Walter Unterpirker, BSc

Anatomical Landmark Localization for an Automatic Multi-Factorial Age Assessment System

MASTER'S THESIS

to achieve the university degree of
Diplom-Ingenieur

Master's degree programme
Telematics

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. DI. Dr.techn. Horst Bischof
Institute for Computer Graphics and Vision, University of Technology, Graz

Advisor

DI. Dr.techn. Martin Urschler
Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz

Graz, Austria, September 2015

Abstract

Anatomical landmark localization in medical images has gained an increasing research interest in the last years. One reason is that many subsequent medical image-processing algorithms benefit from an accurate and reliable preceding automatic localization step. One important application which is considered in this work, is the automated biological age assessment of humans. This is based on the ossification and mineralization process of various anatomical structures. For this thesis, these structures are acquired by non-invasive and ionizing radiation free magnetic resonance imaging. A first step towards such an automated age assessment system is to locate the age-relevant anatomical structures.

In this work, Random Regression Forests (RRFs) are explored in more detail to locate structures at hand-bones, wisdom teeth and clavicle bones. Firstly, a geodesic weighting scheme for hand-bone localization is proposed. This is based on the underlying idea that closer and simultaneously less shape-varying structures to an anatomical landmark contribute more to an accurate localization. In a second contribution, the appearance of landmarks are directly incorporated into the *RRF* framework. Thus allowing to increase the confidence of a correct landmark estimation. Due to strongly varying appearance and shapes within medical images a final contribution investigates the idea of using restricted image information around landmarks. Subsequently, the anatomical variations at the landmarks themselves are explored in more detail by the *RRF*.

Keywords. Random Forest, Hough Forest, Landmark Localization, Third Molars, Wisdom Teeth, Clavicle, Hand-Bones, Automatic Age Assessment, Biological Age Assessment, Machine Learning

Kurzfassung

Lokalisierung von anatomischen Strukturen in medizinischen Bildern hat in den letzten Jahren stark an Forschungsinteresse zugenommen. Ein Grund dafür ist, dass viele nachfolgende medizinische Bildverarbeitungsalgorithmen von einer genauen und zuverlässigen automatischen Lokalisierung als Initialisierungsschritt profitieren. Eine wichtige Anwendung, welche in dieser Arbeit als Motivation dient, ist die automatische Altersschätzung von Menschen. Diese basiert auf den Ossifikations- und Mineralisierungsprozessen verschiedener anatomischen Strukturen, welche zum Beispiel durch nicht-invasive und nicht-ionisierende Methoden wie die Magnetresonanztomographie aufgenommen werden. Ein erster Schritt in Richtung solch eines Systems zur automatischen Altersschätzung ist das Lokalisieren von diesen altersrelevanten Strukturen.

In dieser Arbeit werden Random Regression Forests (RRF) für die Lokalisierung von Hand-Knochen, Weisheitszähnen und Schlüsselbeinknochen untersucht. An erster Stelle wird ein geodätisches Gewichtungsschema für die Hand-Knochen Lokalisierung vorgestellt. Weiters werden zwei Methoden untersucht, wie das Aussehen einer zu lokalisierenden anatomischen Struktur direkt in das RRF System eingebettet werden kann. Darausfolgend sollen die zu lokalisierenden Strukturen mit einer höheren Genauigkeit gefunden werden.

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used.

The text document uploaded to TUGRAZonline is identical to the present master's thesis dissertation.

Date	Signature

Acknowledgments

At first, I would like to express my highest gratitude to my advisor Martin Urschler. With his encouragement, expertise and helpful discussions he supported me a lot during the whole work. Furthermore, this thesis would have been impossible to do without him. Further, I wish to thank Prof. Horst Bischof for being my supervisor.

Next, I would like to thank my colleagues from the Medical Image Processing group, Thomas Ebner, Darko Stern and Kerstin Hammernik who supported and shared their knowledge with me. Furthermore, I express my gratitude to all my friends who stayed on my site during my whole study.

Above all, my biggest thanks go to my loving parents, my siblings and especially my nephews and nieces who always supported and encouraged me.

This work was partly supported by the Austrian Science Fund (FWF):
P 28078-N33.

Contents

List of Figures

1 Introduction

Contents

Most living beings retrieve visual information via the visual cortex system. This information is immediately processed after the acquisition to make a decision, based on a complex construction of neurons and synapses. Nowadays, mostly computers are used to process information which is gathered from cameras and other sensor technologies. Especially in medical imaging, modalities like Magnetic Resonance (MR), Computed Tomography (CT), Positron Emission Tomography (PET), ultrasound imaging, etc. are often used to acquire images from inner body parts and have gained a lot of research interest [26]. This medical information is important and helpful, since it can be used to recognize anomalies and diseases. Further, surgical interventions can be planned in a more sophisticated way. However, handling all this information is tedious and complex. Therefore, the field of computerized Medical Image Processing (MIP) has gained a lot of research interest in the last decades.

MIP is the process of analyzing and (pre-)processing any kind of medical images for the reason of supporting medical experts by providing important information and tools for a specific kind of problem. For instance, with a content based image retrieval system [53], information can be pre-filtered to retain only the important one, thus reducing the complexity of the data. In addition to the retrieval system, a computer-assisted diagnosis or visualization system can be used to visualize the before retrieved important information in a clear and convenient way. For instance, with a head-mounted display [14], a medical expert can be supported by providing additional information to a patient during a surgical intervention. A different sub-area of *MIP* is described within this thesis, namely the

automatic localization of anatomical landmarks.

1.1 Landmark Localization and Applications

Landmarks are notice- and recognizable features which occur frequently at a specific position within images [59]. For example, this can be a remarkable position within a photo of a landscape or an unique, re-occurring pattern in noise. Further specializing these features to medical applications leads to the focus of this work, namely medical landmarks. They can be seen as certain locations at structures of interest within the human body. Examples are wisdom teeth (also known as third molars), finger tips, points at highest curvature of clavicles, hand-bone joints, organs, knees, etc. Such landmarks have often been manually located by medical experts which is highly cost- and labor-intensive. Moreover, manually searching for such landmarks in a huge amount of datasets can lead to errors due to negligence caused by monotony and other factors like tiredness.

In the last years, the need of a reliable landmark localization system has become very high since landmarks are often used as a preceding step for further (automatic) processing algorithms. For instance, in the field of image registration [5, 37, 42], it is necessary that landmarks are always recognized at the same positions within images, showing the same object. Based on the landmarks, one use of image registration is to re-align and fuse all the information of the same object of images from different modalities, thus allowing to combine information from different perspectives [51]. Further, distorted images which are taken over time can be registered to see temporal changes of the brain [69]. Localization of landmarks can also be used to eliminate the need for a manual seed point selection in image segmentation algorithms [26, 36], or to generate point and intensity distribution models of shapes and objects, e.g. by active shape [8] or active appearance models [7]. Further, landmarking is often used in object recognition, detection and regression challenges. For example in cephalometric image analysis [48], certain positions at the head or skull are localized. These landmarks can then be used for treatment planning systems [16, 41].

As shown in all these examples, landmark localization is an essential first step for many different subsequent image algorithms. This thesis is focused on a fully automated anatomical landmark localization approach for an automatic age assessment system by extending the approaches in [24, 25] and investigating them more deeply.

1.2 Age Assessment

Anatomical landmark localization is an important initialization step for an automatic age assessment system and a worth-while goal to research on. For that, first certain anatomical points must be located and afterwards surrounding structures must be analyzed since they contain information about the age. This process is illustrated in Fig. 1.1. Examples for such anatomical locations are third molars, hand-bones or joints between them, clavicles,

knees, etc. An estimated age is also known as biological age and is often compared to the chronological age in clinical medicine for different purposes. Some examples are determining growth disorders due to prenatal issues or genetic defects [32], estimation of final heights of young children [32, 72], etc.

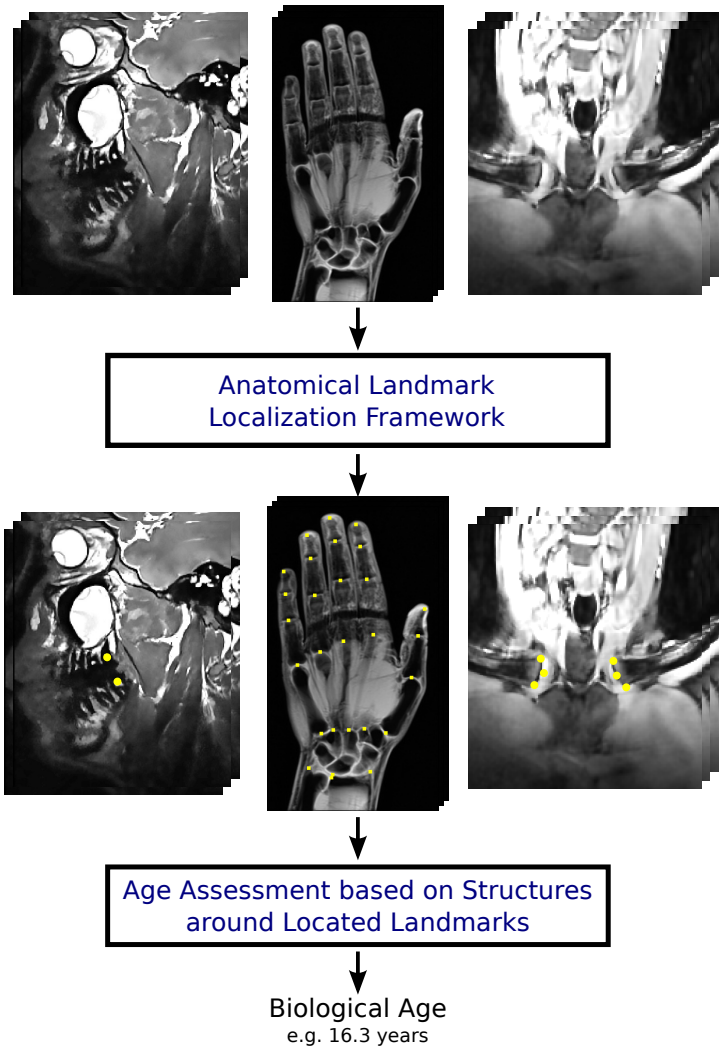


Figure 1.1: Overview of a fully automated multi-factorial age assessment system.

Beside the clinical applications age assessment is also a crucial part in legal medicine, e.g. in sports [50, 75] or when dealing with young asylum seekers [38]. Some persons who escape from their countries due to war or other reasons are not able to identify themselves with passports or identification documents. For that reason, the Ludwig Boltzmann Institute for Clinical Forensic Imaging in Graz performs research on a fully automatic multi-factorial age assessment system using non-invasive and ionizing free *MR* data. *MR* imaging is used since people are not exposed to ionizing radiation, which is the case in

the often used x-ray imaging. Previous and ongoing research of age assessment has been done by the study group on forensic age diagnostics (AGFAD) [64]. They suggest several approaches to assess the age from humans using several parts of the body without having the need for a chronological age [62, 63]. For instance, these parts can be clavicle, teeth or hand. Basically, these structures are exposed to a certain process during maturation. For example, bones underly an ossification process starting from the birth until up to a certain age when the development finishes. A similar behavior is also known for teeth which is based on mineralization, i.e. mineralization of third molars. These processes are described in the next sections in more detail.

1.2.1 Hands

Hands and wrist-bones are often used for age assessment, since the ossification process can be seen very clearly over the time. In 21 tubular and 8 carpal bones in the wrist, the process lets the bones start growing from birth on up to 19 years [32]. Figure 1.2 illustrates on the right a sketch of the main characteristics of the ossification process in long elongated bones, e.g. at fingers or at clavicle bones. While such bones are maturing they consist of diaphysis, metaphysis and ephiphysis whereby metaphysis and epiphysis are separated through a growth plate. During the maturation process, epiphysis and diaphysis start growing accordingly to the ossification process.

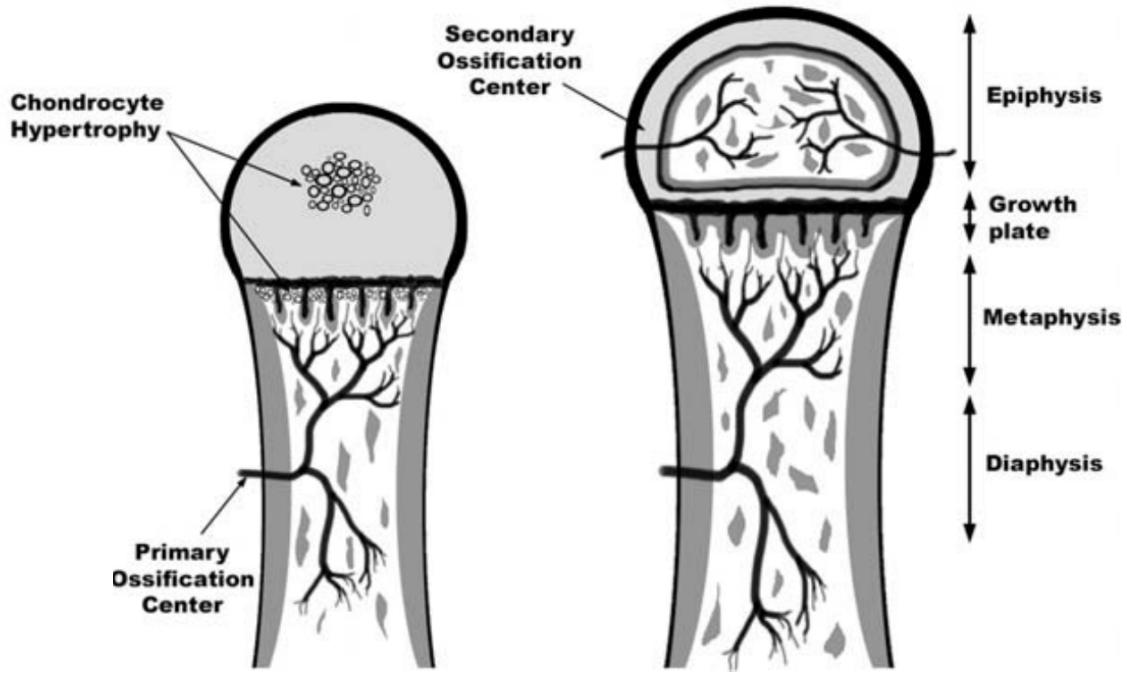


Figure 1.2: Ossification process of elongated bones (Source [32]).

This ossification process is proceeding in the primary and secondary ossification cen-

ter [32]. While the primary ossification center exists also for non-elongated bones like the flat carpals, the secondary ossification center only exists in the cartilage at the ends of longitudinal bones. The diaphysis starts growing from the primary ossification center, as illustrated in Fig. 1.2 on the left. The epiphysis on the other hand appears and grows up to a certain age due to the secondary center, see Fig. 1.2 on the right. Over time bones get larger and some of them will coalesce with other bones.

However, not all bones start and complete this ossification process at the same age. Therefore, a short overview for hand bone growth based on the six different stages according to [32] will be discussed now:

- Stage 1: Up to an age of around 14 months (males) and 10 months (females), age assessment is primarily done by looking at two carpals, namely hamate and capitate and the radius bone. See Fig. 1.3 (a).
- Stage 2: Then, up to 2 years of age for female and 3 years for male the ossification of the epiphysis of four different bones can be used, i.e. of the distal, middle and proximal phalanges and some metacarpals. See Fig. 1.3 (b).
- Stage 3: In the pre-puberty age, up to 7 and 9 years for female and male respectively, the ossification of the epiphysis at the distal, middle and proximal phalanges bones, is very strong. See Fig. 1.3 (c).
- Stage 4: Further, analyzing the epiphysis of distal and middle phalanges can be done up to an age of 13 years (female) and 14 years (male). See Fig. 1.3 (d).
- Stage 5: In the late puberty, bones begin to fuse, i.e. the distal phalanges, metacarpals, proximal phalanges and middle phalanges. See Fig. 1.3 (e).
- Stage 6: After that, fusion of the ulna and radius from females and males up to 17 and 19 years shows differences. See Fig. 1.3 (f).

With hand-bones it is feasible to distinguish biological ages of human up to approximately 18 years. However, many countries apply laws which are based on three age thresholds, namely 14, 18 or 21 years [20]. Hence other bones have to be considered to estimate ages above 18 years, i.e. clavicles.

1.2.2 Clavicles

The clavicle bones are among the last bones finishing ossification. This makes them a good additional source for age assessment [43, 62]. According to the research in [66], five stages can be used to differentiate roughly between different age groups (see Fig. 1.4 for an overview):

- Stage 1: No ossification process is ongoing and therefore the ossification center has not been ossified.

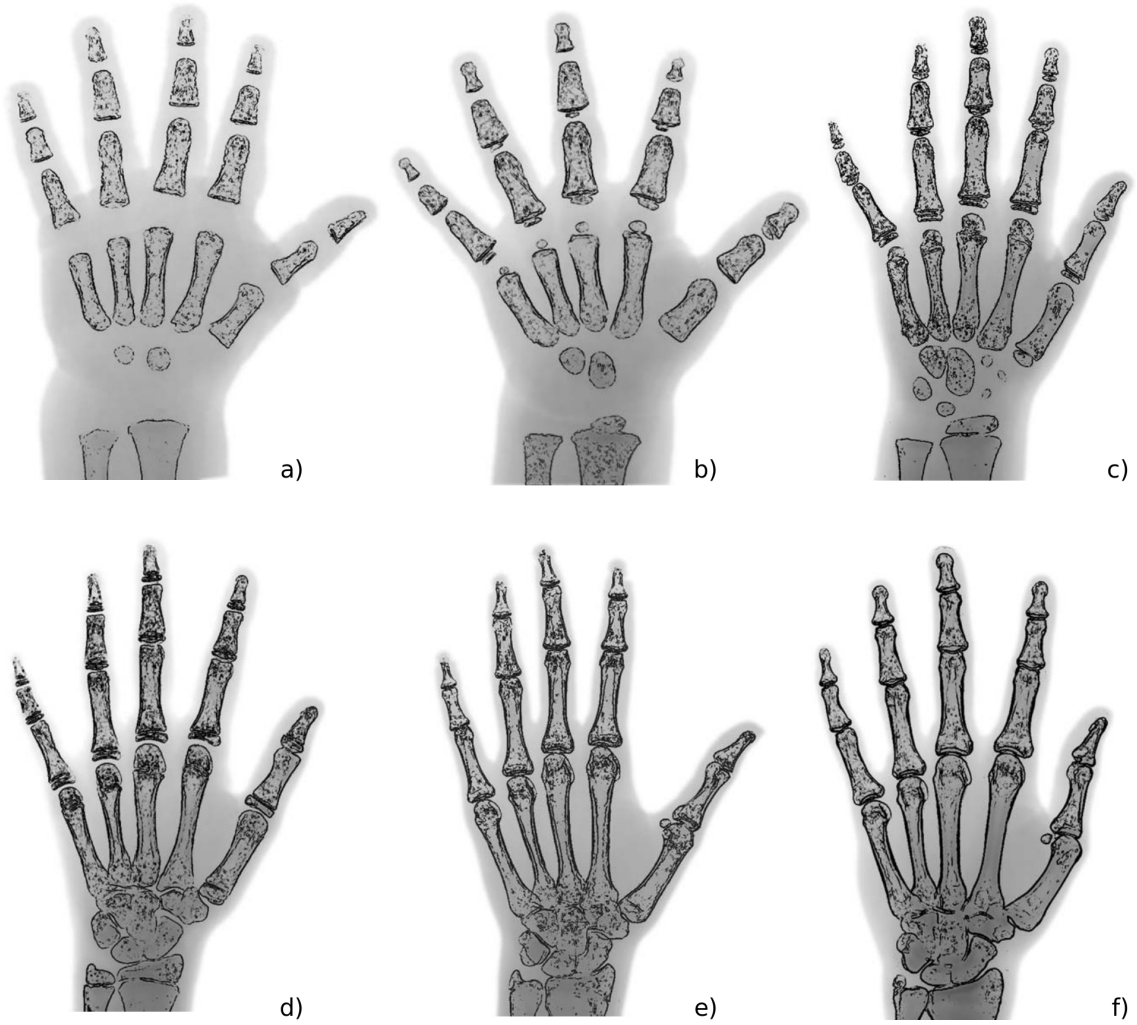


Figure 1.3: Overview of different ossification stages of hand-bones (Source [32]). Stages 1-3 in (a-c) and stages 4-6 in (d-f).

- Stage 2: Ossification starts at the ossification center but no ossification can be seen in the epiphyseal cartilage.
- Stage 3: Partly ossification of the epiphyseal cartilage.
- Stage 4: Full ossification of the epiphyseal cartilage, but scars between cartilage and clavicle are still visible.
- Stage 5: No scars are visible since the epiphyseal cartilage coalesces together with the clavicle.

Combining bones of hands and clavicles leads to a powerful age assessment from which ages up to more than 21 years can be distinguished. However, to achieve a reliable and

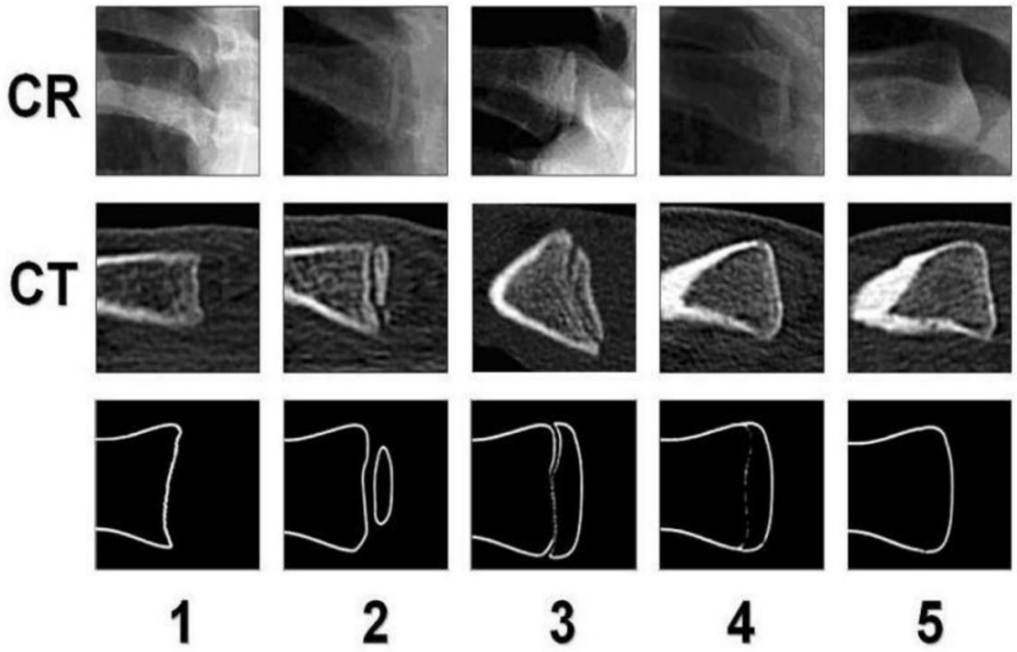


Figure 1.4: Overview of different ossification stages of clavicles (Source [64]).

accurate age assessment, other anatomical structures can be used to support the accuracy of the prediction, i.e. third molars.

1.2.3 Third Molars

Mineralization stages of third molars can be taken to support the age assessment accuracy up to about 20 years [54]. Similar as in the previous examples for hands and clavicles, formation stages of teeth have been described in previous literature. For instance, third molars can be grouped into eight stages (Demirjian) as described next according to [15] and [45]:

- Stage A: Mineralization started but mineralized points have not been fused yet. See Fig. 1.5 (a).
- Stage B: Fusion of the calcified points started, illustrated in Fig. 1.5 (b).
- Stage C: Ongoing process of the dentinal deposit, formation of the crown started and first appearance of a curved pulp chamber. See Fig. 1.5 (c).
- Stage D: The cementoenamel-junction is reached by the crown and a trapezoidal shape of the pulp chamber can be seen, depicted in Fig. 1.5 (d).
- Stage E: The crown is smaller than the root lengths and the radicular-bifurcation is ongoing which can be seen in Fig. 1.5 (e).

- Stage F: Crown reaches the same length as roots which have funnel-shaped ends, illustrated in Fig. 1.5 (f).
- Stage G: Canals of the root are parallel with still opened apical root end. See Fig. 1.5 (g).
- Stage H: Closed root apex, as illustrated in Fig. 1.5 (h).

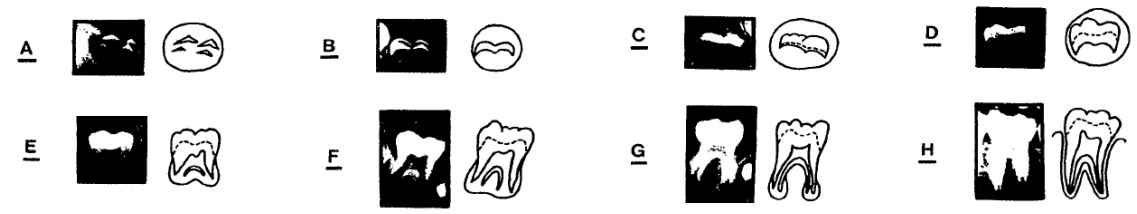


Figure 1.5: Overview of different Demirjian mineralization stages of molars (Source [15]).

1.2.4 Age Assessment

Having defined stages and regions from where the age can be assessed, methods to estimate the age from a radiographic image will be discussed. Most of the existing methods rely basically on atlas matching. There, an object without known age is compared to a set of objects with associated ages. The most similar match in terms of appearance and shape between objects with unknown and known ages, results in an age prediction. For hand and clavicle bones, atlas matching can be done either using the whole object, e.g. images of the hand, or by taking only subparts of the objects, e.g. single bones or structures as previously discussed. Comparing the whole object at once is also referred to as the method of Greulich-Pyle [35], whereas the latter is done by Tanner-Whitehouse [71]. With Greulich-Pyle method an age from a whole hand can be determined directly. In contrast to this the Tanner-Whitehouse method assigns a score to each single structure, i.e. each bone or each tooth. Afterwards, the scores are combined to retrieve a final age assessment. This scoring method has been adopted in [15], such that it can be used also for teeth. However, until now the most research is focused on using x-ray images. Only a few publications exist which involve *MR* techniques to assess the age [22, 23, 70, 73, 75].

Still one question remains: Can the skeletal development and regarding age from different ethnical groups be compared? Many studies are currently considering this question and compare the differences of cultures. For example see [55, 63, 65].

For each of the age assessment applications mentioned above, the localization and comparison based on some methods is up to now mostly done manually by medical experts which have to spend a lot of time doing that for all the different bones and organs. A further challenge of manually estimating age is that the assessed age among medical experts

themselves vary (inter-observer variability) and also the estimation of the same image may vary over time by the same expert (intra-observer variability). Therefore and due to the use in legal and clinical medicine it would be of great benefit to have a fully automatic age assessment system. Such a system could take as input an image of the necessary age assessment structures, i.e. scans of hands, clavicles and teeth/head and as result an assessed age. An important first step for such a system is the localization of anatomical points at the organs and bones.

1.3 Contribution

This work focuses on extending and investigating more deeply the Random Regression Forest (RRF) landmark localization algorithm of [25]. Therefore, a geodesic distance measure for the weighting approach of [25] for *MR* hand images is introduced. This geodesic distance measure lets less varying anatomical structures and simultaneously structures closer to a landmark contribute more to the correct prediction of that landmark. Further, two different feature and voxel-selection strategies are developed which incorporate more directly the appearance at landmarks. For instance, long-distance on-landmark features are introduced to capture the image information at the landmarks directly. This approach use the whole image information whereas in a second method only closer structures to a landmark are used to train an *RRF*.

These approaches are evaluated on three medical datasets which are acquired from young people within an age range of about 13 to 24 years. These are 3D *MR* volumes of the head to localize third molars, upper part of the chest for localization of landmarks at the sternal extremity of the clavicle bones and left hand volumes to localize hand and wrist bones. These datasets will be discussed in more detail later on in chapter 5.

1.4 Outline

Chapter 2 shows a short overview of different related research on landmark localization, starting with general landmark localization methods based on mathematical descriptions up to machine learning techniques which have been strongly researched over the last years. In chapter 3, an introduction of the Random Forest (RF) framework developed in [3] is shown. This algorithm has seen a lot of research interest in the last decades due to its simple structure and rather easy adaption to different datasets. In chapter 4, an *RRF* framework is developed to locate point landmarks built upon the basics of chapter 3. A novel vote-weighting method for the projected 2D *MR* hand images is introduced based on the variation of fingers. Chapter 5 discusses technical details of the proposed *MR* datasets which are evaluated in chapter 6. There, the advantages and disadvantages of the novel ideas for the given datasets are shown and discussed by some experiments. A conclusion and an outlook of future research directions are presented in chapter 7.

2 Localization

Contents

Localization of landmarks in general is a difficult task, both manually and automatically. In the beginning of Medical Image Processing (MIP), anatomical points were often manually selected for further processing which is time consuming. Therefore, several computerized (semi-)automatic approaches for landmark localization have been developed. Such algorithms can be used to process a huge amount of data within a short time. Further, many algorithms need a reliable landmark localization approach as a pre-processing step, e.g. for registration [61], segmentation [49], etc. However, the great variety and change in shape, contrast, color, modalities and other properties make it difficult to develop such an algorithm.

In this chapter a short overview of the most important research in the field of medical image feature extraction and landmark localization is given. Firstly, approaches which work with basic algorithms like corner detectors based on gradients and other processing pipelines, is discussed. Afterwards, landmarking based on registration algorithms is shown. Further, methods based on generative modeling and discriminative learning are going to be discussed. Finally, an overview of combinations of the methods is given, to demonstrate how well such approaches can work together.

2.1 Localization based on Low Level Descriptions

Mathematical descriptions have been developed to automate the process of feature point extraction [59]. Many of them make use of image gradients or other mathematical operators and depend directly on the distribution of intensities in images. Further, geometrical descriptions like curvature extrema, saddle points, crossing of lines and much more have been researched. Algorithms based on such models, may need post- or pre-interaction or refinement steps from a human operator to achieve reliable landmark localizations.

The authors in [58] describe a few corner detection methods for automatically locating landmarks in 3D volumes. For instance, image gradients and Hessian matrices are used to locate certain points in human brain Magnetic Resonance (MR) volumes. A 3D differential corner detector which is one of these methods described in [58] is used as a landmark localization step in a semi-automatic image registration approach in [60]. The corner detector basically finds high variations of intensities. Due to noise and similar structures with high intensity variations this approach leads to multiple false-positive landmark candidates. To filter these candidates and to find a final correct landmark candidate, a human-computer interaction is performed in three steps. First of all, the user selects a region of interest in which next to it the corner detector is applied. At the end the user selects the best landmark candidates for the further image registration process. This results in a decrease of error made by the corner detector and therefore in better registration results. However, the cost of processing interruptions due to the semi-automatic approach is high. In general such mathematical corner detection methods often rely strongly on local appearance information. Little changes in local appearances may result in high variations of the results. The main benefit of such basic corner detectors are their fast computation time. Therefore, such algorithms are still used frequently.

Another more domain specific feature extraction method for age assessment is presented in [56]. In their work, they have given a radiograph database consisting of wrist images from left hands. As a first step these images are pre-processed by first correcting the orientation. Afterwards, the images are thresholded according to the background to achieve a better and easier distinction between foreground (hand) and background. After the preprocessing step, the image is binarized in an automatic way. This results in low values at air or background and high values at phalanges or bones. From these binarized images, first finger tips and then whole fingers are detected. Further processing using an interest region detector yields to an extraction of the metaphyseal and epiphyseal regions within the fingers which can be used to assess an age.

However, such approaches are very specific and cannot be applied easily to more general datasets. Therefore, the need for approaches such as statistical shape models, image registration methods or machine learning algorithms, has become very high. Although such algorithms may also need some manual initialization process, they are more general in their behavior.

2.2 Registration based Localization

One higher level approach for organ localization or segmentation is image registration, i.e. atlas based methods [57, 68]. An atlas is a database consisting of one or more given fixed images with associated labels, e.g. single points or bigger objects like organs. Localization can be achieved by deforming a new unseen (moving) image with unknown position of landmarks to match one or several fixed images with known landmark coordinates. This deformation can be done by different registration algorithms [82].

Such a registration algorithm is a powerful method for localization, but the challenges lie in variations of shape and appearance in images, i.e. fingers in hand images. Handling this big variation can be done by different approaches as listed next:

- Linear or rigid registration
- Non-linear, elastic or non-rigid registration

Linear registration is a deformation type which matches the moving image to the fixed image globally, e.g. by translation, rotation or scale of the whole object. For example, a whole hand will be registered as it is but single fingers stay at the relative position to each other. Non-linear registration on the opposite, warps also finer variations. For instance individual fingers are deformed to match other poses. This yields to more precise results but needs computationally expensive algorithms. In general, many of those image registration algorithms consist of following steps, in which each single step is already a difficult challenge to solve [82]:

1. Initial feature detection in moving and fixed image
2. Feature matching between moving and fixed image
3. Model estimation: One wants to find a transformation T such that a fixed image F matches best to a moving image M. This is done by optimizing a fitting measure or cost-function C:
$$T_{best} = \underset{T}{\operatorname{argmin}}\ C(F, M) \tag{2.1}$$
4. Transformation and re-sampling of the moving image

Having found a transformation T_{best}, landmark or organ positions can be transformed with the same transformation from the fixed to the moving image coordinate frame. This results in a final landmark prediction. Robustness of such atlas based object or landmark localization can further be improved by matching a moving image to multiple fixed atlas images or vice-versa [40]. In such an approach a final landmark or organ prediction can be found by combining the multiple received landmark predictions. For instance, the combination can depend on the transformation-costs.

A slightly different approach of registration based landmark localization is proposed in [27]. Two types of 3D anatomical point landmarks are extracted, namely curvature extrema and saddle structures from the human head. The algorithm is based on registering self-designed 3D deformable models to a normal intensity based image. The models consist of a parametric form of an ellipsoid and a hyperboloid. In an optimization task the model fitting has been formulated using an edge based fitting procedure with respect to the model parameters (rotation, scale and translation). As initialization a semi-automatic approach and an ellipsoid as initial object is used for the models.

2.3 Localization based on Generative and Discriminative Models

Another way of defining the feature point detector and landmark localization task is described in this section. For instance, the probability that a landmark or point is located at a certain position l in an image I, is $p(l|I)$ [47]. Several ways exist to calculate $p(l|I)$:

1. Discriminative type - Classification or Regression: $p(l|I)$ is either directly estimated or regressed by multiple predictions
2. Generative type: $p(l|I)$ is generated from previous data: $p(l|I) \propto p(I|l)$

Ad 1 - Classification: Using the discriminative classification type for landmark or object localization, a classifier is trained to differentiate between classes in an image, i.e. the probability $p(l|I)$ can be directly estimated. Applying such a classifier onto a whole new image yields to a pixel-wise labeling of the image, e.g. binary or multi-class classification or segmentation task. Another way how such a classification task can be seen is to learn a function that maximizes the probability for correct classification or vice-versa, minimizes the misclassification probability [81]. Training of such a classifier to retrieve a good labeling or localization can be a difficult task. For example, training-samples used for training have to be carefully defined for each class.

Ad 1 - Regression: Instead of directly estimating the probability $p(l|I)$ it is also possible to aggregate votes from many landmark predictions. This is done by using the whole or a subset of the image information. For instance, each pixel in an image votes to a probable landmark position for a class. The position which retrieves the most votes can be used as a landmark estimation.

Predicting a certain position can be done by using the information from the whole image (global context information), i.e. anatomical structures from the whole image are able to vote to a landmark. On the other hand this can also be done by using only local information from certain sub-regions of the image (local context information), i.e. anatomical structures within a certain range around a landmark are allowed to predict a

position [24]. Each of these methods has benefits and drawbacks. For instance, using the whole image information can yield to an inaccurate localization result. Pixels and structures farther away from a landmark position may have no knowledge about the variations of the landmark to where they vote. Further, using local image information can lead to multiple landmark predictions if similar repeating patterns occur.

Beside the contextual information another differentiation can be done by marginal-space learning approaches [44, 77, 78]. They were recently researched in addition to the common full-space learning approaches. The main difference between these two approaches is either that all (non-)linear transformations in the images are learned at once or the search-space is constrained and iteratively more parameters are added. For example, first search for best translation and afterwards search iteratively for other parameters, e.g. rotation, scale, etc.

Ad 2 - Generative Models: Landmark prediction using generative models can be done by creating a model based on the statistics of the distribution of already given landmarks from training-images, e.g. shape or appearance models [7, 8]. A learned model can be fitted to new images of the same type by varying parameters such that a criterion is minimized. If the learned shape with known landmarks is correctly fitted to the new image, the landmark positions can directly be read and used for predictions of the new image. However, generative models like Active Shape Model (ASM), Statistical Shape Model (SSM) or also Markov Random Field (MRF) are more-often used to refine an already given landmark prediction from other models, e.g. discriminative models.

2.3.1 Localization based on Statistical and Active Shape Models

A generative landmark prediction model based on *SSM* can be built by first generating a mathematical model. This model describes shape variations between images of the same type, e.g. the movement of fingers or the variations within brains over several people. Therefore, the following steps have to be performed:

1. Remove Similarity Transformation: The landmarks of the shapes of the different training images must be aligned into a common coordinate frame, e.g. using Procrustes analysis [34].

2. Model the distribution which results in a parameterized model:

$$\boldsymbol{x} = \hat{\boldsymbol{x}} + \boldsymbol{P}\ \boldsymbol{b} \tag{2.2}$$

   where $\hat{\boldsymbol{x}}$ is the mean shape, $\boldsymbol{P}$ are eigenvectors and $\boldsymbol{b}$ are the shape model parameters. For instance after applying a Principal Component Analysis (PCA) to the data, each eigenvector reflects a certain variation in the training data, e.g. movement of the thumb. The first eigenvector in $\boldsymbol{P}$ controls the largest variations in the training

data, whereas the last eigenvector controls the smallest variations. To reduce the model complexity, small variations are often dismissed by taking only a subset t of all the eigenvectors:

$$\boldsymbol{x_{appr}} \approx \hat{\boldsymbol{x}} + \boldsymbol{P_t}\ \boldsymbol{b_t} \tag{2.3}$$

The next part is to create an algorithm which varies the modeled point distribution model to fit it to new shapes, i.e. by an *ASM*. Therefore, following steps have to be performed:

1. Set the previously generated *SSM* to an initial position X^i ($i = 0$ at initialization).
2. Search near the previously found or initial position X^i for best new local matches for each data-point in the model. Some examples how this can be done are discussed in the following examples of this section.
3. Update the *SSM* parameters to the newly found best local matches.
4. Repeat previous two steps, until a task-dependent cost function converges.

One of the main challenges of using an *ASM* is the initialization process and defining a good convergence criterion. If the initialization is too inaccurate the *ASM* may converge into a wrong local minimum and therefore in a wrong shape position. Furthermore, if a new shape is not covered by the combination of the training data, it is unlikely that the trained model matches to the new shape. Especially this challenge occurs in medical imaging since it is difficult to create and get a big and good training set.

Further improvements have been developed by incorporating also the change of appearance over training data. This leads to the Active Appearance Model (AAM) [7] in which additionally to the shape the intensity variations are modeled.

2.3.1.1 Example: Cephalometric Image Analysis

One medical application in which *ASMs* are used is cephalometric analysis which deals with the measurement of the head. In [39] relationships between points on major structures in head images have been modeled. To build the *SSM*, points are manually annotated by experts which can be seen in Fig. 2.1. Each of the training datasets are then aligned accordingly to the annotations using Procrustes analysis. Afterwards, a *PCA* is applied to all points of the shape to model the main shape distribution in a lower dimensional space by retaining all large variations.

Having built the *SSM*, an *ASM* approach is used to fit the model to new images. Therefore, first the learned mean shape is overlaid to the new image. Afterwards, a local search around each point of the initial *SSM* is performed to find the next local match which fits best. This is done by finding the strongest edge in the vicinity of each modeled data-point and then calculating the direction vectors to this edge. These vectors are combined with the *PCA* which yields to a new temporary fitted shape model. This

procedure is iteratively applied until a convergence criterion is reached and a final fitted shape is retrieved. See Fig. 2.2 for the initialization of the shape on the left hand side and results after matching the shape model on the right hand side. However, the matching of such a model depends strongly on the initialization of the mean shape, the local search around the shape points and on the variation of the new target image. If one of these steps fails or the new image has a high unusual variation the shape model will converge to a false position and further processing algorithms may fail.

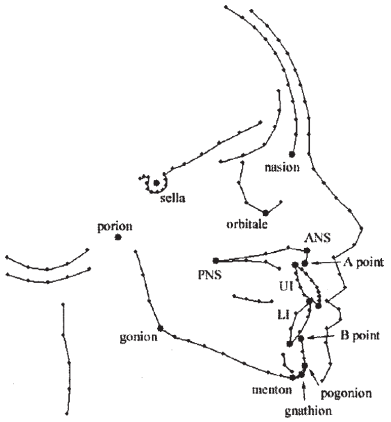


Figure 2.1: Cephalometric shape template to build an *ASM* (Source [39]).

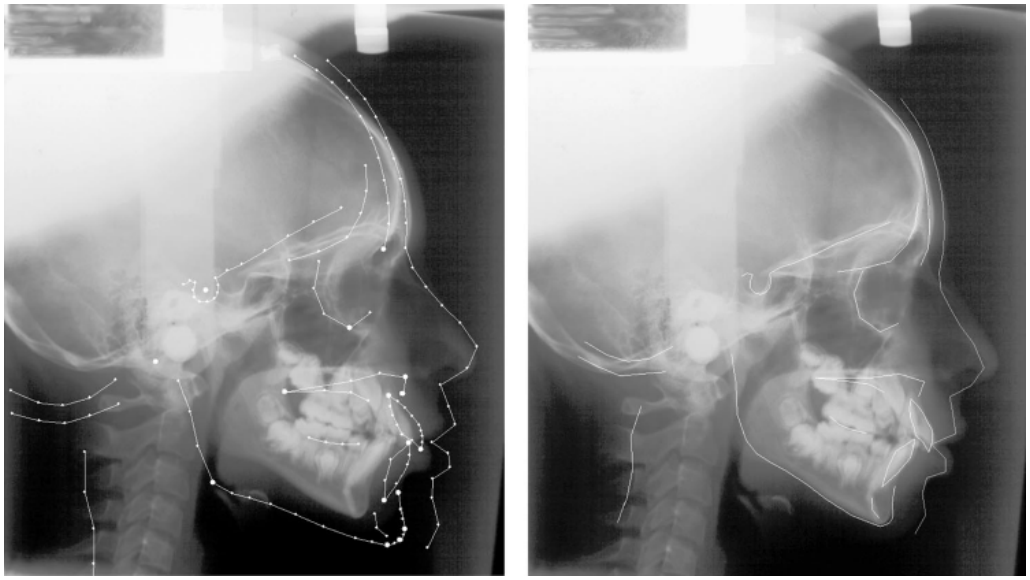


Figure 2.2: Cephalometric shape matching of the previously modeled shape to an unseen image (left) and results after the matching algorithm (right) (Source [39]).

2.3.1.2 Example: Anatomical Hand Landmarking with Top-Down Patch Regression

In [18], statistics of landmarks in hands are incorporated into a landmark localization algorithm which is based on the appearance of images. The algorithm basically consists of a training and testing part.

Training: During the training part a so called regression codebook is generated. This codebook consists of squared sized image patches randomly placed inside a certain range around landmark positions to capture the local appearance. Such patches are generated at different image scales s_0 to s_n which results into multiple codebooks, one for each landmark l and scale s. During patch extraction relative distances from the patch center to all landmark positions which are captured by this patch are stored in the codebook. For example, at scale s_0 (the whole image) all landmarks are captured whereas at the highest scale s_n only one landmark is visible. See Fig. 2.3 for an illustration of the codebook generation. This leads to high dimensional codebooks which are reduced by *PCA* by keeping 90 % of the variance. Since only the appearance around landmarks within a certain range is used it is necessary to capture the global landmark configuration of the hands. This is done by a point distribution model and a lossless *PCA* model.

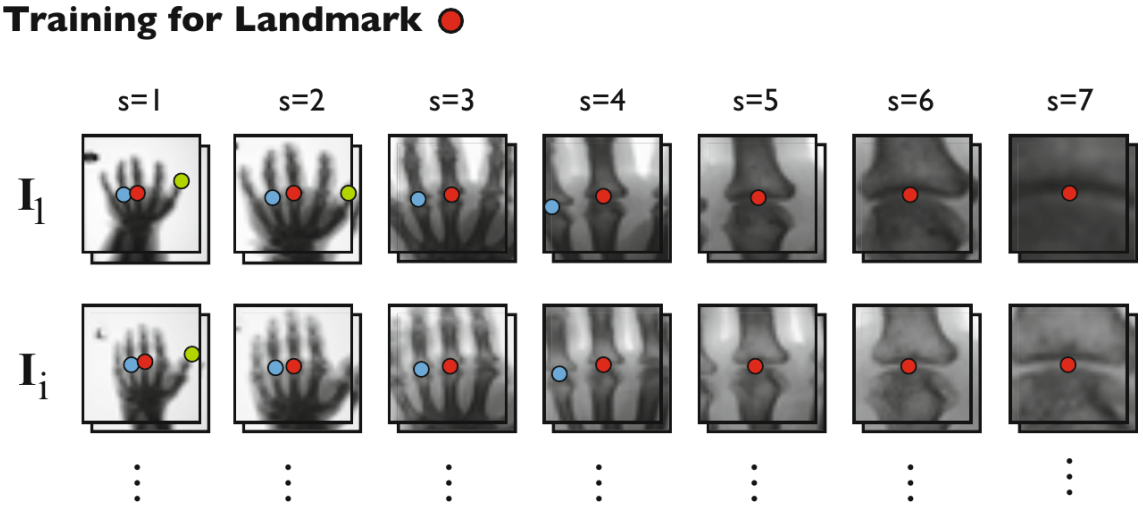


Figure 2.3: Multilevel patch codebook generation example for a single landmark (red): Overview along different scales (Source [18]).

Testing: To find the learned landmarks the image is first analyzed at lowest scale s_0 which corresponds to the whole image. Since no landmarks are known at scale s_0, for each landmark a patch is extracted around the center of the image. Each extracted patch is projected to the *PCA* space. Afterwards, the generated codebook is searched at the

actual scale for the most similar patch using a nearest neighbor search algorithm. The patch comparison is done for each extracted patch of each landmark. This results in multiple landmark predictions for one landmark from different patches. Therefore, a final landmark prediction is generated by calculating the median landmark position from all found predictions. This procedure is done recursively over all scales which results into finer landmark predictions when increasing the search-scale. An overview of the codebook search can be seen in Fig. 2.4. Since similar structures in the hand appear multiple times at different locations, landmarks can converge to wrong positions. Therefore, the learned statistical model is used to regularize landmark candidates at scales s_0 to s_{n-3}. Since at lower scales a single landmark is detected, solely the local appearance information is used.

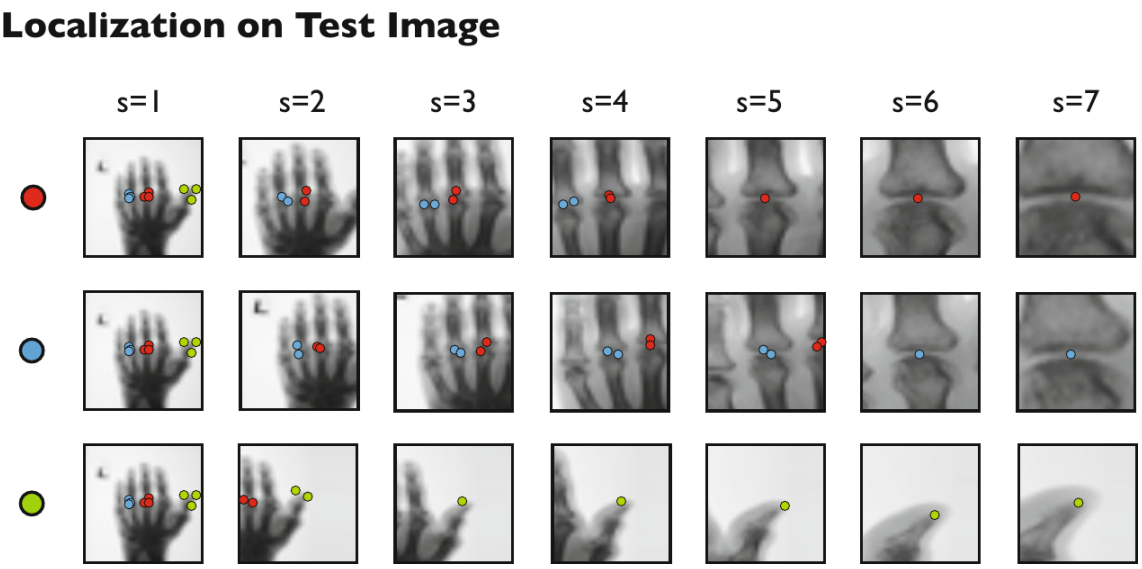


Figure 2.4: Localization of three landmarks (red, blue and green) in a new unseen image using a multilevel patch codebook (Source [18]).

2.3.2 Localization using Markov Random Fields

Another possibility to incorporate the shape information are *MRF*. In recent landmark localization literature, they are often put on the top of a local appearance based landmark localization. There, they are used as a discrete optimization task to regularize multiple found candidate positions. In principle, an *MRF* formulation consists of nodes and undirected edges between the nodes, i.e. an undirected graph. For instance, in terms of landmark regularization, nodes correspond to landmark positions and edges to a connection between two nodes. For each node and each edge potentials or weights are assigned which depend on the performed task. In an optimization task the best combination of all node potentials and edge potentials is searched. However, *MRF* are complex and computationally costly to solve if the undirected graph contains too many edges and nodes. On

the other hand, by carefully defining which connections are modeled, an *MRF* can be a powerful regularization tool. Next, some examples which use *MRFs* will be discussed.

2.3.2.1 Example: Anatomical Hand Landmarking using Local Appearance and Markov Random Field Regularization

Local symmetry interest point and Harris corner detectors are used in [17] to retrieve hand and vertebrae landmark candidate locations. To explore the symmetry attributes in medical hands or vertebrae images, a gradient vector flow field is applied to a gray value image which is used for a symmetry detector. This yields to local responses at strong symmetry based appearances, as illustrated in Fig. 2.5. To find the best landmark configurations, an *MRF* is used. As edge information for the *MRF*, the edge length between nodes and relative angles between edges are used. To incorporate also the retrieved quality of landmark candidates, node potentials are modeled by some measurement between model and target point descriptors.

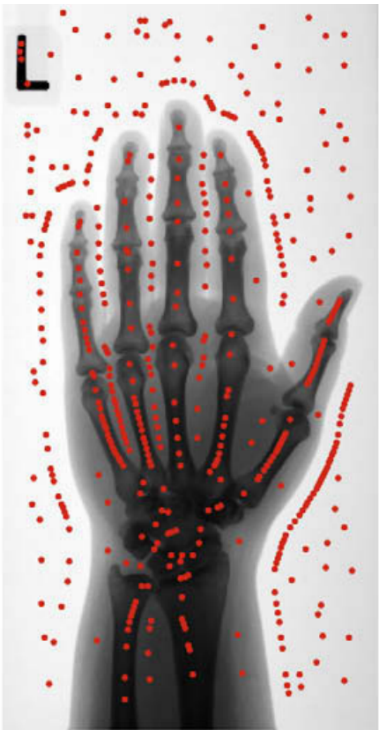


Figure 2.5: Detector responses from the local symmetry based detector (Source [18]).

2.3.3 Localization based on Discriminative Models

Recently, discriminative approaches like Random Forest (RF) [3], boosting [28], etc. have seen a lot of research interest. They are well suited for object classification, localization

and regression tasks which is shown in some examples in this section.

Boosting: In boosting algorithms [28], many weak decision rules are combined such that they form a very strong and accurate one. Training samples are used within a training process which consists of several rounds. In each round, a weak learner or hypothesis is generated according to the training samples at this round. These weak learners are further weighted according to a certain error they make. This error is calculated w.r.t. the used distribution of training samples. Generating and combining many of these weak learners yields to a very strong machine learning framework.

Random Forests: In this work another learning technique called *RF* is used which was originally described in the seminal work of [3]. During training, random sub-samples of the training data are used to train several randomized trees. A tree consists of split, leaf and root split nodes. At each root and split node, the random sub-sampled training data reaching this node is split into a left and right subset according to a feature-test and a threshold. Such a feature-test consists of several parameters which are determined within a node optimization task, e.g. by optimization of a measure based on the information gain. The best combination of parameters is stored in the root or split node and is applied to the data reaching the node. This yields to a feature-response of the data which can be thresholded, thus splitting up the training data into two subsets. The node optimization task is iteratively applied in each split node until one or multiple pre-defined stopping criteria are reached. Examples are a maximal tree depth or a minimum number of training-samples within a node. At this point, a leaf node is created in which a prediction model based on the training data is stored. During a testing stage, new data is pushed through the trained *RF* and the randomly trained decision trees. According to the stored feature-tests and thresholds, they end up in leaf nodes. The previously stored predictions are accumulated from the leaf nodes which lead to a prediction for the testing data.

2.3.3.1 Example: Boosting for Regression Tasks

In [79], an image based boosting algorithm formulated as a regression task is used for three difficult challenges, i.e. localization, age estimation from face images and detection. The boosting approach generates weak learners which analyze and keep only relevant local appearance information in multiple rounds. In each round, a weak learner is weighted by a factor such that a loss-function is minimized. After multiple rounds, one single strong learning framework is created to handle the large variations in the used datasets. The loss function objective of [79] is further optimized in [80] and more sophisticated weak learners are developed using an image based boosting ridge regression algorithm.

2.3.3.2 Example: Tooth Detection using Random Forest Classification

In [6] the authors make use of classification *RF* to automatically detect dental landmarks in 3D Cone Beam Computed Tomography (CBCT) volumes. An example volume is illustrated in Fig. 2.6. They observed that the dental landmarks are roughly located at the same absolute position in their volumes. This prior knowledge is used to train and apply an *RF* classifier only at this pre-calculated region, which is illustrated in Fig. 2.7.

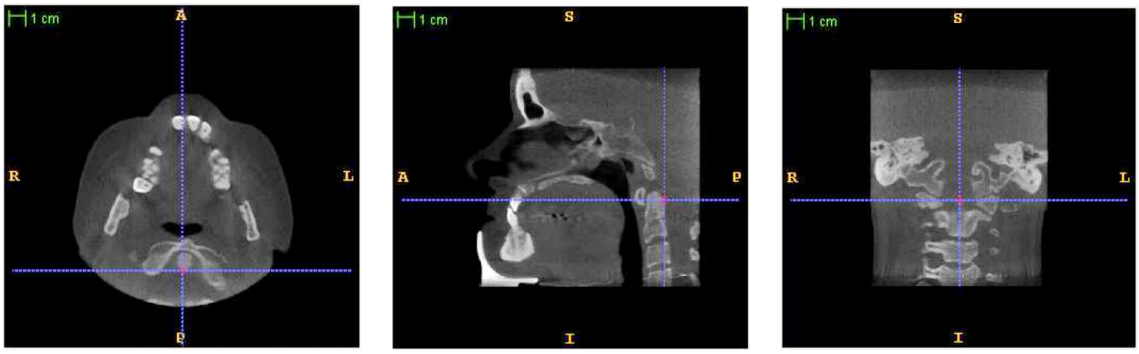


Figure 2.6: Head volumes for dental landmark localization from different field of views (Source [6]). Left: axial, middle: sagittal and right coronal view.

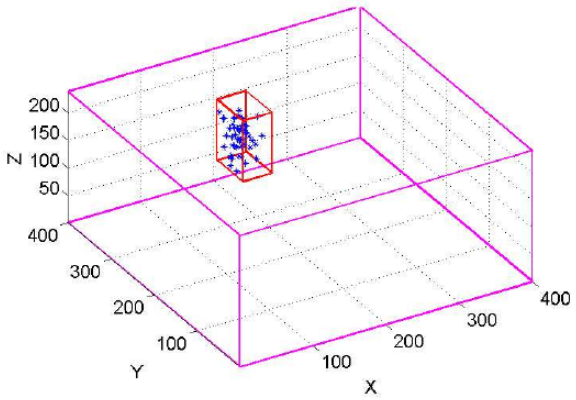


Figure 2.7: Absolute search position constraint for dental landmark detection in [6] (Source [6]).

To train the *RF* classifier, positive training samples within a certain range around a given landmark position are chosen, i.e. voxels inside a range smaller than 2 mm around a landmark. Voxels which are farther away than 4 mm from landmarks are labeled as negative training instances. Afterwards, an *RF* is trained with these labels. To generate a weak learner in split nodes, they have used gradient and intensity features based on a random number and position of neighboring voxels around a currently processed voxel.

Localization of teeth in a new unseen image I is done by classifying all voxels within the

previously specified restricted area where the teeth probably lie. This is done by pushing these voxels through the forest, which end up in several leaf nodes. From these leaf nodes, the learned probabilities to which class a voxel v belongs is read out. Averaging over all predictions from each tree per voxel, yields then to a final result, i.e. to which class/tooth c a voxel v belongs to: $p(c|I, v)$. However, such probabilities must be thresholded to get only responses at teeth locations which is a quite difficult challenge. Averaging over all positions of the final remaining responses, leads to a final landmark position.

The authors compared their *RF* method with a boosting framework and retrieved better results in average, minimal and maximal Euclidean distance using the *RF* approach. However, choosing a region as positive instances in the training process yields to an imprecise localization of the teeth, due to defining multiple positive landmark candidates for one landmark.

2.3.3.3 Example: Organ and Vertebrae Localization using Random Forest Classification

An *RF* classification approach for center localization of organs is proposed in [11]. For each voxel a label is assigned to the training data which yields to a multi-class *RF* approach, i.e. labels for heart, left eye, left kidney, etc. This is done similar as in [6], by using voxels in a small range around ground truth annotation as positive samples and farther away voxels as negative/background instances. Further, different labels are used for the positive samples from different organs. To capture the whole image information, context-rich long-distance Haar-like features relative to a voxels position are used. For instance, one or two cuboid boxes with a random size and placed at a random distance relative to a training voxels position are used to create feature responses for training voxels. This is done by calculating the mean intensity or gradient values of voxels which lie inside such a random cuboid feature box.

Locating organs in a new unseen image is done by pushing each voxel of the image through the *RF*. This results in voxel-wise probabilities for a certain class which are thresholded. Afterwards, remaining voxel positions for a class are averaged which yields to a final landmark prediction for a certain organ. With such a classification task, also the presence of an organ can be detected. For their dataset this is helpful since images are cropped at different positions, thus a various number of organs occur in the images.

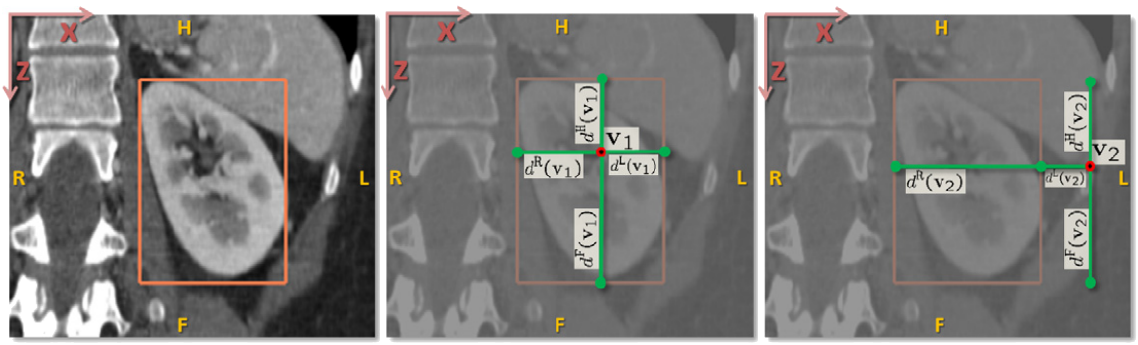
An *RF* classification approach is also used for vertebrae localization in pathological spine Computed Tomography (CT) scans in [33] by additionally developing a sparse to dense labeling approach.

2.3.3.4 Example: Random Forest Regression for Organ Localization

The same authors of [11] circumvent the problem of voxel-wise labeling organs by using an *RF* regression approach in [9]. Firstly, bounding boxes around different organs are created, each of them aligned on the axis $[X, Y, Z]$ as illustrated in Fig. 2.8 (a). Borders of

the bounding boxes of one organ are treated as a continuous 6-dimensional vector which depicts the absolute position in a volume. During training, all voxels from a volume are then pushed through trees in which split nodes are optimized according to the same features as in [11]. Basically, voxels with similar local appearance are grouped by features and a variance minimizing information gain criterion, until they end up in leaf nodes. According to positions of voxels reaching the same leaf, continuous relative distances to all bounding box borders are stored in the leaf nodes.

When testing a new unseen volume, voxels are pushed through the trained *RF* which end up in leaf nodes. All relative distances to the bounding boxes of all organs over all trees are read out and an average bounding box position for each organ is calculated. A schematic overview of the voting of two voxels v_1 and v_2 to the bounding boxes, can be seen in Fig. 2.8 (b-c).



(a) **(b)** **(c)**

Figure 2.8: Bounding boxes around an organ (a) and example voting from voxels v_1 and v_2 to the axis aligned bounding boxes (b and c) (Source [9]).

However, since organs vary a lot in their shape and mostly have no cuboid shape, bounding boxes for organ localization lead only to a rough detection. Further, also organs which are absent, will be located with this approach.

2.3.3.5 Example: Point Landmark Localization using a Two-Step Random Regression Forest Framework

The authors of [25] improved the previous bounding-box approach of [9] by two ideas. Localizing single point landmarks at hand-bones in 3D *MR* volumes and additionally training and applying a second Random Regression Forest (RRF) to get more accurate results, as illustrated in Fig. 2.9.

The first *RRF* uses context-rich and long-distance features, similar as in [9]. These features are combined with a variance minimizing information gain criterion to group similar voxels in their appearance and position in the volume. From voxels $\boldsymbol{v_i} = [v_{\{x,i\}}, v_{\{y,i\}}, v_{\{z,i\}}]$

which reach the same leaf nodes, relative distances $\boldsymbol{d}_c(\boldsymbol{v}) = \boldsymbol{l_c} - \boldsymbol{v_i}$ to each landmark position $\boldsymbol{l_c} = [l_{\{x,c\}}, l_{\{y,c\}}, l_{\{z,c\}}]$ per class c are calculated. These relative distances are stored in the leaf nodes as continuous histograms. During the testing part, histograms from leaf nodes in which testing voxels end up are used to retrieve a final landmark estimate by accumulating over all predictions.

In a first *RRF*, the authors use voxels from a whole volume to estimate all landmarks. Since voxels may only vote very uncertain to a landmark farther away, they added a weighting term which penalizes long distance votes. To improve on the localization performance, a more locally second *RRF* is trained restricted to voxels from estimated landmark positions from the first *RRF*.

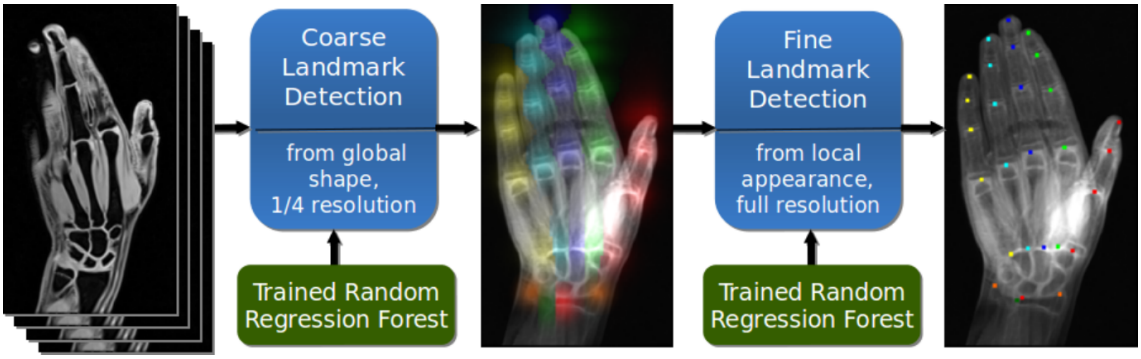


Figure 2.9: Overview of the *RRF* approach for localizing hand-bone point landmarks (Source [25]).

However, although they achieve good localization results in precision of accuracy, two challenges occur: a) estimated landmarks switch between locations with similar shape, which is due to an inaccurate first *RRF* and b), voxels farther away are not able to capture the information to where they vote. Because of the latter it can happen that the first forest predicts landmarks in background/air locations. A locally trained second *RRF* can therefore easily converge to completely meaningless wrong landmark positions or can even switch to similar but false landmark locations.

2.4 Combining Localization Approaches

As previously shown, several approaches are developed for landmark or bounding box localization tasks, all of them with benefits and drawbacks. Researchers also try to combine the benefits of multiple approaches to handle the drawbacks of each one. In this chapter a few of such combinations will be discussed using some examples.

2.4.1 Example: Automatic Teeth Detection and Classification with Shape Models and Support Vector Machine

In [21], a teeth detection and classification pipeline is developed as depicted in Fig. 2.10. Firstly, the maxillary bone is segmented by a combination of automatic bone thresholding and an *SSM*. See Fig. 2.10(a-c) (yellow segmentation and contour lines). Having retrieved the position of the maxillary bone, a predefined teeth region model is placed on the bottom of this segmentation. This teeth region model is partitioned by 15 separation planes by iteratively solving a cost function, see Fig. 2.10(c). At the end, each subregion should consist of one or no tooth which is classified by a Support Vector Machine (SVM) [74]. By placing such a separated subregion to the teeth, it is feasible to label and extract each tooth.

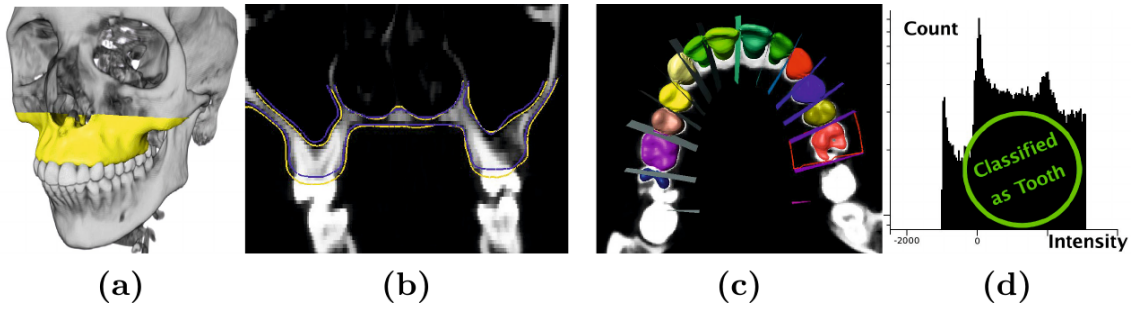


Figure 2.10: Image processing pipeline for tooth classification and detection (Source [21]).

2.4.2 Example: Kidney Localization and Segmentation based on Random Forests and Deformable Models

In [13], bounding boxes around kidneys are detected by a similar approach as in [11]. Additionally, a second step is added in which for each kidney one local forest is trained as a refinement step for centering the bounding boxes more precisely, as illustrated in Fig. 2.11.

After kidney bounding boxes are located, an *RF* classifier is trained to assign to each voxel in the bounding boxes a probability of kidney appearance. These probabilities are then used for a template deformation algorithm similar as in [52], to segment the kidneys based on elliptical shape deformations.

2.4.3 Example: Shape Model Matching using Random Forest Regression

In [47], a multi-step regression *RRF* combined with an *SSM* as regularization step is used to locate landmarks for different datasets.

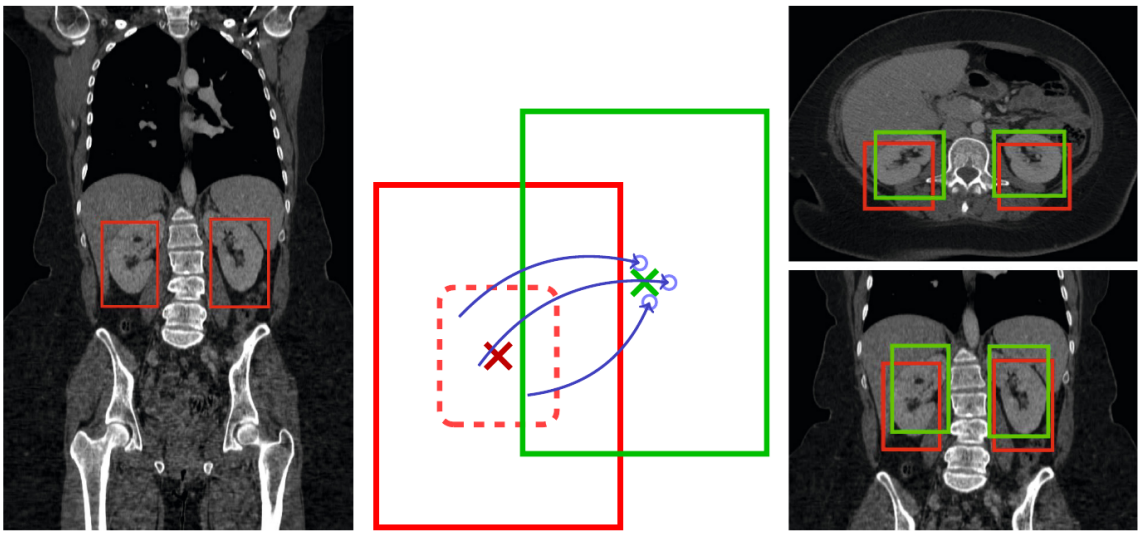


Figure 2.11: Kidney localization: (a) rough bounding box localization, (b) bounding box center refinement step, and (c) results of rough (red) and refined (green) bounding box localizations (Source [13]).

To locate one landmark, a single *RRF* is trained similar to Hough-Forests (HFs) [29, 30]. The main contribution of *HFs* is that multiple extracted image patches from the whole image votes for a probable object localization. For instance, different parts of one object votes for the center of the whole object in the image. The summation of all votes of all extracted small image patches are accumulated in a Hough image in which the maxima indicates the most probable position of an object. Further, in the original *HF* work two information gain criteria are used for splitting image patches after node optimization. One information gain consists of the purity of class-labels, i.e. split among foreground classes and background class. The second information gain splits among different uncertainties of relative localizations between data-points and landmarks.

The information gain used in [47] tries to group similar feature patches according to their appearance and relative distances to landmarks. Testing a new unseen image, patches are randomly selected from the images and pushed through the forest which end up in certain leaf nodes. The landmark predictions stored in the leaf nodes are then accumulated in a 2D histogram. An example for facial landmark localization is illustrated in Fig. 2.12.

However, the locally trained forest can yield to multiple responses for the same landmark in the whole image, as illustrated in Fig. 2.13. Therefore, a point distribution model is incorporated as a refinement step, i.e. *SSM*. The algorithm is described next for hand radiographs, which is similar to the proposed hand datasets in chapter 5 in this work:

1. Locate points at the palm (four) and at each base at the fingers (five), using the *HF*.

2. Initialize the pose of a previously generated statistical model, based on the nine located hand points.

3. Apply another trained *HF* for each landmark to voxels around the landmarks of the initialized shape.

4. Fit the initialized model to the newly found responses, by using an iterative constrained local model matching [12].

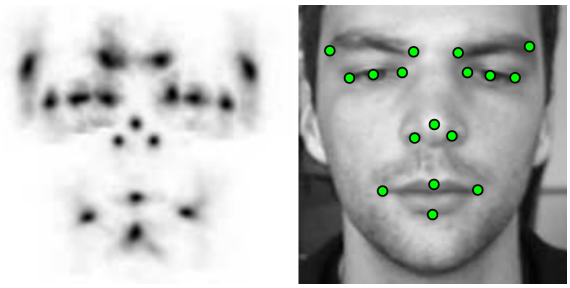


Figure 2.12: Exemplary *HFs* results: 2D image histogram on the left (all landmark predictions overlaid) and the according 17-point model on the right (Source [47]).

To show the generality of this *RF-SSM* combination, they evaluated the same approach, to locate points in cephalometric images [48] and to segment the proximal femure in radiographs of the pelvic [49].

2.4.4 Example: Localization of Anatomical Objects using Random Forests and Discrete Optimization

Another interesting idea is presented in [19]. A multi-step random forest approach is performed, followed by an *MRF* discrete optimization step as a regularization part to handle multiple landmark candidates for one landmark.

Firstly, an *RF* classifier with extremely randomized trees [31] is trained, using voxels in a small range around each landmark. Therefore, these voxels are labeled according to the landmark which should be classified, i.e. $[1, 2, ..., L]$. Additionally, random background voxels with different label $L + 1$ are added. Since such a classifier yields to inaccurate localization results, one *RRF* per each landmark is trained as a local refinement step using voxels in a small range around the landmark.

After training, landmarks are obtained by classifying voxels with the *RF* which results in $L + 1$ probability maps, one for each class (incl. background class). These probability maps are thresholded. Afterwards, the voxels at the positions from the remaining high probabilities are sent through the *RRF* according their associated class from the classification forest. Thus a more accurate localization is retrieved. Since the classifier and the

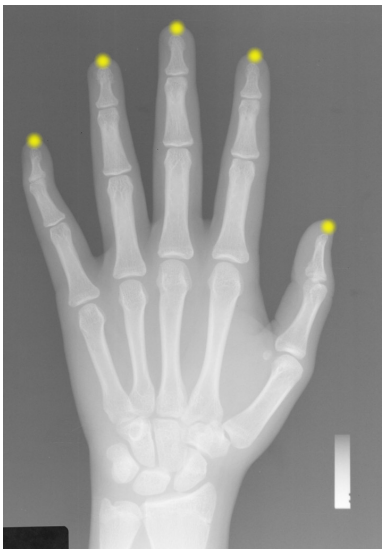


Figure 2.13: Example of multi-responses from the *HF*. A *HF* is trained on local appearance of one finger tip. Applying this forest to the whole image results in multiple responses at each finger tip due to similar local structures.

regressor are trained very locally, multiple responses occurred on similar locations which are regularized by an *MRF*.

For this reason, first a non-maxima suppression is performed which yields to one or multiple landmark candidate positions per each landmark. Next, an *MRF* is used as a discrete optimizer which tries to find the optimal configuration over all landmark candidates by minimizing a cost-function $\mathcal{C}$:

$$\mathcal{C}(M) = \sum_{\forall l \in L} \mathcal{P}(l, M(l)) + \sum_{\forall e \in E} \mathcal{G}(e, M(e)) \tag{2.4}$$

Therefore, a graph topology, unary terms $\mathcal{P}$ and binary terms $\mathcal{G}$ are modeled. The graph topology models the dependencies between landmarks. For instance, landmarks which are near to each other might be more associated than landmarks placed farther away to each other. On the other hand, the unary term contains the confidence of a landmark candidate position, i.e. the probability obtained from the *HF*. Binary terms $\mathcal{G}$ are used to model the weights among the graph topology. In [19] an automatic approach is developed which derived a topology as depicted in Fig. 2.14.

2.5 Conclusion

Many approaches for landmark localization have been developed. Methods which work on interest point detectors or mathematical and geometrical descriptions seem to be fast, but often result in many false positive predictions which have to be post-processed by heuristic

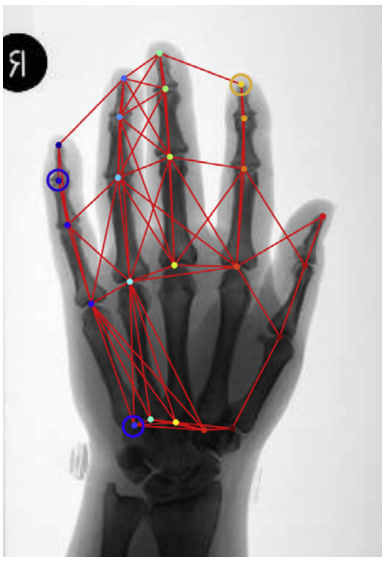


Figure 2.14: Automatically derived *MRF* graph topology (Source [19]).

methods. However, dealing with a huge number of false positives, the computation time may be increased drastically.

Next, more sophisticated methods which work on whole images have been developed, i.e. image registration methods. These methods are often used in medical images to register scans of different image modalities but can also be used for landmark localization. However, for matching a moving image to a fixed image, the appearance and shape of the moving image has to be similar to the fixed image. If the difference is too large, the warping process of image registration may fail. Also a good initialization is an important key to achieve a reliable registration result, but is challenging. This must be done manually or by some alternative methods.

In recent years, approaches which incorporate knowledge about previously seen training images are often used, like boosting or *RF*. One aim of such algorithms is to learn, which features or characteristics can be used to locate landmarks correctly. However, the lack of data in medical images makes it challenging to cover all different variations in pose and appearance. To support such machine-learning algorithms, different statistical models are often used as a regularization step to improve the accuracy, i.e. *SSM* or *MRF*.

3 Random Forests

Contents

In the last decades, machine learning methods have seen a lot of research interest since they can be rather easily adapted to different challenges. One of these machine-learning algorithms are Random Forests (RFs) [3]. As listed in the previous literature review, *RFs* have been successfully applied to different landmark localization challenges.

An *RF* builds up on simpler structures, i.e. decision trees. Decision trees are inspired by Classification and Regression Trees (CARTs) of [4] in which the authors developed the concepts of supervised decision trees and their applications in regression and classification tasks. For example, the aim of a classification task is to assign an input variable a *discrete* label, e.g. binary and multi-class categorization. Regression on the other hand is used to predict a *continuous* output variable for a given input, e.g. a probability or a distance. For these tasks the input variables on which is learned have an associated label. This is known as supervised learning. In contrast to supervised learning, other learning techniques like semi-supervised or unsupervised learning exist in which given input variables do not necessarily have an associated known label. However, this work is based on supervised learning using *RFs*.

Therefore, a short overview of building a supervised *RF* is given in the next sections based on the work and structure in [10] and [24]. Firstly, the basic idea of decision trees is introduced followed by adding randomization techniques and ensembles of trees which lead to *RFs*. Finally, this knowledge is used to build a more sophisticated point landmark localization task with Random Regression Forests (RRFs) in chapter 4.

3.1 Decision Trees

A decision tree is the basic element of an RF. It is structured as a binary graph illustrated in Fig. 3.1 in which progressively simple feature-tests are used to solve a highly complex problem. These feature-tests are generated and stored starting at a root node in one decision tree. Applying the feature-tests to the incoming input variables which are also called data-points from the complex problem yield to feature-responses for each of the data-points. According to a binary decision from the feature-test the responses are thresholded, thus splitting up the incoming data-points into a left and right sub-set which are forwarded to a left or right child node. Applying iteratively several feature-tests in split nodes along a path through the decision tree, small subsets of data-points end up in different leaf nodes in which an estimation is stored, i.e. a leaf prediction model based on the associated labels to the data-points.

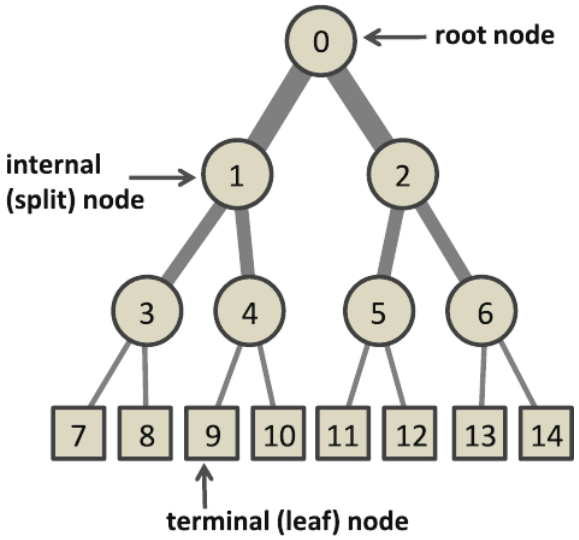


Figure 3.1: Basic binary decision tree structure (Source [10]).

3.1.1 Example

As an example in Fig. 3.2 a decision tree is built to decide whether a picture is taken in an outdoor or an indoor place [10]. Therefore, several simpler questions (feature-tests) are applied iteratively at the root and several split nodes to solve the complex one. According to the responses from the feature-tests, the complex question is forwarded either to a left or right child node. Applying iteratively several feature-tests leads to an increase in confidence of a correct answer. After a few depth levels the complex problem reaches a leaf node in which an answer is stored according to previously learned similar problems during the training process.

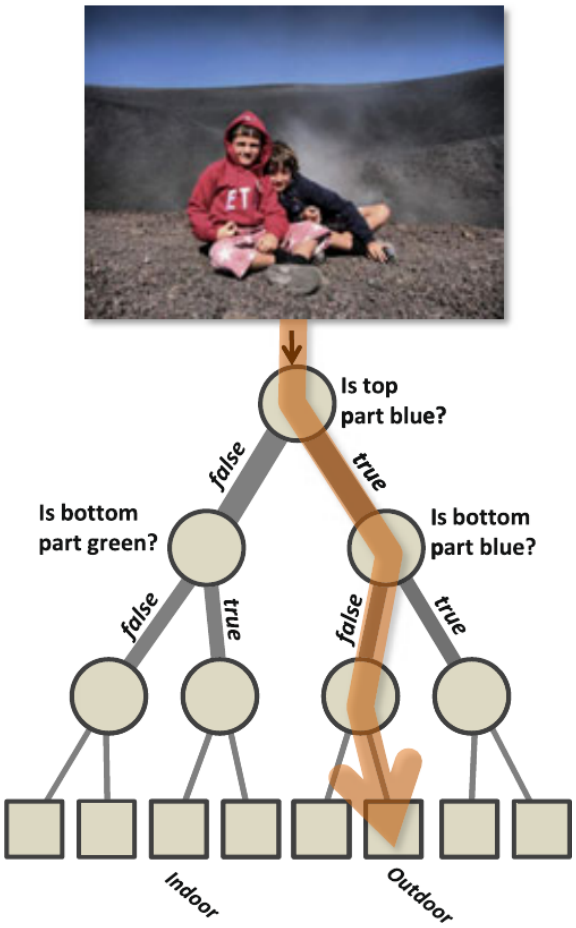


Figure 3.2: Example of solving the highly complex question: *"Is the picture acquired indoor or outdoor"* using a hand-crafted decision tree (Source [10]).

3.1.2 Training

The training process consists of creating feature-tests in root- and split nodes which splits training data-points also called training samples along the depth of the decision tree. At a certain depth or at a defined stopping criterion, training samples reach leaf nodes according to the previously applied feature-tests. There, a final prediction from the labeled training samples is stored. For instance in Fig. 3.2, in the leaf nodes a class c which belongs to indoor or outdoor place is stored. Creating feature-tests can be done manually as depicted in the example in Fig. 3.2.

However, since decision trees are mostly used in highly complex problems, feature-tests can be generated automatically using a set of training samples and associated prediction-labels. This training-set is denoted as $\mathcal{S}$ and consists of several data-points or training samples $\boldsymbol{v}$ with a varying number of features. Examples for such features in computer vision applications are pixel color, position of a pixel or the gradient value of it.

For each root- and split node several feature-tests $h(\boldsymbol{v}, \boldsymbol{\Phi})$ with split-parameters $\boldsymbol{\Phi}$ from a parameter set $\mathcal{G}$ are generated. The split-parameters consist of a feature-response function $f_r(\boldsymbol{v}, \boldsymbol{\theta})$, feature-parameters $\boldsymbol{\theta}$ used in f_r and a threshold τ. For each feature-response function a response for a data-point is calculated. According to the thresholded τ the data-points are split into a left and right subset according to following formulation:

$$h(\boldsymbol{v}, \boldsymbol{\Phi}) = [\tau < f_r(\boldsymbol{v}, \boldsymbol{\theta})] \tag{3.1}$$

In each split node j one *best* feature-test with its split-parameters is chosen which *best* splits up the incoming set $\mathcal{S}_j$ of data-points. These parameters are obtained using following node optimization formulation in which an information gain measure is optimized:

$$\boldsymbol{\Phi}_j = \{\boldsymbol{\theta}_j, f_r, \tau_j\} = \underset{\boldsymbol{\Phi}_j \in \mathcal{G}}{\operatorname{argmax}} \mathcal{I}(S_j, \boldsymbol{\Phi}_j) \tag{3.2}$$

Since in each split node a different subset of all training samples arrive, different *best* split-parameters are learned in each node. This splitting procedure is iteratively applied along all split nodes in all depth levels of a tree. After reaching a certain stopping criterion instead of a split node, a leaf node is created. Typical examples for such stopping criteria are a maximum depth of a tree or a minimum number of training samples within a split node. At such a leaf node, a prediction $p(\cdot|\boldsymbol{v})$ from the labeled data-points which reach this leaf node can be stored and used in a testing-stage. The structure of the label $\cdot$ depends on the type of the task, e.g. classification (discrete, class c) or regression (continuous, multivariate variable $\boldsymbol{y}$).

3.1.3 Testing

Predicting an unknown label for a new testing-set from the same type as the training-set is done by pushing testing data-points $\boldsymbol{v}$ from the testing-set $\mathcal{S}$ through the hierarchically trained tree. Starting from the root node and further in each split node j the previously learned best feature-tests $h(\boldsymbol{v}, \boldsymbol{\Phi}_j)$ are applied to the data-points reaching that node according to equation 3.1. Thus each data-point is sent either to the left or to the right sub-tree. At a certain depth a data-point reaches a leaf node from which the previously stored prediction model $p(\cdot|\boldsymbol{v})$ can be used for a prediction. Since one prediction might be very uncertain a final prediction can be made by combining all the predictions from all data-points, e.g. by summation of all predictions.

3.2 Random Decision Trees

To improve the generalization and the computational performance of decision trees, randomness can be injected during the node optimization task. This is done by using a subset $\mathcal{G}_j$ of the whole parameter set $\mathcal{G}$ for a split node j, also known as randomized node optimization [10]:

$$\mathbf{\Phi}_j = \{\boldsymbol{\theta}_j, f_r, \tau\} = \underset{\mathbf{\Phi}_j \in \mathcal{G}_j}{\operatorname{argmax}} \mathcal{I}(S_j, \mathbf{\Phi}_j) \tag{3.3}$$

3.3 Random Forests: Ensemble of Random Decision Trees

A tree which is trained on all given training data-points is likely to be over-trained thus the tree might lead to over-fitting. As a consequence the learned model does not generalize well to new data. A countermeasure against this behavior is bootstrap-aggregating or shortly bagging [2]. With this approach multiple random decision trees are trained, each with a random subset of the data which helps to decrease the variance and correlation among the trees. The remaining step is the task of combining several leaf-prediction models $p_t(\cdot|\boldsymbol{v})$ from the different learned trees t. This can be done by several approaches, e.g. averaging over all leaf prediction models, taking the median result or multiplying all results according to some normalization factor. This ensemble technique is called *RFs*.

3.4 Conclusion

This chapter has discussed the principles of supervised *RFs*, that are decision trees, randomized node optimization and ensembles of multiple random decision trees. In the next chapter 4 these basics are further developed to a point localization method. Therefore, an *RRF* is used to predict continuous distances to obtain landmark positions based on the appearance information from images.

4 Random Regression Forests

Contents

In this chapter Random Forests (RFs) are adapted to supervised Random Regression Forests (RRFs) which predict continuous outputs. First the basics of *RRFs* are discussed. Afterwards a continuous formulation for anatomical point localization is developed based on the prior works of [10] and [24]. Further, novel ideas to improve the localization performance are introduced, i.e. a novel type of vote weighting and a novel feature-generation approach.

4.1 Random Regression Forests

RRFs are used to predict non-linear continuous models $p(\boldsymbol{y}|\boldsymbol{v})$, i.e. a continuous multi-variate output $\boldsymbol{y}$ which depends on the multi-variate input-variables also called data-points $\boldsymbol{v}$. *RRFs* are ensembles of multiple random regression trees. Each regression tree estimates a complex non-linear model by combining several simpler models, i.e. linear predictions models [9]. An example of the training and testing stage of an *RRF* is illustrated in Fig. 4.1. For the training-stage input training data-points $\boldsymbol{v}$ (x-coordinates) and associated continuous outputs $\boldsymbol{y}$ (y-coordinates; black dots) are given, illustrated in Fig. 4.1(a). An *RRF* is used to describe their distribution by ensembles of simple linear models in the node optimization tasks. Figures 4.1(b-d) illustrate how the *RRF* predicts the continuous outputs of the training data-points over several depths. Ensembles of simpler models are fitted through the training data-points describing their distribution better with increasing depth. Afterwards, this trained *RRF* can be used to predict a continuous output variable

for a new input data-point. For instance in Fig. 4.1(a), an unknown y-coordinate should be predicted for a new given input x-coordinate (red dot).

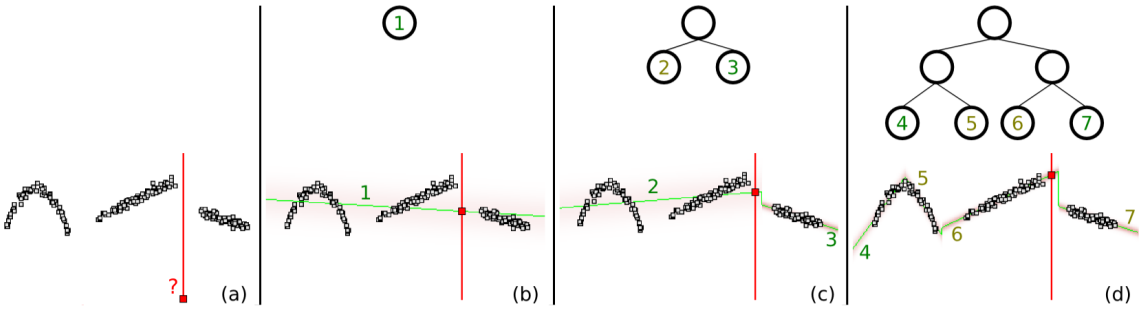


Figure 4.1: An illustration of how simpler models are combined to solve a non-linear regression problem using *RRFs*. (a) shows training data-points (x-coordinates) with associated continuous outputs (y-coordinates) which are marked as black dots. The non-linear model is approximated by simpler models with increasing depth in (b-d). The aim of creating this prediction model is to estimate a y-coordinate from a new unseen testing data-point at a certain x-coordinate (red dot in (a)). Results from the estimation of the red dot y-coordinate over depth can be seen along (b-d).

4.2 Localization of Anatomical Landmarks

By defining the non-linear prediction models as continuous distances, *RRFs* can be used to locate anatomical point landmarks using image voxels as data-points, as illustrated in Fig. 4.2. Therefore, the distances are defined as the relative distance between a voxel $\boldsymbol{v} = v_{\{x,y,z\}}$ and a landmark position $\boldsymbol{l} = l_{\{x,y,z\}}$ in each axis separately, similar as in [24]:

$$\boldsymbol{d} = \boldsymbol{v} - \boldsymbol{l} \tag{4.1}$$

During training an *RRF*, voxels from multiple training volumes with similar local appearance and distance to a landmark should be grouped together. This is done by splitting the voxels into left and right subsets in each split-node according to good feature-tests. The feature-tests are based on the local appearance information around the position of voxels. Thus similar voxels from the whole training set end up in the same leaf nodes in which the relative distances $\boldsymbol{d}$ are stored.

In a testing stage voxels from a testing volume of the same type as the training volumes are pushed through the trained forest. Thus they end up in different leaf-nodes according to the learned feature-tests. The previously stored relative distances can be used to estimate a landmark based on the absolute position of the voxel. This can be seen as letting a voxel vote for an absolute position based on the relative distance voting vector.

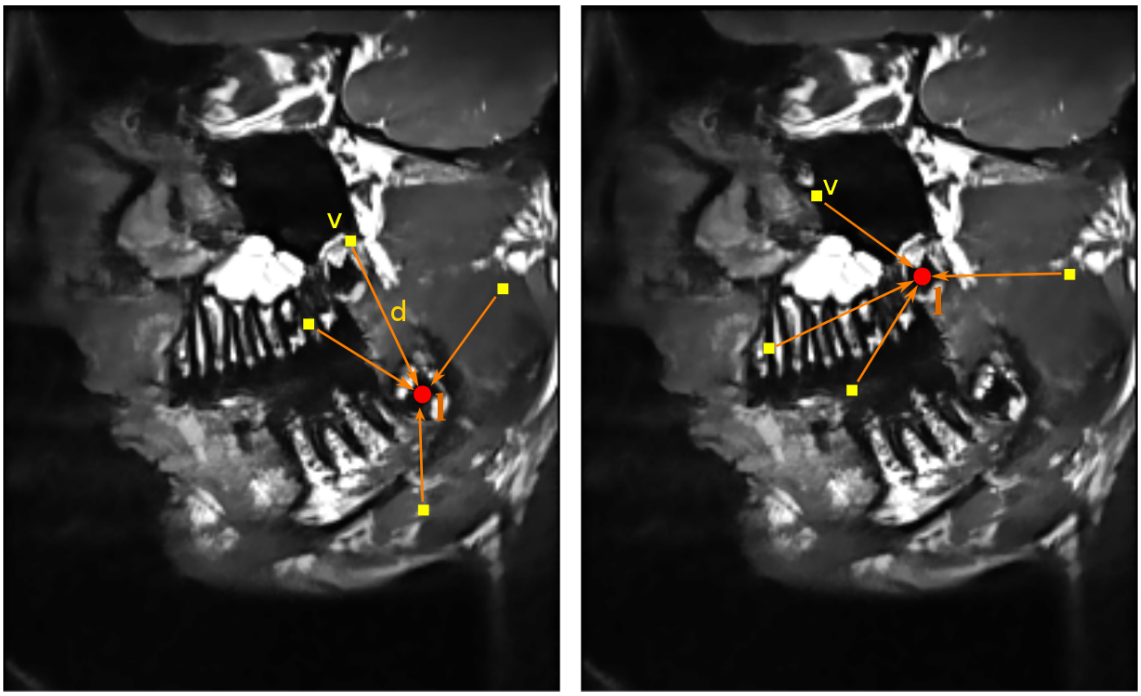


Figure 4.2: Illustration of sample voxels (yellow points) which vote for anatomical point-landmarks (red circles).

4.2.1 Training

To train the point localization *RRF*, t random subsets of all voxels are selected from training volumes. With these subsets t regression trees are trained, thus reducing the correlation between them. Two types of voxel-selection are explored in this work, namely *global* and *local* voxel selection around landmarks. The idea behind the global scheme is that voxels from anywhere in the image are used to predict a landmark position. On the other hand using the local voxel selection scheme only voxels within a certain range around a landmark are used to train a forest. Thus they explore the structure near to the landmarks in more detail than the global approach.

Having selected subsets of voxels as training data-points, multiple regression trees are trained independently and in parallel which is a huge benefit of *RFs*. As explained in chapter 3, in each split node a *best* feature-test (or split-function) is learned according to the randomized node optimization task.

4.2.1.1 Randomized Split Node Optimization

In the split node optimization task an incoming data-point set $\mathcal{S}$ consisting of voxels $\boldsymbol{v}$ from images is split into a left and right subset. The idea behind this splitting for landmark localization is that dissimilar voxels are split up and similar voxels are forwarded to the

same child nodes. The similarities of voxels are explored by their local appearance which is captured using appearance based features. These features are later explained in more detail. Voxels which are located at roughly the same distance and direction from a certain landmark are likely to have the same local appearance.

This work explores two different feature-generation approaches based on geometrical feature-parameters $\boldsymbol{\theta}$. Firstly, completely randomized features are used which depend on the local appearance information around the position of voxels similar as in [9, 11, 25]. Further, a novel feature-generation approach which creates features at the landmark positions is introduced, thus capturing the appearance at locations to where voxels vote.

Random feature generation: Haar-like features are used within the feature response function f_r to describe a voxels occurrence within a volume, similar as in [9, 11, 24]. Therefore, one ($b = 0$) or two ($b = 1$) random boxes are created randomly relative to a voxels position. Afterwards, for each box the mean intensity value of all voxels from the image captured by the box is calculated. If two boxes are used the difference between the two mean values is used. This approach is described by equation 4.2.

$$f_r(\boldsymbol{v}, \boldsymbol{\theta}) = \frac{1}{\mathcal{Q}_1} \sum_{\boldsymbol{q} \in Q_1} I(\boldsymbol{q} + \boldsymbol{v}) - b \cdot \frac{1}{\mathcal{Q}_2} \sum_{\boldsymbol{q} \in Q_2} I(\boldsymbol{q} + \boldsymbol{v}) \tag{4.2}$$

The boxes $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are generated randomly within a maximum range $\boldsymbol{p} = p_{\{x,y,z\}}$ relative to a voxels position. Further, the geometry of a box is created randomly, i.e. a random side-length up to maximum size of $\boldsymbol{s} = s_{\{x,y,z\}}$ in each dimension. $\boldsymbol{q}$ denotes a position inside a box $\mathcal{Q}$ and $I(\boldsymbol{x})$ the intensity at a certain position $\boldsymbol{x}$.

On-landmark feature generation: As described above the displacement distances $\boldsymbol{p}$ are traditionally chosen randomly inside a certain range to capture the local appearance information. A novel idea which is investigated in this work is to capture the appearance information directly at landmarks. For instance, instead of generating the feature-responses only locally around the voxels $\boldsymbol{v}$, they are also placed to where the voxels vote to. Thus, knowledge about the appearance information of landmarks is directly incorporated during training. This idea is further referred to as On-Landmark Feature Generation (OLFG) whereas the random approach is denoted as Random Feature Generation (RFG). An illustration of the differences between the *RFG* and the *OLFG* method is depicted in Fig. 4.3. To model these new features, the distance $\boldsymbol{p}$ is fixed to a relative distance between a random voxel $\boldsymbol{v}_R$ within a split-node to the landmark on which the *RRF* is actually trained. Next, the random boxes are created relatively to this probable landmark position. According to this idea voxels which have the landmark at the same relative distance as the voxel $\boldsymbol{v}_R$ should be grouped together.

As not every feature-parameter $\boldsymbol{\theta}$ and threshold τ is good for splitting the voxels in a split-node, the best combination of these two parameters is obtained in the node optimization task as described next.

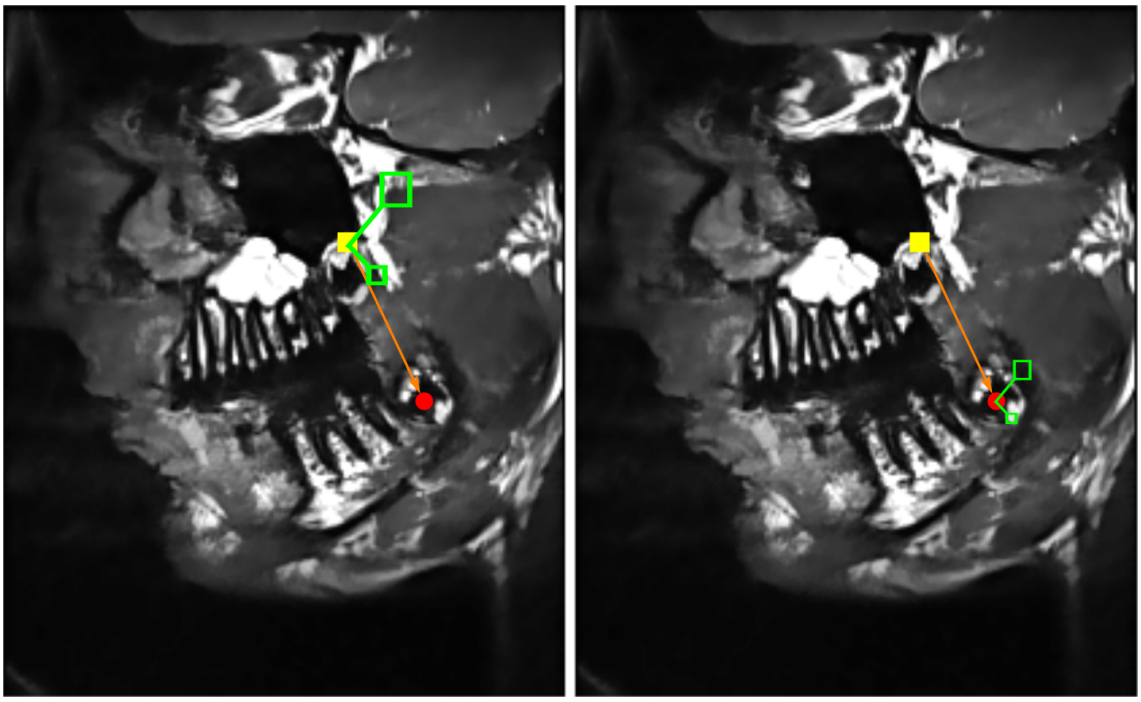


Figure 4.3: Illustration of the difference between the two feature generation methods. On the left: Standard *RFG* around a voxel (yellow point) which votes to a landmark (red dot). Only the local appearance is captured to describe the voxels position within a volume. On the right: *OLFG* approach to capture the information a voxel votes for.

Selecting the best features: In a root- or split-node j, a pool of feature-parameters $\boldsymbol{\theta}_j$ and for each of them several thresholds τ are randomly generated. According to equation 3.1 an incoming set of voxels $\mathcal{S}_j$ is split into a left and right subset $\mathcal{S}_j^L$ and $\mathcal{S}_j^R$. Due to equation 3.3, a best combination of feature-parameters and thresholds is chosen and stored using following information gain measure $\mathcal{I}$ for regression objectives [10, 24]:

$$\mathcal{I}(\mathcal{S}_j, \boldsymbol{\theta}_j, f_r, \tau) = \mathcal{H}(\mathcal{S}_j) - \sum_{c \in [L,R]} \frac{|\mathcal{S}_j^c|}{|\mathcal{S}_j|} \mathcal{H}(\mathcal{S}_j^c) \tag{4.3}$$

$|S_j|$ denotes the number of voxels in a set S_j. The information gain measure $\mathcal{I}$ is based on the Shannon entropy $\mathcal{H}$ and is calculated on the incoming and split sets $\mathcal{S}_j$, $\mathcal{S}_j^L$ and $\mathcal{S}_j^R$:

$$\mathcal{H}(\mathcal{S}) = \sum_{\boldsymbol{v} \in \mathcal{S}} log\left(|\boldsymbol{\Sigma}(\boldsymbol{v})|\right) \tag{4.4}$$

The covariance matrix $\boldsymbol{\Sigma}$ models the certainty of landmark estimations from voxels in a set $\mathcal{S}$ to a landmark. In other words, since the information gain is maximized according to equation 3.3 voxels within a split-node are split up such that the entropies $\mathcal{H}(\mathcal{S}^{\{L,R\}})$ for the subsets $\mathcal{S}_j^L$ and $\mathcal{S}_j^R$ are minimized. For instance, the more voxels in a set $\mathcal{S}$ estimate

the landmark at a similar position, the lower is the entropy.

An example for the similarity grouping based on features of voxels along the depth can be seen in the y-histograms h_y in Fig. 4.4. In the root-node in Fig. 4.4(a) two completely contrary d_y coordinates exist due to the different voxel positions in the volume, $d_{y,\{1,2\}}$ and $d_{y,3}$. The illustrated *RRF* tries to find best split features to describe these two rather different positions. Some examples are depicted with the green boxes with size $\boldsymbol{s}$ at a certain relative distance $\boldsymbol{p}$ to the voxels position. Having found best feature-tests to describe their local appearance and therefore their local position, the two voxels at the bottom are forwarded to the same child node on the right. The voxel which is located at the upper half in the volume is forwarded to the left child node, thus grouping similar voxels. The results of splitting are illustrated in Fig. 4.4(b-c).

4.2.1.2 Leaf Prediction Model

Voxels $\boldsymbol{v}$ which fall into the same leaf node l_t in a tree t are likely to have the same local appearance and are located roughly at the same relative position to a landmark within a volume. Therefore, the relative distances $\boldsymbol{d}$ of equation 4.1 from all these voxels to the landmark are stored in a prediction model in the leaf nodes. Similar as in [24] this prediction model consists of three 1D distance histograms $h_{\{x,y,z\}}(l_t(\boldsymbol{v}))$ in which all relative distances are stored in each axis separately. Some 2D example histograms are illustrated in Fig. 4.4.

4.2.2 Testing

The aim to train *RRFs* is that they can be used to predict landmarks in new medical volumes. Therefore, voxels $\boldsymbol{v} \in \mathcal{S}$ from a set of pixels $\mathcal{S}$ from a testing volume are pushed through the previously trained trees $t \in T$. The learned splitting features in root- and split-nodes are applied to the voxels, thus forwarding them either to left or right child-nodes until a certain leaf node is reached. For each voxel and for each tree relative distance histograms $h_{\{x,y,z\}}(l_t(\boldsymbol{v}))$ are retrieved. A final landmark prediction is computed by combining these histograms which can be done by several methods. In this work the histogram aggregation which was successfully applied in [24] is compared to [47] in which each voxel votes directly into an image space accumulator using a 3D histogram.

4.2.2.1 Histogram Accumulation

The histogram accumulation type proceeds as follows [24]:

$$H_{\{x,y,z\}} = \frac{1}{T} \cdot \frac{1}{\sum_{\boldsymbol{v} \in \mathcal{S}} w(\boldsymbol{v})} \sum_{t \in T} \sum_{\boldsymbol{v} \in \mathcal{S}} w(\boldsymbol{v}) \cdot h_{\{x,y,z\}}(l_t(\boldsymbol{v})) \tag{4.5}$$

The histograms $h_{\{x,y,z\}}(l_t(\boldsymbol{v}))$ are accumulated for each axis independently and result in three final separated 1D histograms $H_{\{x,y,z\}}$ for a landmark. A final landmark prediction

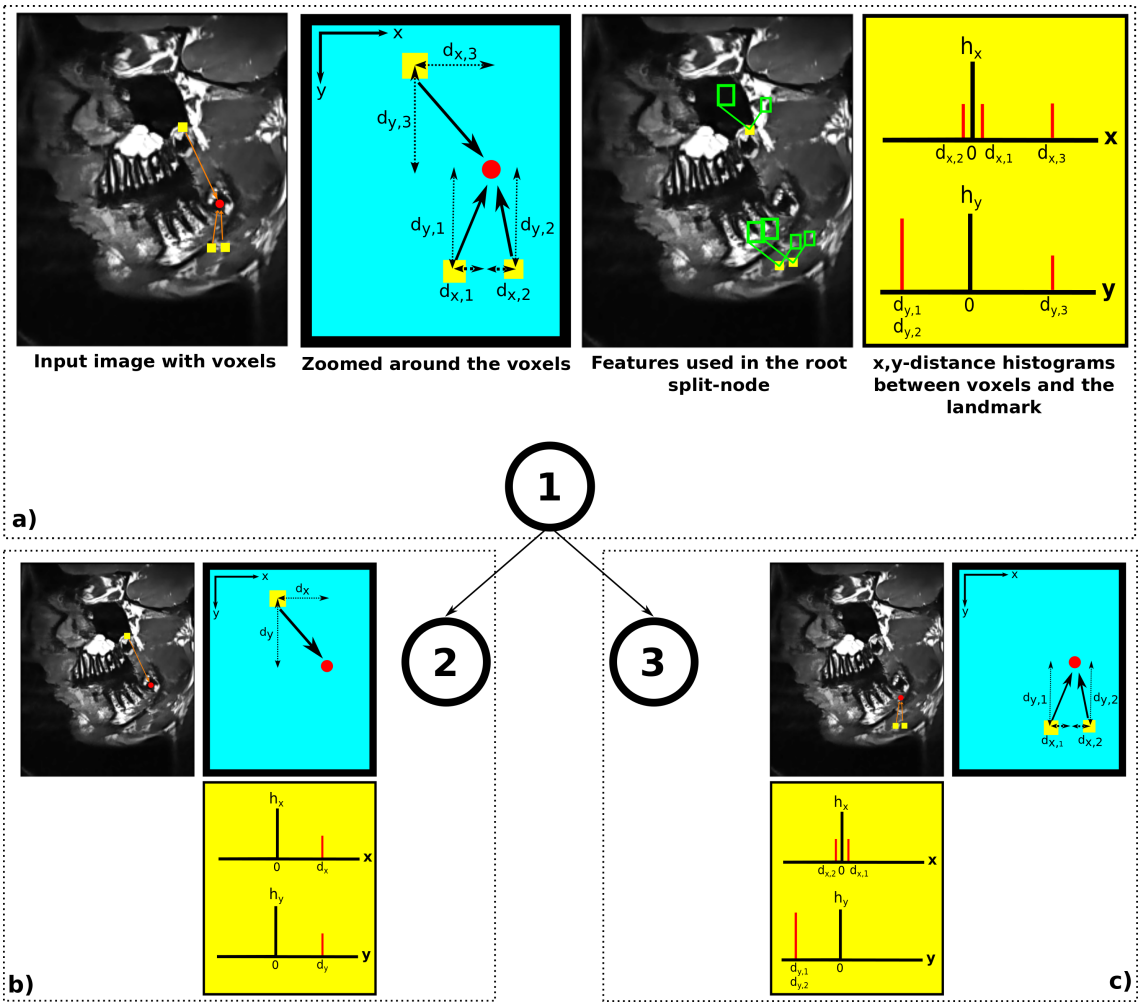


Figure 4.4: Illustration of sample voxels (yellow squares) which are trained to locate anatomical point-landmarks (red circles). (a) shows one training image and selected voxels (left), a zoomed view to the voting vectors and generated Haar-like feature-parameters used for the splitting. On the right, distances $\boldsymbol{d}$ between the voxels and the landmark are illustrated in two $1D$ histograms, one for each axis. According to the variance minimizing information gain measure in the split-node optimization, voxels from the bottom and from the top, are split since this leads to the best variance minimization in the histograms. (b) and (c) depict leaf nodes and the stored one dimensional distance histograms which models the distances from all voxels in that leaf node to the landmark position.

can be made by finding the maximum in each axis of $H_{\{x,y,z\}}$. In [24], additionally a weighting factor w was added which weights a vote from a voxel during testing. This weighting factor is discussed later in more detail.

4.2.2.2 Image Space Accumulation

For each voxel a probable landmark position is retrieved by finding the maxima values directly in $h_{\{x,y,z\}}$. This single prediction is used to vote for the best position in a 3D histogram also called image space accumulator. If all voxels have voted for a certain position the maximum within the image space accumulator is searched which indicates the most probable landmark position. In this work the maxima values of $h_{\{x,y,z\}}$ are used since they showed the best results for the point-voting scheme in [24]. Several other approaches exist to retrieve this single prediction, e.g. using the median values of $h_{\{x,y,z\}}$, calculating the mean value positions, etc. Similar as in the histogram accumulation type, votes to the image space accumulator can be weighted with a certain weight $w(\boldsymbol{v})$ which is discussed in the next section.

4.3 Vote Weighting Approaches

4.3.1 Euclidean Distance

In [24] votes from voxels $\boldsymbol{v}$ to a landmark $\boldsymbol{l}$ are weighted according to their relative Euclidean distance d, i.e. the length of the voting vector as follows:

$$w(\boldsymbol{v}) = \exp(-\alpha \cdot ||d||) \tag{4.6}$$

α is a tuning parameter which controls the strength of the weighting term. For instance voxels which vote to a landmark far away are weighted less than voxels voting for a near landmark if α becomes larger. This approach shows promising improvements in [24] in contrast to equally weighting all votes ($\alpha = 0$).

However, voxels which vote with a higher weight to near landmarks can also be a disadvantage. Think of a data-point $\boldsymbol{v}$ which votes from one finger tip to another one in hand volumes as illustrated in Fig. 4.5 on the left. Fingers are varying the most as observed in [24]. Thus these votes might be very uncertain for this dataset but an Euclidean weighting term would emphasize votes from such data-points. Therefore, a novel distance measure for hand images which makes use of underlying structures is introduced, i.e. replacing the Euclidean distance with a geodesic metric as depicted in Fig. 4.5 on the right. This geodesic metric is based on the idea of [46]. There, the authors assume that the shortest path between two points according to the geodesic distance contains rich information about the underlying structures between these two points. For instance, the geodesic distances between points of the same organ are very small whereas points between different organs have a larger distance since different tissues have to be passed.

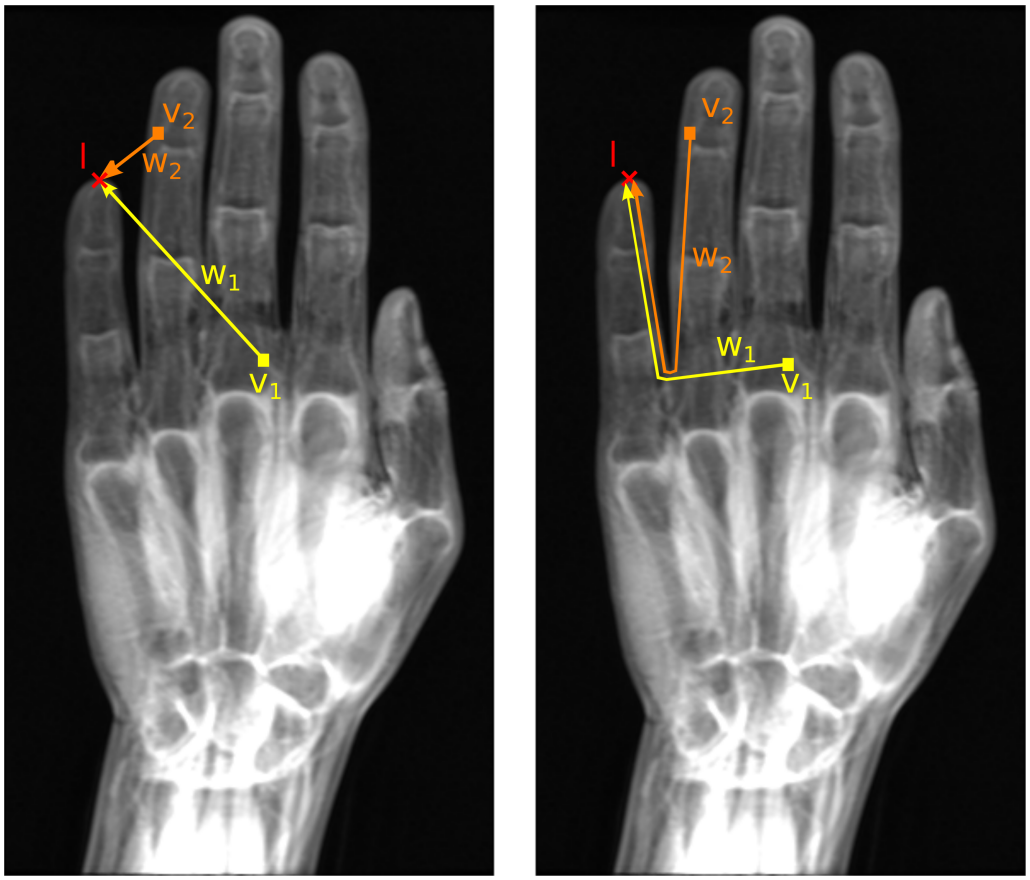


Figure 4.5: Illustration of the vote weighting approaches using an Euclidean distance of [24] (left) and a geodesic distance (right). Left: A voxel $\boldsymbol{v}_1$ far away from the landmark $\boldsymbol{l}$ is lower weighted with w_1 than a voxel $\boldsymbol{v}_2$ closer to the prediction of the landmark. Note that $\boldsymbol{v}_2$ has more influence than $\boldsymbol{v}_1$ to the landmark prediction. Right: The weighting factor depends on the underlying structures of the voting-vector, i.e. voting weight-vectors cannot jump between fingers and have to follow a certain path. Therefore, $\boldsymbol{v}_2$ has lower influence to the prediction as $\boldsymbol{v}_1$.

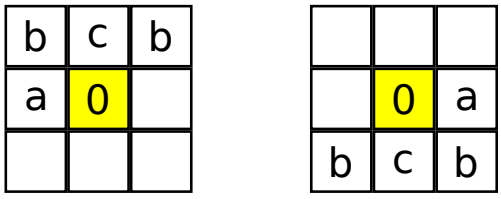
4.3.2 Geodesic Distance

In contrast to the Euclidean distance which is the distance between a voxels position and the position to where it votes, the geodesic metric needs a more sophisticated method to be calculated. Assume that foreground and background segmentations are given additionally to the 2D Magnetic Resonance (MR) images. Foreground segmentations indicate the inner part of the hand whereas background indicates the background of an image. For each voxel position a geodesic distance map to any other voxel position is calculated to get a distance between a testing voxel and the position to where it votes. This is done by using an approximation of the geodesic distance metric based on the sequential algorithm in [1]. In the next two subsections a detailed overview of how this geodesic distance calculation is achieved, is given.

4.3.2.1 Distance transformation

To calculate the distance transformation, a forward and backward mask have to be defined which are illustrated in Fig. 4.6. These masks are applied to an image consisting of one feature pixel with initial value zero and non-feature pixels with initial values of ∞ at the beginning [1]. An example of an image from which the distance transformation is calculated is illustrated in Fig. 4.7(a).

The distance computation starts by overlaying the forward mask pixel-wise with the center of the mask from left to right and from top to bottom of the image in a first forward iteration. The value of each distance in the mask (0, a, b and c) is added to the pixel value of the image below them. The actual processed pixel in the image is then replaced by the minimum of these four sums. Afterwards, a backward iteration using the background-mask is applied from right to left and bottom to top to the binary image.



(a) Fwd. mask **(b)** Bwd. mask

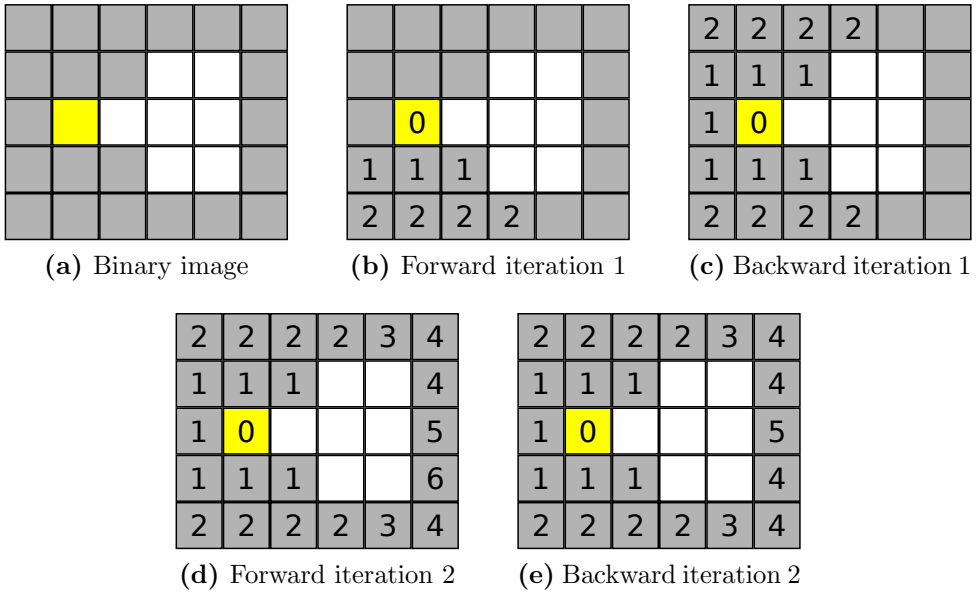
Figure 4.6: Forward and backward masks for the distance transformation.

(a) Binary image **(b)** Forward iteration **(c)** Backward iteration

Figure 4.7: Computation of the distance transformation using the mask-values $a = b = c = 1$. - are indicating initial values of ∞ and $*$ indicates the start feature pixel (Source [1]).

This distance transformation has to be applied several times iteratively in case of more complex images, i.e. the hand images. The toy example in Fig. 4.8(a) illustrates what happens if the distance transform is solely applied once to the foreground-pixels, marked as one yellow feature pixel and gray squares. If the distance transform is only applied once,

some distances are not assigned to a few foreground-pixels as depicted in Fig. 4.8(b-c). Therefore, the distance transform must be at least applied a second time to cover also the missing foreground-pixels, as illustrated in Fig. 4.8(d).



(a) Binary image **(b)** Forward iteration 1 **(c)** Backward iteration 1

(d) Forward iteration 2 **(e)** Backward iteration 2

Figure 4.8: Toy example which shows the distance transformation applied multiple iterations.

The next two steps describe how this distance transformation is applied to retrieve geodesic distance maps from the 2D hand images. Two different cases exist, based on whether a voxel which votes to a landmark is located inside the foreground-segmentation or inside the background-segmentation.

4.3.2.2 Creating Geodesic Distances Maps of Voxels Inside the Foreground Segmentation

From a position of a testing voxel $\boldsymbol{v}$ which is located inside the foreground-segmentation a two-step approach is used to calculate the distance to any other voxel. First, the distance transformation is applied to all voxels within the foreground-segmentation several times using the position of $\boldsymbol{v}$ as the starting feature position. Therefore, the values a, b and c are set to the voxel spacing $s_{x,y}$ of the images as follows: $a = s_x$, $c = s_y$ and $b = \sqrt{s_x^2 + s_y^2}$.

This results in a distance map within the foreground segmentation as illustrated in Fig. 4.9(a). The blue cross indicates the starting feature position. Dark values indicate a low distance to the starting position. The brighter the values become, the larger is the distance to the starting position.

Next, the distances within the background segmentation are calculated based on the geodesic distance values at the border of foreground distance map. To punish a path

through the background, a weight β for the masks is introduced as follows: $a = \beta \cdot s_x$, $c = \beta \cdot s_y$ and $b = \sqrt{(\beta \cdot s_x)^2 + (\beta \cdot s_y)^2}$. A final distance map is illustrated in Fig. 4.9(b).

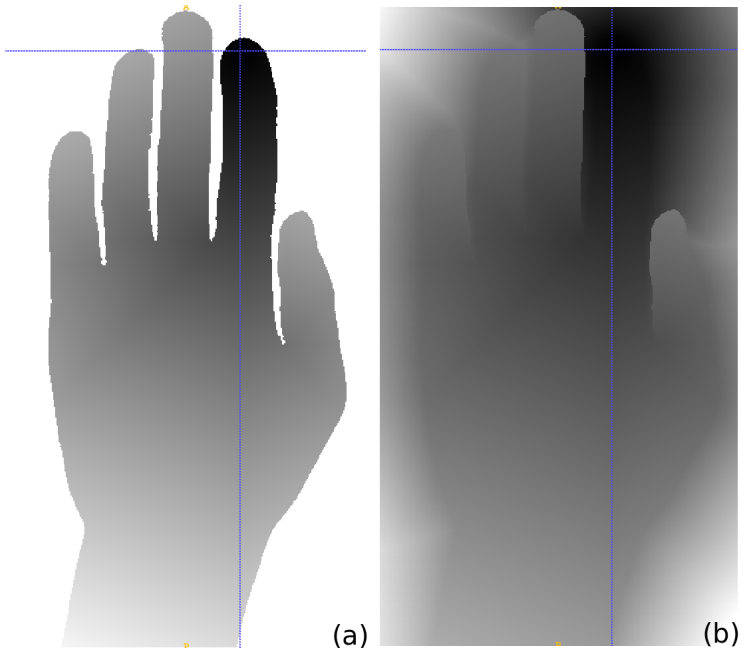


Figure 4.9: Illustration the geodesic distance map creation starting from the point at the finger tip. Dark values indicate a low distance to the starting position. The brighter the values become, the larger is the distance to the starting position.

4.3.2.3 Creating Geodesic Distances Maps of Voxels Inside the Background Segmentation

In contrast to step described above, three steps have to be performed if a testing voxel votes from the background-segmentation to a certain position. The first and second step are similar as the previously described one, but conversely calculated, i.e. first creating the distances for the background and then for the foreground segmentation. An illustration of this step is depicted in Fig. 4.10(a-b).

Using only these two steps leads to large distances in the background far away from the starting position as illustrated as white areas in Fig. 4.10(b). To allow paths going through the hand, a third step is performed which re-calculates all possible background-segmentation distances again based on the distances of the borders of the foreground-segmentation. As a final background-distance map the pixelwise minimum of either the first or the second background distance map are used. Note the differences between Fig. 4.10(b) and Fig. 4.10(c) in the area on the bottom right.

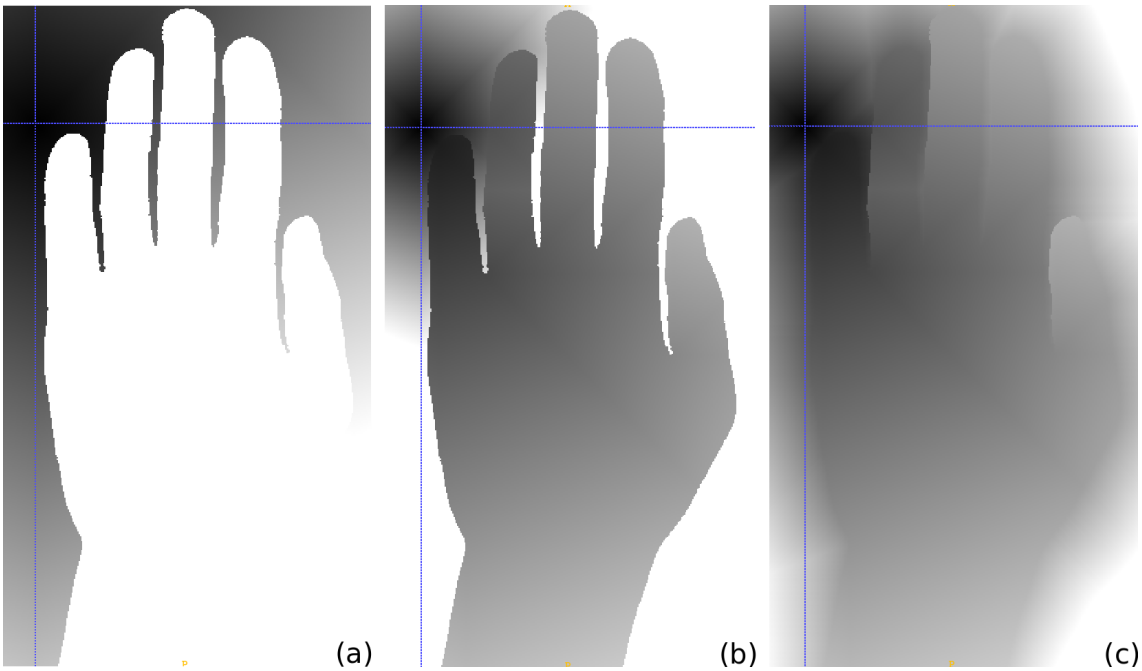


Figure 4.10: Illustration the geodesic distance map creation starting from the point at the blue-cross outside the hand. The pixelwise minimum of (a) or (c) is used to create the final distance map, shown in (c).

4.3.3 Weighting

During testing, first the distance maps for all voxels are created. Afterwards, the distance d is read out at the position of a probable landmark position where a voxels votes for. This distance d is applied to equation 4.6 which results in a weighting for the vote of a voxel.

4.4 Conclusion

This chapter has shown how *RRFs* can be used for anatomical point localization tasks. The regression model, objective functions and feature generation methods have been developed accordingly and a histogram method for storing landmark predictions in leaf-nodes has been introduced. In addition, a novel vote weighting distance measure based on a geodesic metric has been proposed with the underlying idea, that votes should be weighted according to a path through the underlying structures. Further, the drawback of long-distance votes in terms of having no knowledge to which a certain voxel votes has been encountered by means of a new feature generation method. In chapter 6, these ideas are evaluated and compared to previous methods on the set of studied datasets described in chapter 5.

5 Dataset

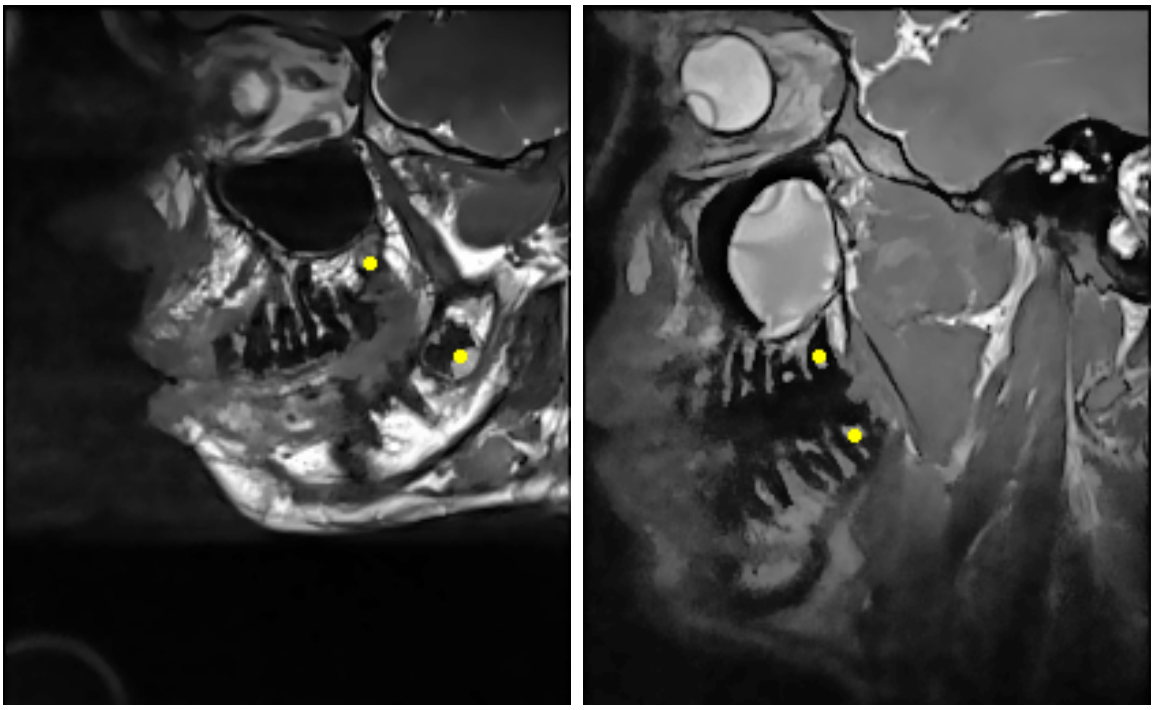
Contents

Three different types of datasets were provided by the Ludwig Boltzmann Institute for Clinical Forensic Imaging. This data is part of an ongoing volunteer study collecting Magnetic Resonance (MR) image scans of male Caucasian subjects with known age between 13 and 24 years. The dataset consists of 3D *MR* volumes of the head to localize third molars, upper part of the chest for localization of the clavicle bone and left hand volumes to localize bone joints. Anatomical points were annotated by medical experts.

5.1 3D MR Head Volumes for Third Molar Localization (MRTM)

276 *MR* volumes showing one half of the head with two annotations in the center of the third molars are given. In some subjects one or even all of the third molars are missing. Therefore, the annotation was placed on an estimated molar position by a medical expert. Challenges in this dataset vary from different volume cropping due to strong translations, varying structures within the head and brain and global rotations. Furthermore artifacts and noise appear frequently due to metallic objects like braces and scanner artifacts. Each of the *MR* volumes was acquired by a Proton-Density Weighted Turbo Spin Echo Sequence (PD-TSE). The voxel size per dimension of the volumes varies little around the mean value of $0.59 \times 0.59 \times 1$ mm per voxel with a size of around $208 \times 256 \times 28$ voxels. Figure 5.1 illustrates extracted 2D slices from 3D *MR* volumes acquired from 13 and 20 year old volunteers. Yellow dots mark the points of interest which should be automatically

located using Random Regression Forest (RRF). This dataset is further referred to as 3D MR Third Molar (MRTM) dataset.

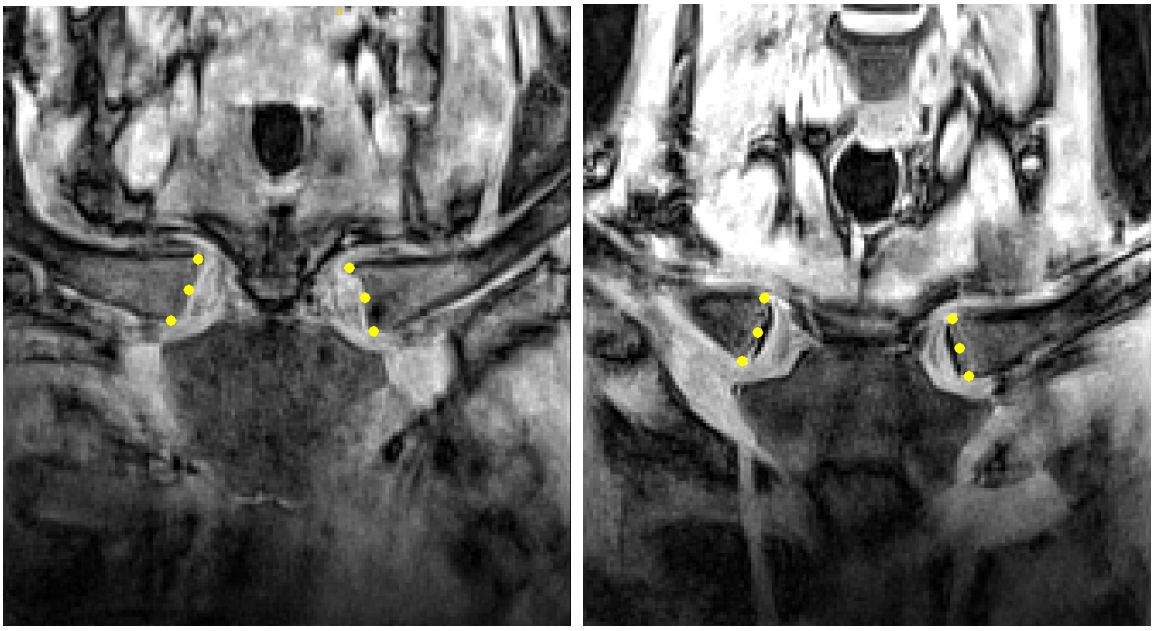


(a) 13 years **(b)** 20 years

Figure 5.1: Example 2D slices from the 3D *MR* head volumes showing 2 yellow marked point-landmarks at third molars.

5.2 3D MR Upper Chest Volumes for Clavicle Localization (MRC)

Landmark localization at the clavicles was performed on 28 *MR* volumes from the upper chest. Examples are illustrated in Fig. 5.2. The acquisition of the volumes was performed using a spacing of 0.89 $\times$ 0.89 $\times$ 0.99 mm per voxel and 168 $\times$ 192 $\times$ 44 voxels in x, y and z dimension. Landmarks were placed on the top, middle and bottom location at the sternal extremity of each clavicle in each volume. This results in 6 landmarks for each 3D volume. Each clavicle has a strong variation in rotation, intensities and shape which makes this small dataset challenging to process. Further, the continuous movement of volunteers due to breathing led to changes in translation and rotation of the inner parts in the chest. This dataset is further referred to as 3D MR Clavicle (MRC) dataset.

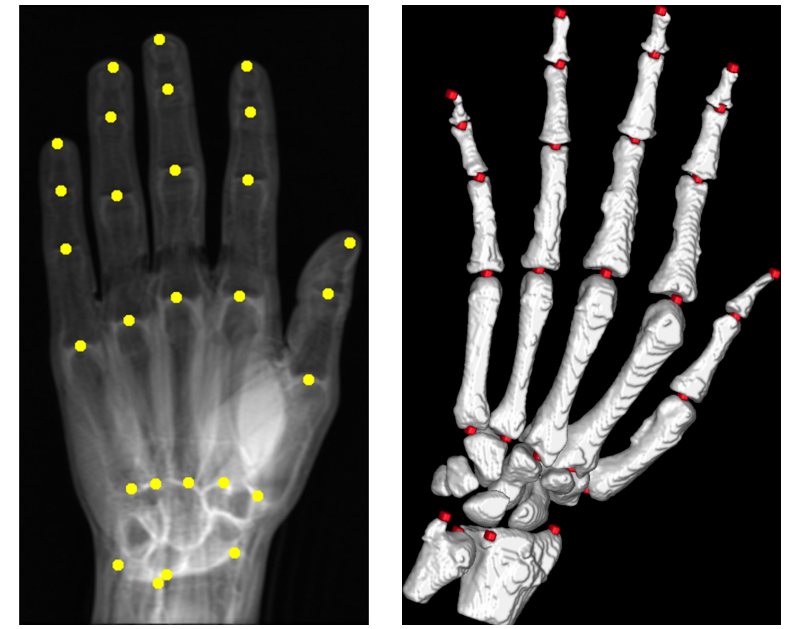


(a) 13 years **(b)** 20 years

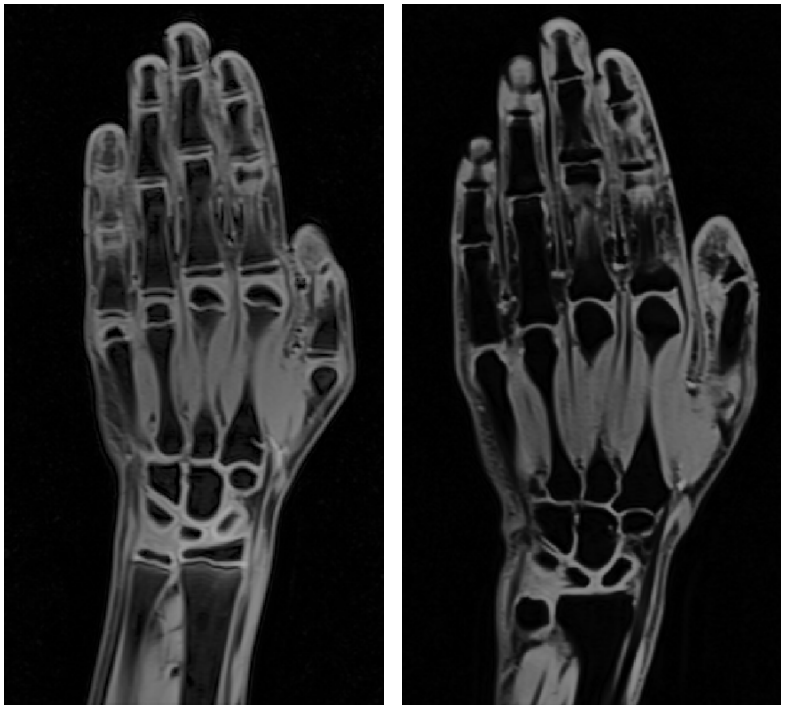
Figure 5.2: Example 2D slices from the 3D *MR* upper chest volumes with 6 yellow marked point-landmarks.

5.3 3D MR Left Hand Volumes for Hand-Bone Localization (MRH)

Left hand *MR* volumes to locate landmarks placed between hand and on wrist bones were acquired from 60 volunteers. Examples are illustrated in Fig. 5.3. This dataset with an average size of around 288 $\times$ 512 $\times$ 72 voxels per volume was acquired with a voxel spacing of 0.47 $\times$ 0.47 $\times$ 0.9 mm per voxel. A hand is located roughly at the center of the acquired volume. The challenges for this dataset are the strong variations of fingers since they are the most flexible part of the hand. The volunteers vary them during acquisition although a heavy weight is placed on the hands to reduce this movement. This might be due to inconvenient positions during the whole scan process which lasts approximately 10-15 minutes. The 3D MR Hand (3D-MRH) dataset is also projected along one axis to get a 2D MR Hand (2D-MRH) dataset of size 288 $\times$ 512 $\times$ 1 voxels. This dataset is used for evaluations with the geodesic distance metric. Further, foreground (hand) and background (no-hand) segmentations are made by a threshold based approach which is refined manually to split fingers which have been grown together.



(a) **(b)**



(c) 13 years **(d)** 19 years

Figure 5.3: (a) shows 28 annotations on a projected 2D *MR* image whereas (b) illustrates these annotations on a 3D bone segmentation of a left-hand volume (Source [24]). Some example 2D slices from the 3D *MR* left hand volumes are depicted in (c) and (d).

6 Experiments

Contents

In this chapter, the ideas of chapter 4 are evaluated, namely the geodesic distance metric, the on-landmark feature generation method and the use of restricted image information for training. In more detail, following evaluations are made:

1. In section 6.1, the vote weighting approach using the geodesic distance metric (Geodesic Distance Vote Weighting (GDVW)) is compared to the Euclidean distance metric (Euclidean Distance Vote Weighting (EDVW)) of [24]. This evaluation is performed on the 2D MR Hand (2D-MRH) dataset.

2. In a second evaluation, the On-Landmark Feature Generation (OLFG) approach is investigated in section 6.2 and compared to the traditional Random Feature Generation (RFG) method. The idea of the *OLFG* approach is originally introduced for the 3D MR Hand (3D-MRH) data. However, it is also studied for the other 3D MR Third Molar (MRTM) and 3D MR Clavicle (MRC) datasets.

3. In a third evaluation in section 6.3, forests are trained for all 3D Magnetic Resonance (MR) datasets using only voxels in a small range around the landmark positions, i.e. similar to Hough-Forest (HF) [29, 30] but including also long-distance features as will be explained later on. This approach is further referred to as Globally-Constrained Hough-Forest (GCHF). This experiment is distinct from to the second evaluation of section 6.2 in which voxels from the whole image are used to train a Random Regression Forest (RRF). Another reason to experiment with this approach is the

smaller anatomical variation near to a landmark. Structures farther away might vary stronger in rotation and translation related on the landmark position. Therefore they can predict the landmark only very roughly. Further, appearance variations at these remote like structures might also negatively influence a landmark prediction.

All the evaluations are quantitatively comparing the Euclidean distance between the found and annotated landmark position in mean ± standard deviation in millimeter. Further, the number of estimated landmark positions greater than a threshold are noted as outliers in percent and the absolute number in braces.

The evaluations of the *3D-MRH* and *MRTM* datasets are done using a three-fold cross-validation setup. For all the parameter optimizations one run of the three-fold crossvalidation setup is performed. For the *MRC* dataset a leave-three-out cross-validation is performed due to the small number of data. The parameter optimization is performed using four runs of the cross-validation. For all evaluations, per each split node 100 features and 10 random thresholds for each feature are generated within the node optimization task.

6.1 Geodesic Distance Evaluation

The evaluation of the *GDVW* approach is performed on the *2D-MRH* dataset. First, the two parameters α and β are optimized. Based on the best parameter setup the tree depth and number of trees are evaluated. Afterwards, landmarks are localized on all images in a three-fold cross-validation. The results are compared to the *EDVW* approach of [24].

6.1.1 Parameter Optimization of α and β

In the *EDVW* approach of [24] the parameter α controls how strong votes from pixels to a landmark are weighted according to their voting-distance. For instance a pixel which votes to a landmark far away has a smaller weight compared to a pixel which votes to a closer landmark. In the *GDVW* method, α is a tuning parameter which controls the overall influence of the votes. β on the other hand controls the weighting of the distances between neighboring background pixels. α and β are evaluated empirically within the following range:

- α: 0.05 up to 0.30 in 0.05 steps
- β: 0.5 up to 3.5 in 0.5 steps, depending on a certain α value

The parameters depicted in Tab. 6.1 are used to train the *RRF* which is afterwards used to optimize α and β.

Results and Discussion:
The results for the *GDVW* parameter optimization can be seen in Tab. 6.2. Note that the α parameter favors a value of 0.05 since the errors in mean and standard deviation are

Parameter	Value
Number of trees	8
Maximal depth	15
Maximal feature size	25 mm
Maximal feature distance	35 mm

Table 6.1: *RRF* parameter setup for *GDVW* α and β optimizations.

		$\boldsymbol{\beta}$						
		0.5	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**
$\boldsymbol{\alpha}$	**0.05**	3.1 ± 2.3	2.8 ± 1.9	2.7 ± 1.9	2.7 ± 1.9	$\mathbf{2.6 \pm 1.8}$	2.7 ± 1.8	2.7 ± 1.8
	0.10	2.9 ± 2.2	2.8 ± 2.3	2.7 ± 2.3	2.7 ± 2.3	2.7 ± 2.2	2.7 ± 2.2	2.7 ± 2.3
	0.15	3.1 ± 2.5	2.9 ± 2.6	2.8 ± 2.4	2.9 ± 2.5	2.8 ± 2.4	2.8 ± 2.5	2.9 ± 2.5
	0.20	3.2 ± 2.6	3.0 ± 2.6	3.0 ± 2.6	3.1 ± 3.2	3.1 ± 3.0	3.1 ± 2.7	3.2 ± 3.2
	0.25	3.6 ± 4.0	3.4 ± 3.9	3.5 ± 4.4	3.1 ± 3.6	3.3 ± 3.7	3.5 ± 4.5	3.6 ± 5.1
	0.30	3.8 ± 3.9	3.5 ± 4.0	3.6 ± 5.0	3.6 ± 5.0	3.5 ± 4.4	3.4 ± 4.4	3.6 ± 5.0

Table 6.2: Results of varying α and β values for the *GDVW* parameter optimization.

$\boldsymbol{\alpha}$						
0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**
5.7 ± 5.0	$\mathbf{2.7 \pm 1.9}$	2.7 ± 2.3	2.8 ± 2.4	3.0 ± 2.6	3.1 ± 2.7	3.3 ± 3.3

Table 6.3: Results of varying α values for the *EDVW* parameter optimization which is later on compared to the *GDVW* approach.

almost the smallest among different β values. The more α increases, the higher the error in mean and standard-deviation becomes in almost every α-β combination. This is due to more noisy estimations since the localization results depend on less pixels the higher α becomes. On the other hand, the errors at $\alpha = 0.05$ and β between 0.5 up to 2.5 decrease. An extended β evaluation for $\alpha = 0.05$ is illustrated in Fig. 6.1. It seems that large β values greater than around 2.5 do only have small influence to the localization performance which might be due to saturation. Votes to landmarks from background pixels become negligible if β increases. Therefore, the parameter combination of $\alpha = 0.05$ and $\beta = 2.5$ has been chosen as optimal. Note that $\alpha = 0$ is evaluated in Tab. 6.3 since with this setup no weighting scheme is used at all, i.e. all votes are weighted equally.

Results of the parameter optimization for the *EDVW* method is illustrated in Tab. 6.3. An α value of 0.05 seems to be promising for further evaluations. Higher values lead to noisy estimations since the prediction belongs to less voxels the higher α becomes as also observed in [24]. Note that the highest errors are retrieved using no weighting term ($\alpha = 0$). It seems that pixels which vote to a landmark farther away have a negative influence on the overall localization performance of the *RRF*, as also observed in [24].

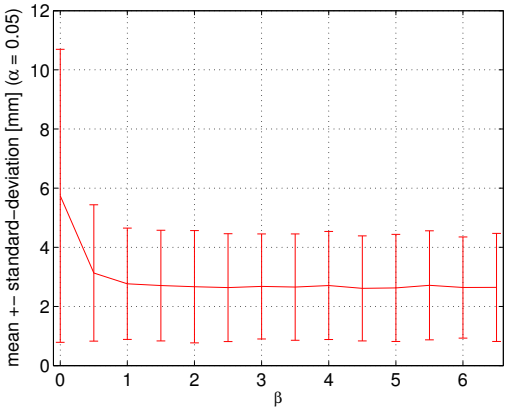


Figure 6.1: Extended results of varying β values with fixed $\alpha = 0.05$ for the *GDVW* parameter optimization.

6.1.2 Evaluation of Tree Depth and Number of Trees

In this section the tree-depth and the number of trees are evaluated to find a good parameter setup for a cross-validation. An *RRF* with a varying depth up to 20 using a constant number of eight trees is evaluated to retrieve the best depth. To evaluate the best number of trees, an *RRF* up to 10 trees is evaluated in steps of two using a depth of 20.

Results and Discussion:
Figure 6.2(a) shows results for the depth evaluation. It seems that the deeper trees are, the better is the localization performance. Further, no over-fitting can be observed. The mean and standard deviation errors decrease fast at the beginning and change only little at higher depths. Due to these results for all other experiments a depth of 20 is used. Larger depths are neglected since training time and memory consumption increase exponentially the deeper the trees are trained.

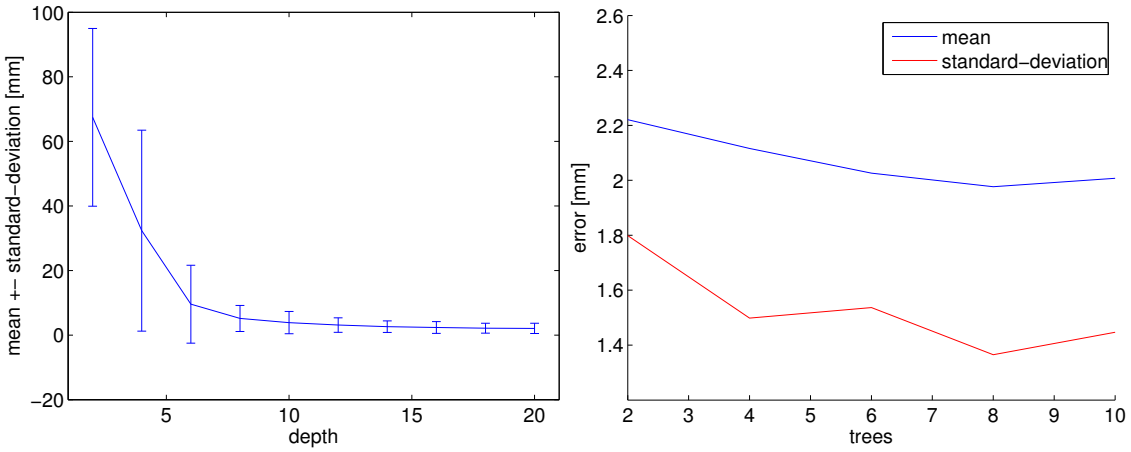
Increasing the number of trees on the other hand seems to have only small influence on localization performance as opposed to tree-depth. This is illustrated in Fig. 6.2(b) and might be the case because the average prediction of multiple pixels is used. Therefore, eight trees per forests are chosen for further experiments since the error at *RRFs* with more than eight trees stays approximately the same.

6.1.3 Cross-Validation

To evaluate the whole *2D-MRH* dataset, a three-fold cross validation is performed using the previously optimized *RRF* parameters. Following experiments are made:

1. Comparing *RRFs* with the *EDVW* and *GDVW* approach

2. Comparison between the histogram accumulation and the image space accumulation



(a) Depth evaluation **(b)** Tree evaluation

Figure 6.2: Results of tree depth and number of trees parameter optimization for the *GDVW* evaluation.

6.1.3.1 Euclidean Distance vs. Geodesic Distance

Figure 6.3(b) and Tab. 6.4 show results comparing *EDVW* and *GDVW*. The y-axis of Fig. 6.3(b) depicts the percentage of landmark estimations (normalized between zero and one) which are located within the Euclidean error in millimeter on the x-axis. Both methods seem to be equally good and only differ in minor deviations. A qualitative comparison of these two approaches is illustrated in Fig. 6.4. Investigations show that the largest outliers are 14 mm away from the correct annotation in both approaches, which is a good result and may be handled by a second more locally *RRF* as done in [24].

Type	**Error**	**# Outlier ≥ 10 mm**
GDVW	1.95 ± 1.48	0.18 % (3/1680)
EDVW	1.92 ± 1.44	0.12 % (2/1680)

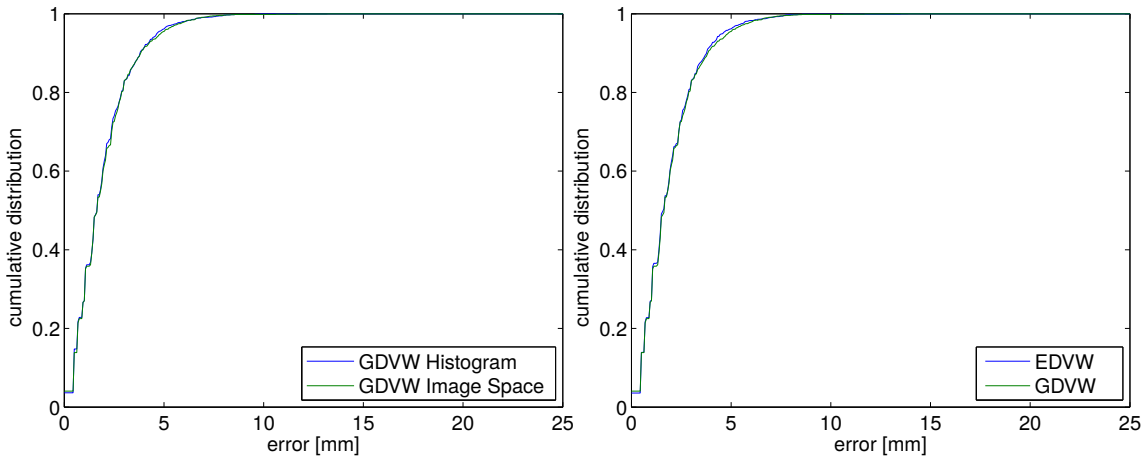
Table 6.4: Results of comparing the *EDVW* to the *GDVW* approach.

The results lead to the conclusion that votes from fingers to landmarks on other fingers and votes to landmarks farther away do not contribute much to the localization performance. This assumption is based on the results of both approaches (*EDVW* and *GDVW*):

- **Impact of *EDVW*:** Assume that the locations on finger tips should be predicted using only voxels from hand-palm. Pixels which are described by local appearance at the palm of the hand cannot cover the strong variations at fingers precisely. On the other hand, pixels near to a predicted landmark have a stronger influence than

pixels farther away. This behavior is observed and tackled by the *EDVW* approach in [24]. However, in most images the fingers are close to each other. Therefore, votes from fingers which vote to landmarks on another finger have a high contribution to the prediction due to the *EDVW* method. By using the *GDVW* method, such votes between different fingers have a lower contribution as described next.

- **Impact of *GDVW*:** The *GDVW* approach does not lead to any improvements compared to *EDVW*, but it also has no disadvantages. Pixels from one finger which vote to landmarks on a different finger seem to have a low influence to the localization performance since it makes no difference to weight them good (*EDVW* approach; see before) or worse (*GDVW* method). This might be due to the strong variation of fingers. Another aspect is that the features, which are generated locally around pixels, already incorporate the variation at the fingers. The same is true for those pixels which are located at the landmarks. This might explain that votes from pixels at fingers which vote to another finger can be neglected.

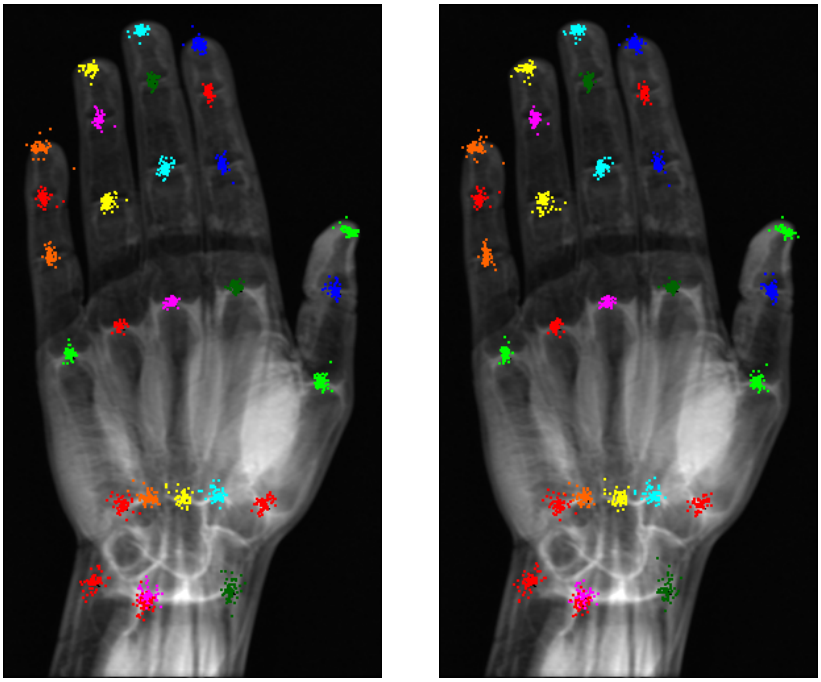


(a) Accumulator comparison **(b)** Distance measure comparison

Figure 6.3: Results of comparing the histogram to the image-space accumulation in (a), and the *EDVW* to the *GDVW* approach in (b).

6.1.3.2 Histogram vs. Image Space Accumulation

Results of comparing the histogram with the image space accumulators using the optimized *GDVW* approach are depicted in Tab. 6.5 and Fig. 6.3(a). The image space accumulation seems to be as good as the histogram accumulation for this dataset. This might be because of the large depth during training and due to the small amount of data from the 2D dataset. In general, the larger the depth, the less voxels reach the same leaf nodes.



(a) *EDVW* **(b)** *GDVW*

Figure 6.4: Qualitative comparison of the *EDVW* and the *GDVW* vote weighting approaches.

Thus the stored distance histograms $h_{\{x,y,z\}}$ in the leaf nodes become very peaky and have a small variation.

Type	Error
***GDVW* Histogram**	1.93 ± 1.45
***GDVW* Image Space**	1.95 ± 1.48

Table 6.5: Results of comparing the histogram to the image-space accumulation approach using *GDVW*.

6.1.4 Conclusion

Overall it can be said, that the *GDVW* approach has no benefits but also no drawbacks compared to the *EDVW* method of [24] in terms of localization accuracy and precision. The only drawback is that for *GDVW*, fore- and background segmentations have to be generated. However, this is a clear drawback since generating segmentations is in general a difficult task. Therefore, this approach is not further adapted to the *3D-MRH* dataset.

6.2 On-Landmark Feature Generation using Whole Image Information

In this section, the *OLFG* method is compared to *RFG* approach on all 3D datasets. The training process of an *RRF* is modified such that starting from a certain depth d_{start}, the forest generates 50% random (*RFG*) and 50% on-landmark features (*OLFG*) of the maximum allowed number of features in each split-node. Further, for all the evaluations the weighting term of [24] is used since it showed promising results for hand-datasets. Therefore, d_{start} and α for the weighting term are optimized for each dataset separately to gain the best parameter-setup.

6.2.1 Hand Dataset

First the 3D hand volumes are evaluated using the parameters as listed in Tab. 6.6. The *OLFG* approach is introduced starting at a depth of $d_{start} = 10$ according to the results of the evaluation in Fig. 6.5. Since it is the same dataset as in [24] and further similar behavior is observed in chapter 6.1, the number of trees, the maximum depth and α are chosen to be similar as in [24].

Parameter	Value
Number of trees	8
Maximal depth	16
Maximal feature size	25 mm
Maximal feature distance	35 mm
Maximal feature size on landmarks	10 mm
α	0.1
OLFG starting depth d_{start}	10

Table 6.6: *RRF* parameter setup for the *3D-MRH* dataset.

6.2.1.1 Results and Discussion

A three-fold cross-validation based on the optimized parameters shows the results in Tab. 6.7 and Fig. 6.6. The *OLFG* method shows the best result of 3.49 ± 2.93 in mean and on the number of outliers compared to the *RFG* approach. The standard-deviation shows only minor and negligible differences. In Tab. 6.8, a detailed analysis is depicted on which parts of the hand the *RRF* makes the most errors. These results show that the improvements of the accuracy come from each part of the hand, i.e. the mean-error can be improved at more stable structures like the radius, ulna, carpometacarpals as well as at the stronger varying interphalangeal joints (along fingers) and finger-tips .

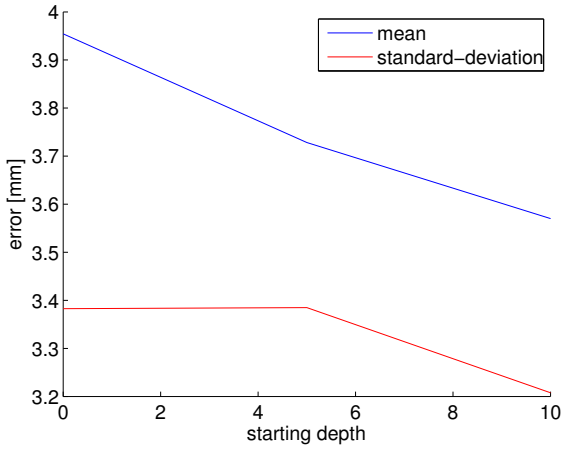
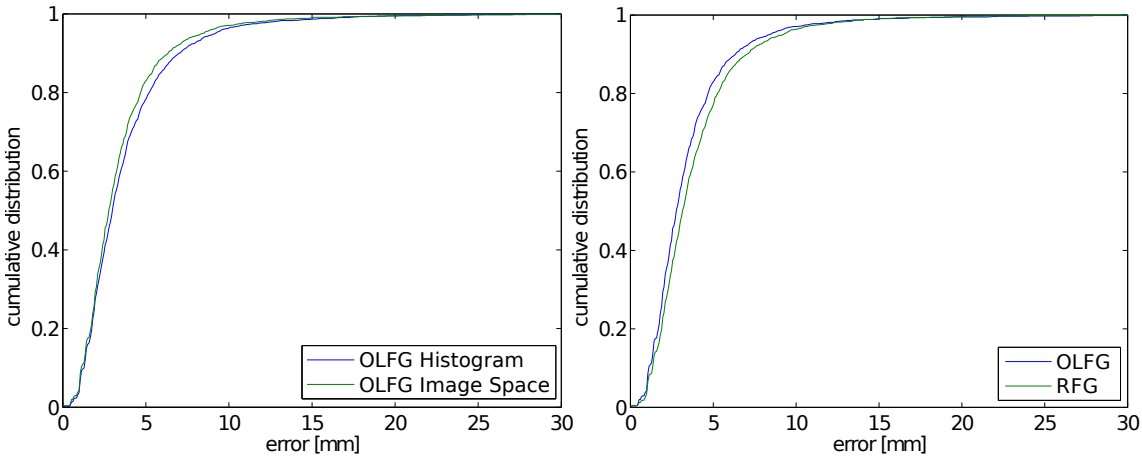


Figure 6.5: Optimization results of the *OLFG* starting depth d_{start} for the *3D-MRH* dataset.

Type	**Error**	**# Outlier $\geq$ 10 mm**
RFG	3.85 ± 2.88	3.45 % (58/1680)
OLFG	3.49 ± 2.93	2.80 % (47/1680)

Table 6.7: Results of comparing the *RFG* to the *OLFG* approach for the *3D-MRH* dataset.



(a) Accumulator Comparison **(b)** Feature Generation Comparison

Figure 6.6: Results of comparing the histogram and the image-space accumulation approaches in (a), and the *RFG* to the *OLFG* approach in (b) for the *3D-MRH* dataset.

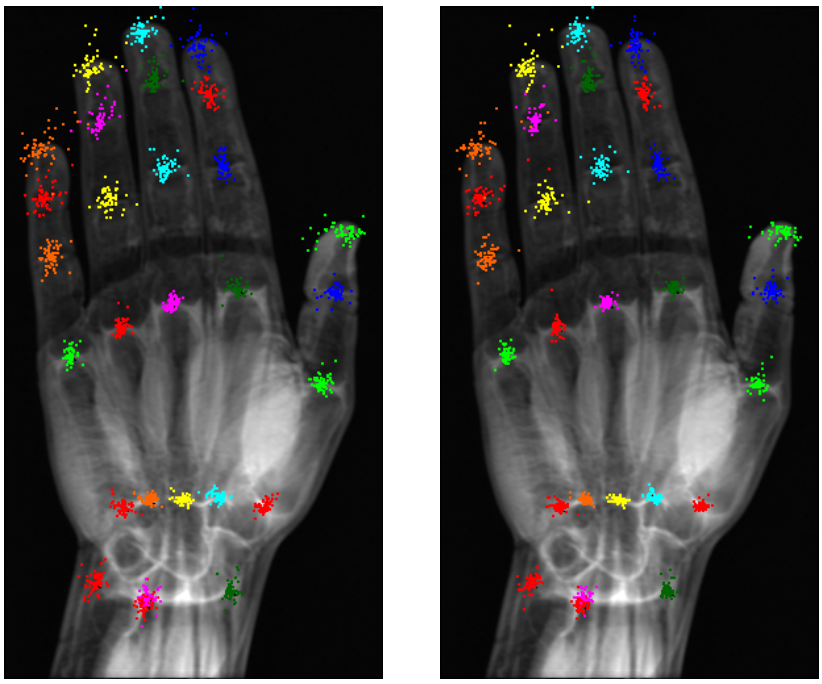
On the other hand, the precision of the *OLFG* approach which is modeled by the standard-deviation is only for the interphalangeal joints more worse than the *RFG* method. One possible explanation for this behavior is that these similar looking landmarks are very close to each other. Therefore, the forest might be very uncertain if it captures the correct

Type	RUC	MC	PJ	FT
RFG	3.09 ± 1.75 (1)	2.97 ± 1.76 (3)	4.02 ± 2.72 (15)	5.43 ± 4.17 (39)
OLFG	2.61 ± 1.47 (2)	2.64 ± 1.73 (2)	3.90 ± 3.03 (16)	4.89 ± 4.14 (27)

Table 6.8: Detailed results of RUC (radius, ulna and carpometacarpal), MC (metacarpal), PJ (interphalangeal joints) and FT (finger-tips) for 3D hand- and wrist-bone localization (Number of outliers are depicted within the braces).

landmark position with the *OLFG* method or a false one which looks similar.

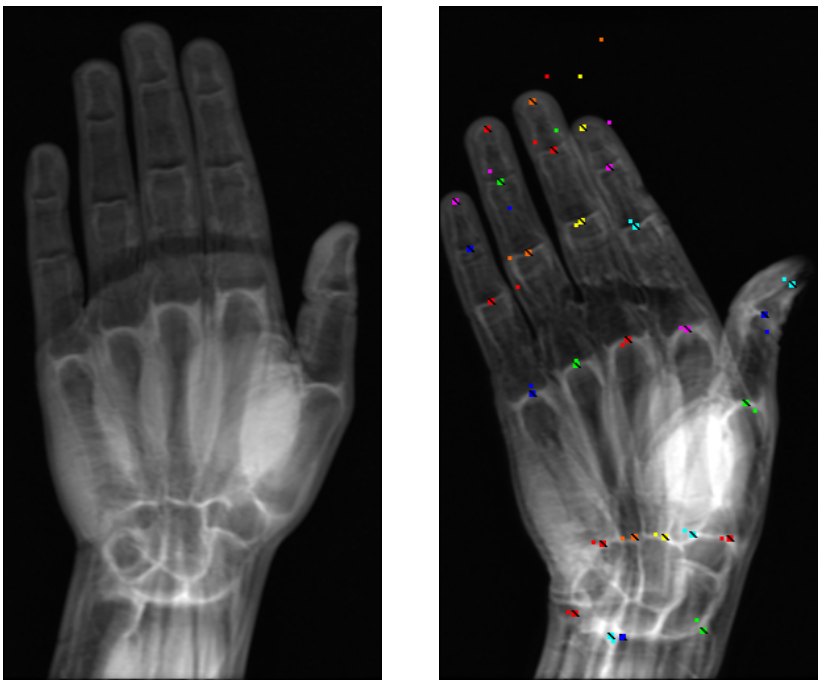
A qualitative comparison between the *OLFG* and the *RFG* approach is depicted in Fig. 6.7. There, the errors between the estimated and the annotated position are plotted into one hand-image. The results of the novel *OLFG* approach are slightly more concentrated to a landmarks position than the results of the *RFG* method. Further, in Fig. 6.8(b) an example is illustrated for which both *RRF* approaches lead to wrong localization results. This is due to the training-data which does not contain such strong shape variations. One shape example which frequently appear in the data is depicted in Fig. 6.8(a).



(a) *RFG* **(b)** *OLFG*

Figure 6.7: Qualitative comparison of the random and the on-landmark feature generation approach for the *3D-MRH* dataset.

Another interesting observation is depicted in Fig. 6.5. It seems the later the *OLFG* is introduced during training, the better the localization performance in mean and standard



(a) Typical shape of most images

(b) Testing image with strongly varying shape

Figure 6.8: In (a) a typical image is illustrated which appear frequently in the *3D-MRH* dataset. In (b) localization results for a testing-image with strong variations at the fingers is depicted.

deviation becomes. One reason for this behavior is that at low depths the voxels have not been located themselves within the image very precisely. At low depths, the long-distance on-landmark features will be generated for the most voxels outside the volumes which makes only little sense. At higher depths, the *RRF* first groups the voxels from the images based on their local appearance. Thus grouping them into small sub-groups of voxels which are located near to each other. The on-landmark features for voxels of these sub-groups might be generated more-often at the landmark positions.

Another comparison is done between the histogram and image-space voting scheme which can be seen from Fig. 6.6(a). The image-space voting scheme seems to be favorable to the histogram voting scheme. The summation of the errors during the histogram accumulation lowers the localization performance since it is the main difference between these two approaches.

Since the results do not show a huge improvement on the *3D-MRH* dataset, a small experiment on the *2D-MRH* dataset is performed. The same parameter settings as for the optimized geodesic distance comparison in section 6.1 are used to train an *RRF*. Results are depicted in Fig. 6.9 and Tab. 6.9 and show that the *OLFG* approach improves the results a lot. The error in precision and accuracy and further the number of outliers can

be decreased significantly. This leads to the assumption that for the *3D-MRH* dataset the localization task is much more challenging than for the *2D-MRH* datasets. This is a legitimate assumption due to the reduced dimensionality the forest has to deal with. It could be that for the *3D-MRH* dataset, *RRF* with a much higher depth must be trained which is only partly possible due to the huge memory consumption. A qualitative comparison for the *2D-MRH* dataset between the *OLFG* and the *RFG* approach is depicted in Fig. 6.10. Similar as for the *3D-MRH* dataset, the results are concentrated more to the correct landmark position with the *OLFG* approach. Due to the good results on the *2D-MRH* dataset, one future-work might be to investigate on localizing using projected images first and then add the remaining third dimension.

Type	**Error**	**# Outlier $\geq$ 10 mm**
2D *RFG*	1.92 ± 1.44	0.12 % (2/1680)
2D *OLFG*	1.19 ± 0.97	0.06 % (1/1680)

Table 6.9: Results of comparing the *RFG* to the *OLFG* approach for projected 2D hand- and wrist-bone localization.

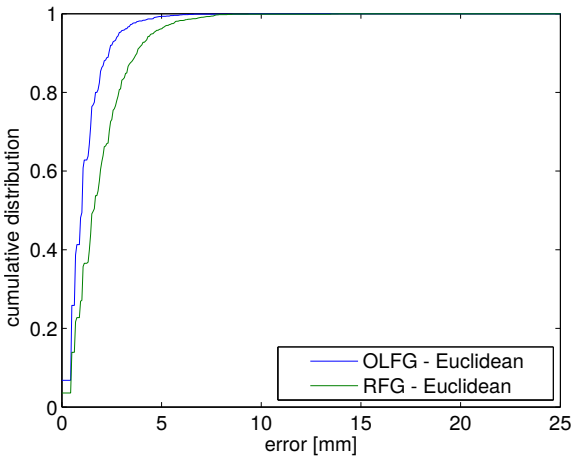
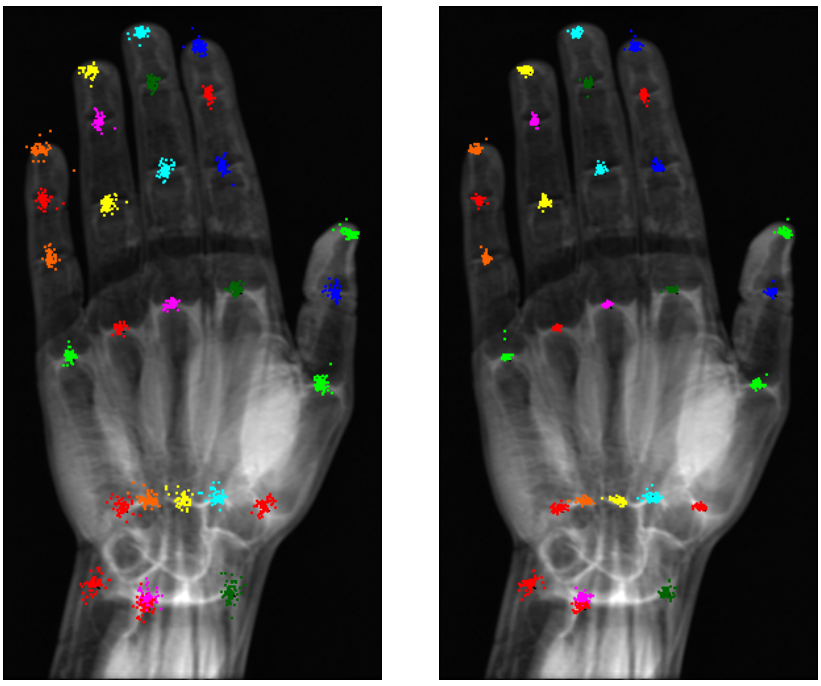


Figure 6.9: Graphical results of comparing the *RFG* to the *OLFG* approach for *2D-MRH* dataset.

6.2.1.2 Conclusion

Overall it can be said that the *OLFG* approach can be used to improve on accuracy and to decrease the number of outliers for the 3D *MR* hand dataset. Especially these two errors are important for a multi-forest approach, i.e. of [24, 25]. The less outliers and the more accurate a first *RRF* is, the more probable a second *RRF* converges to the correct landmarks.



(a) *RFG* **(b)** *OLFG*

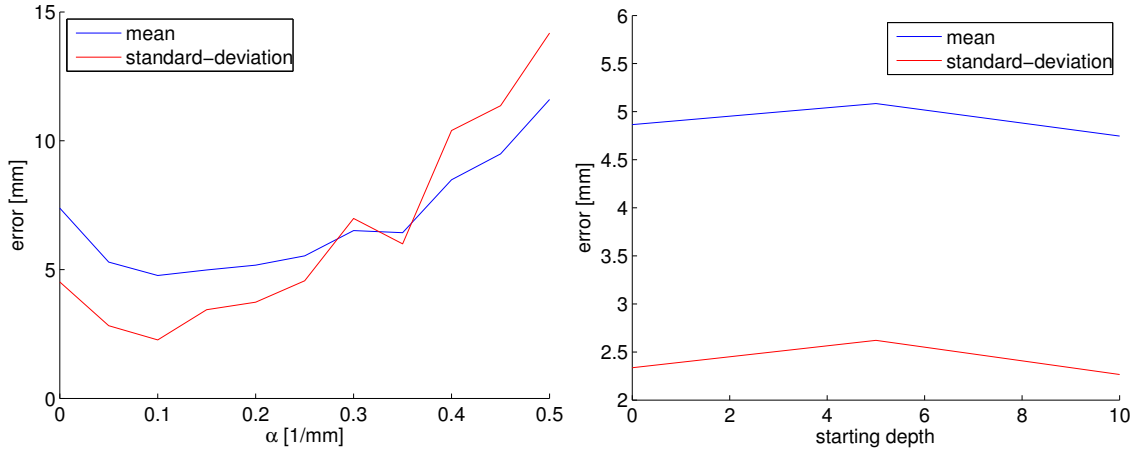
Figure 6.10: Qualitative comparison of the *RFG* and the *OLFG* approach for the *2D-MRH* dataset.

6.2.2 Tooth Dataset

The *MRTM* dataset is evaluated using the parameters according to Tab. 6.10. The α parameter is tuned according to results illustrated in Fig. 6.11(a). It can be seen that an α of 0.1 is optimal in sense of mean and standard deviation. The starting depth d_{start} seems to have no influence on the localization performance for the *OLFG* approach as depicted in Fig. 6.11(b). Therefore and due to the results of the *3D-MRH* evaluation, $d_{start} = 10$ is chosen for a three-fold cross-validation which is evaluated next. Further, the depth and the number of trees behave similar to the *2D-MRH* dataset as illustrated in Fig. 6.12(a-b).

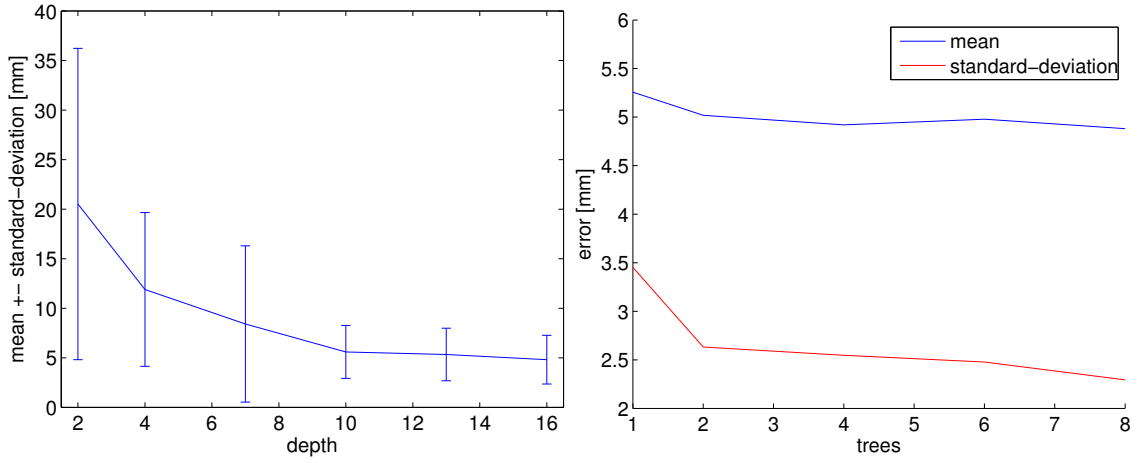
6.2.2.1 Results and Discussion

From Fig. 6.13(b) and Tab. 6.11 no significant difference between the *RFG* and *OLFG* approach can be seen. See also Fig. 6.14 for qualitative comparisons. The results show that the *OLFG* method do not help to improve the localization performance which might be due to several reasons. One reason might be that third molars at which the landmarks are placed are very small and difficult to capture with the *OLFG* method. This is due to the strongly varying shape and appearance information in the volumes. For instance, in



(a) **(b)**

Figure 6.11: Results of parameter optimization of the weight voting α parameter in (a). Results of the *OLFG* starting depth d_{start} parameter optimization for third molar localization in (b).



(a) Depth evaluation **(b)** Tree evaluation

Figure 6.12: Results of tree depth and number of trees parameter optimization for third molar localization.

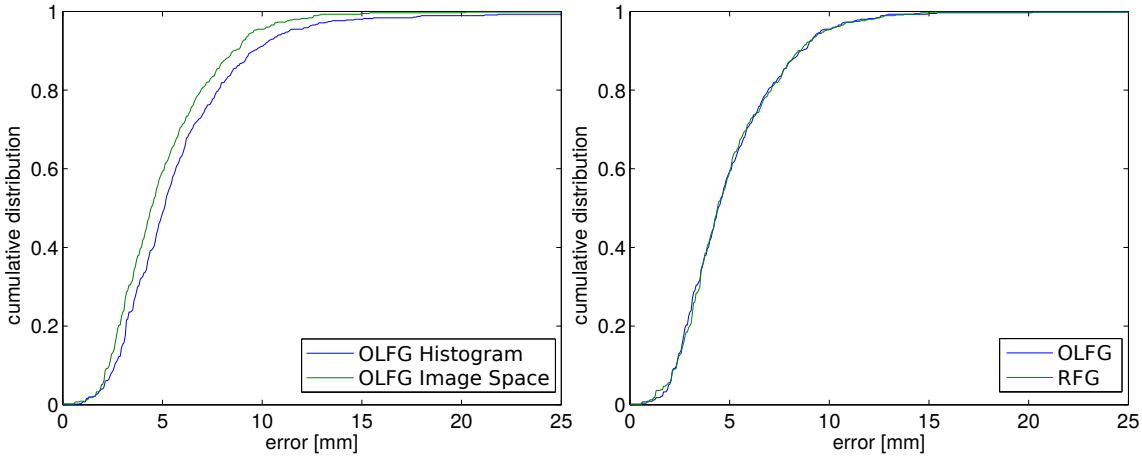
Parameter	Value
Number of trees	8
Maximal depth	16
Maximal feature size	15 mm
Maximal feature distance:	20 mm
Maximal feature size on landmarks	8 mm
α	0.1
OLFG starting depth d_{start}	10

Table 6.10: *RRF* parameter setup for 3D third molar localization.

the *3D-MRH* dataset most structures where the landmarks are placed are clearer visible and distinguishable than in the *MRTM* dataset.

Type	Error	# Outlier ≥ 7 mm
RFG	5.06 ± 3.01	20.47 % (113/552)
OLFG	5.07 ± 3.09	19.38 %(107/552)

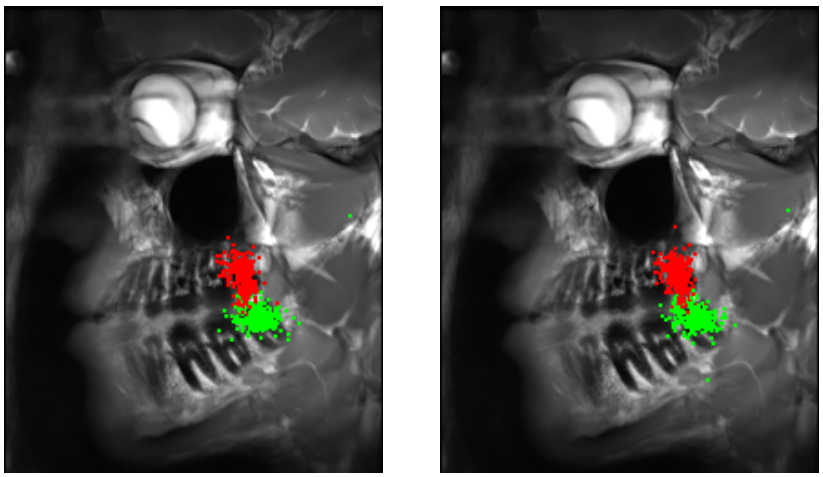
Table 6.11: Results of comparing the *RFG* to the *OLFG* approach for third molar localization.



(a) Accumulator Comparison **(b)** Feature Generation Comparison

Figure 6.13: Results of comparing the histogram and the image-space accumulation approaches in (a), and the *RFG* to the *OLFG* approach in (b) for third molar localization.

Deeper investigations of what the *RRF* learns, deliver some important observations as described next. One behavior is that the *RRF* tends to learn and most accurately localize landmarks based on the shape of the head, i.e. nose, chin, etc. In a few cases of occluding shapes due to strong translations of the head during *MR* acquisition or due to artifacts caused by braces, the *RRF* votes for locations which are far away from annotations. A next



(a) *RFG* **(b)** *OLFG*

Figure 6.14: Qualitative comparison of the random and the on-landmark feature generation approach for the *MRTM* dataset.

observation is that from locations at soft-tissues like the brain, the forest votes inaccurately to a landmark position. This might be due to the strong variations. However, the *RRF* using the voxels from the whole volume tends to vote for the correct position to the right of the *second* molar, if a *third* molar is missing. All these observations might indicate that the *RRF* does not learn that much from the structures at the tooth region which is the most stable part in these volumes. This assumption is strengthened by the experiments and better results in chapter 6.3 in which only voxels within a certain range around the landmarks are used to train an *RRF*.

According to Fig. 6.13(a) the image-space voting scheme is better suited for this dataset than the histogram voting scheme, similar as in the hand-dataset. The accumulation of all variations in the distance histograms seems to negatively influence the localization performance.

6.2.2.2 Conclusion

It seems that the *OLFG* method has no or a small negative influence to the localization for this dataset. This might be because of the strong variations in shape of the volumes and also due to difficult to capture appearance of the tiny third molars.

6.2.3 Clavicle Dataset

The *OLFG* method is also applied to the new *MRC* dataset which actually consists only of a few annotated volumes at the end-date of this thesis. This makes them a challenging datasets since they contain only very few data to learn from. In contrast to the *3D-MRH*

and *MRTM* datasets, the clavicles are evaluated using a leave-three-out cross-validation, i.e. training on (N-3) of the N volumes and testing on the remaining three in multiple iterations. The parameter setup is depicted in Tab 6.12 based on optimization of α and d_{start} depicted in Fig. 6.15(a-b). The number of trees and the tree depth have not been evaluated empirically since this dataset is actually not discriminative enough.

Parameter	Value
Number of trees	8
Maximal depth	16
Maximal feature size	20 mm
Maximal feature distance	30 mm
Maximal feature size on landmarks	10 mm
α	0.1
OLFG starting depth d_{start}	10

Table 6.12: *RRF* parameter setup for 3D clavicle localization.

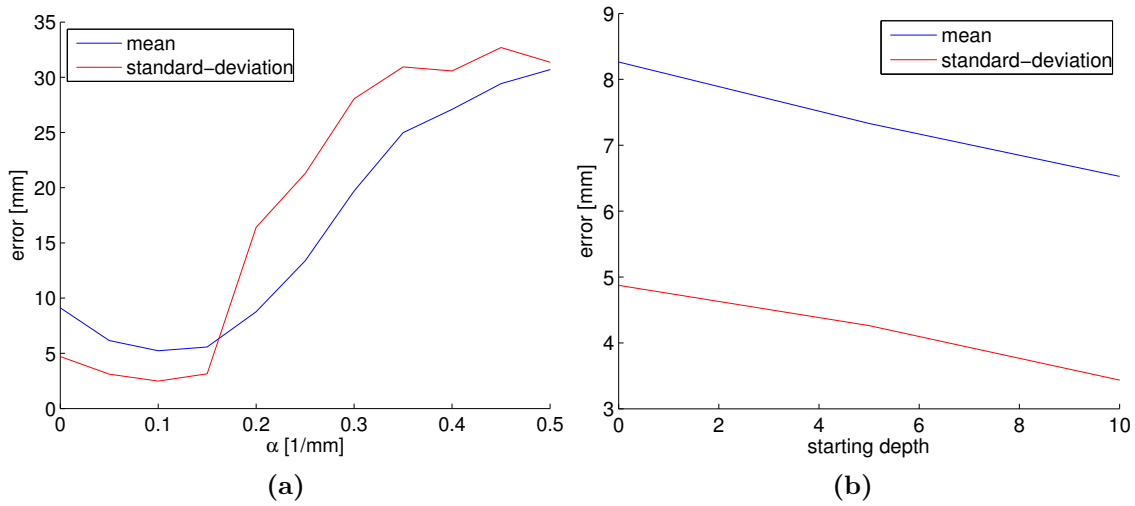


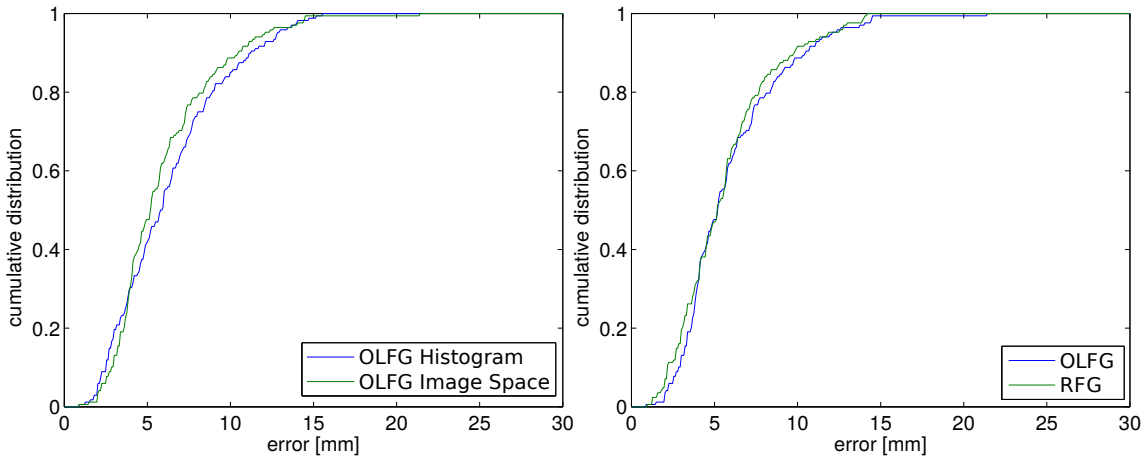
Figure 6.15: Results of parameter optimization of the weight voting α parameter in (a). Results of the *OLFG* starting depth d_{start} parameter optimization for clavicle localization in (b).

6.2.3.1 Results and Discussion

From Fig. 6.16(b) and Tab. 6.13 it can be seen that this dataset is one of the most challenging ones due to the large errors. The main reasons are the small amount of data available and the low resolution of the *MR* acquisition.

Type	Error	# Outlier ≥ 10 mm
RFG	5.59 ± 2.93	8.33 % (14/168)
OLFG	5.93 ± 3.16	11.31 % (19/168)

Table 6.13: Results of comparing the *RFG* to the *OLFG* approach for clavicle localization.



(a) Accumulator Comparison **(b)** Feature Generation Comparison

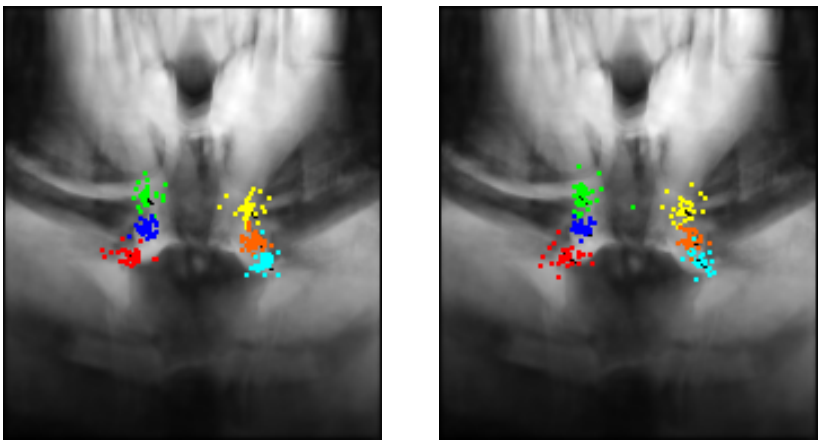
Figure 6.16: Results of comparing the histogram and the image-space accumulation approaches in (a), and the *RFG* to the *OLFG* approach in (b) for clavicle localization.

As also can be seen, the *OLFG* approach cannot help to improve the results of the *RFG* method which is illustrated qualitatively in Fig. 6.17. It is difficult to give a reason for that behavior due to the small dataset. The most probable assumption is, that based on the small amount of training data the strong variations at the landmarks are not captured sufficiently enough. The elongated clavicles tend to have a rotation within each volume and the appearance in size vary at the sternal extremities on which the landmarks are placed.

However, the *OLFG* method might help to improve the localization result if more data is available. The landmarks are placed on relatively clear structures similar to the hand images. On the other hand, this improvement might be very small since it has also shown only minor improvements on the 3D *MR* hands.

6.2.4 On-Landmark Feature Generation Conclusion

In this section, the *OLFG* approach is compared to the *RFG* method. In case of the *3D-MRH* and *2D-MRH* datasets, the *OLFG* method seems to improve on the overall localization performance. Regarding to the results of the *MRTM* and *MRC* dataset, no improvement has been found. Especially for the *MRC* dataset it is difficult to say which



(a) *RFG* **(b)** *OLFG*

Figure 6.17: Qualitative comparison of the *RFG* and the *OLFG* approach for the *MRC* dataset.

parameters of the forest lead to the worse localization performance. For the *MRTM* dataset the *OLFG* approach seems not to improve the localization performance but also not to worsen them drastically. As observed during this thesis, the forests tend to learn strongly from the shape, i.e. the mouth, chin, nose, etc. Therefore the question raises what happens if only voxels within a certain range of a landmark are used for training. For instance, learning how the third molars and their neighboring teeth looks like in more detail. This may lead to improve the localization performance since the forest concentrates only on this information during training. This idea is evaluated within the next section.

6.3 Training RRFs using Restricted Image Information

The previous section 6.2 makes use of the whole image information, i.e. all voxels are used to train an *RRF* to predict a certain landmark position. In this section an approach is investigated which uses only voxels within a certain region around landmarks to train an *RRF*, i.e. an *HF*. Thus learning and covering more of the structures at the landmarks themselves. Therefore, also more local random features are used. However, since this leads to multiple landmark candidates for one landmark prediction, also more global information is incorporated during training which is later explained in more detail. This novel approach is further referred to as *GCHF*.

A first simple experiment on the *MRTM* dataset shows what happens if voxels from the whole images and only from a small region around landmarks are used for training.

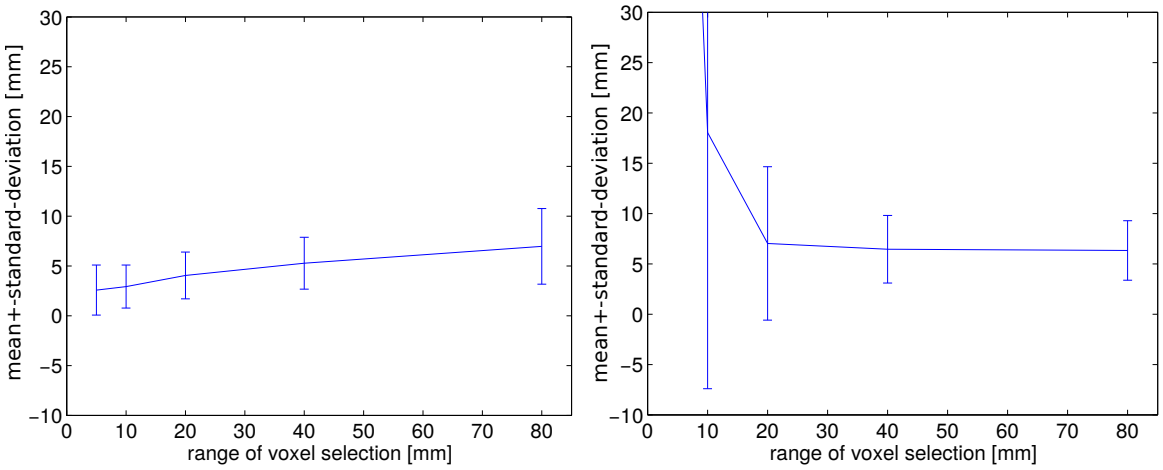
6.3.1 Toy Example

In this experiment, a decreasing subset of voxels within a certain range around the landmarks are used to train an *RRF* while leaving each other forest parameter unchanged. Since the landmarks can vary in their position, all voxels from the volumes are used to predict the landmarks during testing. Results of this experiment can be seen in Fig. 6.18(a-b). Voxels have been chosen within a radius of 80, 40, 20, 10 and 5 millimeter around the landmarks. Note that 80 millimeters cover almost the entire volume. Further, the image-space accumulation and the histogram accumulation is compared.

As illustrated in Fig. 6.18(a), decreasing the range from 80 to 10 mm improves the localization performance significantly. At a range of 5 mm the standard-deviation becomes little higher which might be due to the small amount of voxels used for training, i.e. the *HF* has to few information to learn from. However, the error only decreases for the image-space accumulation scheme. Using the histogram accumulation approach, the error increases as depicted in Fig. 6.18(b).

This behavior can be explained with the 2D toy example illustrated in Fig. 6.19. Assume that two voxels vote for position $(x_1,\ y_1)$ and three other voxels to three different positions $(x_2,\ y_2)$, $(x_2,\ y_3)$, $(x_2,\ y_4)$. The histogram accumulation approach sums up the axes independently which yield to the histograms h_x and h_y. Afterwards, these two histograms are searched for the maximum which results in the landmark estimation at position $(x_2,\ y_1)$ although no voxel has voted for this position. This behavior occurs if multiple landmark candidates at various positions are located by an *HF* which is trained by a few voxels around a landmark. This drawback is circumvented by the image-space accumulation approach since voxels vote directly into the 2D/3D accumulator, as illustrated with the red and black dots in Fig. 6.19.

In the next sections, the proposed 3D datasets are evaluated using voxels around landmarks for training. During testing, the image-space accumulator and all voxels from a volume are used to predict landmarks.



(a) Image-Space accumulation **(b)** Histogram accumulation

Figure 6.18: Localization results for restricting the range of voxel selection around landmarks.

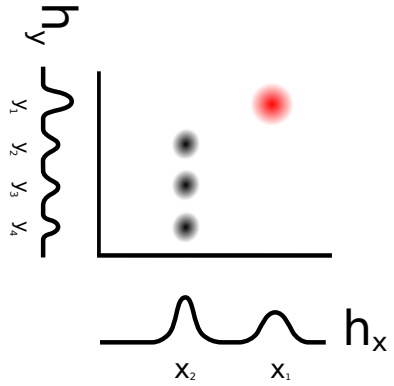


Figure 6.19: Toy example which illustrates the drawback of the histogram accumulation scheme over the image-space accumulation.

6.3.2 Tooth Dataset

This section evaluates the *MRTM* dataset using the parameter setup depicted in Tab. 6.14. The voxel selection range of 10 mm has been chosen since it showed a good performance in the results of Fig. 6.18(a).

During the evaluation of experiments it was shown, that an *HF* which is trained using only very small and local features locates the last molar which is visible during testing. For instance, if a third molar is missing the *HF* always switches to the neighboring second molar to the left. Therefore, in addition to the local small features, long-range features are introduced which start at a certain depth during training, i.e. the *GCHF* generates 50 % local and 50 % long-range features. This approach is further referred to as *GCHF*.

Parameter	Value
Number of trees	8
Maximal depth	18
Voxel selection range	± 10 mm
Maximal size of local features	7 mm
Maximal distance of local features	10 mm
Maximal size of long-distance features	20 mm
Maximal distance of long-distance features	30 mm
Starting depth of long-distance features (see Tab. 6.15)	5

Table 6.14: *RRF* parameter setup for third molar localization using *GCHF*.

Starting depth	Error	# Outlier ≥ 10 mm
HF	2.65 ± 3.02	6.52 %
0	2.97 ± 3.54	3.80 %
5	$\mathbf{2.67 \pm 2.25}$	**3.26 %**
10	2.67 ± 2.61	4.35 %
15	2.60 ± 2.72	5.43 %

Table 6.15: Results of parameter optimization for introducing long-distance features starting at certain depths for third molar localization.

The results of the starting-depth evaluation are illustrated in Tab. 6.15 which shows the best results with long-range features starting at a depth of five. Note the smaller number of outliers which is due to the long-range features. Next, a three-fold cross-validation is performed to test each volume.

6.3.2.1 Results and Discussion

Localizing third molars using particular selected voxels around the landmark yields to a localization error of 2.92 ± 2.49 mm in mean and standard deviation, as illustrated in Fig. 6.20 and Tab. 6.16. To make a comparison with the *RRF* which use the whole image information the results from section 6.2.2 are added. They are marked as *RRF* whereas the results from this approach are noted as *GCHF*. Note also that the parameters have slightly changed. However, as shown in the introducing example, the most influence is due to the restricted range of voxels used for training.

Type	Error	# Outlier ≥ 7 mm
GCHF	2.92 ± 2.49	6.88 % (38/552)
RRF	5.06 ± 3.01	20.47 % (113/552)

Table 6.16: Results of the local *GCHF* approach for third molar localization. The local approach is compared to the best global *RRF* approach.

It can be seen that using voxels near a landmark results in a significant improvement of

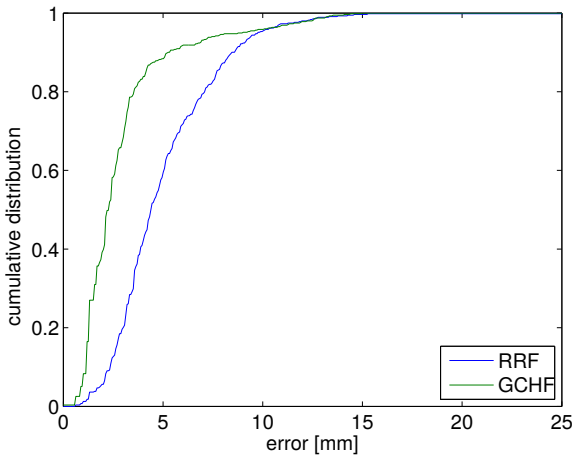


Figure 6.20: Results of the *GCHF* approach for third molar localization. The local approach is compared to the best global *RRF* approach.

Forest type	Visible teeth	2	1	0
GCHF	**Error**	2.7 ± 2.2	3.9 ± 3.2	6.1 ± 2.7
	# Outlier	4.3 % (20/470)	17.7 % (11/62)	35 % (7/20)
RRF	**Error**	5.1 ± 3.2	4.8 ± 2.0	4.1 ± 1.5
	# Outlier	21.5 % (101/470)	17.7 % (11/62)	5 % (1/20)

Table 6.17: Detailed results for the teeth localization approaches using the *GCHF* and global *RRF*. Outliers are denoted as landmark estimations which are farther than 7 mm away from the annotation of the medical expert.

the localization performance in accuracy and precision. This strengthens the assumption and conclusion from section 6.2.2, that the varying structures within the head tend to vote imprecisely and inaccurately to landmarks and therefore destroys the overall localization performance although the Euclidean weighting term is used.

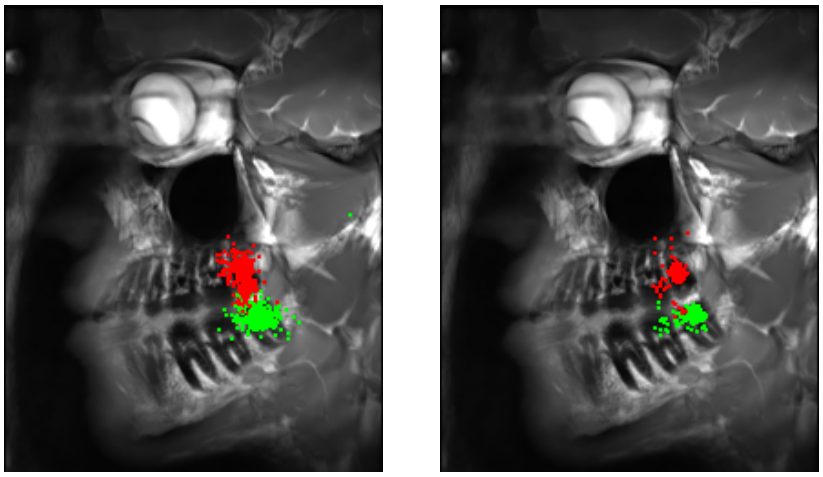
Investigating the results from the local *RRF* from Tab. 6.17 in more detail shows that the worst results come from missing third molars. Deeper investigations of the outliers with errors larger than 7 mm show that 13 of 18 outliers (7+11) switch to the neighboring second molars or are pulled towards them. The remaining few outliers are the top third molars. They are located to the right of the second molars but are predicted above the possible estimated position from the medical expert. This might be due to growing top third molars which are located at this position in the training-sets. As can be seen, the forest is able to predict 33 of 51 (62/2+20) missing teeth within a range of 7 mm.

The outliers from the volumes which contain all third molars are mainly due to two reasons. First, 12 of the 20 outliers have just started growing and therefore are in an early development stage. A possible explanation which is observed is that the data contains only few of such young staged third molars. The remaining outliers are caused by third molars

which are rotated up to 90 degrees or have a strong unnatural translation or appearance.

Comparing the results from Tab. 6.17 of the more local *GCHF* and the global *RRF* which use the whole image information yields to more conclusions. The results from the global *RRFs* is worse if all teeth are visible within the volumes. This is mainly due to the reasons mentioned in section 6.2.2. However, it seems that some outliers in which all teeth or one tooth is missing can be fixed by the global approach but at the cost of reduced accuracy. This is since the global *RRF* does not emphasize the tooth region as much as the more local approach.

See also Fig. 6.21 for a qualitatively comparison between the global *RRF* and the local *GCHF* approach. Note that the results of the *GCHF* approach are more precise than the *RRF* approach.



(a) *RRF* **(b)** *GCHF*

Figure 6.21: Qualitative comparison of the global *RRF* and the local *GCHF* approach for the *MRTM* dataset.

6.3.2.2 Conclusion

Constraining the area from where voxels are selected for training seems to have a huge positive influence for the *MRTM* dataset. The precision and accuracy can be significantly improved by using the more fixed and lower varying structures at the tooth region. However, since still landmark predictions often switch to neighboring second molars some kind of subsequent image processing has to be performed to get rid of the remaining outliers for a reliable fully automated age assessment system.

6.3.3 Hand Dataset

Hand volumes contain multiple similar structures, e.g. finger tips. Therefore, it is the most challenging dataset in terms of switching of landmarks to wrong positions. To locate hand landmarks, an *HF* is trained using local features, i.e. the range has been chosen to cover roughly the width of a finger. To incorporate also some global shape information, at a certain depth long-range features are generated in addition to the local ones. Therefore, the setup depicted in Tab. 6.18 is used to train a *GCHF*.

Parameter	Value
Number of trees	8
Maximal depth	18
Voxel selection range	9 mm
Maximal size of local features	3 mm
Maximal distance of local features	10 mm
Maximal size of long-distance features	30 mm
Maximal distance of long-distance features	50 mm
Starting depth of long-distance features (see Fig. 6.22)	5

Table 6.18: *RRF* parameter setup for hand- and wrist-bone localization using *GCHF*.

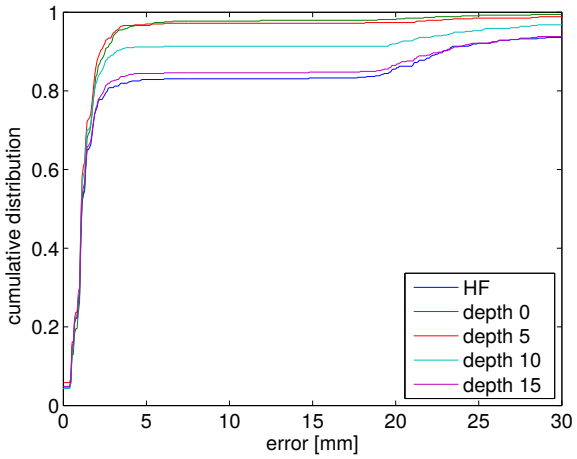


Figure 6.22: Results of parameter optimization for introducing long-distance features starting at certain depths for hand- and wrist-bone localization.

6.3.3.1 Results and Discussion

The results for the hand are depicted in Fig. 6.23 and Tab. 6.19. Additionally, the results from section 6.2.1 denoted as *RRF* are compared to the new results, although again the parameters have slightly changed.

From Fig. 6.23 it can be seen, that the local *GCHF* approach is by far better than the

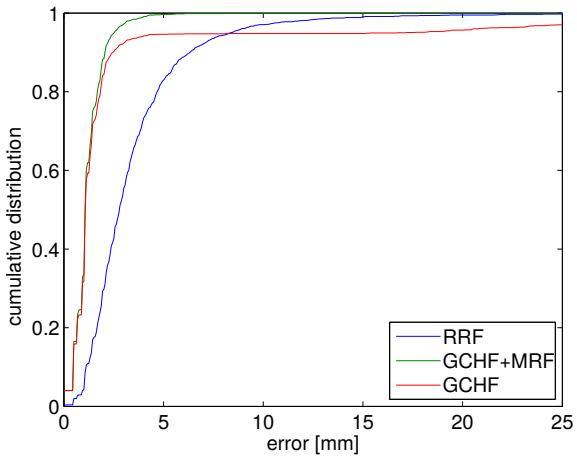


Figure 6.23: Results of the local *GCHF* approach for hand- and wrist-bone localization. Additionally, a Markov Random Field (MRF) is applied to the responses of the *GCHF* approach. All results are compared to the best global *RRF* approach.

Type	**Error**	**# Outlier $\geq$ 10 mm**
GCHF	2.91 ± 8.63	4.94 % (83/1680)
RRF	3.49 ± 2.93	2.80 % (47/1680)
GCHF + MRF	1.23 ± 0.86	0.06 % (1/1680)

Table 6.19: Results of the local *GCHF* approach for hand- and wrist-bone localization. Additionally, an *MRF* is applied to the responses of the local approach. All results are compared to the best global *RRF* approach.

global *RRF* approach up to a localization percentage of around 95 % which is mainly due to the restricted number of voxels. For instance 95 % of all the landmarks are correctly located using the local approach. However, the remaining estimations are located at similar looking positions with the image which limits the overall localization performance of the *GCHF* approach. Therefore, the global approach might be preferable since it locates each landmark within a smaller error than the local approach. In this case, a more local second forest on the top of the global *RRF* approach can help to improve further in precision and accuracy which is shown in [24]. For instance the author of [24] was able to predict the anatomical landmarks for this dataset up to an error of 1.4 $\pm$ 1.5 mm with only six outliers greater than 10 mm.

Since the *GCHF* is trained on more local image information of landmarks, it yields to multiple landmark predictions for each landmark. Therefore a non-maxima suppression is applied to the response images from the *GCHF* with a window-size of 12 mm which is roughly the width of a finger. From the non-maxima suppression, the 10 highest responses for each landmark are chosen to be possible landmark candidates. Since one of these landmark candidates belongs to the correct position, a generative model is used to find the best overall landmark configuration. Therefore an *MRF* is built and used to model

connections between certain landmarks. This discrete optimization model is developed and discussed in appendix C. The results which are retrieved from the *MRF* on top of the local *GCHF* are denoted as *GCHF+MRF*.

It can be seen that the overall performance drastically improves, which is due to fixing almost each outlier from the local *GCHF*. However, still one landmark estimation shows an error greater than 10 mm. This outlier is illustrated in Fig. 6.24. Due to an unnaturally bent little finger, the *GCHF* does not result in a response at this particular position. Therefore, also the *MRF* results in a wrong localization.

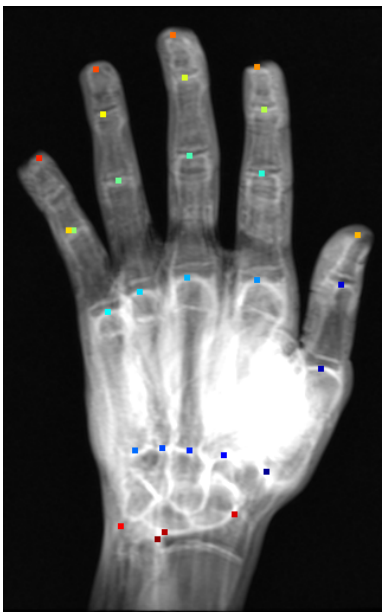


Figure 6.24: Landmark on the little finger for which no response exists from the *RRF*.

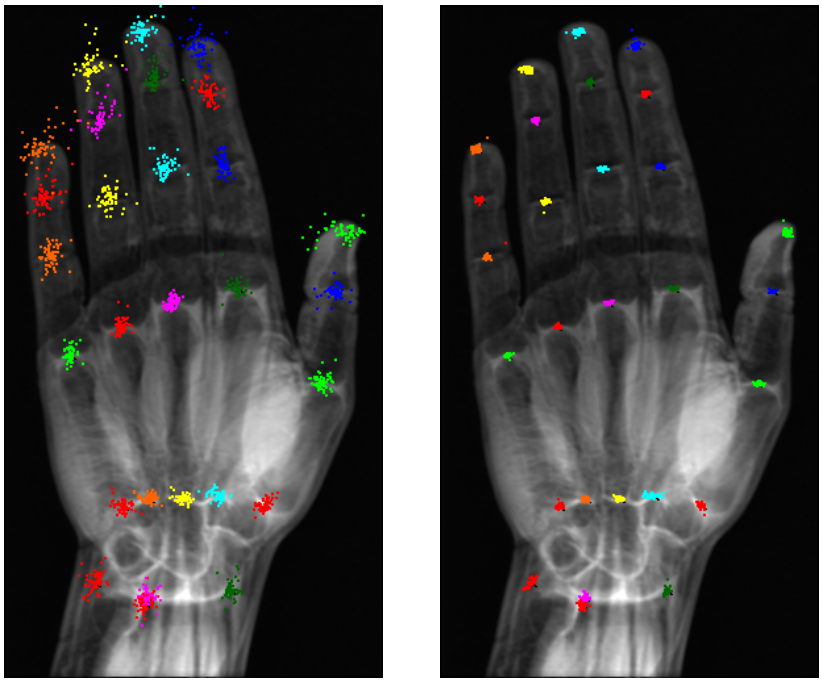
A qualitative comparison between the global *RRF* and the local *GCHF+MRF* approach is depicted in Fig. 6.25.

6.3.3.2 Conclusion

The local *GCHF* seems to work for the *3D-MRH* dataset in a restricted way. The precision and accuracy can only improve for around 95 % of all landmarks. However, since all these landmarks are globally constrained, a generative model can be used to improve on the remaining outliers. This is done by using an *MRF* which results in superior results.

6.3.4 Clavicle Dataset

Clavicle volumes have a quite clear and almost unique structure at the landmark positions. However, the strong variation and the small dataset are the main challenges. A *GCHF* is trained using the parameters depicted in Tab. 6.20.



(a) *RRF* **(b)** *GCHF+MRF*

Figure 6.25: Qualitative comparison of the global *RRF* and the local *GCHF+MRF* approach for the *3D-MRH* dataset.

Parameter	Value
Number of trees	8
Maximal depth:	18
Voxel selection range:	± 10 mm
Maximal size of local features:	8 mm
Maximal distance of local features:	20 mm
Maximal size of long-distance features:	20 mm
Maximal distance of long-distance features:	40 mm
Starting depth of long-distance features (see Tab. 6.21):	5

Table 6.20: *RRF* parameter setup for clavicle localization using *GCHF*.

Starting depth	Error	# Outlier ≥ 10 mm
HF	4.25 ± 2.45	3.57 %
0	4.50 ± 2.44	5.36 %
5	$\mathbf{4.10 \pm 2.35}$	**1.79 %**
10	3.96 ± 2.40	2.38 %
15	5.82 ± 9.94	7.74 %

Table 6.21: Results of parameter optimization for introducing long-distance features starting at certain depths for clavicle localization.

6.3.4.1 Results and Discussion

A leave-three-out cross-validation results in the errors depicted in Fig. 6.26 and Tab. 6.22. It can be seen that the error and the number of outliers decrease if only local information is used. The remaining three outliers are all predicted near to the landmark and do not converge to other positions like in the last two datasets. Further, the results illustrate the impact of the low number of training-data due to the weak localization performance as also depicted qualitatively in Fig. 6.27. This might be also due to the low resolution of the *MR* volumes of around one millimeter voxel-size in each dimension.

Type	**Error**	**# Outlier ≥ 10 mm**
GCHF	4.10 ± 2.35	1.79 % (3/168)
RRF	5.59 ± 2.93	8.33 % (14/168)

Table 6.22: Results of the local *GCHF* approach for clavicle localization are compared to the best global *RRF* approach.

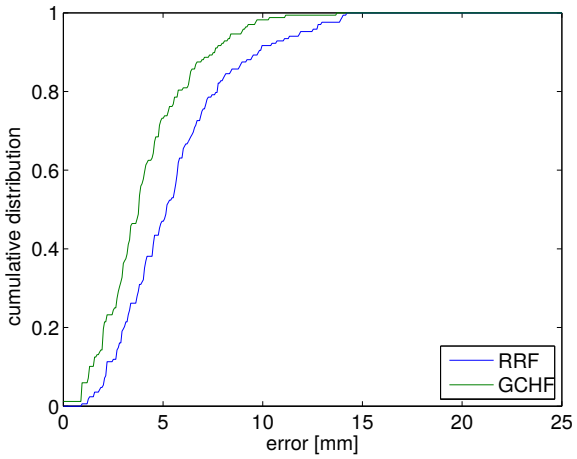
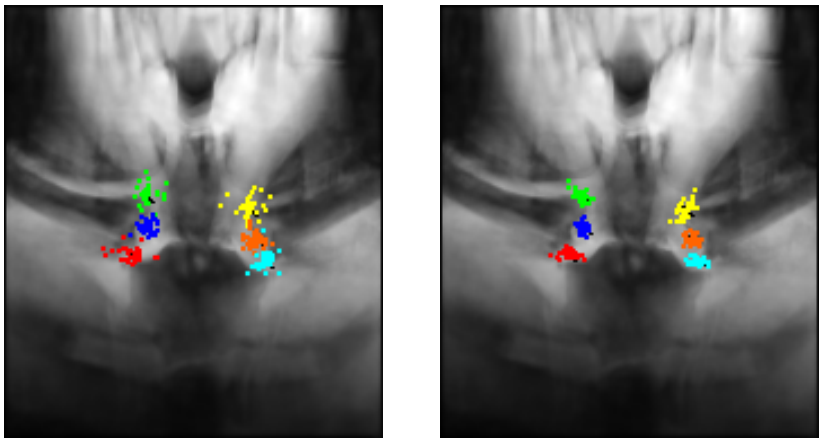


Figure 6.26: Results of the local *GCHF* approach for clavicle localization are compared to the best global *RRF* approach.

However, since the experiments of the last two dataset show significant improvements, this behavior is expected also for this dataset as soon as more annotated data is available.

6.3.4.2 Conclusion

Since this dataset contains too few data to locate landmarks, it is difficult to give any clear conclusions. The *GCHF* approach in this section works for the other two datasets quite well and therefore it is also assumable that increasing the number of training data leads to a strong improvement on the localization results for this dataset. However, one parameter which may prevent to reach an accuracy and precision of one or two millimeter is the low resolution of the volumes, as depicted in chapter 5.



(a) *RRF* **(b)** *GCHF*

Figure 6.27: Qualitative comparison of the global *RRF* and the local *GCHF* approach for the *MRC* dataset.

7 Conclusion and Outlook

Contents

7.1 Conclusion

The Ludwig Boltzmann Institute for Clinical Forensic Imaging in Graz performs research on human age assessment based on the non-invasive and ionizing radiation free Magnetic Resonance (MR) technique. Therefore, *MR* volumes of left hands, third molars and clavicles can be used. Since it takes a long time to analyze them manually, a worth-while goal is to develop an automatic age assessment system. This work focuses on part of such a system, namely the automatic localization of anatomical landmarks which are placed at structures from which the age can be determined. Therefore, the Random Regression Forest (RRF) framework based on the work of [24] has been investigated more deeply.

Firstly, a geodesic metric for the vote-weighting approach for the 2D MR Hand (2D-MRH) dataset has been developed. This approach is based on the idea that votes from voxels should be weighted based on the underlying structures between the voxels position and to the position to where it votes. The results led to the conclusion, that long-distance votes and votes from fingers to landmarks on other fingers only have a small influence to an overall reliable localization result. This conclusion is also strengthened by the third experiment of the Globally-Constrained Hough-Forests (GCHFs).

Second, a novel feature-generation approach has been introduced which analyzes surrounding structures to which a voxel votes to, i.e. the On-Landmark Feature Generation (OLFG) method. This approach is investigated since in related methods the voxels have no knowledge to where they vote. The *OLFG* shows better results for the 3D MR Hand (3D-MRH) and *2D-MRH* datasets compared to an *RRF* with only randomly trained fea-

tures. For the 3D MR Third Molar (MRTM) dataset no improvements have been found, most likely due to strong anatomical variations in the medical images and the small molars.

A third contribution of this work is the research on using more local information around landmarks to train a *GCHF* which shows promising results. Using the whole image information to train an *RRF* seems to yield to inaccurate and imprecise localizations. However, since the *GCHF* is locally trained, it results in multiple landmark candidates which led to wrong landmark predictions. Therefore, global information is incorporated with long-range features instead of directly using the whole image information. Nevertheless, this still yields to a few wrong landmark localizations especially for the *3D-MRH* dataset which in most cases are fixed with a Markov Random Field (MRF).

In general, one has to think carefully about when to use local or global approaches. This is mainly based on the underlying dataset. In the work of [24] the biggest vulnerability is the first global *RRF* localization step to locate landmarks in the *3D-MRH* dataset. If this step fails, a second more local *RRF* is hardly able to refine this localization to the correct position. On the other hand, also the *GCHF* predicts landmarks at wrong positions. But since the *GCHF* yields to multiple landmark candidates, regularization steps which model the global landmark configuration can be used to find a final good solution. Therefore and due to the results, a locally trained approach incorporating also global information with an additional regularization step might be the better option for the *3D-MRH* dataset in future. For the *MRTM* dataset the *GCHF* is preferable over a global *RRF* approach. Due to strong varying shapes and appearances in the dataset a global *RRF* leads to imprecise results. It seems to be better that the *RRF* concentrates more directly at the structures at third molars and their neighborhood. For the 3D MR Clavicle (MRC) dataset it is difficult to say which approach might be better due to the small amount of data given. But since the results of the *GCHF* approach for the *3D-MRH* and the *MRTM* datasets are promising, it is likely that the *GCHF* also improves on results at the *MRC* dataset compared to a global *RRF* approach.

Altogether and as an outcome of the results, the more locally trained *RRF* methods seem to be better suited for a practically relevant usage. Investigating them in more detail and incorporating other techniques for fixing the global landmark configuration may be the better option to go.

7.2 Outlook

During this work, several new topics were raised which might be worth to look into in more detail. For instance, results on the *2D-MRH* dataset using the whole image information showed a superior localization performance. Therefore, estimating a landmark first in 2D images and afterwards in the third dimension, might lead to improved results. Also the idea of using local surrounding structures of landmarks for training seems to be promising for further research. Especially, the incorporation of more global information to the local surround structures seems to be favorable over using only local image information.

A

List of Acronyms

2D-MRH	2D MR Hand
3D-MRH	3D MR Hand
AAM	Active Appearance Model
ASM	Active Shape Model
CART	Classification and Regression Tree
CBCT	Cone Beam Computed Tomography
CT	Computed Tomography
EDVW	Euclidean Distance Vote Weighting
GCHF	Globally-Constrained Hough-Forest
GDVW	Geodesic Distance Vote Weighting
HF	Hough-Forest
MAP	Maximum A Posteriori
MIP	Medical Image Processing
MR	Magnetic Resonance
MRC	3D MR Clavicle
MRF	Markov Random Field
MRTM	3D MR Third Molar
OLFG	On-Landmark Feature Generation
PCA	Principal Component Analysis
PD-TSE	Proton-Density Weighted Turbo Spin Echo Sequence
PET	Positron Emission Tomography
RF	Random Forest
RFG	Random Feature Generation
RRF	Random Regression Forest
SSM	Statistical Shape Model
SVM	Support Vector Machine

List of Publications

Automatic Third Molar Localization from 3D MRI using Random Regression Forests

Walter Unterpirker, Thomas Ebner, Darko Stern and Martin Urschler
In: *Medical Image Understanding and Analysis - MIUA 2015 (19th Annual Conference)*
July 15-17, 2015, University of Lincoln, Lincoln, UK
Accepted for poster presentation

Abstract: Radiological age estimation of living subjects from MR images has recently become very popular. Besides skeletal ossification this can be done using the mineralization status of wisdom teeth. To support potential automatic age estimation, an important preliminary step is a reliable and automatic localization of the wisdom teeth. Therefore, we propose a random regression forest framework to localize third molars, which is capable to predict landmarks up to an error of 3.55 ± 2.62 mm in mean and standard deviation in a challenging 3D MRI dataset.

C

Markov Random Field

Contents

In this appendix, an *MRF* algorithm is developed to regularize multiple landmark candidate positions retrieved from the *GCHF* approach. The *MRF* is only used and specialized for hand volumes since there multiple false positive landmarks are located.

An *MRF* is an undirected graph containing nodes and modeled edges between some or all of them. Each edge and each node has to be modeled. This consists of generating a connected graph structure between several nodes, calculating weights for each edge and assigning weights to each node. The modeling can be done either manually which is done in this work or automatically as done for example in [19]. Having defined the graph and weights, a new over-determined configuration of nodes can be solved by using the *MRF* to find the best overall configuration based on the built model. For this work following formulation is used to retrieve a final landmark configuration from multiple possible retrieved landmark candidates:

$$\mathcal{C}(M) = \sum_{\forall l \in L} \mathcal{P}(M_l) + \sum_{\forall e \in E} \mathcal{G}(M_{n,i}, M_{n,j}) \tag{C.1}$$

Unary terms $\mathcal{P}$ model the node confidence value according to the matches M_l, i.e. node weights. Contrary to the unary terms, binary terms $\mathcal{G}$ define the strength of the relationship between two nodes or matches $M_{n,i}$ and $M_{n,j}$, i.e. edge weights.

C.1 Graph Topology

A manual graph topology as depicted in Fig. C.1 was used to address the challenge of finding the best overall landmark configuration for the *3D-MRH* dataset.

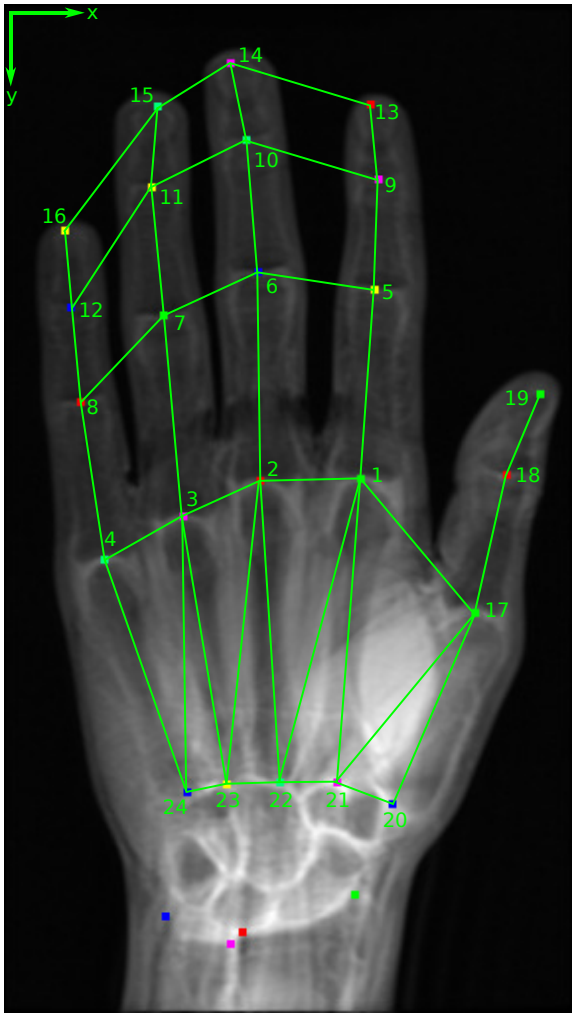


Figure C.1: Hand-crafted *MRF* graph topology.

C.2 Node Weights $\mathcal{P}$

All nodes are weighted according to the responses from the *GCHF*, normalized between 0 and 1. Further, maximal 10 landmark candidates per each landmark are used.

C.3 Edge Weights $\mathcal{G}$

Edges between two nodes $M_{n,i}$ and $M_{n,j}$ at position $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ are weighted according to a statistical distribution built from the training-data. Therefore, first distances between all nodes shown in Fig. C.1 are calculated in terms of mean Euclidean distance μ and standard-deviation σ. Afterwards, a Gaussian distribution is used to model the edge weights based on the statistics as followed:

$$p(M_{n,i}, M_{n,j}) = exp\left(-\frac{1}{2} \cdot \left(\frac{\mu - ||\boldsymbol{c}_i - \boldsymbol{c}_j||}{\sigma}\right)^2\right) \tag{C.2}$$

In general, all connections are modeled by setting the edge weights to $\mathcal{G}(M_{n,a}, M_{n,b}) = p(M_{n,i}, M_{n,j})$. However, taking only the distances between two nodes into account, interchanges between nodes can happen since no direction conditions are modeled until yet. Therefore, a hard-sorting in x- and y-coordinates is incorporated using following conditions:

- X-Sorting between nodes along x-coordinates $c_{i,x}$ with i in one of following sets: $[1, 2, 3]$, $[5, 6, 7]$, $[9, 10, 11]$ and $[13, 14, 15]$, according to Fig. C.1:

$$\mathcal{G}(M_{n,a}, M_{n,b}) = \begin{cases} p(M_{n,i}, M_{n,j}) & \text{if } \boldsymbol{c}_{i+1,x} + 4mm \leq \boldsymbol{c}_{i,x} \\ 0 & \text{else} \end{cases}$$

- Y-Sorting of nodes along y-coordinates $c_{i,y}$ with i in one of following sets: $[1, 5, 9]$, $[2, 6, 10]$, $[3, 7, 11]$ and $[4, 8, 12]$ according to Fig. C.1:

$$\mathcal{G}(M_{n,a}, M_{n,b}) = \begin{cases} p(M_{n,i}, M_{n,j}) & \text{if } \boldsymbol{c}_{i+4,y} \leq \boldsymbol{c}_{i,y} \\ 0 & \text{else} \end{cases}$$

- Y-Sorting of nodes along y-coordinates of the thumb $c_{i,y}$ with i in: $[17, 18]$ according to Fig. C.1:

$$\mathcal{G}(M_{n,a}, M_{n,b}) = \begin{cases} p(M_{n,i}, M_{n,j}) & \text{if } \boldsymbol{c}_{i+1,y} \leq \boldsymbol{c}_{i,y} \\ 0 & \text{else} \end{cases}$$

C.4 Solving Markov Random Fields

To find the best overall configuration of the modeled graph structure, the Maximum A Posteriori (MAP) estimation is calculated which is equivalent to maximize the energy:

$$\mathcal{C}^* = \underset{M}{argmax}\ \mathcal{C}(M) = \underset{M}{argmax}\ \left(\sum_{\forall l \in L} \mathcal{P}(M_l) + \sum_{\forall e \in E} \mathcal{G}(M_{n,a}, M_{n,b})\right) \tag{C.3}$$

However, in general the exact solution or inference of MAP is $\mathcal{NP}$-hard. Therefore, several approximations exist but may converge in local minima. In this work a sum-product message passing also known as belief propagation [76] is used to solve the graph within the framework of [67][1].

The belief propagation algorithm iteratively passes so called messages between connected nodes over edges. At first, all messages are initialized to 1, resembling a totally uninformative message. From a node n_1 a message is passed to a connected node n_2 in which this message is updated by the knowledge of n_2 and the edge information between these two nodes. Thus the message from n_1 to n_2 is updated. Afterwards this message is propagated to another connected node n_3 of n_2. Note that a message from n_1 which is updated by the knowledge of n_2 is *not* back propagated to n_1. Applying this message passing algorithm to all connected nodes, an overall *best* approximated configuration can be found which can lead to the global optimum in some cases.

[1] `http://www.cs.ubc.ca/~schmidtm/Software/UGM.html` (online; last accessed on September 29, 2015)

Bibliography

[1] Borgefors, G. (1986). Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344 – 371. (page 45, 46)

[2] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140. (page 35)

[3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. (page 9, 20, 21, 31)

[4] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press. (page 31)

[5] Brown, L. G. (1992). A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376. (page 2)

[6] Cheng, E., Chen, J., Yang, J., Deng, H., Wu, Y., Megalooikonomou, V., Gable, B., and Ling, H. (2011). Automatic dent-landmark detection in 3D CBCT dental volumes. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 6204–6207. (page 22, 23)

[7] Cootes, T., Edwards, G., and Taylor, C. (2001). Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685. (page 2, 15, 16)

[8] Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1995). Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59. (page 2, 15)

[9] Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., and Siddiqui, K. (2013). Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical Image Analysis*, 17(8):1293 – 1303. (page 23, 24, 37, 40)

[10] Criminisi, A. and Shotton, J. (2013). *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media. (page 31, 32, 33, 34, 37, 41)

[11] Criminisi, A., Shotton, J., and Bucciarelli, S. (2009). Decision forests with long-range spatial context for organ localization in CT volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 69–80. (page 23, 24, 26, 40)

[12] Cristinacce, D. and Cootes, T. (2008). Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054 – 3067. (page 28)

[13] Cuingnet, R., Prevost, R., Lesage, D., Cohen, L., Mory, B., and Ardon, R. (2012). Automatic detection and segmentation of kidneys in 3D CT images using random forests.

In Ayache, N., Delingette, H., Golland, P., and Mori, K., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012*, volume 7512 of *Lecture Notes in Computer Science*, pages 66–74. Springer Berlin Heidelberg. (page 26, 27)

[14] Cutolo, F., Parchi, P. D., and Ferrari, V. (2014). Video see through AR head-mounted display for medical procedures. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 393–396. (page 1)

[15] Demirjian, A., Goldstein, H., and Tanner, J. M. (1973). A new system of dental age assessment. *Human Biology*, 45(2):pp. 211–227. (page 7, 8)

[16] Denolf, P., Vanderveken, O. M., Marklund, M., and Braem, M. J. (2015). The status of cephalometry in the prediction of non-CPAP treatment outcome in obstructive sleep apnea patients: a literature review. *Sleep Medicine Reviews*. (page 2)

[17] Donner, R., Langs, G., Micusik, B., and Bischof, H. (2010). Generalized sparse MRF appearance models. *Image and Vision Computing*, 28(6):1031 – 1038. (page 20)

[18] Donner, R., Menze, B., Bischof, H., and Langs, G. (2013a). Fast anatomical structure localization using top-down image patch regression. In Menze, B., Langs, G., Lu, L., Montillo, A., Tu, Z., and Criminisi, A., editors, *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, volume 7766 of *Lecture Notes in Computer Science*, pages 133–141. Springer Berlin Heidelberg. (page 18, 19, 20)

[19] Donner, R., Menze, B. H., Bischof, H., and Langs, G. (2013b). Global localization of 3D anatomical structures by pre-filtered hough forests and discrete optimization. *Medical Image Analysis*, 17(8):1304 – 1314. (page 28, 29, 30, 91)

[20] Dünkel, F., Van Kalmthout, A., and Schüler-Springorum, H. (1997). Entwicklungstendenzen und Reformstrategien im Jugendstrafrecht im europäischen Vergleich. Forum Verlag Godesberg. (page 5)

[21] Duy, N., Lamecker, H., Kainmueller, D., and Zachow, S. (2012). Automatic detection and classification of teeth in CT data. In Ayache, N., Delingette, H., Golland, P., and Mori, K., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 609–616. Springer Berlin Heidelberg. (page 26)

[22] Dvorak, J., George, J., Junge, A., and Hodler, J. (2007a). Age determination by magnetic resonance imaging of the wrist in adolescent male football players. *British Journal of sports medicine*, 41(1):45–52. (page 8)

[23] Dvorak, J., George, J., Junge, A., and Hodler, J. (2007b). Application of MRI of the wrist for age determination in international U-17 soccer competitions. *British Journal of sports medicine*, 41(8):497–500. (page 8)

[24] Ebner, T. (2014). Localization of hand bones from 3D magnetic resonance images for bone age estimation. Master's thesis, Graz University of Technology. (page 2, 15, 31, 37, 38, 40, 41, 42, 44, 45, 54, 55, 56, 57, 59, 60, 61, 62, 66, 80, 85, 86)

[25] Ebner, T., Stern, D., Donner, R., Bischof, H., and Urschler, M. (2014). Towards automatic bone age estimation from MRI: Localization of 3D anatomical landmarks. In *Medical Image Computing and Computer-Assisted Intervention*. MICCAI 2014. (page 2, 9, 24, 25, 40, 66)

[26] Fitzpatrick, J. M. and Sonka, M., editors (14 June 2000). *Handbook of Medical Imaging, Volume 2: Medical Image Processing and Analysis*. SPIE, the international society for optics and photonics. (page 1, 2)

[27] Frantz, S., Rohr, K., and Stiehl, H. (2000). Localization of 3D anatomical point landmarks in 3D tomographic images using deformable models. In Delp, S., DiGoia, A., and Jaramaz, B., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2000*, volume 1935 of *Lecture Notes in Computer Science*, pages 492–501. Springer Berlin Heidelberg. (page 14)

[28] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139. (page 20, 21)

[29] Gall, J. and Lempitsky, V. (2009). Class-specific hough forests for object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1022–1029. (page 27, 55)

[30] Gall, J., Yao, A., Razavi, N., Van Gool, L., and Lempitsky, V. (2011). Hough forests for object detection, tracking, and action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2188–2202. (page 27, 55)

[31] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42. (page 28)

[32] Gilsanz, V. and Ratib, O. (2005). *Hand bone age: a digital atlas of skeletal maturity*. Springer Science & Business Media. (page 3, 4, 5, 6)

[33] Glocker, B., Zikic, D., Konukoglu, E., Haynor, D., and Criminisi, A. (2013). Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In Mori, K., Sakuma, I., Sato, Y., Barillot, C., and Navab, N., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 262–270. Springer Berlin Heidelberg. (page 23)

[34] Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):pp. 285–339. (page 15)

[35] Greulich, W. W. and Pyle, S. I. (1959). Radiographic atlas of skeletal development of the hand and wrist. *The American Journal of the Medical Sciences*, 238(3):393. (page 8)

[36] Hammernik, K., Ebner, T., Stern, D., Urschler, M., and Pock, T. (2015). Vertebrae segmentation in 3D CT images based on a variational framework. In Yao, J., Glocker, B., Klinder, T., and Li, S., editors, *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lecture Notes in Computational Vision and Biomechanics*, pages 227–233. Springer International Publishing. (page 2)

[37] Han, D., Gao, Y., Wu, G., Yap, P.-T., and Shen, D. (2014). Robust anatomical landmark detection for MR brain image registration. In Golland, P., Hata, N., Barillot, C., Hornegger, J., and Howe, R., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014*, volume 8673 of *Lecture Notes in Computer Science*, pages 186–193. Springer International Publishing. (page 2)

[38] Hjern, A., Brendler-Lindqvist, M., and Norredam, M. (2012). Age assessment of young asylum seekers. *Acta Paediatrica*, 101(1):4–7. (page 3)

[39] Hutton, T. J., Cunningham, S., and Hammond, P. (2000). An evaluation of active shape models for the automatic identification of cephalometric landmarks. *European Journal of Orthodontics*, 22(5):499–508. (page 16, 17)

[40] Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., and van Ginneken, B. (2009). Multi-atlas-based segmentation with local decision fusion - application to cardiac and aortic segmentation in CT scans. *Medical Imaging, IEEE Transactions on*, 28(7):1000–1010. (page 13)

[41] Johnston, C., Burden, D., Kennedy, D., Harradine, N., and Stevenson, M. (2006). Class {III} surgical-orthodontic treatment: A cephalometric study. *American Journal of Orthodontics and Dentofacial Orthopedics*, 130(3):300 – 309. (page 2)

[42] Joseph V. Hajnal, Derek L.G. Hill, D. J. H., editor (2001). *Medical Image Registration*. CRC Press. (page 2)

[43] Kellinghaus, M., Schulz, R., Vieth, V., Schmidt, S., and Schmeling, A. (2010). Forensic age estimation in living subjects based on the ossification status of the medial clavicular epiphysis as revealed by thin-slice multidetector computed tomography. *International Journal of Legal Medicine*, 124(2):149–154. (page 5)

[44] Kelm, B. M., Wels, M., Zhou, S. K., Seifert, S., Suehling, M., Zheng, Y., and Comaniciu, D. (2013). Spine detection in CT and MR using iterated marginal space learning. *Medical Image Analysis*, 17(8):1283 – 1292. (page 15)

[45] Knell, B., Ruhstaller, P., Prieels, F., and Schmeling, A. (2009). Dental age diagnostics by means of radiographical evaluation of the growth stages of lower wisdom teeth. *International Journal of Legal Medicine*, 123(6):465–469. (page 7)

[46] Kontschieder, P., Kohli, P., Shotton, J., and Criminisi, A. (2013). Geof: Geodesic forests for learning coupled predictors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 65–72. (page 44)

[47] Lindner, C., Bromiley, P., Ionita, M., and Cootes, T. (2014). Robust and accurate shape model matching using random forest regression-voting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1–1. (page 14, 26, 27, 28, 42)

[48] Lindner, C. and Cootes., T. (2015). Fully automatic cephalometric evaluation using random forest regression-voting. *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI) 2015 Grand Challenges in Dental X-ray Image Analysis - Automated Detection and Analysis for Diagnosis in Cephalometric X-ray Image, 2015.* (page 2, 28)

[49] Lindner, C., Thiagarajah, S., Wilkinson, J., Consortium, T., Wallis, G., and Cootes, T. (2013). Fully automatic segmentation of the proximal femur using random forest regression voting. *Medical Imaging, IEEE Transactions on*, 32(8):1462–1472. (page 11, 28)

[50] Malina, R. (2011). Skeletal age and age verification in youth sport. *Sports Medicine*, 41(11):925–947. (page 3)

[51] Mattes, D., Haynor, D., Vesselle, H., Lewellen, T., and Eubank, W. (2003). PET-CT image registration in the chest using free-form deformations. *Medical Imaging, IEEE Transactions on*, 22(1):120–128. (page 2)

[52] Mory, B., Somphone, O., Prevost, R., and Ardon, R. (2012). Real-time 3D image segmentation by user-constrained template deformation. In Ayache, N., Delingette, H., Golland, P., and Mori, K., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 561–568. Springer Berlin Heidelberg. (page 26)

[53] Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1 – 23. (page 1)

[54] Olze, A., Reisinger, W., Geserick, G., and Schmeling, A. (2006). Age estimation of unaccompanied minors: Part ii. dental aspects. *Forensic Science International*, 159, Supplement(0):S65 – S67. International IOFOS Symposium on Forensic Odontology 2006 and 3rd International Conference on Reconstruciton of Soft Facial Parts 2006. (page 7)

[55] Olze, A., Schmeling, A., Taniguchi, M., Maeda, H., van Niekerk, P., Wernecke, K.-D., and Geserick, G. (2004). Forensic age estimation in living subjects: the ethnic factor in

wisdom tooth mineralization. *International Journal of Legal Medicine*, 118(3):170–173. (page 8)

[56] Pietka, E., Gertych, A., Pospiech, S., Cao, F., Huang, H., and Gilsanz, V. (2001). Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal roi extraction. *Medical Imaging, IEEE Transactions on*, 20(8):715–729. (page 12)

[57] Ranjan, S. (2011). Organ localization through anatomy-aware non-rigid registration with atlas. In *Applied Imagery Pattern Recognition Workshop (AIPR), 2011 IEEE*, pages 1–5. (page 13)

[58] Rohr, K. (1997). On 3D differential operators for detecting point landmarks. *Image and Vision Computing*, 15(3):219 – 233. (page 12)

[59] Rohr, K. (2001). *Landmark-Based Image Analysis: Using Geometric and Intensity Models*. Kluwer Academic Publishers. (page 2, 12)

[60] Rohr, K., Stiehl, H. S., Sprengel, R., Buzug, T. M., Weese, J., and Kuhn, M. (2001). Landmark-based elastic registration using approximating thin-plate splines. *Medical Imaging, IEEE Transactions on*, 20(6):526–534. (page 12)

[61] Salah, A. and Akarun, L. (2006). 3d facial feature localization for registration. In Gunsel, B., Jain, A., Tekalp, A., and Sankur, B., editors, *Multimedia Content Representation, Classification and Security*, volume 4105 of *Lecture Notes in Computer Science*, pages 338–345. Springer Berlin Heidelberg. (page 11)

[62] Schmeling, A., Grundmann, C., Fuhrmann, A., Kaatsch, H.-J., Knell, B., Ramsthaler, F., Reisinger, W., Riepert, T., Ritz-Timme, S., Rösing, F., Rötzscher, K., and Geserick, G. (2008). Criteria for age estimation in living individuals. *International Journal of Legal Medicine*, 122(6):457–460. (page 4, 5)

[63] Schmeling, A., Olze, A., Reisinger, W., and Geserick, G. (2004a). Forensic age diagnostics of living people undergoing criminal proceedings. *Forensic Science International*, 144(2-3):243 – 245. (page 4, 8)

[64] Schmeling, A., Prieto, J. L., Landa, M. I., and Garamendi, P. M. (2011). *Forensic age estimation in unaccompanied minors and young living adults*. INTECH Open Access Publisher. (page 4, 7)

[65] Schmeling, A., Reisinger, W., Loreck, D., Vendura, K., Markus, W., and Geserick, G. (2000). Effects of ethnicity on skeletal maturation: consequences for forensic age estimations. *International Journal of Legal Medicine*, 113(5):253–258. (page 8)

[66] Schmeling, A., Schulz, R., Reisinger, W., Mühler, M., Wernecke, K.-D., and Geserick, G. (2004b). Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. *International Journal of Legal Medicine*, 118(1):5–8. (page 5)

[67] Schmidt, M. (2007). Ugm: A matlab toolbox for probabilistic undirected graphical models. (page 94)

[68] Shimizu, A., Ohno, R., Ikegami, T., Kobatake, H., Nawano, S., and Smutek, D. (2007). Segmentation of multiple organs in non-contrast 3D abdominal CT images. *International Journal of Computer Assisted Radiology and Surgery*, 2(3-4):135–142. (page 13)

[69] Smith, S. M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P., Federico, A., and Stefano, N. D. (2002). Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17(1):479 – 489. (page 2)

[70] Stern, D., Ebner, T., Bischof, H., Grassegger, S., Ehammer, T., and Urschler, M. (2014). Fully automatic bone age estimation from left hand mr images. In Golland, P., Hata, N., Barillot, C., Hornegger, J., and Howe, R., editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014*, volume 8674 of *Lecture Notes in Computer Science*, pages 220–227. Springer International Publishing. (page 8)

[71] Tanner, J. M., Whitehouse, R., Marshall, W., Healty, M., and Goldstein, H. (1975a). Assessment of skeleton maturity and maturity and prediction of adult height (TW2 method). (page 8)

[72] Tanner, J. M., Whitehouse, R. H., Marshall, W. A., and Carter, B. S. (1975b). Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height. *Archives of Disease in Childhood*, 50(1):14–26. (page 3)

[73] Urschler, M., Grassegger, S., and Stern, D. (2015). What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents. *Annals of Human Biology*, 0(0):1–10. PMID: 26313328. (page 8)

[74] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA. (page 26)

[75] Vieth, V., Schulz, R., Brinkmeier, P., Dvorak, J., and Schmeling, A. (2014). Age estimation in U-20 football players using 3.0 tesla MRI of the clavicle. *Forensic Science International*, 241(0):118 – 122. (page 3, 8)

[76] Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in mrfs. *Advanced Mean Field Methods-Theory and Practice*, pages 229–240. (page 94)

[77] Zheng, Y. and Comaniciu, D. (2014). Constrained marginal space learning. In *Marginal Space Learning for Medical Image Analysis*, pages 79–101. Springer New York. (page 15)

[78] Zheng, Y., Georgescu, B., and Comaniciu, D. (2009). Marginal space learning for efficient detection of 2D/3D anatomical structures in medical images. In Prince, J., Pham, D., and Myers, K., editors, *Information Processing in Medical Imaging*, volume 5636 of *Lecture Notes in Computer Science*, pages 411–422. Springer Berlin Heidelberg. (page 15)

[79] Zhou, S., Georgescu, B., Zhou, X. S., and Comaniciu, D. (2005). Image based regression using boosting method. In *Tenth IEEE International Conference on Computer Vision, 2005.*, volume 1, pages 541–548 Vol. 1. (page 21)

[80] Zhou, S., Zhou, J., and Comaniciu, D. (2007). A boosting regression approach to medical anatomy detection. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007.*, pages 1–8. (page 21)

[81] Zhou, S. K. (2014). Discriminative anatomy detection: Classification vs regression. *Pattern Recognition Letters*, 43(0):25 – 38. (page 14)

[82] Zitová, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000. (page 13)