing diagnostic measurements based on predefined landmark locations. This process suffers from low image quality, superimposition due to the lateral capturing position, and ill-defined landmarks. As a result, the automatic localization of these landmark positions greatly benefits from the inclusion of knowledge about their spatial relation. The thesis presented a fully-automatic approach which uses the segmentation and numbering concept to predict the position of 19 landmarks in cephalometric images. The coupled shape model was extended to include the landmark locations as single point items in order to exploit not only their spatial relation to each other but also their relation to the shapes of head and skull outlines in the image. It was initially placed onto the imaged based on a semantic segmentation of the head (generated by the CNN) and then adapted using gradient image features. A landmark-specific random forest-based regression was used to detect the final landmark positions in a region of interest centered at the predicted location.

The landmark localization approach was evaluated on a test set of 250 images and achieved a success detection rate of 76.04 % in the clinically relevant 2.0 mm accuracy range. Using the spatial relation to the skull and head outline via the coupled shape model successfully constrained the search space

of possible landmark locations to a small region of interest. The trained landmark-specific Hough Forests however struggled to robustly detect the position of ill-defined landmarks based on their local appearance within the predicted region of interest. In conclusion, the presented approach was able to outperform some recently presented methods [HMK*21, PŠBU19], but is nonetheless roughly 8 % behind the best performing approaches in this area [KSP*20, LWH*16].

8.2. Future work

The approaches presented in this thesis offer several directions for future research. This section outlines some of the open issues and proposes some ideas how to tackle them. The discussion addresses each of the three application areas in a separate subsection.

8.2.1. Teeth segmentation & numbering in panoramic radiographs

One of the open issues is that the discussed initialization dependence greatly impacts the achievable accuracy in case of incorrect initial model placements. Improving the robustness of the model initialization process is therefore a high priority. This could include calculating multiple potential initial positions, running the adaption for each and selecting the (optimal) final result based on the achieved energy minimization. Another option could be to utilize separate models for upper and lower jaw. While this should allow for an easier initialization of each model, the lost spatial relation between both dental arches could cause new challenges for difficult cases (e.g. multiple missing teeth). Another topic is to explore ways to handle a larger number of missing teeth. The current idea behind the iterative nature of the adaptation steps is to move individual model items to their correct position by relying on neighboring teeth. This concept is not applicable for a large number of missing teeth and larger changes to the approach might be required. One possible idea, that could also increase the overall segmentation accuracy, is to combine the coupled shape model with a neural network for a different segmentation and numbering approach. A suitable network could be trained to directly predict the configuration of the CSM for a given input image – this would exploit the potential of CNNs while still keeping the learned statistical information about tooth shape and spatial relation encoded in the CSM. The inclusion of wisdom teeth is a further topic for improvement. While the CSM can easily be modified to be trainable including wisdom teeth, the model initialization process would require some changes to handle the variability in the presence of (individual) wisdom teeth.

8.2.2. Image-based 3-D teeth reconstruction

The biggest issue for the silhouette-based teeth reconstruction is the sub-optimal contour extraction performance in the lateral (left and right) views. Because the teeth outlines provide the only reference for the reconstruction process, an accurate object level segmentation is required for good reconstruction performance. The variation in viewing direction present for these views needs to be better

handled during the model initialization and adaptation. While user-provided references (to better estimate the camera position or initialize the model) could alleviate some problems, this would break the fully-automatic processing pipeline. Instead, such references should be automatically determined from either the binary mask or the photograph. Other CNN architectures like Mask-RCNN could also be used to provide object-based information for the model initialization/adaptation or replace the contour extraction entirely – provided that the numbering information is accurate enough. Recent approaches in image-based reconstruction focus more on CNN-based methods [HLB21]. While most techniques deal with images that contain a single object, multi-object reconstruction in the presence of occlusion is still an interesting future research topic that fits the presented approach. In this context, the deformation-based reconstruction offers the possibility to explore options to determine the optimal model configuration via a CNN-based method while still relying on a silhouette-based loss function and the (numbered) teeth contours extracted from the photographs. Another issue is that the contours from five photographs provide only limited information about the teeth. The accuracy of the reconstruction could potentially be improved by extracting further information from the occlusal views (maxilla and mandible views) that helps to better estimate the rotation and tilt of (frontal) teeth. However, such information is not easy to retrieve because of the discussed appearance properties of teeth and would therefore require additional research to offer robust results.

8.2.3. Cephalometric landmark localization

The main issue for the cephalometric landmark localization approach is an inaccurate position refinement. Although 92.94 % of predicted locations are within 10.0 mm around the true location (98.7 % are within 15 mm), the refinement using landmark-specific Hough Forest provides unsatisfactory localization results. Despite the greatly reduced search space, the decision trees struggle to encode the local landmark appearance. While extensive parameter tuning as performed by Lindner et al. [LWH*16] could improve the performance of the Random Forest-based regression, the presented approach offers additional options. A possible solution could be to convert the two-step approach into an iterative optimization that alternates between prediction and refinement steps. Utilizing the confidence measure of the location hypotheses (generated by the Hough Forest), weak locations (with low confidence) could be further constrained based on strong localizations of neighboring landmarks. The CSM adaptation could also be extended to use the location hypotheses as bottom-up information and actively adapt the landmarks such that the optimal model configuration directly provides the final landmark locations. The landmark refinement process could also be changed to employ convolutional neural networks to determine the locations based on the predicted positions, similar to the approach of Kim et al. [KSP*20]. Again, the CSM adaptation could also incorporate these locations to further enforce the spatial relations between landmarks and other structures.

Bibliography

[AA15] AMER Y. Y., AQEL M. J.: An Efficient Segmentation Algorithm for Panoramic Dental Images. *Procedia Computer Science 65* (2015), 718–725. 8, 9, 25, 32, 33

[ACR*18] ANDREA T., CONGCONG W., RAFAEL P., FAOUZI A. C., AZEDDINE B., BJØRN E., JAKOB E. O.: Validation of stereo vision based liver surface reconstruction for image guided surgery. In *2018 Colour and Visual Computing Symposium (CVCS)* (2018), IEEE, pp. 1–6. 48

[ADA99] ADA: *Current Dental Terminology Third Edition (CDT-3)*. American Dental Association, 1999. 20

[AFSEM14] ABDELREHIM A. S., FARAG A. A., SHALABY A. M., EL-MELEGY M. T.: 2D-PCA Shape Models: Application to 3D Reconstruction of the Human Teeth from a Single Image. In *Medical Computer Vision. Large Data in Medical Imaging*. Springer International Publishing, 2014, pp. 44–52. 3, 11, 50, 51

[AIX17] ARIK S. Ö., IBRAGIMOV B., XING L.: Fully automated quantitative cephalometry using convolutional neural networks. *Journal of Medical Imaging 4*, 1 (2017), 014501. 158

[AK13] AJAZ A., KATHIRVELU D.: Dental biometrics: Computer aided human identification system using the dental panoramic radiographs. In *2013 International Conference on Communication and Signal Processing* (2013), IEEE, pp. 717–721. 3

[Als18] ALSMADI M. K.: A hybrid Fuzzy C-Means and Neutrosophic for jaw lesions segmentation. *Ain Shams Engineering Journal 9*, 4 (2018), 697–706. 2, 9, 34

[ASKT18] ALI M., SON L. H., KHAN M., TUNG N. T.: Segmentation of dental X-ray images in medical imaging using neutrosophic orthogonal matrices. *Expert Systems with Applications 91* (2018), 434–441. 34

[BIV*19] BREHLER M., ISLAM A., VOGELSANG L., YANG D., SEHNERT W., SHAKOOR D., S. DEMEHRI M. D., SIEWERDSEN J. H., ZBIJEWSKI W.: Coupled active shape models for automated segmentation and landmark localization in high-resolution CT of the foot and ankle. In *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging* (2019), Gimi B., Krol A., (Eds.), vol. 10953, International Society for Optics and Photonics, SPIE, pp. 125–130. 13, 41, 56

[BJT*20] BI H., JIANG Y., TANG H., YANG G., SHU H., DILLENSEGER J.-L.: Fast and accurate segmentation method of active shape model with Rayleigh mixture model clustering for prostate ultrasound images. *Computer Methods and Programs in Biomedicine 184* (2020), 105097. 40, 56

[BM92] BESL P. J., MCKAY N. D.: Method for registration of 3-D shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures* (1992), Schenker P. S., (Ed.), vol. 1611, International Society for Optics and Photonics, SPIE, pp. 586–606. 149

[BM93] BEUCHER S., MEYER F.: The morphological approach to segmentation: the watershed transformation. *Mathematical morphology in image processing 34* (1993), 433–481. 36

[Bre01] BREIMAN L.: Random Forests. *Machine Learning 45*, 1 (2001), 5–32. 164, 169

[BTVG06] BAY H., TUYTELAARS T., VAN GOOL L.: SURF: Speeded Up Robust Features. In *European conference on computer vision* (2006), Springer, Springer Berlin Heidelberg, pp. 404–417. 49

[Can86] CANNY J.: A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8*, 6 (1986), 679–698. 35

[CE97] COLLINS D. L., EVANS A. C.: Animal: Validation and Applications of Nonlinear Registration-Based Segmentation. *International Journal of Pattern Recognition and Artificial Intelligence 11*, 08 (1997), 1271–1294. 40

[CET01] COOTES T. F., EDWARDS G. J., TAYLOR C. J.: Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 6 (2001), 681–685. 13, 39

[CLP*21] CHUNG M., LEE J., PARK S., LEE M., LEE C. E., LEE J., SHIN Y.-G.: Individual tooth detection and identification from dental panoramic X-ray images via point-wise localization and distance regularization. *Artificial Intelligence in Medicine 111* (2021), 101996. 10, 104

[CPW*19] CHEN K., PANG J., WANG J., XIONG Y., LI X., SUN S., FENG W., LIU Z., SHI J., OUYANG W., ET AL.: Hybrid Task Cascade for Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), IEEE, pp. 4974–4983. 105

[CT*04] COOTES T. F., TAYLOR C. J., ET AL.: Statistical models of appearance for computer vision, 2004. 114

[CTCG92] COOTES T. F., TAYLOR C. J., COOPER D. H., GRAHAM J.: Training Models of Shape from Sets of Examples. In *BMVC92* (1992), Hogg D., Boyle R., (Eds.), Springer London, pp. 9–18. 39, 60

[CTCG95] COOTES T. F., TAYLOR C. J., COOPER D. H., GRAHAM J.: Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding 61*, 1 (1995), 38–59. 13, 39, 55

[CZL*19] CHEN H., ZHANG K., LYU P., LI H., ZHANG L., WU J., LEE C.-H.: A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific Reports 9*, 1 (2019). 10, 46, 47

[CZL*21] CHEN Q., ZHAO Y., LIU Y., SUN Y., YANG C., LI P., ZHANG L., GAO C.: MSLPNet: multi-scale location perception network for dental panoramic X-ray image segmentation.

Neural Computing and Applications (2021). 45, 75

[DA00] DUNCAN J. S., AYACHE N.: Medical image analysis: progress over two decades and the challenges ahead. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 1 (2000), 85–106. 30

[Dic45] DICE L. R.: Measures of the Amount of Ecologic Association Between Species. *Ecology 26*, 3 (1945), 297–302. 146

[Dof00] DOFKA C.: *Dental terminology*. Delmar, Albany, 2000. 123

[Dow56] DOWNS W. B.: Analysis of the dentofacial profile. *The Angle Orthodontist 26*, 4 (1956), 191–212. 155

[EF17] ESFANDIARKHANI M., FORUZAN A. H.: A generalized active shape model for segmentation of liver in low-contrast CT volumes. *Computers in Biology and Medicine 82* (2017), 59–70. 40, 56

[EGSEB11] ELNAKIB A., GIMEL'FARB G., SURI J., EL-BAZ A.: *Medical Image Segmentation: A Brief Survey*. Springer New York, 2011, pp. 1–39. 30

[EJ19] ELIZABETH J. R., JULIET S. E.: A Survey on Various Segmentation Methods in Medical Imaging. *International Journal of Emerging Trends in Engineering Research 7*, 11 (2019), 1–5. 29, 30

[EUR04] EUROPEAN COMMISSION: *Radiation Protection 136: European guidelines on radiation protection in dental radiology*, 136 ed. Directorate-General for Energy and Transport Directorate H, 2004. 26

[FEA*13] FARAG A., ELHABIAN S., ABDELREHIM A., ABOELMAATY W., FARMAN A., TASMAN D.: Model-Based Human Teeth Shape Recovery from a Single Optical Image with Unknown Illumination. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging* (2013), Springer, Springer Berlin Heidelberg, pp. 263–272. 12, 51

[FKW*19] FAN B., KONG Q., WANG X., WANG Z., XIANG S., PAN C., FUA P.: A Performance Evaluation of Local Features for Image-Based 3D Reconstruction. *IEEE Transactions on Image Processing 28*, 10 (2019), 4774–4789. 49

[GKK*13] GRÁFOVÁ L., KAŠPAROVÁ M., KAKAWAND S., PROCHÁZKA A., DOSTÁLOVÁ T.: Study of edge detection task in dental panoramic radiographs. *Dentomaxillofacial Radiology 42*, 7 (2013), 20120391. PMID: 23640989. 9, 35

[GL13] GALL J., LEMPITSKY V.: Class-Specific Hough Forests for Object Detection. In *Decision Forests for Computer Vision and Medical Image Analysis*, Criminisi A., Shotton J., (Eds.). Springer London, 2013, pp. 143–157. 164

[Gow75] GOWER J. C.: Generalized procrustes analysis. *Psychometrika 40*, 1 (1975), 33–51. 61, 138

[Gre15] GREENBERG A. M.: Digital technologies for dental implant treatment planning and guided surgery. *Oral and Maxillofacial Surgery Clinics 27*, 2 (2015), 319–340. 2

[GYR*11] GALL J., YAO A., RAZAVI N., VAN GOOL L., LEMPITSKY V.: Hough Forests for Object Detection, Tracking and Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 11 (2011), 2188–2202. 164, 169

[HAGM14] HARIHARAN B., ARBELÁEZ P., GIRSHICK R., MALIK J.: Simultaneous Detection and Segmentation. In *Computer Vision – ECCV 2014* (Cham, 2014), Fleet D., Pajdla T., Schiele B., Tuytelaars T., (Eds.), Springer International Publishing, pp. 297–312. 80

[Har05] HARRIS E. F.: Tooth-Coding Systems in the Clinical Dental Setting. *Dental Anthropology Journal 18*, 2 (2005), 43–49. 20

[HGDG17] HE K., GKIOXARI G., DOLLAR P., GIRSHICK R.: Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), IEEE. 43, 105

[HIHY16] HASAN M. M., ISMAIL W., HASSAN R., YOSHITAKA A.: Automatic segmentation of jaw from panoramic dental X-ray images using GVF snakes. In *2016 World Automation Congress (WAC)* (2016), IEEE, pp. 1–6. 9, 38

[HJCH19] HWANG J.-J., JUNG Y.-H., CHO B.-H., HEO M.-S.: An overview of deep learning in the field of dentistry. *Imaging Science in Dentistry 49*, 1 (2019), 1. 45

[HKR93] HUTTENLOCHER D. P., KLANDERMAN G. A., RUCKLIDGE W. J.: Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence 15*, 9 (1993), 850–863. 147

[HLB21] HAN X.-F., LAGA H., BENNAMOUN M.: Image-Based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence 43*, 5 (2021), 1578–1604. 11, 49, 52, 53, 139, 140, 179

[HLKH17] HAFEEZ J., LEE S., KWON S., HAMACHER A.: Image Based 3D Reconstruction of Texture-less Objects for VR Contents. *International Journal of Advanced Smart Convergence 6* (2017), 9–17. 48

[HM09] HEIMANN T., MEINZER H.-P.: Statistical shape models for 3D medical image segmentation: a review. *Medical Image Analysis 13*, 4 (2009), 543–563. 56, 103, 120

[HMK*21] HWANG H.-W., MOON J.-H., KIM M.-G., DONATELLI R. E., LEE S.-J.: Evaluation of automated cephalometric analysis based on the latest deep learning method. *The Angle Orthodontist 91*, 3 (2021), 329–335. 159, 172, 178

[Hux76] HUXLEY T. H.: The nature of the craniofacial apparatus of Petromyzon. *Journal of anatomy and physiology 10*, Pt 2 (1876), 412–429. 155

[HvGS*09] HEIMANN T., VAN GINNEKEN B., STYNER M. A., ARZHAEVA Y., AURICH V., BAUER C., BECK A., BECKER C., BEICHEL R., BEKES G., BELLO F., BINNIG G., BISCHOF H., BORNIK A., CASHMAN P. M. M., CHI Y., CORDOVA A., DAWANT B. M., FIDRICH M., FURST J. D., FURUKAWA D., GRENACHER L., HORNEGGER J., KAINMÜLLER D., KITNEY R. I., KOBATAKE H., LAMECKER H., LANGE T., LEE J., LENNON B., LI R., LI S., MEINZER H.-P., NEMETH G., RAICU D. S., RAU A.-

M., VAN RIKXOORT E. M., ROUSSON M., RUSKO L., SADDI K. A., SCHMIDT G., SEGHERS D., SHIMIZU A., SLAGMOLEN P., SORANTIN E., SOZA G., SUSOMBOON R., WAITE J. M., WIMMER A., WOLF I.: Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. *IEEE Transactions on Medical Imaging 28*, 8 (2009), 1251–1265. 145

[IANJ15] INDRASWARI R., ARIFIN A. Z., NAVASTARA D. A., JAWAS N.: Teeth segmentation on dental panoramic radiographs using decimation-free directional filter bank thresholding and multistage adaptive thresholding. In *2015 International Conference on Information & Communication Technology and Systems (ICTS)* (2015), IEEE, pp. 49–54. 9, 33

[ILPV14] IBRAGIMOV B., LIKAR B., PERNUS F., VRTOVEC T.: Automatic Cephalometric X-Ray Landmark Detection by Applying Game Theory and Random Forests. In *Proc. ISBI Int. Symp. on Biomedical Imaging* (2014), pp. 1–8. 157

[ILPV15] IBRAGIMOV B., LIKAR B., PERNUS F., VRTOVEC T.: Computerized cephalometry by game theory with shape-and appearance-based landmark refinement. In *Proceedings of International Symposium on Biomedical imaging (ISBI)* (2015). 157

[Ire10] IRELAND R.: *A Dictionary of Dentistry*. Oxford University Press, Oxford New York, 2010. 19

[IS15] IOFFE S., SZEGEDY C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning* (2015), Bach F., Blei D., (Eds.), vol. 37 of *Proceedings of Machine Learning Research*, PMLR, pp. 448–456. 43

[ISO16] ISO: *ISO 3950:2016: Dentistry — Designation system for teeth and areas of the oral cavity*, 4 ed. International Organization for Standardization, Geneva, Switzerland, 2016. 20

[Jac01] JACCARD P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat 37* (1901), 547–579. 80, 146

[JCCS19] JOSKOWICZ L., COHEN D., CAPLAN N., SOSNA J.: Inter-observer variability of manual contour delineation of structures in CT. *European Radiology 29*, 3 (2019), 1391–1399. 30

[JFR*18] JADER G., FONTINELI J., RUIZ M., LIMA K., PITHON M., OLIVEIRA L.: Deep Instance Segmentation of Teeth in Panoramic X-Ray Images. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (2018), IEEE, pp. 400–407. 3, 10, 43, 44, 76, 104

[JKW16] JUNG F., KNAPP O., WESARG S.: Automatic Segmentation of Structures in CT Head and Neck Images using a Coupled Shape Model. *The MIDAS Journal - Head and Neck Auto Segmentation Challenge* (2016). 59

[JMDW17] JUNG F., MEDEA B.-R., DAISNE J.-F., WESARG S.: Automatic Sentinel Lymph Node Localization in Head and Neck Cancer Using a Coupled Shape Model Algorithm. In

Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures (2017), Cardoso M. J., Arbel T., Luo X., Wesarg S., Reichl T., González Ballester M. Á., McLeod J., Drechsler K., Peters T., Erdt M., Mori K., Linguraru M. G., Uhl A., Oyarzun Laura C., Shekhar R., (Eds.), Springer International Publishing, pp. 133–140. 13, 41, 56

[JMF99] JAIN A. K., MURTY M. N., FLYNN P. J.: Data clustering: a review. *ACM Computing Surveys 31*, 3 (1999), 264–323. 114

[JSK*14] JUNG F., STEGER S., KNAPP O., NOLL M., WESARG S.: COSMO – Coupled Shape Model for Radiation Therapy Planning of Head and Neck Cancer. In *Clinical Image-Based Procedures. Translational Research in Medical Imaging* (Cham, 2014), Springer International Publishing, pp. 25–32. 13, 41, 56, 59, 138

[Kir13] KIRSCHNER M.: *The probabilistic active shape model: From model construction to flexible medical image segmentation*. PhD thesis, Technische Universität Darmstadt, 2013. 56, 66, 68, 138, 147

[KKJ*20] KIM C., KIM D., JEONG H., YOON S.-J., YOUM S.: Automatic Tooth Detection and Numbering Using a Combination of a CNN and Heuristic Algorithm. *Applied Sciences 10*, 16 (2020), 5624. 10, 46, 47

[KPIB19] KOCH T. L., PERSLEV M., IGEL C., BRANDT S. S.: Accurate Segmentation of Dental Panoramic Radiographs with U-NETS. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (2019), IEEE, pp. 15–19. 10, 43, 64, 75, 104

[KS00] KUTULAKOS K. N., SEITZ S. M.: A Theory of Shape by Space Carving. *International Journal of Computer Vision 38*, 3 (2000), 199–218. 139

[KSP*20] KIM H., SHIM E., PARK J., KIM Y.-J., LEE U., KIM Y.: Web-based fully automated cephalometric analysis by deep learning. *Computer Methods and Programs in Biomedicine 194* (2020), 105513. 155, 159, 172, 178, 179

[KWT88] KASS M., WITKIN A., TERZOPOULOS D.: Snakes: Active contour models. *International Journal of Computer Vision 1*, 4 (1988), 321–331. 37

[Lau94] LAURENTINI A.: The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence 16*, 2 (1994), 150–162. 49, 52, 140

[LC15] LINDNER C., COOTES T. F.: Fully automatic cephalometric evaluation using random forest regression-voting. In *IEEE International Symposium on Biomedical Imaging* (2015), Citeseer. 157

[LGN09] LIRA P. H. M., GIRALDI G. A., NEVES L. A. P.: Panoramic dental X-Ray image segmentation and feature extraction. In *Proceedings of V workshop of computing vision, Sao Paulo, Brazil* (2009). 10, 39, 40

[LGN13] LIRA P. H. M., GIRALDI G. A., NEVES L. A. P.: Using the mathematical morphology and shape matching for automatic data extraction in dental X-ray images. In *IX Workshop de Visao Computacional (WVC2013)* (2013). 32

[LHH*14] LIN P. L., HUANG P. Y., HUANG P. W., HSU H. C., CHEN C. C.: Teeth segmentation of dental periapical radiographs based on local singularity analysis. *Computer Methods and Programs in Biomedicine 113*, 2 (2014), 433–445. 9

[LL08] LAI Y. H., LIN P. L.: Effective segmentation for dental X-ray images using texture-based fuzzy inference system. In *International Conference on Advanced Concepts for Intelligent Vision Systems* (Heidelberg, Berlin,, 2008), Springer, pp. 936–947. 9

[LLH10] LIN P. L., LAI Y. H., HUANG P. W.: An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information. *Pattern Recognition 43*, 4 (2010), 1380–1392. 9, 37

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014* (Cham, 2014), Fleet D., Pajdla T., Schiele B., Tuytelaars T., (Eds.), Springer International Publishing, pp. 740–755. 43

[Low04] LOWE D. G.: Distinctive Image Features from Scale-Invariant Keypoints. *International journal of computer vision 60*, 2 (2004), 91–110. 49

[LQQ*18] LIU S., QI L., QIN H., SHI J., JIA J.: Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), IEEE, pp. 8759–8768. 105

[LSD15] LONG J., SHELHAMER E., DARRELL T.: Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), IEEE, pp. 3431–3440. 41

[LSSL12] LI H., SUN G., SUN H., LIU W.: Watershed algorithm based on morphology for dental X-ray images segmentation. In *2012 IEEE 11th International Conference on Signal Processing* (2012), vol. 2, pp. 877–880. 9, 36

[LWH*16] LINDNER C., WANG C.-W., HUANG C.-T., LI C.-H., CHANG S.-W., COOTES T. F.: Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms. *Scientific Reports 6* (Sept. 2016), 33581. 155, 156, 158, 172, 178, 179

[LWP*15] LEE R. J., WEISSHEIMER A., PHAM J., GO L., DE MENEZES L. M., REDMOND W. R., LOOS J. F., SAMESHIMA G. T., TONG H.: Three-dimensional monitoring of root movement during orthodontic treatment. *American Journal of Orthodontics and Dentofacial Orthopedics 147*, 1 (2015), 132–142. 2

[Mah36] MAHALANOBIS P. C.: On the generalized distance in statistics. In *Proceedings of the National Institute of Science of India* (1936), vol. 12, National Institute of Science of India, pp. 49–55. 67

[MBP*21] MINAEE S., BOYKOV Y. Y., PORIKLI F., PLAZA A. J., KEHTARNAVAZ N., TERZOPOULOS D.: Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. 41, 78

[MD11] MODI C. K., DESAI N. P.: A simple and novel algorithm for automatic selection of ROI for dental radiograph segmentation. In *2011 24th Canadian Conference on Electrical and Computer Engineering(CCECE)* (2011), IEEE, pp. 000504–000507. 9, 36

[MEA*14] MOSTAFA E., ELHABIAN S., ABDELRAHIM A., ELSHAZLY S., FARAG A.: Statistical morphable model for human teeth restoration. In *2014 IEEE International Conference on Image Processing (ICIP)* (2014), IEEE, IEEE, pp. 4285–4288. 12, 51

[MOM12] MONTAVON G., ORR G. B., MÜLLER K.-R. (Eds.): *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, 2012. 64

[MP97] MOGHADDAM B., PENTLAND A.: Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*, 7 (1997), 696–710. 67

[MSV95] MALLADI R., SETHIAN J. A., VEMURI B. C.: Shape modeling with front propagation: a level set approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17*, 2 (1995), 158–175. 37

[MT96] MCINERNEY T., TERZOPOULOS D.: Deformable models in medical image analysis: a survey. *Medical Image Analysis 1*, 2 (1996), 91–108. 55, 105

[Na'17] NA'AM J.: Edge Detection on Objects of Medical Image with Enhancement Multiple Morphological Gradient Method. *Proceeding of the Electrical Engineering Computer Science and Informatics 4*, 1 (2017), 1–7. 9

[NAM05] NOMIR O., ABDEL-MOTTALEB M.: A system for human identification from X-ray dental radiographs. *Pattern Recognition 38*, 8 (2005), 1295–1305. 32

[NAM07] NOMIR O., ABDEL-MOTTALEB M.: Human identification from dental X-ray images based on the shape and appearance of the teeth. *IEEE Transactions on Information Forensics and Security 2*, 2 (2007), 188–197. 32

[NAM08a] NOMIR O., ABDEL-MOTTALEB M.: Fusion of Matching Algorithms for Human Identification Using Dental X-Ray Radiographs. *IEEE Transactions on Information Forensics and Security 3*, 2 (2008), 223–233. 2, 32

[NAM08b] NOMIR O., ABDEL-MOTTALEB M.: Hierarchical contour matching for dental X-ray radiographs. *Pattern Recognition 41*, 1 (2008), 130–138. 32

[NNH*21] NISHITANI Y., NAKAYAMA R., HAYASHI D., HIZUKURI A., MURATA K.: Segmentation of teeth in panoramic dental X-ray images using U-Net with a loss function weighted on the tooth edge. *Radiological Physics and Technology 14*, 1 (2021), 64–69. 10, 44, 104

[NYD16] NEWELL A., YANG K., DENG J.: Stacked Hourglass Networks for Human Pose Estimation. In *European conference on computer vision* (2016), Springer, Springer International Publishing, pp. 483–499. 159

[OEC17] OECD: *Health at a Glance 2017*. 2017. 1

[OT09] OLIVEIRA F. P. M., TAVARES J. M. R. S.: Matching contours in images through the use of curvature, distance to centroid and global optimization with order-preserving con-

straint. *Computer Modeling in Engineering & Sciences 43(1)* (2009), 91–110. 61

[Ots79] OTSU N.: A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics 9*, 1 (1979), 62–66. 32

[PNdS20] PADILLA R., NETTO S. L., DA SILVA E. A. B.: A Survey on Performance Metrics for Object-Detection Algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)* (2020), IEEE, pp. 237–242. 80

[PŠBU19] PAYER C., ŠTERN D., BISCHOF H., URSCHLER M.: Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical Image Analysis 54* (2019), 207–219. 158, 172, 178

[PXP00] PHAM D., XU C., PRINCE J.: A Survey of Current Methods in Medical Image Segmentation. *Annual Review of Biomedical Engineering 2*, 1 (2000), 315–37. 6, 29

[RAA*18] ROTH G. A., ABATE D., ABATE K. H., ABAY S. M., ABBAFATI C., ET AL.: Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet 392*, 10159 (2018), 1736–1788. 1

[RAH*14] RAZALI M. R. M., AHMAD N. S., HASSAN R., ZAKI Z. M., ISMAIL W.: Sobel and Canny Edges Segmentations for the Dental Age Assessment. In *2014 International Conference on Computer Assisted System in Health* (2014), IEEE, pp. 62–66. 9, 35

[RAMZI14] RAZALI M. R. M., AHMAD N. S., MOHD ZAKI Z., ISMAIL W.: Region of adaptive threshold segmentation between mean, median and otsu threshold for dental age assessment. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)* (2014), pp. 353–356. 32

[RC19] RAHAMAN H., CHAMPION E.: To 3D or Not 3D: Choosing a Photogrammetry Workflow for Cultural Heritage Groups. *Heritage 2*, 3 (2019), 1835–1851. 48

[RF18] REDMON J., FARHADI A.: Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018). 159

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (2015), Navab N., Hornegger J., Wells W. M., Frangi A. F., (Eds.), Springer International Publishing, pp. 234–241. 42, 43, 57, 58, 63

[RHGS17] REN S., HE K., GIRSHICK R., SUN J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 6 (2017), 1137–1149. 43, 44, 47

[RIA*17] RAZALI M. R. M., ISMAIL W., AHMAD N. S., BAHARI M., ZAKI Z. M., RADMAN A.: An Adaptive Thresholding Method for Segmenting Dental X-Ray Images. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 9*, 4 (2017), 1–5. 32

[RRC*82] RICKETTS R. M., ROTH R. H., CHACONAS S. J., SCHULHOF R. J., ENGEL G. A.: *Orthodontic diagnosis and planning ... their roles in preventive and rehabilitative dentistry*, vol. 1. Denver: Rocky Mountain Data Systems, 1982. 155

[RRKB11] RUBLEE E., RABAUD V., KONOLIGE K., BRADSKI G.: ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision* (2011), pp. 2564–2571. 49

[RRN13] RAD A. E., RAHIM M. S. M., NOROUZI A.: Digital Dental X-Ray Image Segmentation and Feature Extraction. *Indonesian Journal of Electrical Engineering and Computer Science 11*, 6 (2013), 3109–3114. 9, 37

[RRN14] RAD A. E., RAHIM M. S. M., NOROUZI A.: Level Set and Morphological Operation Techniques in Application of Dental Image Segmentation. *International Journal of Computer and Information Engineering 8*, 4 (2014), 182–185. 9, 38

[SBVS20] SPEIDEL S., BODENSTEDT S., VASCONCELOS F., STOYANOV D.: Chapter 29 - Interventional imaging: Vision. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, Zhou S. K., Rueckert D., Fichtinger G., (Eds.), The Elsevier and MICCAI Society Book Series. Academic Press, 2020, pp. 721–745. 49

[SCKM20] SIVAGAMI S., CHITRA P., KAILASH G. S. R., MURALIDHARAN S. R.: UNet Architecture Based Dental Panoramic Image Segmentation. In *2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)* (2020), pp. 187–191. 10, 45, 104

[SEG17] SALEHI S. S. M., ERDOGMUS D., GHOLIPOUR A.: Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In *Machine Learning in Medical Imaging* (2017), Wang Q., Shi Y., Suk H.-I., Suzuki K., (Eds.), Springer International Publishing, pp. 379–387. 64

[SFNA04] SAID E., FAHMY G. F., NASSAR D., AMMAR H.: Dental x-ray image segmentation. In *Biometric Technology for Human Identification* (2004), vol. 5404, International Society for Optics and Photonics, pp. 409–418. 5, 25

[SJW14] STEGER S., JUNG F., WESARG S.: Personalized articulated atlas with a dynamic adaptation strategy for bone segmentation in CT or CT/MR head and neck images. In *Medical Imaging 2014: Image Processing* (2014), Ourselin S., Styner M. A., (Eds.), vol. 9034, International Society for Optics and Photonics, SPIE, p. 90341I. 59, 67

[SKRB13] SUBRAMANIAM P., KUMAR K., RAMAKRISHNA T., BHADRANNA A.: Bone regeneration with plasma-rich-protein following enucleation of traumatic bone cyst. *European journal of dentistry 7*, 03 (2013), 377–381. 25

[SKW12] STEGER S., KIRSCHNER M., WESARG S.: Articulated atlas for segmentation of the skeleton from head & neck CT datasets. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on* (2012), IEEE, pp. 1256–1259. 59, 62

[Smi78] SMITH A. R.: Color gamut transform pairs. *ACM Siggraph Computer Graphics 12*, 3 (1978), 12–19. 115, 116

[SN12] SWIFT J. Q., NELSON W. J.: The Nature of Third Molars: Are Third Molars Different than Other Teeth? *Atlas of the Oral and Maxillofacial Surgery Clinics 20*, 2 (2012), 159–162. 8

[SNFA06] SAID E. H., NASSAR D. E. M., FAHMY G., AMMAR H. H.: Teeth segmentation in digitized dental X-ray films using mathematical morphology. *IEEE Transactions on Information Forensics and Security 1*, 2 (2006), 178–189. 6

[SOP18] SILVA G., OLIVEIRA L., PITHON M.: Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert Systems with Applications 107* (2018), 15–31. 2, 10, 30, 32, 43, 44, 47, 75, 77, 104

[Sor48] SORENSEN T. A.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar. 5* (1948), 1–34. 146

[SPOP20] SILVA B., PINHEIRO L., OLIVEIRA L., PITHON M.: A study on tooth segmentation and numbering using end-to-end deep neural networks. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (2020), pp. 164–171. 10, 75, 104

[SS74] SCOTT J. H., SYMONS N. B. B.: *Introduction to dental anatomy*. Churchill Livingstone, 1974. 8, 22

[SSD20] SULTANA F., SUFIAN A., DUTTA P.: Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey. *Knowledge-Based Systems 201-202* (2020), 106062. 41

[ST16] SON L. H., TUAN T. M.: A cooperative semi-supervised fuzzy clustering framework for dental X-ray image segmentation. *Expert Systems with Applications 46* (2016), 380–393. 9, 30, 33, 34

[Ste53] STEINER C. C.: Cephalometrics for you and me. *American Journal of Orthodontics and Dentofacial Orthopedics 39*, 10 (1953), 729–755. 155, 157

[Ste13] STEGER S.: *Registrierung und Segmentierung von Lymphknoten aus multimodalen Zeitreihen im Kopf-Hals-Bereich*. PhD thesis, Technische Universität, 2013. 30

[Tel04] TELEA A.: An Image Inpainting Technique Based on the Fast Marching Method. *Journal of Graphics Tools 9*, 1 (2004), 23–34. 116

[TF88] TERZOPOULOS D., FLEISCHER K.: Deformable models. *The Visual Computer 4*, 6 (1988), 306–331. 39

[TFW17] TIAN Y., FAN B., WU F.: L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). 49

[TNB*16] TIKHE S. V., NAIK A. M., BHIDE S. D., SARAVANAN T., KALIYAMURTHIE K. P.: Algorithm to Identify Enamel Caries and Interproximal Caries Using Dental Digital Ra-

diographs. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)* (2016), pp. 225–228. 2

[TTB*19] TUZOFF D., TUZOVA L., BORNSTEIN M., KRASNOV A., KHARCHENKO M., NIKOLENKO S., SVESHNIKOV M., BEDNENKO G.: Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology 48*, 4 (2019), 20180051. 10, 104

[TWT*03] TSAI A., WELLS W. M., TEMPANY C., GRIMSON E., WILLSKY A. S.: Coupled Multi-shape Model and Mutual Information for Medical Image Segmentation. In *Information Processing in Medical Imaging*. Springer Berlin Heidelberg, 2003, pp. 185–197. 13, 40, 56

[TWT*04] TSAI A., WELLS W., TEMPANY C., GRIMSON E., WILLSKY A.: Mutual information in coupled multi-shape model for medical image segmentation. *Medical Image Analysis 8*, 4 (2004), 429–445. 40

[WA82] WEBB J. A., AGGARWAL J. K.: Structure from motion of rigid and jointed objects. *Artificial Intelligence 19*, 1 (1982), 107–130. 140

[WBG*16] WU C., BRADLEY D., GARRIDO P., ZOLLHÖFER M., THEOBALT C., GROSS M. H., BEELER T.: Model-based teeth reconstruction. *ACM Transactions on Graphics 35*, 6 (2016), 1–13. 11, 12, 49, 50, 52, 53, 114

[WCKE18] WANG C., CHEIKH F. A., KAANICHE M., ELLE O. J.: Liver surface reconstruction for image guided surgery. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling* (2018), Fei B., Robert J. Webster I. I. I., (Eds.), vol. 10576, International Society for Optics and Photonics, SPIE, pp. 576–583. 48

[WHL*16] WANG C.-W., HUANG C.-T., LEE J.-H., LI C.-H., CHANG S.-W., SIAO M.-J., LAI T.-M., IBRAGIMOV B., VRTOVEC T., RONNEBERGER O., FISCHER P., COOTES T. F., LINDNER C.: A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis 31* (2016), 63–76. 2, 4, 5, 25, 41, 155, 156, 157, 165, 166, 171

[WJN*21] WIRTZ A., JUNG F., NOLL M., WANG A., WESARG S.: Automatic model-based 3-D reconstruction of the teeth from five photographs with predefined viewing directions. In *Medical Imaging 2021: Image Processing* (2021), Landman B. A., Išgum I., (Eds.), vol. 11596, International Society for Optics and Photonics, SPIE, pp. 198–212. 59, 109

[WLW20] WIRTZ A., LAM J., WESARG S.: Automated Cephalometric Landmark Localization using a Coupled Shape Model. *Current Directions in Biomedical Engineering 6*, 3 (2020), 56–59. 59, 156

[WM16] WU J., MAHFOUZ M. R.: Robust x-ray image segmentation by spectral clustering and active shape model. *Journal of Medical Imaging 3*, 3 (2016), 1–8. 40, 56

[WMW18] WIRTZ A., MIRASHI S. G., WESARG S.: Automatic Teeth Segmentation in Panoramic X-Ray Images Using a Coupled Shape Model in Combination with a Neural Network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (2018),

Frangi A. F., Schnabel J. A., Davatzikos C., Alberola-López C., Fichtinger G., (Eds.), Springer International Publishing, pp. 712–719. 59, 75

[WWW18] WIRTZ A., WAMBACH J., WESARG S.: Automatic Teeth Segmentation in Cephalometric X-Ray Images Using a Coupled Shape Model. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 194–203. 59, 75

[WWZ20] WANG Z., WANG E., ZHU Y.: Image segmentation evaluation: a survey of methods. *Artificial Intelligence Review* (2020). 78

[YYMS09] YASUNORI E., YUKIHIRO H., MAKITO Y., SADAAKI M.: On semi-supervised fuzzy c-means clustering. In *2009 IEEE International Conference on Fuzzy Systems* (2009), IEEE, pp. 1119–1124. 9, 33

[ZLG*20] ZHAO Y., LI P., GAO C., LIU Y., CHEN Q., YANG F., MENG D.: TSASNet: Tooth segmentation on dental panoramic X-ray images by Two-Stage Attention Segmentation Network. *Knowledge-Based Systems 206* (2020), 106338. 75

[ZLNW18] ZHAO T., LI S., NGAN K. N., WU F.: 3-D reconstruction of human body shape from a single commodity depth camera. *IEEE Transactions on Multimedia 21*, 1 (2018), 114–123. 48

[ZLS11] ZHENG S.-X., LI J., SUN Q.-F.: A novel 3D morphing approach for tooth occlusal surface reconstruction. *Computer-Aided Design 43*, 3 (2011), 293–302. 11, 51

[ZWZ*20] ZHANG H., WU C., ZHANG Z., ZHU Y., LIN H., ZHANG Z., SUN Y., HE T., MUELLER J., MANMATHA R., ET AL.: Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955* (2020). 105

A. Publications and Talks

The thesis is partially based on the following publications and talks:

A.1. Publications

WMW18 Wirtz A., Mirashi S. G., Wesarg S.: *Automatic Teeth Segmentation in Panoramic X-Ray Images Using a Coupled Shape Model in Combination with a Neural Network*. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 (2018), Springer International Publishing, pp. 712–719.

WWW18 Wirtz A., Wambach J., Wesarg S.: *Automatic Teeth Segmentation in Cephalometric X-Ray Images Using a Coupled Shape Model*. In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, 2018, pp. 194–203.

WLW20 Wirtz A., Lam J., Wesarg S.: *Automated Cephalometric Landmark Localization using a Coupled Shape Model*. Current Directions in Biomedical Engineering 6, 3 (2020), 56–59.

WJN*21 Wirtz A., Jung F., Noll M., Wang A., Wesarg S.: *Automatic model-based 3-D reconstruction of the teeth from five photographs with predefined viewing directions*. In Medical Imaging 2021: Image Processing (2021), vol. 11596, International Society for Optics and Photonics, SPIE, pp. 198–212.

B. Supervising Activities

The following list summarizes the student bachelor, diploma and master thesis supervised by the author. The results of these works were partially used as an input into the thesis.

B.1. Master Thesis

1. Wambach, J. "Vollautomatische Segmentierung der Zähne in Fernröntgenseitenbildern unter Verwendung statistischer Formmodelle", Master Thesis, Hochschule Darmstadt, 2018
2. Mirashi, S. G. "Model-based Segmentation of the Teeth in Panoramic Radiograph Images", Master Thesis, Technische Universität Darmstadt, 2021

B.2. Bachelor Thesis

1. Singer, M. "Vollautomatische Initialisierung von Coupled-Shape-Modellen mithilfe von Deep Neural Networks", Bachelor Thesis, Technische Universität Darmstadt, 2021
2. Lam, J., "Vollautomatische Analyse von Fernröntgenseitenbildern unter Berücksichtigung von geometrischen Lagebeziehungen der Merkmalspunkte", Bachelor Thesis, Hochschule Darmstadt, 2019