Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on the robust learning from incomplete and spurious data

Deliverable number	*D2.4*
Dissemination level	*Public*
Delivery data	*due 28.2.2013*
Status	*Final*
Authors	*Georg Langs, René Donner, Dimitrios Markonis, Matthias Dorfer, Henning Mueller*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Executive Summary

This deliverable describes methodology for the computational learning from medical imaging data developed in the KHRESMOI project. Learning is a central part of the project, since it forms the basis for indexing and representing the data to guide the retrieval process. Research on learning methodology is performed throughout the entire project, and the present document summarizes the current status of development. We present three directions in which learning is relevant. Firstly, the recognition and localization of anatomical structures in imaging data is a necessary prerequisite for anatomy specific retrieval, since it allows the adaptation of descriptors to specific organs and contexts. Initial results show that it yields high localization accuracy for anatomical structures. Secondly, reference spaces, or *atlases* are necessary to summarize and index across large populations. They rely on the anatomical, or physiological correspondence among individual data. Experimental evaluation shows that we can build atlases from heterogeneous clinical imaging data. Thirdly, classification approaches contribute categorial information to the retrieval process. Experiments show that they make filtering, and re-ranking of results possible. In this deliverable we present methods tackling these three lines, and report experimental results that high-light current capabilities, and limitations of the methods.

Table of Contents

List of Figures

Notation

$\mathbf{I}_i$	Image or volume with index i. If it is 2D or 3D data will become clear from the context.
$\mathbf{I}_i \in \mathbb{R}^2$	2D data such as images.

Abbreviations

CMC	carpometacarpal
CPU	Central Processing Unit
CT	Computed Tomography
DIP	distal interphalangeal
GPU	Graphics Processing Unit
LBP	Local Binary Patterns
MCP	metacarpophalangeal
MR	Magnetic Resonance
MRF	Markov Random Field
PACS	Picture archiving and communication system
PCA	Principal Component Analysis
PIP	proximal interphalangeal

1 Introduction

Workpackage 2 of KHRESMOI aims at developing and implementing methodology for large scale biomedical image retrieval. Biomedical images, and in particular those images generated in a clinical radiology setting carry certain properties different to arbitrary imaging data.

The most prominent feature of radiology data, is that it stems from a limited domain, the human body (instead of e.g., all vacation destinations on this world). Furthermore, the variability and differences in appearance that are relevant for retrieval, are not necessarily the dominant features in the data. For example, while there is an obvious difference between a liver and a lung, presenting arbitrary lungs to a radiologist, who is searching for a particular lung disease, is not helpful. Instead, retrieval should result in examples exhibiting a disease that matches the query case. Typically the characteristics of diseases are subtle compared to differences among anatomical regions. To summarize,

- Radiological data images the human body or sections of the human body.
- It is crucial to take anatomical context into account, when searching for cases with similar disease based on appearance.

Given these characteristics certain families of learning approaches become relevant. They exploit the ability to generalize across data, to use additional meta information such as the imaging modality, and to further organize indices beyond simple similarity measures, and search results.

1. **Recognition and localization** Methods for the localization of anatomical structures identify organs in imaging data, estimate the position of anatomical landmarks, and segment individual structures. They assign individual voxel labels that indicate the anatomical structure. During retrieval, this is used to identify the anatomical context of a region of interest for which the user searches comparable cases. The majority of previous work concentrates on the localization of specific regions, for which approximate locations are known. Corresponding to the use case needs, in the course of the project we developed methodology that does not need any initialization but localizes organs in images in a fully automated manner.

2. **Reference spaces and atlases** Since the imaging data shows different parts of the human body, we can learn a mapping of each data to a reference space, a so–called atlas. A methodology that solves this mapping was developed in specific contexts, such as neuroimaging. Atlas approaches that encompass the entire body, and are learned from clinical data are novel, and have been developed within Khresmoi.

3. **Categorial information and classifiers** Medical images are acquired in a variety of imaging technologies, or *modalities* (CT, MRI, xray, etc.) that can be extended to an even larger number of image types in the medical literature (graphs, flow charts, etc.). This meta information can be used to filter search results and show only diagnostic images that constitute 99% of the visual information needs. Whereas tabbed browsing using radiology modalities we developed a novel hierarchy of image types and filtering out the unwanted modalities corresponds to a clearly expressed user needs and has to our knowledge never been proposed for biomedical images.

In the following three sections we describe the methods developed and implemented in the project. Part of this work has been published already, and we indicate those sections with the corresponding references in the text.

2 Anatomical Structure Localization

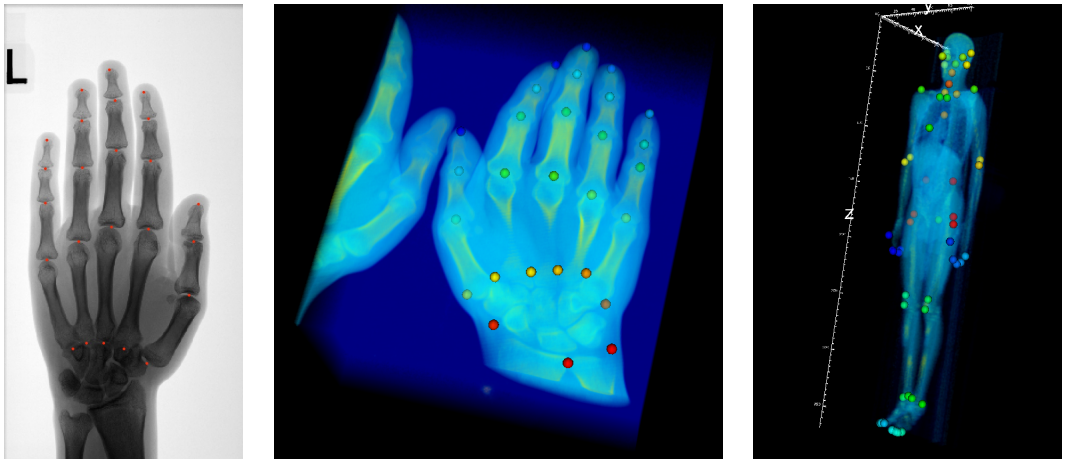
Note that this section is partly published in [22] and [21]. The accurate localization of anatomical landmarks in medical imaging data is a challenging problem, due to rich variability and frequent ambiguity of their appearance. In this section we propose two generic approaches for global landmark localization in 2D and 3D medical imaging data.

The detection of landmarks is relevant in different contexts. Anatomical landmarks can serve as basis for morphometric measurements in anatomical structures for clinical applications such as the assessment of spinal flexibility [40], knee alignment angles [35, 27], joint space narrowing in rheumatoid arthritis [42, 56], or fracture detection in osteoporosis diagnosis [57]. It can facilitate navigation through large medical imaging data by pointing to relevant locations, and is necessary if subsequent image analysis algorithms need initial location estimates as for instance segmentation algorithms, or accurate identification of anatomical landmarks for localized follow-up analysis. In computer aided diagnosis it can serve as a means to identify target regions for further analysis [17].

Segmentation approaches such as Level-Sets [10], Active Shape and Appearance Models [8, 9], Graph Cuts [4], or Random Walks [33] typically require at least a coarse initial localization, either by user input or by an appropriate algorithm to set either seed points (Graph Cuts, Random Walks) or to initialize contour positions [49]. Registration based segmentation approaches such as those applied in brain imaging (e.g., FreeSurfer [28]) typically assume coarse alignment across individuals. While this is the case in adult neuroimaging studies it does not hold for studies involving other anatomical structures, or fetal brain imaging data [16]. To make atlas based approaches applicable to large scale studies on other anatomical structures, prior automatic alignment is necessary, for which landmarks offer a feasible solution. Landmark localization can also be regarded as a form of semantic parsing [61] when point-wise rather than regional information is required. Typically, landmarks are identified by anatomy-specific algorithms that first obtain coarse location estimates for selected landmarks – sometimes aided by manual interaction – and then apply a domain specific model for their refinement. Among the reasons for the difficulties are noise (including local and global intensity changes), cluttered image data (overlapping structures in 2D projections, highly structured background in 3D organ segmentation), and anatomical structures that exhibit a high degree of similarity (e.g., fingers or vertebrae).

We propose an algorithm that copes with these challenges and offers a general approach to accurately localize landmarks without initialization or the need for subsequent refinement.

The method first generates very accurate position hypotheses for landmarks in a global search, and then disambiguates the landmarks based on their spatial configuration. A global prefiltering classification with local Hough Forests yields the hypotheses, while the disambiguation is performed by solving a Markov Random Field (MRF). The method starts from local image information, and includes the global, geometrical information of the anatomical structure in the graph matching step.



(a) (b) (c)

Figure 1: Examples from the three data sets employed in this paper. a) Hand radiographs, b) high resolution hand CTs and c) whole body CTs. The objective of the proposed method is to localize the depicted anatomical landmarks in an unseen target image or volume. Published in [22].

The proposed approach performs global search. It learns optimal landmark detectors and features from the training data. It adapts the topology of the shape model to the anatomical structure based on the variability of the training examples, and aggregated appearance evidence for stable and accurate landmark location estimates. Due to the relatively low number of candidates for each landmark it scales well to large 3D data.

2.1 State of the Art

Several related approaches to anatomical structure localization exist in recent literature. They mainly differ regarding the type of semantic representation that is obtained to describe the image data. We distinguish between approaches that either 1) indicate the *positions* of individual landmarks, 2) result in *model parameters* which describe the position and shape of the object or 3) provide *voxel-wise labels* for different organs.

Localizing anatomical landmarks using the *positions* of selected interest points has been the objective of [23] and [20]. The methods employ interest point detectors (such as symmetry maxima, Harris corners) to find candidates for the individual model landmarks, and perform discrete optimization of a graph matching problem to obtain the final localization. The model landmarks are thus restricted to positions where these interest points are reliably detected. The methods in [3, 59] and [18] employ learnt interest point detectors to be able to obtain candidate points which directly represent anatomical landmarks. In both cases, the final representation consists of the matched model graph vertices (i. e. the landmarks) in the space of the target image or volume. To balance the low accuracy of the initial landmark estimates, [3, 59] employ a subsequent refinement step *after* matching the model. While the investigated anatomical structure (spine) is embedded in a 3D data set, the landmarks are all contained in a 2D subspace. In contrast to this, [18] matches a 3D geometric model to candidates obtained as mean-shift

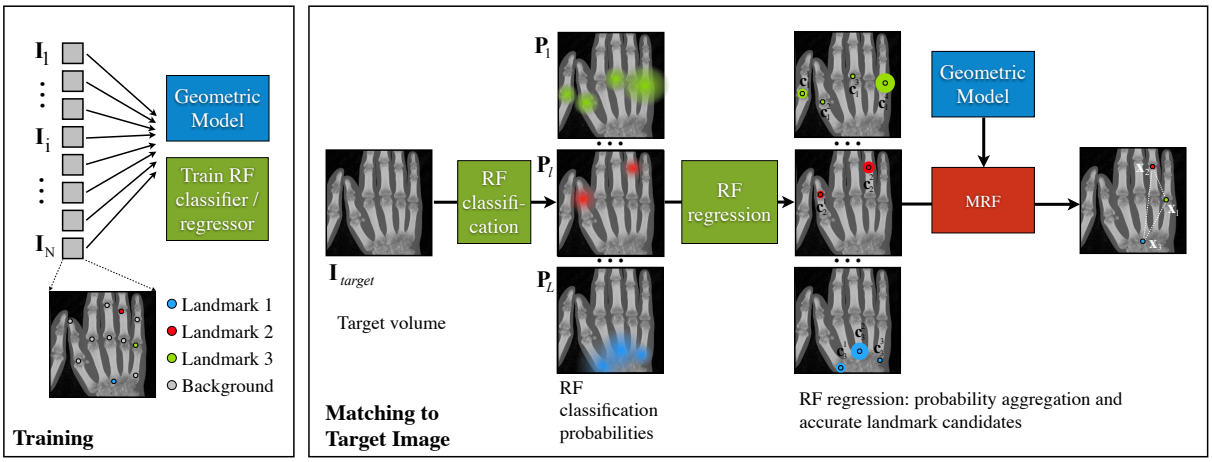


Figure 2: Outline of the proposed anatomical structure localization approach. Published in [22].

cluster centers of 3D probability volumes resulting from the classification of CT data using Haar-like wavelets and Random Forests.

In related work [63] predicts joint positions in 2D depth images by classification and mean-shift clustering of the resulting labeling. Due to the characteristics of the data set, a disambiguating, final optimization step is not required. Seifert et al. [61] parse whole body CT data in a hierarchical fashion, to detect larger scale organs. To reduce the complexity of the task this parsing is performed on axial slices. They first search for one salient slice in each dimension and consequently only localize landmarks within these slices. While substantially speeding up the localization this only works for objects which are rather large in respect to the volume size, as all the objects have to be visible in at least one of the three slices.

The idea of using ensemble regressors to estimate model parameters by having parts of the image vote for positions in the parameter space has been very successfully employed outside of the domain of medical imaging in the form of Hough forests [29, 63]. A first application to the localization of organs in thorax CTs has been been proposed in [11]. They train Random Regression Forests on Haar-like long-range features to predict the position and size of bounding boxes. An extension using Hough ferns was presented in [55] to predict the bounding boxes of multiple organs at once in full-body MR data. Working on axial slices of CT scans, [69] estimates bounding boxes through a boosted learning scheme and the combination of the independent axial predictions to obtain a localization in 3D.

The task of assigning *pixel-wise labels* to segment entire organs or organ structures has been approached by [12] and [51] using Random Forest classification. The work in [51] extends the Auto-Context [66] idea to include intermediate decisions within a tree to speed up the classification, which also incorporates information about the relative spatial positions of the objects.

Relying on stochastic optimization instead of ensemble classification or regression, Marginal Space Learning [68] tries to find the parameters of a bounding box or a parametric and data-driven *shape model* [8] to localize and segment anatomical structures. In contrast to standard Particle Filters [14], this is not performed on all parameter dimensions at once, but rather the number of searched parameter dimensions is increased after each convergence. Iterative ap-

proaches have been proposed to cope with repetitive structures such as the spine [39].

2.2 Limitations of Existing Methods

The existing approaches exhibit a number of limitations if the goal is to accurately identify and precisely localize anatomical landmarks.

Local search leads to local minima especially in the case of repetitive anatomical structures Approaches without a global cost function which optimizes the matching of the model to the target volume cannot cope with the high degree of repetitive sub-parts existing in anatomical structures. Different anatomical sub-parts may show a very similar local appearance pattern, and landmark detectors based on classification may yield highly inter-changeable responses requiring further disambiguation. For this reason, a global search and robust disambiguation of possible landmark configurations is necessary.

A fixed graph structure cannot represent arbitrary anatomical structures optimally Previously proposed methods for landmark-based localization employ a limited geometric model using simple mean $\pm$ standard deviation length distributions on the model's edges [20]. To effectively deal with the topological and geometrical richness of human anatomy, non-parametric geometric models that adapt to the structure based on the training data are preferable.

Manageable numbers of accurate candidates To achieve high accuracy during global search, the candidate points have to be precise even before disambiguation. This is not possible with purely classifier-based approaches. Furthermore, the number of candidates must be low in order to make the matching of large models feasible.

The alternative of using only Hough Forests, i. e. letting all voxels predict a landmark's position as in [11, 55] leads to low accuracy predictions: while the long range, global predictions are required to indicate the region in the image where the landmark is located, only the local predictions indicate the landmark's position with high accuracy. A pre-filtering step (as proposed in this work) or an iterative, course-to-fine Hough Regression scheme is required to cope with this limitation.

Lastly, to be of practical relevance, the resulting approach should not require a delicate estimation of algorithm parameters or model topology, while model training as well as the matching of the model to the target volume should be fast.

Contributions We prose two distinct approaches to deal with the challange of the localization of anatomical structures in 2D and 3D medical image data:

Pre-filtered Hough Forests and Discrete Optimization, Sec. 2.3 An algorithm representing a global search that tackles scalability, accuracy and adaptability of representation. We demonstrate that a combined classification and regression approach to candidate point estimation yields highly accurate positions. The candidate detector exhibits high specificity, reducing the number of candidates per landmark while increasing their likelihood and thus substantially reducing the computational complexity of the graph matching, which is performed using global optimization. The use of state of the art ensemble methods ensures fast training and prediction times. The graphical model topology is derived automatically from the data. Finally, the parameters of the system generalize well, and were identical for all experiments reported in this paper.

Top-down Image Patch Regression, Sec. 2.4 We present a simple, fast method for the global, accurate localization of anatomical structures in 2D/3D data based on an appearance codebook, and location predictors that capture sub-configurations of a landmark set. It demonstrates that a top-down nearest neighbor matching strategy of image patches drastically reduces the number of required feature computations and yields localization results comparable to the state of the art.

2.3 Global Localization of 3D Anatomical Structures by Pre-filtered Hough Forests and Discrete Optimization

The approach, first published in [22], is divided into a training phase and a localization phase as shown in Fig. 2. During training the algorithm constructs a geometric model by estimating the mutual predictive accuracy between landmark positions, trains a random forest classifier for predicting landmark locations and a random regression forest (*Hough Forests*) for refining position estimates.

During localization (Fig. 2, right) Random Forest classifiers pre-filter the target volume and Hough Forests aggregate the classification probabilities into accurate landmark candidates. Subsequently an MRF that encodes the mutual location information of the landmark configuration is solved to disambiguate landmarks with multiple candidates.

The following sections detail the training of the Random Forest classifiers and regressors, and the formulation of the MRF to perform the model matching.

2.3.1 Landmark Candidate Pre-filtering – Random Forest Classification

We train the algorithm with a set of N training images or volumes $\mathbf{I}_i$ each containing corresponding landmark annotations. The annotations are coordinates $\mathbf{l}_l^i$, $l \in \{1,\ldots,L\}$ of L landmarks of the anatomical structure in question, and are present in each of the training volumes. From the annotated examples we learn their local appearance for the classification and regression, and their spatial relations for the geometric model.

First we train a classifier for the detection of landmark candidates. For each landmark l and each training volume $\mathbf{I}_i$ all voxels $\mathbf{v}$ within a small radius $|\mathbf{v}-\mathbf{l}_l^i| \leq 2$ are considered as positive samples for the landmark l. This results in a classification task with $L+1$ classes, with the background class consisting of a random subset of the voxels not labeled as a landmark.

We capture the local image information at each voxel v through a vector of gray value differences between the voxel and F=100 randomly chosen voxels in its vicinity, similar to [44] and [3], resulting in a feature vector $\mathbf{f}_v$. Using intensity differences to model local appearance aids with dealing with low-frequency brightness changes. The random offsets are taken from a Gaussian distribution with standard deviation $\sigma_{features} = 10$. A different set of F random offsets is used for each decision tree. We train the random forest using $n_{trees} = 32$ extremely randomized trees as an $L+1$ class classifier. We chose extremely randomized trees [32] as they are similar to conventional random forests [64] in performance and accuracy, while they are at the same time extremely fast to train and simple to implement.

During localization each voxel in the target volume $\mathbf{I}_{target}$ is classified, and each tree votes for a class for each voxel in the test volume. Normalizing these votes yields $L+1$ class conditional probability volumes $\mathbf{P}_l$ for the target volume. The training and prediction is performed on

down-sampled volumes according to a downscaling factor δ, which brings considerable performance benefits for both training and localization (compare Fig. 5). We improve the localization accuracy in a following step – through the aggregation of candidate probabilities with Hough Regression Forests which are trained on the full resolution data.

2.3.2 Accurate Landmark Candidates by Probability Aggregation – Particle Hough Forests

The classification results indicate those regions of the image or volume that are likely to contain one of the landmarks. We want to condense this information into precise and discrete landmark candidate positions, weighted with the aggregated probability information. This is achieved by training Hough regressors [29] for each landmark on the full-resolution data, which learn to vote for the exact position of a landmark for the voxels within a local neighborhood of that landmark. We extend this idea by using the Hough Forest in a particle-like probability aggregation.

To train the Hough regressors, we obtain features/offset pairs as follows: For each landmark l we take voxels v_l around l from all training images. For every v_l we calculate the offset $\delta\mathbf{x}_v$ between its spatial position and the location of the corresponding landmark l: $\delta\mathbf{x}_v = (\delta x_v, \delta y_v, \delta z_v)$; and we extract the local image features $\mathbf{f}_v$ at the position of v_l. Using features and offsets of a large number of voxels as training samples we then train a regressor $\mathcal{R}_l$ – the Hough forest – that is predicting the relative position offset $\delta\mathbf{x}$ using features $\mathbf{f}_v$. As the accuracy of this prediction decreases with increasing distance from l, the regressor is only trained within a local neighborhood of l, i. e. from small sub-volumes centered around each l.

During localization, after the probability maps $\mathbf{P}_l$ have been computed by the classifier, the Hough forest shifts the positions of all voxels with probabilities $p_v^l \in \mathbf{P}_l$ above the threshold $\beta = 0.5 * max(\mathbf{P}_l)$. By updating the position of each of these voxels according to the regressor prediction

$$\begin{aligned} \delta\mathbf{x}_v &= \mathcal{R}_l(\mathbf{f}_v) \\ \mathbf{x}_v &= \mathbf{x}_v + \delta\mathbf{x}_v \end{aligned} \tag{1}$$

it aggregates the probability mass in $\mathbf{P}_l$ towards specific positions in the target volume. Due to noise in the predictions and higher prediction accuracy once the voxels move closer to the true landmark location, this refinement is iteratively repeated $n_{iter} = 3$ times to yield the final voxel positions. Each unique voxel position, i. e. each candidate $\mathbf{c}_l$ gets assigned the accumulated probability mass $p(\mathbf{c}_l) = \sum_{\mathbf{x}_v = \mathbf{c}_l} p_v^l$. Thereby the original probability mass of $\mathbf{P}_l$ is now highly localized. Finally, the number of candidates is reduced by non-maxima suppression with radius $r_{supp} = 10$. The result of this refinement step through Hough regression is a set of highly accurate candidate positions for each landmark. Examples of the refined candidates are shown in Fig. 3(b,c).

2.3.3 Model Localization via Graph Matching

For most landmarks the previous step returns multiple candidate positions. To select a single candidate for each landmark, i. e. to obtain the localization result, we make use of the high level information encoded into a parts-based anatomy model.

Having obtained the landmark candidates $\mathbf{c}_l$ for a given target volume, we perform the matching of the graph of model landmarks onto these candidates. The MRF's task is to se-

lect, for each landmark, one of the landmark's candidates, such that the overall probability of the match is maximized.

While we could build a parts-based model using our anatomical knowledge, or relying on basic geometries arising, for example, from next neighbour triangulation, we choose to actually *learn* the MRF geometry from the annotated training data instead. We construct a Markov Random Field with L nodes $\mathcal{N}_l$, connected by E edges. The aggregated probabilities $p(\mathbf{c}_l)$ for each candidate $\mathbf{c}_l$ are used to derive the unary terms $\mathbf{U}$, while the incorporation of geometric constraints is achieved through the binary terms $\mathbf{B}$. The landmark candidates $\mathbf{c}_l$ corresponding to landmark l are the possible labels for node $\mathcal{N}_l$. The discrete optimization objective function for a match $\mathcal{M}$ of the model to the target volume is thus

$$Conf(\mathcal{M}) = \sum_{l=1\ldots L} \mathbf{U}(l, \mathcal{M}(l)) + \sum_{e=1\ldots E} \mathbf{B}(e, \mathcal{M}(e)). \tag{2}$$

The MRF is defined by the unary and binary terms, and by its topology. In the following we will explain how to derive each of them.

Unary Terms – Aggregated Candidate Probabilities Only the $d = \{1,\ldots,D\}$ candidates with the largest weights $p(\mathbf{c}_l)$ are considered for the MRF. Their weights are normalized such that

$$\sum_{d=1\ldots D} p(\mathbf{c}_l^d) = 1. \tag{3}$$

Each node $\mathcal{N}_l$ of the MRF has at most D labels. The $D \times L$ unary terms $\mathbf{U}$ are thus set such that $\mathbf{U}(d,l) = p(\mathbf{c}_l^d)$. Fig. 3(d) shows the typical, fast drop-off of the candidates' weights.

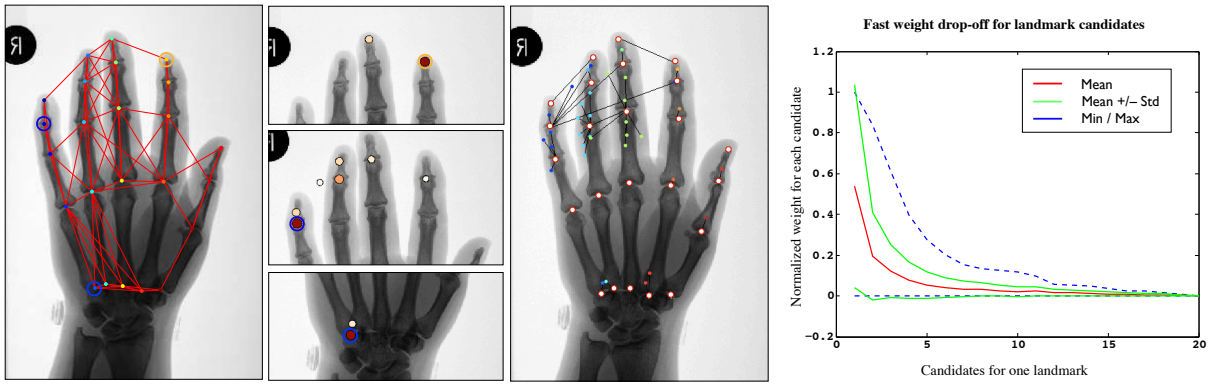
Graph Topology The topology of the MRF is derived automatically from the training data based on the strength of the relationship between landmark positions. Looking at the variations of the geometric distances between the landmarks in the training data, differential entropy is used to estimate the reliability of one landmark's position in predicting another's position. The most predictive landmark pairs are then connected in the topology of the geometric model and thus the MRF. This topology estimation is performed once, during training.

Given two landmarks s,t and their coordinates $\mathbf{l}_s^i, \mathbf{l}_t^i$, in all the N training volumes, their offset distances are $\delta\mathbf{l}_{s,t}^i = \mathbf{l}_t^i - \mathbf{l}_s^i$. We associate each $\delta\mathbf{l}_{s,t}^i$ with a Gaussian distribution $g_{s,t}^i = \mathcal{N}(\delta\mathbf{l}_{s,t}^i, \sigma_{DE})$ and compute the differential entropy

$$h_{s,t}(\mathbf{I}_{target}) = -\int_{\mathbf{I}_{target}} \log k_{s,t}(v) dv \tag{4}$$

by integrating over all voxels v of the volume $\mathbf{I}_{target}$, whereby $k_{s,t}(v) = \sum_i g_{s,t}^i(v)/N$ and σ_{DE} is empirically chosen (identical for all landmarks and data sets). We thus obtain a low entropy $h_{s,t}$ when the landmarks s and t show little variance in their relative positions, or when these relative positions form clusters. The T smallest entropies in

$$e_{1,\ldots,T}^s = \underset{L}{\operatorname{argmin}}\, h_{s,t} \tag{5}$$



(a) Automatically derived topology of the MRF. Among 24 landmarks pairs of mutually predictive landmarks are connected.

(b) Candidates and corresponding weights for 3 selected landmarks (highlighted in (a))

(c) Candidates for all 24 landmarks resulting from regression, and result of MRF disambiguation. To visualize the distribution, black lines link each candidate to the true position of the respective landmark.

(d) On average candidate weights drop quickly after the top ranked candidate.

Figure 3: (a): Topology of the MRF learnt from the 2D radiographs data set (Eq. 2.3.3 and 2.3.3). (b): For 3 landmarks the candidates and their weights are visualized (red is highest weight). (c) MRF disambiguation result for the entire landmark set (red circles). In addition all candidates are shown. To visualized the distribution the candidates are connected to the ground truth. (d): The normalized weights assigned to the candidates, which form the basis for the unary terms of the MRF, exhibit very fast drop-off, i. e. the are typically only 2-3 candidates of interest for each landmark, ensuring fast and robust MRF solutions. Published in [22].

yield the indices e_i^s of the landmarks s is connected to. This specifies the set $\mathcal{E}$ of $e \in \{1,\ldots,E = L \cdot T\}$ unique undirected edges of the geometric model and thus the MRF. An example for the derived topology is depicted in Fig. 3(a). Alternatively, but not investigated in this work, T can be adapted for each landmark s individually by further investigating the distribution of the corresponding entropies $h_{s,t}$.

Binary Terms – Non-parametric Geometric Model Constraints The $D \times D \times E$ binary terms $\mathbf{B}$ of the MRF relate the relative positions of the landmark candidates connected by edges to the spatial information contained in the training set. The distance between two candidates $\mathbf{c}_s^m, \mathbf{c}_t^n$ for the landmarks s and t of an edge $\langle s,t \rangle$, being $\delta\mathbf{c}_{s,t}^{m,n} = \mathbf{c}_t^n - \mathbf{c}_s^m$, is compared to the N edges between s and t in the training set, each modeled by the distribution $\bar{g}_{s,t}^i = g_{s,t}^i / max(g_{s,t}^i)$, $i \in \{1,\ldots,N\}$. The resulting value for the edge $\langle \mathbf{c}_s^m, \mathbf{c}_t^n \rangle$ representing an edge $\langle s,t \rangle$ is thus given by

$$y_{s,t}^{m,n} = max_N(\bar{g}_{s,t}^i) \tag{6}$$

Evaluating this measure for all edges $\mathcal{E}$ which are part of the MRF's topology defines the values of the binary terms

$$\mathbf{B}(s_m, t_n, e) = y_{m,n}^{s,t}. \tag{7}$$

Solving the MRF Having defined the unary and binary terms as well as the topology of the MRF, the MRF's solution, the so called *labeling*, can now be obtained by optimizing Eq. 2:

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmax}} \, Conf(\mathcal{M}) \tag{8}$$

which assigns each of the L model nodes $\mathcal{N}$ to one landmark candidate $\mathbf{c}_l^m$ in the target image, matching the model to the target volume. This yields the final position estimates in the target image for each landmark. While this optimization is in general NP-hard several approximate algorithms exist to obtain a locally optimal solution. Due to the simple structure of the MRF employed in this approach (low number of nodes/landmarks, low number of labels/candidates per node with a quick decrease of confidence for unlikely labels and a topology with few edges) the choice of MRF solver (believe propagation, graph cuts, tree-reweighted message passing, dual decomposition) is of little importance. We employ loopy believe propagation [67] throughout this work.

2.3.4 Experimental Setup

We evaluated the proposed approach on the three separate data sets shown in Fig. 1: 1) 20 hand radiographs, 2) 12 high resolution hand CTs and 3) 20 whole body CTs.

Data set 1: Hand Radiographs $N = 20$ hand radiographs with an average size of 460×260 pixels with a resolution of 0.423mm/pixel were annotated with $L = 24$ landmarks. The landmarks include the five finger tips, as well as the distal interphalangeal (DIP), proximal interphalangeal (PIP), metacarpophalangeal (MCP) and carpometacarpal (CMC) joints for each finger.

Data set 2: Hand CTs The 3D hand CTs have a voxel size of $0.5mm \times 0.5mm \times 0.66mm$ resulting in an average size of $256 \times 384 \times 330$ voxels. They are annotated with the same 24 landmarks as the hand radiographs, with three additional landmarks placed around the carpus at the radiocarpal, radioulnar, and ulnocarpal joints, totaling in $L = 27$.

Data set 3: Whole body CTs 57 landmarks were annotated for the first 20 whole body CTs from the Whole Body Morphometry Project [46]. The volumes have an average size of $512 \times 512 \times 1900$ voxels, with a voxel size of $1.3mm \times 1.3mm \times 1mm$. The landmarks are distributed throughout the entire body, to be able to localize all major body parts. The landmarks in the whole body CT are depicted in Fig. 1(c).

Setup The experiments were run in a leave-one-out cross validation framework, training the Random Forest classifier and the L Hough Forests $\mathcal{R}_l$ and the MRF topology on $N-1$ images / volumes and performing the localization on the remaining image / volume. The main measure of interest for each landmark is the residual distance between the position of the selected candidate and the corresponding ground truth. The Random Forest classifier and the regressors were trained using 32 extremely randomized trees, which were grown to their full depth. The number of connecting edges between landmarks in the MRF topology was set to $L/10$ for each data set. The downscaling factor δ was set to 0.5, 0.5 and 0.2 for the hand radiographs, the hand CTs and

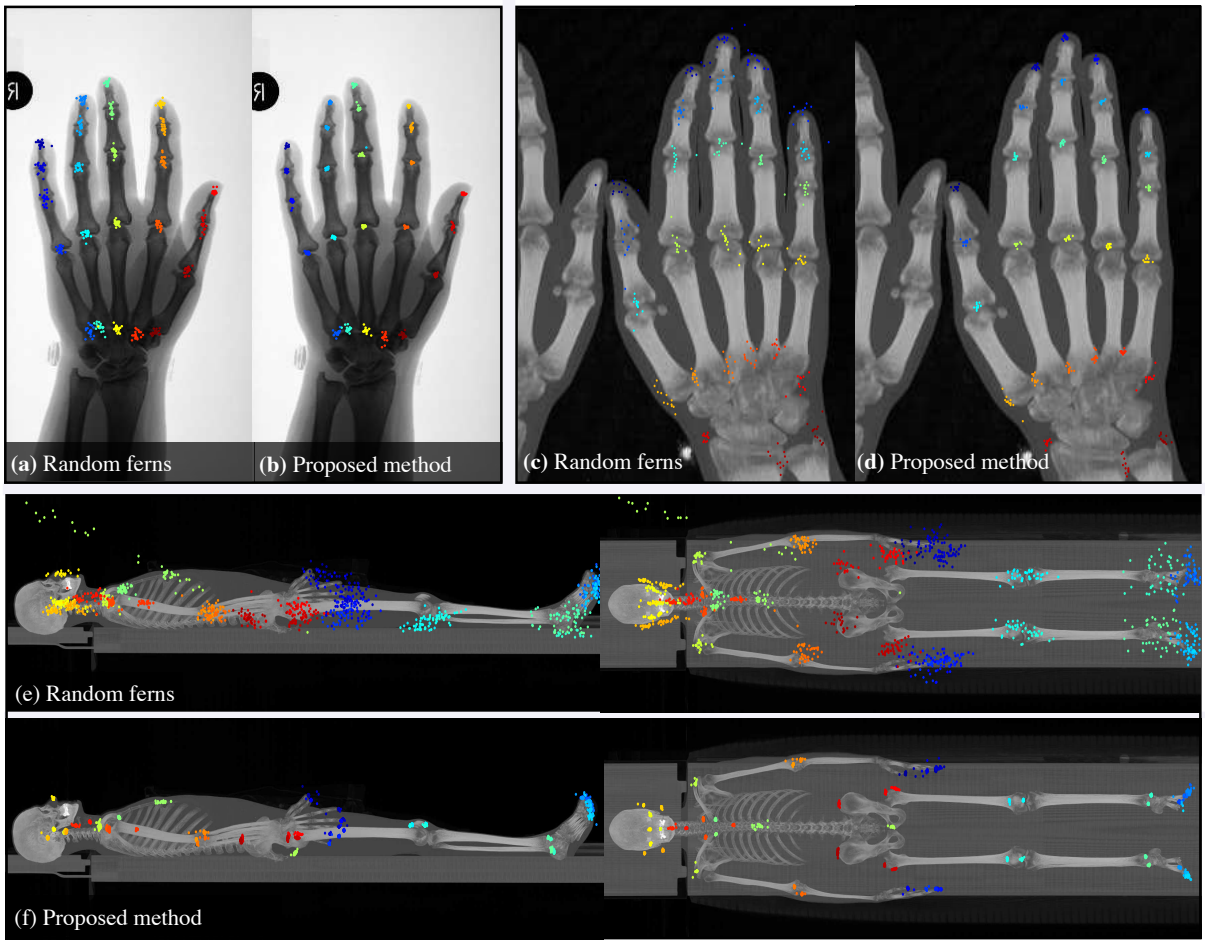


Figure 4: Resulting localizations on the three data sets visualized on one image/volume, for Random Fern Regression [55] (a,c,e) and the proposed pre-filtered Hough Forests (b,d,f). Published in [22].

the whole body CTs, respectively. All other parameter settings are *identical* for the experiments on the three data sets.

Comparison with existing candidate generation approaches Apart from investigating the accuracy of the overall landmark localization, we compared the influence of the proposed candidate probability aggregation (Hough regression) with two recently published approaches for candidate position estimation: mean-shift clustering and simple non-maxima suppression of the classification probabilities.

Mean-shift based Interest Point Generation [63] Mean-shift [6] is a method for the density estimation and cluster analysis of a set of points in a feature space, and was employed in [63] on classification probabilities to obtain landmark positions. Given a d-dimensional dataset $\mathbf{D}$ mean-shift iteratively moves each data point $\mathbf{d}_i$ towards the weighted mean of the data points, according to a specified kernel with bandwidth σ_{MS} around $\mathbf{d}_i$. The process is repeated

until equilibrium is reached, i. e. when no data point is shifted anymore. The candidate positions $\mathcal{C}_l$ for landmark l are thus estimated by employing mean-shift with a Gaussian kernel $\mathcal{N}(0, \sigma_{MS})$ on a point set obtained from $\mathbf{P}_l$ through rejection sampling. As in the present approach, the aggregated probability mass in each cluster is assigned to the cluster center, i. e. the landmark candidate.

Interest Points from Local Non-Maxima Suppression As suggested in [3], the candidates for each landmark class can be chosen by finding local maxima directly from the probability maps. For each probability map $\mathbf{P}_l$, only points above a threshold $T_c * max(\mathbf{P}_l)$ are considered. While [3] employs empirically estimated individual thresholds for every single landmark, we use $T_c = 0.5$ throughout our experiments. To further reduce the number of candidates, non-maxima suppression within the radius σ_n is performed. Note that, in contrast to our approach and the mean-shift approach, no aggregation of probabilities takes place.

Comparison with Random Fern Regression Localization We compare our approach with the method recently proposed in [55], which uses random regression ferns on binary Haar-like features to predict all landmark positions at once. Random ferns are similar to random forests in that they partition the input space into cells and record class membership histograms (classification) or linear models between input and output features (regression) of the contained data points. The partition is obtained by randomly splitting random projections of the input space n_{splits} times, yielding $2^{n_{splits}}$ cell per fern.

The input features employed in [55] are 3 scales of binary Haar-like features, i. e. at each scale the mean intensity 26 boxes with random offsets and sizes are compared to the mean intensity of a box centered on the voxel in question. The target values for each voxel are its relative positions to all L landmarks that are to be predicted, leading to a $n_{dim} \cdot L$ target vector $\delta\mathbf{x}$. For each cell of each fern, a linear regression model is estimated linking the input and output spaces.

During localization, the binary input feature vector is computed for each voxel v, the cell in fern f containing this data point is calculated and the linear model predicts the output feature vector $\delta\mathbf{x}_v^f$. A weighted sum of these output vectors over all ferns, following Gaussian distributions modeling the input data points of each cell, yields $\delta\mathbf{x}_v$. Voxel v thus estimates the landmarks to be found at $\mathbf{x}_v + \delta\mathbf{x}_v$. Integrating these predictions over the entire target volume yields the final landmark positions estimates $\mathbf{x}_l$ for each landmark. To allow for a fair comparison with the proposed method we refine these estimates locally: For each landmark l an additional random fern regressor is trained to predict $\delta\mathbf{x}_l$ in a local neighborhood of l, and $\mathbf{x}_l$ is refined accordingly. All parameters of the method were set to the values reported in [55].

2.3.5 Experimental Results

A. Candidate accuracy – Comparison with Mean-Shift and Non-Maxima Suppression The comparison of the three different strategies evaluated in this paper for the estimation of landmark candidates is shown in Fig. 5. Two important aspects in this evaluation are the time required to perform the Random Forest classification and to solve the Markov Random Field, as these are the dominating factors.

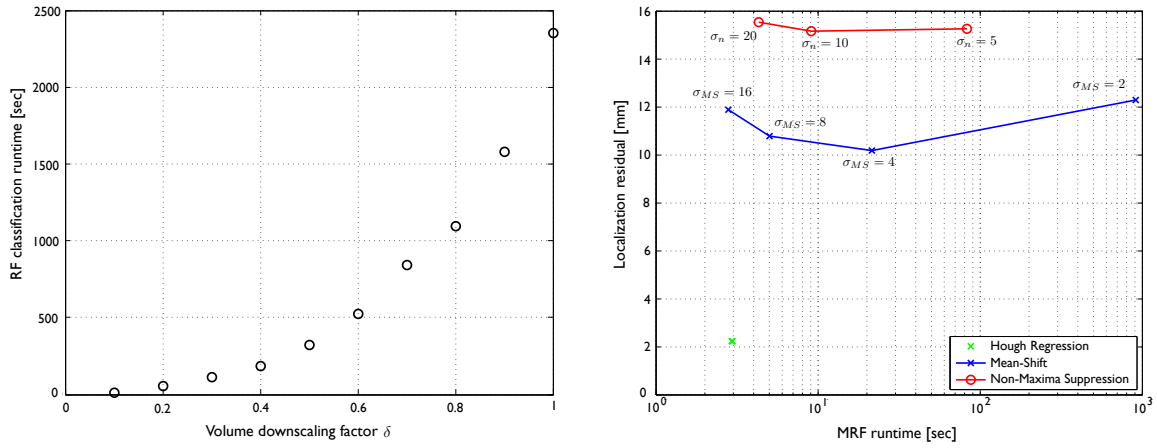


Figure 5: Left: Time required for the classification of one whole body CT depending on the chosen downscaling factor δ. Downscaling the volume is crucial to maintain reasonable run-times. Right: Comparison of the median landmark error and runtime for solving the MRF with landmark candidates generated by the proposed approach, by non-maxima suppression [3], and by mean-shift [63] for whole body CT localization with $\delta = 0.2$. Published in [22].

Fig. 5 (right) shows the median localization residuals of the three landmark candidate estimation schemes Hough Regression, Mean-Shift and Non-Maxima Suppression on a leave-one-out run on the whole body CT data set, with a down-scaling factor of $\delta = 0.2$. First, it can be observed how learnt aggregation in the Hough Regression approach (whose regressors are trained on full resolution data) is able to cope with the strongly down-sampled data used during classification. In contrast to this, Mean-shift and Non-Maxima Suppression cannot exploit this additional information and are thus adversely affected when limited to down-scaled image data.

The iterative Hough predictions performed in our approach exhibit run-times very similar to mean-shift clustering, the main time difference between these methods stems from the time required to solve the MRF. The non-maxima suppression approach does not perform any candidate position refinement, leading to faster candidate generation times at the expense of of poorer accuracy.

The main influence on the MRF runtime comes from the number of candidates per landmark, which depends on the kernel bandwidth σ_{MS} (mean-shift) and the non-maxima suppression radius σ_n, as detailed in Fig. 5 (right). Using a smaller σ_n yields more accurate landmark candidates, but their larger quantity results in considerably slower MRF inference. Albeit much less pronounced, a similar effect can be observed in the mean-shift data. At very small bandwidths ($\sigma_{M2} = 2$) results are slightly worse, presumably due to the fact that the MRF solver gets stuck at a local optimum.

In summary, the results show that the proposed Hough regression approach is able to yield highly accurate localization results while minimizing the runtime of both the classification and the MRF inference. This balance of accuracy and performance results from the landmark-specific aggregation functions which are learnt on the full resolution data.

B. Localization accuracy on the three data sets The results of the leave-one-out evaluation of the landmark localization is presented in Tab. 1 and the visualization in Fig. 4 which shows the localization accuracy for each individual landmark in the three data sets and the two compared methods.

Data set 1: Hand Radiographs For this 2D data set, both the median and mean residual are below 1mm, at 0.8mm and 0.99mm respectively (median / mean error of 2.00 / 2.34 pixels). As we will see below, this level of accuracy is representative for the method, with the other data sets showing similar results. Fig. 4(b) shows the distribution of the localization residual over the individual landmarks. The localization of the CMC joints shows slightly elevated residuals in comparison to the other landmarks. Comparing their positions in Fig. 1 shows that the CMC landmarks are in a region of more cluttered appearance compared to the PIP and MCP joints. The highly symmetric structure of the more distal joints (in both x and y direction) allows the Random Forests to make more precise predictions from the features $\mathbf{f}_v$. We also attribute the larger error in the CMC joints to the fact that the manual annotations are more difficult to perform. The overall localization is very robust, with only 12 outliers (out of $L \cdot N = 480$ localizations) with a residual of more than 6mm, i. e. 97.5% of the localized positions were within 6mm of the ground truth position. No residual was larger than 14mm and only 4 outliers' residuals were larger than 10mm.

Data set 2: Hand CTs The results for the 3D hand CT data set show similar characteristics as the 2D case (Fig. 4(d)). The median / mean residuals of 2.23 and 2.61 voxels correspond to 1.19 and 1.45mm, respectively. Again we see the larger errors for the CMC joints and the additional three joints, which due to the lack of local scale image information are more difficult to localize (and to annotate). For the 24 landmarks with positions comparable to those in the 2D data, 97% of the localized positions were within 6mm of the ground truth position, and only 5 outliers' residuals were larger than 10mm.

Data set 3: Whole body CTs This 3D data set was the most complex data set, with 57 landmarks. The median residual, 2.23 voxels, closely follows the trend set in the other two data sets. Due to a larger number of difficult to localize landmarks, the mean error for this data set is slightly higher at 4.34 voxels. These results correspond to 2.71mm and 5.25mm, respectively.

Looking at the results presented in Fig. 4(f) in more detail, clear trends can be detected in the performance on different body parts. In the landmarks located in the hands motion artifacts in two of the volumes introduce a number of outliers. In the volumes without artifacts, the performance is on a par with the results in the above 2D/3D hand data evaluations. Mismatches between the 3rd and 4th toes can be observed in both feet, in addition to a single instance where a right foot was localized as being a left one. We attribute this to the fact that the geometric model, by connecting each landmark to several of its neighbors, in some cases is overpowered by (sets of) very high candidate probabilities – which can occur due to the high degree of left / right similarity in the body.

Putting these results into perspective, out of the 1140 individual landmarks localizations, only 12 outliers (1.05%) had errors $> 30mm$. We consider the results on the three data sets to clearly demonstrate the ability of the proposed approach to find the landmark positions in

Method	Data set	Residual in Voxels			Residual in mm			p
		Median	Mean	Std	Median	Mean	Std	
Random	Hand Radiographs	3.40	3.76	2.48	1.44	1.59	1.05	
Fern	Hand CTs	4.42	7.43	10.08	2.60	4.16	5.15	
Regression	Full body CTs	37.30	48.53	82.00	43.37	54.80	86.98	
pre-filtered	Hand Radiographs	2.00	2.34	1.95	0.80	0.99	0.82	$< 10^{-19}$
Hough	Hand CTs	2.23	2.61	1.96	1.19	1.45	1.13	$< 10^{-16}$
Regression	Full body CTs	2.23	4.34	11.96	2.71	5.25	15.08	$< 10^{-66}$

Table 1: Localization accuracy: Landmark localization error for Random Fern Regression and for the proposed approach. Published in [22].

the target volume with high accuracy, with the consistent localization of detailed anatomical structures with a median residual of ≈ 2 pixels/voxels, independent of the data set.

C. Comparison with Random Fern Regression Localization Fig. 4(a,c,e) show the results of performing random ferns based localization [55] on the three data sets. In the 2D hand data set the method is able to localize the structure, with only minimal outliers, but overall the localized positions exhibit more deviation from the ground truth than in the proposed method. This is also reflected in the median/mean error of 3.40/3.76 pixels, corresponding to 1.44/1.59 mm. In the 3D case the localization residual increases considerably, with a median/mean error of 4.42/7.43 pixels, which is 2.60/4.16 mm. The localization error affects all parts of the structure equally, i. e. while the initial prediction of $\mathbf{x}_l$ does predict the right parts of the image it seems to fail to do so with high enough accuracy that the local refinement can have impact. This is even more pronounced in the case of the full body CTs (Fig. 4(e)). The size of the individual anatomical structures in a full body CT are, in relation to the size of the volume, comparatively small, putting an even higher burden on good initial position estimates. In this case the median/mean localization error rises to 37.30/48.53 pixels and 43.37/54.80 mm respectively.

D. Run times The run-times of the proposed approach for the localization on three data sets were in the order of 20sec / 90sec / 120sec on a 2009 8-core Xeon MacPro. The implementation was performed in Matlab except for the C-based Random Forests and the loopy belief propagation MRF solver [2]. Solving the MRF only takes < 3 sec, even for the whole body data set, i. e. the vast majority of runtime is spent on evaluation of the random forests. Recent advances in GPU-based Random Forests [62] showed classification runtime for whole body CTs of about 5sec [11]. By exploiting such an implementation in the two steps of low-resolution classification and subsequent refinement the runtime of the random forest evaluation of our approach could be lowered to below 5sec, resulting in a potential total runtime of less than 8sec including the time so solve the MRF.

2.3.6 Discussion

The experimental results demonstrate how a generic approach that performs global search for landmarks can yield high localization accuracy in a variety of medical imaging domains. En-

semble learning techniques enable the algorithm to obtain landmark candidates with sufficiently high specificity to allow for successfully disambiguation of locations via discrete optimization even on large scale data. The localization accuracy can be improved by Hough regression, and the MRF topology can be learnt from training data.

The results clearly show that the spatial accuracy and low number of candidates stems from combined classification and regression. In addition, the incorporation of the geometric information into the MRF is crucial for the disambiguation between the landmark candidates.

The distribution of the probability mass within the candidates depends on both the number of candidates and their specificity. A lower number of candidates is preferable, as it reduces the number of labels of the MRF, while a steep drop-off of the candidates' weights enables the MRF to find the solution in a very short runtime.

Since classification is performed on a per-voxel basis, computation time has a cubic relation to the downscaling factor δ, as measured for the whole body CT in Fig. 5 (left). Therefore, performing initial classification on a down-sampled volume, and then local regression refinement as learnt on full-resolution data results in highly accurate candidates while keeping computational costs low.

The aggregation of the classification probabilities through a learnt regressor instead of mean-shift (or non-maxima suppression), allows to find more accurate candidates while reducing their number, resulting in faster MRF solution as shown in Fig. 5 (right).

The comparison of the proposed approach with the localization method based on random ferns shows the importance of predicting first local model characteristics and to include global information at a later stage. In the random ferns approach, local image features vote for all landmarks in the volume, which are potentially far away from the voting voxel. Using local refinement is thus only successful when the initial prediction of the landmarks is close enough to the actual landmark position, i. e. within the capture range of the local refinement regressor.

The proposed approach on the other hand initially only estimates local parts of the model, i. e. the candidates, and focuses on their spatial accuracy, low number and high specificity. Only then the global structure of the model is exploited to select from these candidates, advancing from the local to the global characteristics of the model.

Approaches such as marginal space learning [68, 39] always work with the full model and do not distinguish between local and global characteristics. They have so far been used for bounding box based localization of anatomical structures, and an extension to focus on individual anatomical landmarks is a alternative approach for future research.

Limitations In contrast to [20], the model is not explicitly rotationally invariant. The geometric model proposed in that work employs a parametric model of spatial relations by encoding mean/standard deviations of both edge lengths and angles between local interest point orientations and neighboring landmarks. But several factors as well as the result from this work seem to indicate that rotation invariance might not be necessary in the medical context. First, the acquisition protocols for the individual organs / modalities ensure very consistent image data with regard to rotations. Second, the relative angles between landmarks when using a non-parametric model and a sufficient number of training examples can describe variations in the data due to the superposition of example geometries. A similar argument is valid for the landmark classifier; in cases with small training sets but considerable amounts of rotation, the training set can be artificially inflated by randomly rotating and perturbing the training data.

The algorithm was implemented on a CPU. Fast implementations of random forests on a GPU exist with speed-ups by factors up to 100 [62], and we expect this to have a substantial impact on the run time of the algorithm, since the classification and regressions forms a large part of computation.

Although it did not occur in the experiments it is possible that no landmark candidate sufficiently close to the true position can be found, and disambiguation can fail. To cope with this so-called dummy labels [24] can be introduced in the MRF, that model absent landmarks.

The proposed model does only use unary and binary terms. In principle higher-order terms can be used to capture spatial relationships among e.g., triplets of landmarks.

Conclusion and Outlook We present an approach for localizing complex, partly repetitive anatomical structures in 3D volumes, which yields highly accurate, robust results with fast runtimes. Based on Random Forests for classification and Hough regression, the method detects landmark candidates in a target volume. The model of the anatomical structure is matched to these candidates by solving a Markov Random Field. The algorithm does not rely on predefined interest point detectors or a manually designed graph topology but learns both from the training data. It achieves high accuracy of the best candidate while at the same time keeping the number of candidates low. The crucial feature of the proposed global localization approach is the combination of Random Forest classifiers and the Hough regressors. They can be trained on full-resolution image data, while the Random Forest classifiers can be applied to strongly down-sampled volumes.

Several areas of research seem promising for future work. In particular, the application of the approach to the localization of structures with occlusions due to the field of view or due to missing parts because of anatomical differences is an important issue for clinical practice. To reduce the amount of work required for the annotation of the training data, a scheme similar to the one presented in [25] could enable the building of models from a single manual annotation. Similarly, building a unified model of the entire human anatomy would allow a matching of the model to any given depicted region. Other variants of Random Forests such as the recently introduced Oblique Random Forests [48] should be investigated. Their oblique splits allow for a higher adaptivity to the data and a better class separation, as well as for correlated features and random offset and scaling that occurs as noise in between observations. The learning paradigm for feature detectors may be particularly helpful for a transfer to MR image data, where such artifacts occur as a result of local and global field inhomogeneities. As a next step, we will evaluate the method on a substantially larger data set.

2.4 Fast Anatomical Structure Localization Using Top-down Image Patch Regression

The approach, first published in [21], is divided into a training phase and a localization phase as shown in Fig. 6 and Fig. 7. During localization a multi-scale codebook of image patches and landmark positions is constructed, which is traversed during the localization phase to obtain increasingly accurate landmark estimates at each scale.

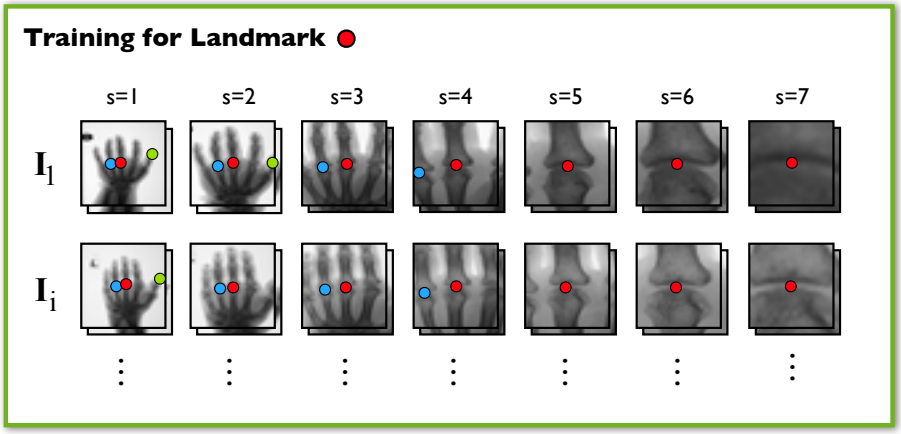


Figure 6: Construction of the regression codebooks during training. For each landmark and scale patches at various offsets and the corresponding relative landmark positions are recorded, using all training images/volumes. Published in [21].

2.4.1 Training – Constructing the landmark regression codebook

The training phase requires a set of N training images or volumes $\mathbf{I}_i$ with corresponding annotations. The annotations represent the coordinates $\mathbf{l}_l^i$ of the $l \in \{1,\ldots,L\}$ landmarks of the anatomical structure in question. Each landmark is present in each of the training volumes.

Codebook construction to connect local appearance and landmark information Our aim is to build multi-scale regression codebooks $\mathcal{C}$ of image patches and corresponding relative landmark positions – one codebook per scale $s \in 1,\ldots,S$ and landmark l. The patches stored in the codebook are extracted around the landmarks with varying offsets and scaling, capturing the typical visual appearance around each landmark. For each patch the positions of all landmarks visible in the patch are recorded, relative to the patch's center. Each of the PN entries in the codebook $\mathcal{C}_{s,l}$ consists of the tuple $\langle \mathbf{P}^p, \mathbf{L}^p \rangle$ of the patch $\mathbf{P}^p$ and the corresponding relative $D \times L$ landmark coordinates $\mathbf{L}^p$ which are visible in the patch. $\mathbf{L}^p$ specifies the coordinates of the landmarks $l \in 1 \ldots L$ relative to the center of the given patch[1]. Landmarks which are outside of the patch are denoted as not visible.

The construction of the codebook proceeds as follows: At the top-most scale $s = 1$ each image or volume is represented by an an-isotropically downscaled miniature of size $m \times m \times m$ (similarly $m \times m$ for images). At each scale s the volume is considered to possess an edge length of $\sqrt{2}(s-1)m$. This re-sampling of the entire image is never actually computed, it simply forms the reference frame for each scale of the codebook generation.

At each scale s, patches $\mathbf{P}$ are extracted from the image or volume data using linear interpolation for each landmark l from all training volumes N. The patches are of size $m \times m \times m$, i. e. at scale $s = 1$ they correspond to the entire image, and for scales $s > 1$ the patches *zoom in* on the landmark, as illustrated in Fig. 6. Parts of patches which would be sampled from outside of

[1]The necessary transformations between image coordinates and patch coordinates are omitted for clarity throughout the text.

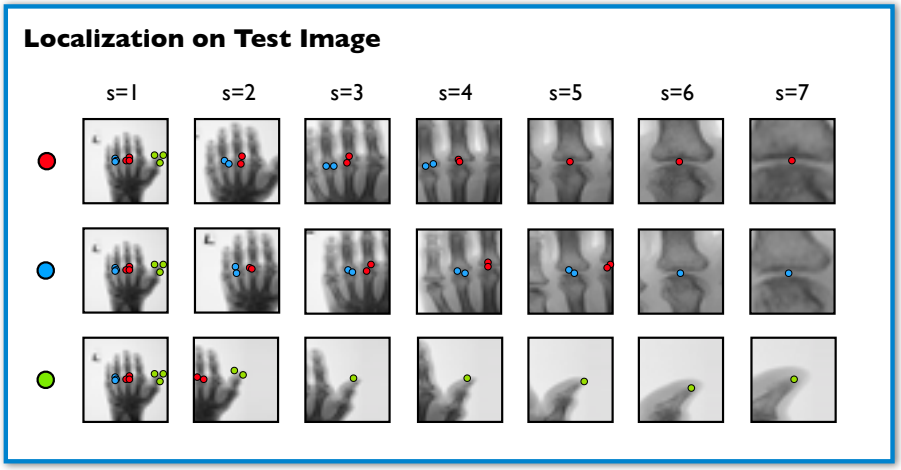


Figure 7: The localization of three landmarks on a test image/volume descends the scale pyramid. At each level regression based on the image patch generates not only a position estimate for the primary landmak, but also for other landmarks visible in the patch. When progressing to a finer scale, for each landmark these estimates vote for the next estimate and center of the finer patch. Published in [21].

the volume are set equal to the closest voxel on the volume's border. The gray values of each patch is normalized to zero mean and unit variance.

To explore the image information in the vicinity of a landmark the entries in the codebook $\mathcal{C}_{s,l}$ at a certain scale s and landmark l, are constructed by extracting several patches around the landmark with, empirically chosen, 7 offsets in the range of $[-6,6]$ voxels for each dimension, along with scaling factors of $\{0.9,1,1.1\}$, resulting in $P = 1029$ patches for one landmark in one training volume at one scale ($P = 147$ for images). To considerably reduce the memory requirements and computational complexity for the codebook lookup, dimensionality reduction of each codebook is performed using PCA, retaining 90% of variance, resulting in PCA coefficients $\mathbf{P}_{PCA}$ and final codebook tuples $\langle \mathbf{P}^p_{PCA}, \mathbf{L}^p \rangle$. This training scheme results in the $S \times L$ regression codebooks $\mathcal{C}_{s,l}$.

Shape model to regularize the localization To be able to regularize the intermediate solutions during the prediction phase, a model of the spatial distribution of the landmarks $\mathbf{s} = \langle \mathbf{l}^i_1, \ldots, \mathbf{l}^i_L \rangle$ in the training data is learned. We compute a point distribution model $\mathcal{S} = \langle \bar{\mathbf{s}}, \mathbf{S} \rangle$ using an eigen-decomposition of the covariance matrix of the training landmarks $\mathbf{l}_l$ as proposed in [8], retaining all eigenvectors and thus the entire shape variance observable in the training set, where the shapes $\mathbf{s}$ in the model can be constructed through a parameter vector $\mathbf{b}$ such that:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{S}\mathbf{b}$$

2.4.2 Localization – Regularized top-down matching

Similar to the training phase the localization is performed in a multi-scale fashion, shown in Fig. 7. The $D \times L$ landmark localization matrix $\mathbf{L}^*_{s=1}$ is initialized with all landmarks starting

Residual in mm	MRF-based graph-matching			Proposed Patch-Regression Method		
	Median	Mean	Std	Median	Mean	Std
Hand Radiographs	0.80	0.99	0.82	0.63	0.77	0.64
Hand CTs	1.19	1.45	1.13	1.43	1.96	1.80

Table 2: Experimental results, localization accuracy in mm: Residual distances of the localization result to the ground truth annotation for the proposed method, in comparison with a state of the art approach. Published in [21].

at the center of the test volume $\mathbf{I}_{target}$. Starting with scale $s = 1$, a patch $\mathbf{P}^l$ for each landmark l is extracted (without additional offsets or scaling variations). The patch is normalized and projected onto the patch PCA model of $\mathcal{C}_{s,l}$, resulting in $\mathbf{P}^l_{PCA}$. The most similar patch p^{l*} in the codebook is found using euclidean nearest neighbor search – leading to the tuple $\langle \mathbf{P}^{l*}_{PCA}, \mathbf{L}^{l*}_p \rangle$ and thus the landmark coordinate predictions $\mathbf{L}^{l*}_p$ as estimated by landmark l. Repeating this codebook lookup for all landmarks yields the $D \times L \times L$ prediction tensor $\mathbf{M_{d,i,j}}$ with position estimates from each landmark i to all landmarks that are visible in the same patch. The median over all predictions j which are not marked as not-visible yields the updated landmark localization matrix $\mathbf{L}^*_s$. This procedure is repeated through all scales, resulting in the final localization result $\mathbf{L}^*_S$.

Shape regularization The position estimates $\mathbf{L}^*_s$ are regularized by projecting them onto the shape PCA model $\mathcal{S}$ and reconstructing them again thereafter. This enforces landmark positions which can be modeled by a linear combination of the shapes observed in the training data. This regularization is performed for scales $s \leq S-3$, to allow for landmark positions which can not be modeled though the shape model at scales $s > S-3$.

2.4.3 Experiments & Discussion

We evaluated the proposed approach on the two separate data sets 1) and 2) as described in Sec. 2.3.4.

Setup The experiments were run using four-fold cross validation, learning the landmark regression codebook on 75% of the N images / volumes and performing the localization on the remaining images / volumes. The main measure of interest for each landmark is the residual distance between the position of the predicted landmark position and the corresponding ground truth. The parameter settings are identical for the experiments on the two data sets, except for the size of the patches: 32×32 in the 2D case and $32 \times 32 \times 32$ for the 3D data. The results are compared with the recently proposed pre-filtered Hough regression Random forests [22], which in turn showed to outperform alternative approaches such as classification-based landmark candidate estimation with graph-based optimization [3] and classification + mean-shift based approaches [63].

Results The results of the evaluation of the landmark localization are presented in Tab. 2, which shows the aggregated localization performance for the two data sets. The accuracy on

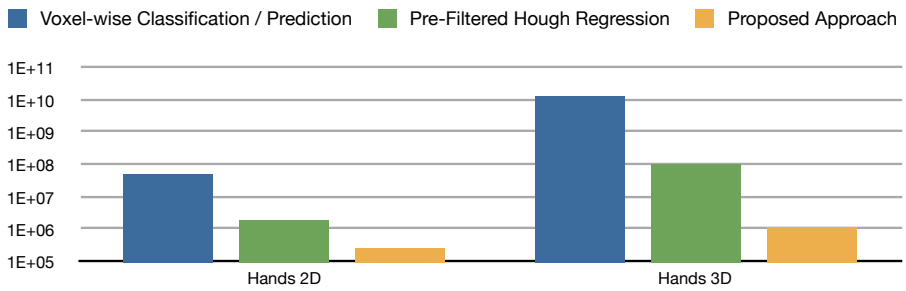


Figure 8: Number of image/volume accesses necessary to compute the features required during the localization phase. Voxel-wise classification / prediction approaches [3, 51] scale with the number of voxels, while pre-filtered Hough regression [22] works on strongly downsampled volumes. In constrast to this, the proposed approach is indepedent of the number of voxels and scales with the number of landmarks. Published in [21].

the 2D radiograph data set is very high with a median residual of 0.63 mm and a mean/std of 0.77/0.64 mm. This result compares favorably with the results reported and methods tested on the same data in [22]. The result on the 3D hand CT data set show a median residual of 1.43 mm and a mean/std of 1.96/1.80 mm. It can be seen that despite a similar median residual, the proportion of localizations with higher error is slighty larger in this case. The run-times of the proposed approach were in the order of 0.6sec for the 2D data set and 4.5sec for the 3D data set on a single core of a 2009 Xeon MacPro. The method was entirely implemented in Matlab - we expect a potential speed-up by a factor of 10 to 100 through a more optimized implementation.

Discussion - Feature computation complexity The main contribution of this work is the demonstration of a feature computation scheme which requires significantly less memory accesses then existing methods.

Voxel-wise classification / prediction approaches such as those proposed in [3, 51] scale with the number of voxels, while pre-filtered Hough regression [22] reduces computational complexity by working on strongly down-sampled volumes. A typical number of 400 memory accesses to compute the classification for a single voxel was assumed in the calculation, corresponding to e. g. 20 individual features in an ensemble of 20 individual classifiers.

In contrast to this, the proposed approach is independent of the number of voxels and only depends on the number of landmarks, with $m \times m \times m$ voxels sampled for the patch at each landmark and scale. The proposed approach thus requires one to four orders of magnitude less image/volume accesses, allowing for fast localization even in unoptimized implementations or cheap commodity hardware.

Conclusion and Outlook We present an approach for localizing complex, partly repetitive anatomical structures in 2D and 3D data. We demonstrate that a top-down nearest neighbor matching strategy of image patches drastically reduces the number of required feature computations and that the prediction of relative landmark positions using codebook regression is feasible.

The results on the two data sets clearly demonstrate the ability of the proposed approach to find the landmark positions in the target volume with accuracy comparable to the state of the art, with the consistent localization of detailed anatomical structures with a median residual of 1.7 to 2.7 pixels/voxels.

We consider the results to be very promising for such a simple method, and will focus on several topics in upcoming work: A detailed analysis of the parameters involved, namely the patch size and the perturbation strategy during codebook generation, as well as approximations of the nearest neighbor search through random subspaces.

3 Learning Reference Spaces

Reference spaces allow us to map location specific information across individuals, by providing maps of the coordinate systems of individual imaging data to a joint reference, such as an average human body. In addition to the reference, there are often matched sets of meta information assigned to positions in the reference space. They are called *atlases*, and an example would be a labeling of an organ in the reference space, that can be mapped to individual examples, by a non-linear coordinate transform.

In the following we describe two approaches that learn reference spaces from medical imaging data. The first focuses on learning a whole body atlas from clinical data. It addresses the specific problem, that each example typically contains only a section of the body, and outlines an approach to create a whole body reference, and atlas from such *fragments*. The second method focuses on the integration of anatomical and physiological, or functional information in the atlas building process. It addresses the question of the appropriate notion of correspondence if we are studying different domains, such as functional neuroimaging data.

3.1 Learning an atlas from clinical imaging data

Note that this work has been published in [26]. Human anatomy exhibits variability across different individuals as well as within single subjects (e.g. over time). It encompasses different locations, sizes or shapes, physiological state, disease characteristics, or tissue differences of organs. Medical imaging based computational anatomy and anatomical atlas construction overcome this variability and establish comparability within large populations in a common reference space referred to as anatomical atlas [37]. The daily routine of hospitals produces hundreds of gigabytes of pathology driven medical imaging data every day [5].

In this section we address the limitation of existing atlases and describe a method for the registration of medical imaging fragments towards a common whole body reference volume. The method is capable to estimate the center as well as the corresponding region of fragments in relation to the whole body reference space. In an iterative procedure these region estimates are used to register fragments to the whole body reference template. Based on the fragment registrations and its inherent information on shape and intensity variation the initial whole body template is updated. This reduces bias with respect to the underlying population of fragments. In the following a problem statement is provided and each of the steps required for fragment to whole body registration as well as fragment based atlas construction are discussed.

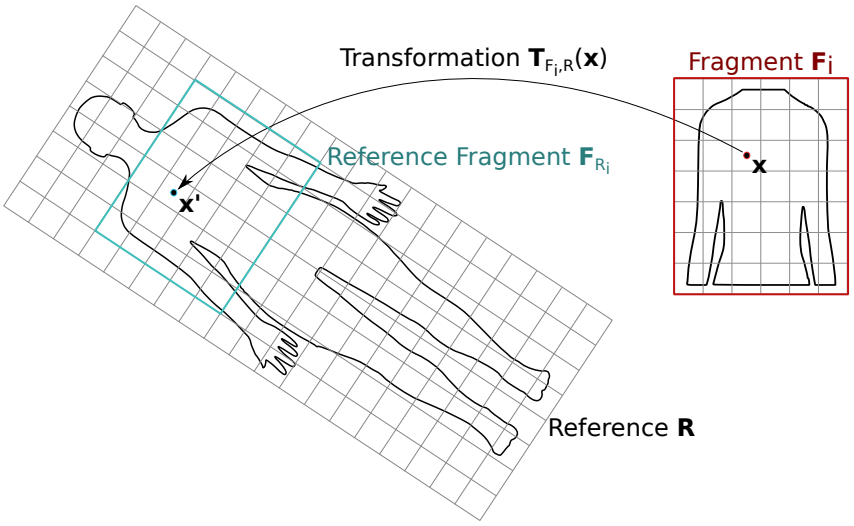


Figure 9: Fragment to whole body reference space registration.

3.1.1 Representing data and problem statement

Given a set of whole body medical imaging volumes $\mathbf{V}_1,...,\mathbf{V}_S$ and a set of medical imaging fragments $\mathbf{F}_1,...,\mathbf{F}_N$ each covering only a limited region of the whole body volumes $\mathbf{V}_s$. The first computation step of the fragment based whole body atlas has the aim to select the whole body volume $\mathbf{V}_s$ showing the lowest bias in relation to the remaining set. This choice is used as initialization for the common reference space $\mathbf{R}$. Further the transformations $\mathbf{T}_{F_i,R}$ registering the fragments $\mathbf{F}_i$ towards the common whole body space $\mathbf{R}$ have to be found. For this purpose the corresponding region $\mathbf{F}_{Ri}$ is determined in the common space $\mathbf{R}$ for each of the fragments $\mathbf{F}_i$ to enable the registration of two fragments containing identical anatomical structures. Figure 9 illustrates the components of the registration problem. The fragment registrations hold information on tissue intensity as well as shape variation and are used to compute an un-biased fragment based update $\mathbf{R}_F$ of the initial whole body template $\mathbf{R}$.

3.1.2 Least biased whole body template selection

This subsection describes methodology for the selection of the least biased whole body reference volume $\mathbf{R}$ from the given set of candidates $\mathbf{V}_1,...,\mathbf{V}_S$. It ensures that bias is kept at a minimum from the first computation step of fragment based atlas construction. Park et. al. published in their work on least biased target selection a method which computes a MDS on a distance matrix constructed from the bending energy in the registration transformations of all pair-wise registrations [54]. In the following paragraph an adapted version of this procedure using registration dissimilarity as cost function is explained.

The first step is the pairwise registration of all candidate volumes $\mathbf{V}_1,...,\mathbf{V}_S$. Based on these registrations the distance (cost) matrix $\mathbf{D}$ is computed as defined in Equation (9). As distance measure serves the registration error after non-rigid registration of all volume pairs $\mathbf{V}_i$ and $\mathbf{V}_j$ in the set.

$$\mathbf{D} = [(d(i,j))] \text{ with } d(i,j) = 1 - SIM(\mathbf{T}_{V_j,V_i}(\mathbf{V}_j),\mathbf{V}_i) \tag{9}$$

SIM holds place for a normalized image similarity measure such as Normalized Cross Correla-

tion (NCC) [13] or Normalized Mutual Information (NMI) [65]. $\mathbf{T}_{V_j,V_i}(\mathbf{V}_j)$ is the registration (transformation) of volume $\mathbf{V}_i$ with volume $\mathbf{V}_j$. Based on the distance matrix a multi dimensional scaling (MDS) is computed [1]. The MDS results in a d-dimensional embedding space $\boldsymbol{\Psi}$ with Euclidean embedding coordinates $\boldsymbol{\psi}_s$ and respective inter-point distances $d_e(\boldsymbol{\psi}_i, \boldsymbol{\psi}_j)$ (L_2-Norm) reflecting the pairwise registration costs $d(i, j)$ of the volumes $\mathbf{V}_i$ and $\mathbf{V}_j$. After the transformation of the volumes into the embedding space $\boldsymbol{\psi}$ the mean volume is computed as the arithmetic mean $\overline{\boldsymbol{\psi}}$ of all embedding coordinates $\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_S$. The least biased initial whole body reference volume $\mathbf{R}$ is defined as the volume $\mathbf{V}_i$ closest to the embedding space center $\overline{\boldsymbol{\psi}}$ [54]. The formal definition of MDS reference selection is provided in Equation (10) where $\boldsymbol{\psi}(\mathbf{V}_i)$ returns the embedding space coordinates $\boldsymbol{\psi}_i$ of whole body volume $\mathbf{V}_i$.

$$\mathbf{R} = arg\,\min_{\mathbf{V}_i} \left(||\boldsymbol{\psi}(\mathbf{V}_i) - \overline{\boldsymbol{\psi}}||_{L2} \right) \tag{10}$$

The result of the described method is an initial version of the whole body reference template $\mathbf{R}$ and used in the remaining steps as a common reference space for fragment to whole body registration.

3.1.3 Fragment center estimation

Given is an input query fragment $\mathbf{F}$ containing only a limited region of the body and the initial whole body reference volume $\mathbf{R}$ selected in the previous step. The first part of fragment to whole body registration is the estimation of the initial position of the fragment center $\mathbf{c}_0$ in relation to the whole body reference $\mathbf{R}$. The fragment center estimation is based on an image retrieval approach proposed by Donner et. al. [19] and consists of the following steps:

1. Construction of a training set containing fragments $\mathbf{F}_j$ with annotated center positions $\mathbf{c}_j$ in a whole body reference volume $\mathbf{R}$.

2. Reduction of the training fragment dimensions to a resolution of $32 \times 32 \times 32$ pixel. The result of this step is an annotated training set of 32^3 voxel miniatures $\mathbf{M}_j$.

3. Reduction of the query fragment resolution corresponding to Step 3. The result is an 32^3 voxel miniature version $\mathbf{M}$ of the input fragment $\mathbf{F}$. Further the pairwise similarities between the input fragment miniature $\mathbf{M}$ and all fragment miniatures $\mathbf{M}_j$ from the trainings set are computed. For similarity computation measures such as NCC and NMI are used.

4. Selection of the k most similar miniatures from the training set and robust estimation of the center position $\mathbf{c}_0$ based on the training set annotations $\mathbf{c}_j$. This center position is assigned to the query fragment $\mathbf{F}$ as initial estimate of the fragment location in relation to the whole body reference $\mathbf{R}$.

Figure 10 provides an overview of the entire center estimation algorithm. The remaining subsection explains robust center estimation in detail.

Given is a set $\mathbf{c}_1, ..., \mathbf{c}_k$ of k estimations for the center of the query fragment $\mathbf{F}$ in $\mathbf{R}$. The estimations originate from the annotations of the k annotated miniatures having the highest similarity with the query miniature $\mathbf{M}$. For a robust center estimation the median and the 50 percent of the estimations closest to the median are used for computation. These are the estimations located in the Inter Quartile Interval IQI $(p_{0.25}, p_{0.75})$ defined by the 25th $(p_{0.25})$ and 75th $(p_{0.75})$

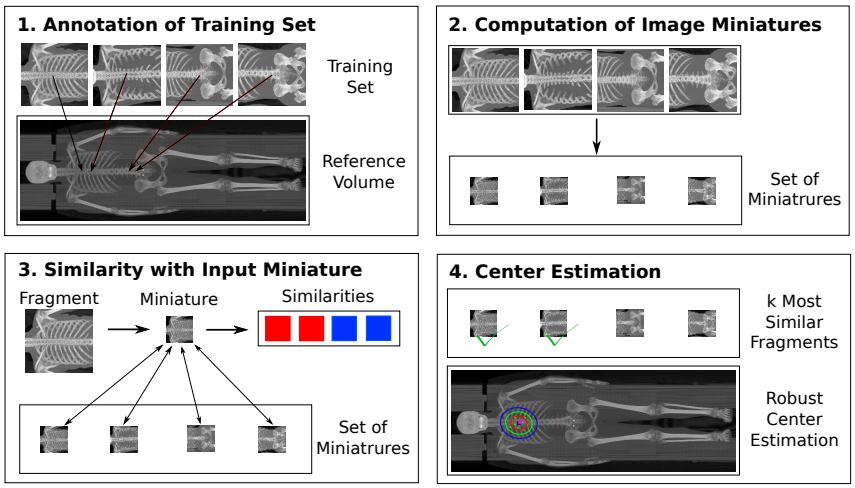


Figure 10: Overview of fragment center estimation.

percentile. Estimations located outside of this interval are not taken into account leading to a point of failure of 50 percent. In Equation (11) the formal definition of the center position estimation $\mathbf{c}_0$ is provided.

$$\mathbf{c}_0 = \frac{1}{k/2} \sum_{c_j \in IQI} \mathbf{c}_j \tag{11}$$

The output of the described method is the estimated center $\mathbf{c}_0$ of the fragment in the common whole body reference space $\mathbf{R}$. This position is used as initialization for the fragment region estimation described in the following subsection.

3.1.4 Fragment region estimation and registration

Once the initial center position $\mathbf{c}_0$ of a fragment $\mathbf{F}$ is located the dimension (corresponding region) of the fragment in the whole body reference volume $\mathbf{R}$ is estimated. In the following an algorithm for the iterative refinement of the fragment center $\mathbf{c}_0$ as well as the fragment region $\mathbf{F}_R$ with respect to the whole body reference $\mathbf{R}$ is described.

The center position is defined by the point vector $\mathbf{c}_0$; the corresponding region by a 3-D bounding box spanned by its corner points $\mathbf{x}_{cor1}$ and $\mathbf{x}_{cor2}$ along one of the diagonals cutting through the bounding box. The region estimation is an iterative procedure based on the following computation steps.

1. In the first step a corresponding reference fragment $\mathbf{F}_R$ around the initial center position $\mathbf{c}_0$ in the reference volume $\mathbf{R}$ is extracted. The region has the same dimensions as the input fragment $\mathbf{F}$. Equation (12) defines the interval $\mathbf{I}_{F_R}$ containing all voxel coordinates of the initial reference fragment $\mathbf{F}_R$.

$$\mathbf{I}_{F_R} = \left(\mathbf{c}_0 - \frac{dim(\mathbf{F})}{2}, \mathbf{c}_0 + \frac{dim(\mathbf{F})}{2} \right) \tag{12}$$

$dim(\mathbf{F}) \in \mathbb{R}^3$ defines the size (dimensions) of the input fragment $\mathbf{F}$. Equation (13) specifies the initial reference fragment $\mathbf{F}_R$ covering similar anatomical structures as the query

fragment $\mathbf{F}$.

$$\mathbf{F}_R = \mathbf{R}(\mathbf{x}_i) \;\; \forall \mathbf{x}_i \in \mathbf{I}_{F_R} \tag{13}$$

2. In step two the input fragment $\mathbf{F}$ is affine registered with the corresponding reference fragment $\mathbf{F}_R$. This registration results in an affine transformation matrix $\mathbf{A}$.

3. The transformation matrix $\mathbf{A}$ holds with its translation, rotation, scaling and shearing parameters the information to update the corresponding region $\mathbf{F}_R$ of the input fragment $\mathbf{F}$ in the whole body reference $\mathbf{R}$. For the region update the two corner coordinates $\mathbf{x}_{cor1} = (1,1,1)$ and $\mathbf{x}_{cor2} = dim(\mathbf{F})$ are transformed by the inverse of $\mathbf{A}$ as described in Equation (14).

$$\mathbf{x}'_{cor\,i} = \mathbf{A}^{-1} \cdot \mathbf{x}_{cor\,i} \;\text{ with }\; i = 1,2 \tag{14}$$

4. The updated corresponding region $\mathbf{F}'_R$ of the fragment is now spanned by the coordinates of the transformed corner points $\mathbf{x}'_{cor1}$ and $\mathbf{x}'_{cor2}$ leading to a replaced reference interval $\mathbf{I}'_{F_R}$. The updated center position $\mathbf{c}'_0$ is computed as the arithmetic mean of the transformed corner points.

The localization procedure is iterated with the updated reference Fragment $\mathbf{F}'_R$ until the region estimation converges.

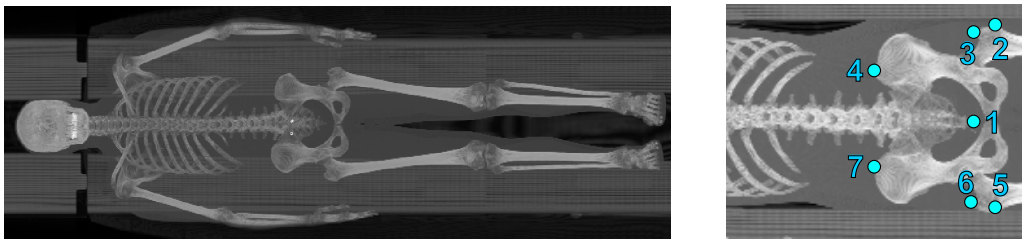
At this stage the corresponding region $\mathbf{F}_R$ of the input fragment $\mathbf{F}$ in the whole body reference volume $\mathbf{R}$ is defined and the fragment is registered to the region non-rigidly B-spline based FFD is used [58],[50].

3.1.5 Fragment based unbiased whole body template update

With the algorithm described in the previous subsections the framework has the ability to register fragments towards a common reference volume $\mathbf{R}$ covering the entire body region. This is applied to register a set of medical imaging fragments $\mathbf{F}_i$ with $i = 1,...,N$ containing body regions such as the head, the thorax or the abdomen with the reference $\mathbf{R}$. The output of these fragment registrations contains:

- The region $\mathbf{I}_{F_{Ri}}$ (see Equation 12) of a query fragment $\mathbf{F}_i$ in the whole body reference $\mathbf{R}$ as well as the corresponding reference fragment $\mathbf{F}_{Ri}$.
- The registration volume $\mathbf{F}'_i$ of the fragment $\mathbf{F}_i$ with $\mathbf{R}$ (in detail the reference fragment $\mathbf{F}_{Ri}$).
- The corresponding deformation $\mathbf{T}_{F_i,R}$ of the registration.

Based on this components the average shape and intensity model proposed by Guimond et. al. [34] is adapted to a fragment based implementation. The result is an unbiased fragment based model $\mathbf{R}_F$ representing the underlying fragments $\mathbf{F}_1,...,\mathbf{F}_N$.



(a) Example of a whole body CT recording $\mathbf{V}_s$.

(b) Abdominal test fragment $\mathbf{F}_i$ with annotated bone landmark positions and indices.

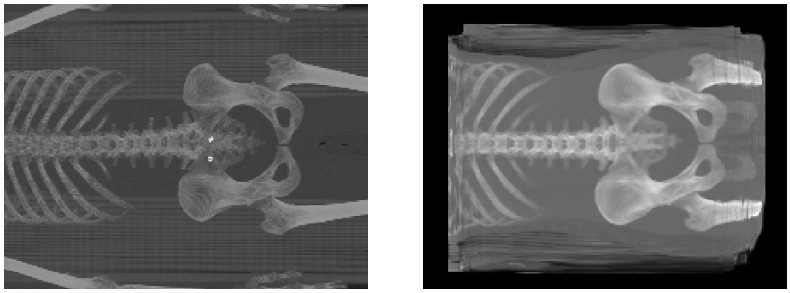
Figure 11: Test data for the evaluation of fragment based atlas construction.

3.1.6 Experimental results

This subsection evaluates the described method for fragment based unbiased shape and intensity model construction. As data serves a set of 20 whole body CT volumes $\mathbf{V}_1, ..., \mathbf{V}_{20}$ and 20 landmark annotated abdominal CT fragments $\mathbf{F}_1, ..., \mathbf{F}_{20}$. Figure 11 shows an example of the whole body CTs as well as the landmark annotated abdominal input fragments as maximum intensity projections. Although abdominal fragments are used in this evaluation, fragments containing arbitrary body regions can be used with the introduced method for fragment based atlas construction. The aim of the experiment is to show that the fragment based whole body reference volume $\mathbf{R}_F$ updated by to the underlying 20 abdominal fragments $\mathbf{F}_i$ improves the representation of the fragments in comparison to the original whole body reference volume $\mathbf{R}$. For evaluation the fragment landmarks (see Figure 11(b)) are transformed using the iterative method for fragment to whole body registration. This produces for each landmark a distribution containing 20 estimations for its locations in the reference spaces $\mathbf{R}$ and $\mathbf{R}_F$ respectively. As evaluation measure the centroid of the distributions as well as the average distance of all landmarks in a distribution to their centroid are computed. If the fragment based model is valid, the average landmark distance is expected to decrease.

Figure 13(a) shows the MDS embedding space $\mathbf{\Psi}$ for least biased initial target selection. Whole body volume $\mathbf{V}_{19}$ is located closest to the MDS embedding center $\overline{\mathbf{\psi}}$ and selected as initialization for the common reference space $\mathbf{R}$. Figure 12(a) shows the abdominal region of the initial least biased whole body reference $\mathbf{R}$. This region is updated by the described method for fragment based whole body atlas construction and shown in Figure 12(b). The atlas represents only regions covered by the given set of fragments $\mathbf{F}_i$ which is in the present case the abdomen.

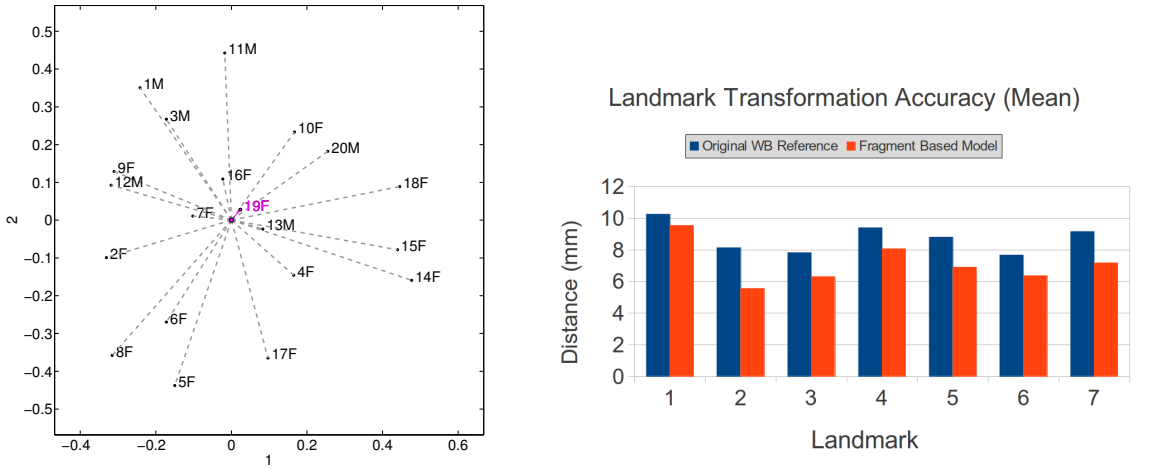
Figure 13(b) illustrates a comparison of the average distance of the landmarks (transformed by fragment to whole body registration) to their centroid before ($\mathbf{R}$) and after fragment based model computation ($\mathbf{R}_F$). The distance to the centroid is decreased with the fragment based model for each of the seven landmarks. The average distance over all seven landmarks decreases from 8,7 mm to 7,1 mm (-1,6 mm). The results show that fragment based model computation is feasible.



(a) Abdominal region of the initial whole body reference volume **R**.

(b) Fragment based average shape and intensity model $\mathbf{R}_F$.

Figure 12: Results of fragment based abdominal shape and intensity model computation.



(a) Selection of the least biased initial whole body reference **R** based on a MDS on the pairwise registration cost.

(b) Comparison of the landmark transformation accuracy of the initial whole body reference **R** and the un-biased fragment based model $\mathbf{R}_F$.

Figure 13: Evaluation of fragment based whole body atlas construction.

3.2 Using anatomical and functional data for atlas learning

Note that this section has been published in [41]. Neuroimaging data is a particularly challenging data, since the relevant structures and observations are made by different modalities. In many scenarios, such as the preparation of neurosurgery, both anatomical and functional characteristics are highly relevant. This is also the case for research questions, where similar structures should be found, and mapped to each other.

In this section we outline the methodology to deal with this data, and focus on the decoupled nature of anatomical and functional data. Neuroimaging data can be seen as one example, and consequently the approach described here, can be applied to other data as well. Note that the underlying basic mathematical tools are widespread in genetic research [45].

Here we first introduce methodology to represent data (e.g. functional data) that is not of inherent spatial nature in a manner, that allows for alignment across subjects, i.e., the matching of corresponding parts in the data. In the present case the parts are functional regions (e.g., language areas) that are present in subjects across the population.

3.2.1 Representing data

Given an fMRI sequence $\mathbf{I} \in \mathbb{R}^{T \times N}$ that contains N voxels, each characterized by an fMRI signal over T time points, we calculate matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ that assigns a non-negative symmetric weight to each pair of voxels (i, j)

$$\mathbf{W}(i,j) = e^{\frac{\langle \mathbf{I}_i, \mathbf{I}_j \rangle}{\varepsilon}}, \tag{15}$$

where $\langle \cdot, \cdot \rangle$ is the correlation coefficient of the time courses $\mathbf{I}_i$ and $\mathbf{I}_j$, and ε is the weight decay. We define a graph whose vertices correspond to voxels and whose edge weights are determined by $\mathbf{W}$ [7, 43]. In practice, we discard all edges that have a weight below a chosen threshold. This construction yields a sparse graph which is then transformed into a Markov chain. We define the Markov transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix such that $d_i = D(i,i) = \sum_j w(i,j)$ is the degree of node i. By interpreting the entries $\mathbf{P}(i,j)$ as transition probabilities, we can define the diffusion distance

$$D_t(i,j) = \sum_{i'=1,\ldots,N} \frac{(\mathbf{P}^t(i,i') - \mathbf{P}^t(j,i'))^2}{\phi(i')} \quad \text{where} \quad \phi(i) = \frac{d_i}{\sum_{i'} d_{i'}}. \tag{16}$$

The distance is determined by the probability of traveling between vertices i and j by taking all paths of at most t steps. The transition probabilities are based on the functional connectivity of node pairs; the diffusion distance integrates the connectivity values over possible paths that connect two points and defines a geometry that captures the entirety of the connectivity structure. This distance is characterized by the operator $\mathbf{P}^t$, the t^{th} power of the transition matrix. The value of the distance $D_t(i,j)$ is low if there is a large number of paths of at most length t steps with high transition probabilities between the nodes i and j.

The diffusion map coordinates $\Gamma = [\gamma_1, \gamma_2, \cdots, \gamma_N]^T$ yield a low dimensional embedding of the signal such that the resulting pairwise distances approximate diffusion distances, i.e., $\|\gamma_i - \gamma_j\|^2 \approx D_t(i,j)$ [53]. They are derived from the right eigenvectors of the transition matrix. In Appendix **??** we show that a diffusion map can be viewed as a solution to a least-squares problem. Specifically, we define a symmetric matrix $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, and let $\mathbf{L}$ be the normalized graph Laplacian

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}^{2t}\mathbf{D}^{-1/2}. \tag{17}$$

The embedding coordinates are then found as follows:

$$\Gamma^* = \underset{\Gamma \in \mathbb{R}^{N \times L}}{\operatorname{argmin}} \sum_{i,j} d_i d_j (\mathbf{L}(i,j) - \gamma_i^T \gamma_j)^2, \tag{18}$$

where L is the dimensionality of the embedding. To simplify notation, we omit t for $\mathbf{L}$ and Γ in the derivations, assuming that all the results are derived for a fixed, known diffusion time.

3.2.2 Rigidly aligning data of multiple subjects

The algorithm requires initial estimates of the latent variables and model parameters. Initialization affects convergence and the quality of the final solution. To align diffusion map Γ_s of subject s to the diffusion map Γ_r of reference subject r, we compute the inter-subject affinities

between the fMRI signals of subjects s and r using Eq. (15) and only keep those with a correlation above the threshold. This step produces a set of M node pairs $\{(i_m, j_m)\}_{m=1}^{M}$, characterized by affinities $\{w_m\}_{m=1}^{M}$. We compute a matrix $\mathbf{Q}$ that minimizes the weighted Euclidean distance between pairs of corresponding embedding coordinates

$$\mathbf{Q}^*_{sr} = \underset{\mathbf{Q}}{\operatorname{argmin}} \left[\sum_{m=1}^{M} w_m ||\mathbf{Q}\Gamma_{si_m} - \Gamma_{rj_m}||^2_{L_2} \right]. \tag{19}$$

We define matrices $\Gamma_{s_m} = [\Gamma_{si_1}, \ldots, \Gamma_{si_M}]^{\mathrm{T}}$ and $\Gamma_{r_m} = [\Gamma_{rj_1}, \ldots, \Gamma_{rj_M}]^{\mathrm{T}}$ [60].

We find initial estimates of $\{\mu_k, \Theta_k, \pi_k\}_{k=1}^{K}$ by fitting a K component Gaussian mixture model to the initial estimates of the atlas embedding coordinates $\{\Gamma_s \mathbf{Q}^*_{sr}\}_{s=1}^{S}$ for a randomly chosen reference subject r.

3.2.3 Non-linear registration

We employ the variational EM algorithm [36] to estimate the parameters of the joint model that represents the entire population to align the entire data of all subjects non-linearly. We approximate the posterior distribution of latent variables $p(\boldsymbol{\Gamma}, \boldsymbol{z}|\boldsymbol{L})$ with a product distribution of the form

$$q(\boldsymbol{\Gamma}, \boldsymbol{z}) = \prod_{s,i} q(\gamma_{si}) q(z_{si}). \tag{20}$$

and minimize the Gibbs free energy

$$\mathcal{F} = \mathbb{E}_q \left[\ln q(\boldsymbol{\Gamma}, \boldsymbol{z}) - \ln p(\boldsymbol{\Gamma}, \boldsymbol{z}, \boldsymbol{L}; \boldsymbol{\mu}, \boldsymbol{\Theta}, \boldsymbol{\pi}) \right], \tag{21}$$

where $\mathbb{E}_q$ indicates the expected value operator with respect to distribution $q(\cdot)$. [41] presents the derivation of the update rules.

3.2.4 Experimental results

Initial results were obtained on a neuroimaging data set. Data was acquired using a 3T GE Signa system (TR=2s, TE=40ms, flip angle=90°, slice gap=0mm, FOV=25.6cm, volume size of $128 \times 128 \times 27$ voxels, voxel size of $2 \times 2 \times 4$ mm^3), and for fMRI acquisition a language paradigm - typical for pre surgery planning - was conducted. Details are reported in [41].

We construct the atlas for all six subjects, the dimensionality of the diffusion map was $L = 20$, diffusion time $t = 2$. To facilitate computation we only keep nodes for which the degree is above a certain threshold. In the experiments reported here we choose a threshold of 100.

The joint representation in the embedding space should allow us to robustly capture the functional structure common to all subjects. In order to validate this, we compare the consistency of clustering structures found in the space of fMRI time courses (*Signal*), a low-dimensional (L=20) PCA embedding of these time courses (*PCA-Signal*), and the low-dimensional (L=20) embedding proposed in this paper. We report results for the initial alignment (*Linear-Atlas*) and the result of learning the joint atlas (*Atlas*). The results are evaluated by comparing clustering on the representations of the individual subjects with the clustering in the joint map. High overlap indicates the ability of the joint representation to capture the characteristics of each individual

accurately. After matching cluster labels, we use the Dice score [15] to measure the consistency between group and subject-specific assignments for each cluster.

In the exemplary results reported here (please refer to [41] for more details) the highest Dice score (0.725) for *Signal* clustering is achieved, with similar values for larger numbers of clusters, *PCA-Signal* space exhibits no noticeable improvement. Initial alignment of the diffusion maps into the *Linear-Atlas* substantially increases the Dice score of the highest ranked clusters for all K, with a maximum value of 0.876. The variational EM algorithm performed using a range of reasonable cluster numbers and further improves the cluster agreement for the top ranked clusters (0.905). These results demonstrate that the representation of fMRI time courses in the low dimensional space of diffusion maps better captures the functional connectivity structure across subjects. Not only are clustering assignments more consistent, but the anatomical characteristics of these clusters also are also more plausible.

4 Classification and Learning in Retrieval

The biomedical open access literature contains millions of images that are in principle accessible, for example in PubMed Central. Searching these images via text can use the full text or figure captions but results in several problems as about 40% of the figures are compound figures that require specific treatment for focused search and another 20% are graphs such as histograms or pie charts. Most graphs are basically never target of search as they are in itself not useful unless to underline facts in a text. By far the largest majority of radiology users only want to get diagnostic images. Do find the modality of an image can use visual features of the images in many cases. Using the caption text and also the semantics of this text via the knowledge base has also shown to significantly improve the modality classification results.

4.1 Modality Classification using noisy training data

Journals in the medical literature contain images of varius types depending on the scope of the journal and the subject of the article. Automatic categorization of image types can help in varius ways when using large image data sets. Radiologists have a preference for searching for specific radiology modadilities as it is described in [47]. Moreover, filtering out graphs and diagrams when searching for images should improve the retrieval performance and efficiency. Part of this section has been published in [30]

While the modality classification techniques used were described in the Khresmoi deliverable document D2.3, some image types in the training data were poorly represented (e.g. the training set containing very few images of that types) hurting the automatic classification accuracy. A way to overcome this problem is to expand the training set by adding (possibly) noisy data [38].

The imageClef2012 dataset that was used in the evaluations contains over 300,000 images of 75'000 articles of the biomedical open access literature. This is a subset from the PubMed Central[2] database containing over one million images. This set of articles contains all articles in PubMed that are open access but the exact copyright for redistribution varies among the

[2] http://www.ncbi.nlm.nih.gov/pmc/

Table 3: Modality classification runs

Run ID	Techniques	Fusion Rule	Training Set	k	Run Type
mc1	BoVW	n/a	original	11	Visual
mc2	BoVW + BoC	combMNZ	original	7	Visual
mc3	BoVW + BoC	combMNZ	visual non-balanced	7	Visual
mc4	BoVW + BoC	combMNZ	visual non-balanced	14	Visual
mc5	BoVW + BoC	combMNZ	visual balanced	7	Visual
mc6	BoVW + BoC + Captions	Reciprocal	original	7	Mixed
mc7	BoVW + BoC + Captions	Reciprocal	mixed non-balanced	7	Mixed
mc8	BoVW + BoC + Captions	Reciprocal	mixed non-balanced	14	Mixed
mc9	BoVW + BoC + Captions	Reciprocal	mixed balanced	7	Mixed

journals. 2000 labels images where split in two sets of 1000 images that served as training and test sets. A more detailed description of the ImageCLEFmed 2012 setup is given in [52].

To achieve the expansion of the training set, training images were used as queries in the full 300,000 images of the ImageCLEFmed 2012 data set and the l highest ranked retrieved images were added as training images into the class of the query image.

Two methods of expanding the training set were examined. In the first expansion technique, s images taken randomly of each class were used as queries. As in the original training set the number of images per class varies from 5 to 50, by choosing $s = 5, l = 20$ we can theoretically obtain a relatively balanced training set (105–150 per class) of 4,100 images. In the second, all the training images were queried, resulting in a larger non–balanced training set. E. g. for $l = 20$ an expanded training set of 21,000 images can be obtained theoretically. In practice, smaller sizes were obtained, mainly because of two reasons: retrieved images that are already contained in the training set were discarded and images retrieved multiple times by query images of different classes were discarded as well.

For a run to qualify as visual, we considered that the expanded training set used in this run needs to be created only by visual means. This means that the queries on the full data set used only visual features for the retrieval. Similarly, this was repeated using mixed (visual and textual) queries for the mixed runs. This resulted in a final number of 5 training sets (2 balanced, 2 non–balanced and the original training set).

A $k-nn$ classifier using weighted voting was used to classify the test images. For the choice of the classifier parameters the results of [31] were taken into account and $k = 11, k = 7$ were used for the visual runs. However, since the non–balanced expanded training set was significantly larger, double the value $k = 14$ was also tested for this case. The inverse of the similarity score of the $k-nn$ images was used to weight the voting.

Table 3 gives the details of the runs evaluated. The bag–of–visual–words (BoVW) and bag–of–colors (BoC) approaches were used for the visual description of the images and are described in the Khresmoi deliverable document D2.3. More information about the fusion rules (combMNZ and Reciprocal) can be found in [52].

Table 4 presents the classification accuracy. The runs *mc8*, *mc6*, *mc7* achieved the three best accuracies in the mixed run category in the imageCLEF2012 challenge. The visual runs

Table 4: Modality classification results

	Visual						Mixed			
Run ID	mc1	mc2	mc3	mc4	mc5	best run	mc6	mc7	**mc8**	mc9
Accuracy (%)	11.1	38.1	41.8	42.2	34.2	69.7	64.2	63.6	**66.2**	58.8

achieved an average performance with the inclusion of BoC as a global descriptor to improving the classification accuracy. It can be observed that the runs *mc4*, *mc8* using the non–balanced expanded training sets and $k = 14$ are outperforming the runs *mc6*, *mc2* that use the original training set. These runs also perform better than the runs *mc3*, *mc7* that use $k = 7$, confirming our hypothesis that using a larger k can improve results.

4.2 2D image retrieval using filtering

As described in the previous section, varius types of images exist in the medical literature. A large amount of these images are graphs, diagrams and other non-diagnostic images. These types of images can be considered as noise when searching for medical images because they do not contain any diagnostic information. Filtering out these images can improve the precision of the retrieval and speed up the system response.

In order to allow for such filtering, all the images of the dataset that is being searched should be mapped to their respective image types. This is a procedure that can be done during index time by using the algorithms presented in the previous section. While the classification accuracy may not be high enough in all classes, the described approach showed good performance in discriminating between diagnostic and non-diagnostic images (95.2% assuming compound image types are always correctly classified and 89.4% when compound images are considered as a third category).

The offline process that happens during indexing time, uses the classifier to assign a modality to each image and save it to the index. In query time, where text and image queries may be present, two filtering steps are taken. First a term filtering is happening so that only images with captions that contain the query keywords are included in the list of results. This step was included after the pilot user tests suggested that text terms, when present, are more important than the low-level visual characteristics. More information can be found in the Deliverable 10.2 document. A second filtering step follows that discards all the non-diagnostic images. Then a reranking of the shortlist takes place, using late fusion of the visual and textual features to result in the final retrieved list. Again, the late fusion rule that is used is the CombMNZ rule and the visual content representations are the Bag–of–Visual–Words and Bag–of–Colors. Figure 14 illustrates the pipelines for the offline and online processes.

5 Conclusion

This deliverable presents methodological approaches for learning from medical imaging data that were developed in the KHRESMOI project. It outlines methodology in three directions

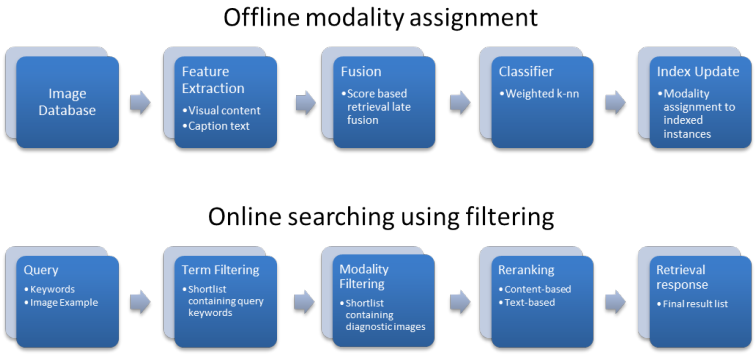


Figure 14: Offline and online processes for filtered retrieval.

that are central to image processing during indexing and during retrieval. First, localization methods are necessary to identify anatomical structures present in the imaging data. Based on this identification and an accurate localization of anatomical landmarks, subsequent processing such as pathology retrieval can deliver anatomy specific results. Secondly, atlas learning methods are necessary to compare and index large amounts of imaging data. The key is the alignment of the data to a common reference frame that allows for location specific indexing, analysis, and retrieval. Lastly classification- and filtering approaches support the image retrieval. As a complement to pure retrieval, they add categorial information such as imaging modalities that can improve the usability of retrieval results.

6 References

[1] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert R. G. Lanckriet, David J. Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. *Journal of Machine Learning Research - Proceedings Track*, 2:11–18, 2007.

[2] B. Andres, J. H. Kappes, U. Koethe, C. Schnörr, and F. A. Hamprecht. An empirical comparison of inference algorithms for graphical models with higher order factors using OpenGM. In *Proc. of DAGM 2010*, pages 353–362, 2010.

[3] Martin Bergtholdt, Jörg Kappes, Stefan Schmidt, and Christoph Schnörr. A Study of Parts-Based Object Class Detection Using Complete Graphs. *IJCV*, 87(1-2):93–117, Mar 2010.

[4] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast Approximate Energy Minimization via Graph Cuts. *TPAMI*, 23(11):1222–1239, 2001.

[5] Andreas Burner, Rene Donner, Marius Mayerhoefer, Markus Holzer, Franz Kainberger, and Georg Langs. Texture bags: Anomaly retrieval in medical images based on local 3d-texture similarity. *Workshop on Medical Content-based Retrieval for Clinical Decision Support at MICCAI 2011*, September 2011.

[6] Yizong Cheng. Mean Shift, Mode Seeking, and Clustering. *TPAMI*, 17(8):790–799, 1995.

[7] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *App. Comp. Harm. An.*, 21:5–30, 2006.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graha. Active Shape Models - Their Training and Application. *CVIU*, 61(1):38–59, January 1995.

[9] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active Appearance Models. *TPAMI*, 23(6):681–685, 2001.

[10] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape. *IJCV*, 72(2):195–215, 2007.

[11] A. Criminisi, J. Shotton, D. Robertson, , and E. Konukoglu. Regression forests for efficient anatomy detection and localization in ct studies. In *Medical Computer Vision 2010: Recognition Techniques and Applications in Medical Imaging, MICCAI workshop*, pages 106–117, 2010.

[12] Antonio Criminisi, Jamie Shotton, and Stefano Bucciarelli. Decision Forests with Long-Range Spatial Context for Organ Localization in CT Volumes. In *Proc. of MICCAI workshop on Probabilistic Models for Medical Image Analysis (MICCAI-PMMIA)*, 2009.

[13] W. R. Crum, T. Hartkens, and D. L. G. Hill. Non-rigid image registration: theory and practice. *The British Journal of Radiology*, 77:140–153, 2004.

[14] Marleen de Bruijne and Mads Nielsen. Shape Particle Filtering for Image Segmentation. In *Proc. MICCAI*, pages 168–175, 2004.

[15] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[16] E Dittrich, T Riklin-Raviv, G Kasprian, PC Brugger, D Prayer, and G Langs. Learning a spatio-temporal latent atlas for fetal brain segmentation. In *Proceedings of MICCAI Workshop on Image Analysis of Human Brain Development*, Aug 2011.

[17] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31:198–211, 2007.

[18] René Donner, Erich Birngruber, Helmut Steiner, Horst Bischof, and Georg Langs. Localization of 3D Anatomical Structures Using Random Forests and Discrete Optimization. In *Proc. MICCAI Workshop on Medical Computer Vision*, pages 86–95, 2010.

[19] René Donner, Sebastian Haas, Andreas Burner, Markus Holzer, Horst Bischof, and Georg Langs. Evaluation of fast 2d and 3d medical image retrieval approaches based on image miniatures. In *Proceedings of the Second MICCAI international conference on Medical Content-Based Retrieval for Clinical Decision Support*, MCBR-CDS'11, pages 128–138, Berlin, Heidelberg, 2012.

[20] René Donner, Georg Langs, Branislav Micusik, and Horst Bischof. Generalized Sparse MRF Appearance Models. *Image and Vision Computing*, 28(6):1031 – 1038, 2010.

[21] René Donner, Björn Menze, Horst Bischof, and Georg Langs. Fast Anatomical Structure Localization Using Top-down Image Patch Regression. In *Proc. MICCAI Workshop on Medical Computer Vision*, 2012.

[22] René Donner, Björn Menze, Horst Bischof, and Georg Langs. Global Localization of 3D Anatomical Structures by Pre-filtered Hough Forests and Discrete Optimization. *Medical Image Analysis, in press*, 2013.

[23] René Donner, Branislav Micusik, Georg Langs, and Horst Bischof. Sparse MRF appearance models for fast anatomical structure localisation. In *Proc. BMVC'07*, 2007.

[24] René Donner, Branislav Mičušík, Georg Langs, Lech Szumilas, Philipp Peloschek, Klaus Friedrich, and Horst Bischof. Object localization based on markov random fields and symmetry interest points. In *Proc. MICCAI*, pages 460–468, 2007.

[25] René Donner', Horst Wildenauer, Horst Bischof, and Georg Langs. Weakly supervised group-wise model learning based on discrete optimization. In *Proc. MICCAI*, 2009.

[26] Matthias Dorfer. A framework for medical-imaging-fragment based whole body atlas construction. Master's thesis, Vienna University of Technology, 2013.

[27] N Fakhrai, P Widhalm, C Chiari, M Weber, R Donner, G Langs, and P Peloschek. Automatic assessment of the knee alignment angle on full-limb radiographs. *European Journal of Radiology*, 74(1):236–240, April 2010.

[28] B. Fischl, D.H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, et al. Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*, 33(3):341–355, 2002.

[29] J. Gall and V Lempitsky. Class-specific Hough Forests For Object Detection. In *Proc. CVPR*, pages 1022–1029, 2009.

[30] Alba García Seco de Herrera, Dimitrios Markonis, Ivan Eggel, and Henning Müller. The medGIFT group in ImageCLEFmed 2012. In *Working Notes of CLEF 2012*, 2012.

[31] Alba García Seco de Herrera, Dimitrios Markonis, and Henning Müller. Bag of colors for biomedical document image classification. In Hayit Greenspan and Henning Müller, editors, *Medical Content–based Retrieval for Clinical Decision Support*, MCBR–CDS 2012. Lecture Notes in Computer Sciences (LNCS), October 2013.

[32] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

[33] Leo Grady. Random Walks for Image Segmentation. *TPAMI*, 28(11):1768–1783, 2006.

[34] A. Guimond, J. Meunier, and J. P. Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77(2):192 – 210, 2000.

[35] SN Issa, D Dunlop, A Chang, J Song, PV Prasad, A Guermazi, C Peterfy, S Cahue, M Marshall abd D Kapoor, K Hayes, and L Sharma. Full-limb and knee radiography assessments of varus-valgus alignment and their relationship to osteoarthritis disease features by magnetic resonance imaging. *Arthritis Rheum*, 57(3):398–406, 2007.

[36] T.S. Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, pages 129–159, 2000.

[37] S. Joshi, B. Davis, B. M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23 Supplement 1:151–160, 2004.

[38] Jayashree Kalpathy-Cramer, Henning Müller, Steven Bedrick, Ivan Eggel, Alba García Seco de Herrera, and Theodora Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.

[39] B. Michael Kelm, S. Kevin Zhou, Michael Suehling, Yefeng Zheng, Michael Wels, and Dorin Comaniciu. Detection of 3D Spinal Geometry Using Iterated Marginal Space Learning. In *Proc. MICCAI Workshop on Medical Computer Vision*, pages 96–105, 2010.

[40] Marie-Eve Lamarre, Stefan Parent, Hubert Labelle, Carl-Eric Aubin, Julia Joncasand Anne Cabral, and Yvan Petit. Assessment of Spinal Flexibility in Adolescent Idiopathic Scoliosis: Suspension Versus Side-Bending Radiography. *Spine*, 34(6):591–597, 2009.

[41] Georg Langs, Danial Lashkari, Andrew Sweet, Yanmei Tie, Laura Rigolo, Alexandra J Golby, and Polina Golland. Learning an atlas of a cognitive process in its functional geometry. *Inf Process Med Imaging*, 22:135–46, 2011.

[42] Georg Langs, Philipp Peloschek, Horst Bischof, and Franz Kainberger. Automatic quantification of joint space narrowing and erosions in rheumatoid arthritis. *IEEE TMI*, 28(1):151–164, Jan 2009.

[43] Georg Langs, Dimitris Samaras, Nikos Paragios, Jean Honorio, Nelly Alia-Klein, Dardo Tomasi, Nora D Volkow, and Rita Z Goldstein. Task-specific functional brain geometry from model maps. In *Proc. of MICCAI*, volume 11, pages 925–933, 2008.

[44] Vincent Lepetit and Pascal Fua. Keypoint Recognition using Randomized Trees. *TPAMI*, 28(9):1465–1479, 2006.

[45] Bo Li, Chun-Hou Zheng, De-Shuang Huang, Lei Zhang, and Kyungsook Han. Gene expression data classification using locally linear discriminant embedding. *Computers in Biology and Medicine*, 40(10):802–810, October 2010.

[46] Mallinckrodt Institute of Radiology Washington University School of Medicine. Whole body morphometry project. http://nrg.wustl.edu.

[47] Dimitrios Markonis, Markus Holzer, Sebastian Dung, Alejandro Vargas, Georg Langs, Sascha Kriewel, and Henning Müller. A survey on visual information search behavior and requirements of radiologists. *Methods of Information in Medicine*, 51(6):539–548, 2012.

[48] Björn Menze, B M Kelm, D N Splitthoff DN, U Koethe, and F Hamprecht. On oblique random forests. In *Proc. of ECML*, pages 453–469, 2011.

[49] Branislav Mičušík and Tomas Pajdla. Multi-label image segmentation via max-sum solver. In *Proc. CVPR*, pages 1–6, 2007.

[50] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. *Comput. Methods Prog. Biomed.*, 98(3):278–284, June 2010.

[51] Albert Montillo, Jamie Shotton, John Winn, Juan Eugenio Iglesias, Dimitri Metaxas, and Antonio Criminisi. Entangled Decision Forests and Their Application for Semantic Segmentation of CT Images. In *Proc. IPMI*, pages 184–196, 2011.

[52] Henning Müller, Alba García Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner Fushman, Sameer Antani, and Ivan Eggel. Overview of the imageclef 2012 medical image retrieval and classication tasks. In *Working Notes of CLEF 2012 (Cross Language Evaluation Forum)*, September 2012.

[53] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis. Diffusion maps-a probabilistic interpretation for spectral embedding and clustering algorithms. *Principal Manifolds for Data Visualization and Dimension Reduction*, pages 238–260, 2007.

[54] H. Park, Peyton H. Bland, Alfred O. Hero, and Charles R. Meyer. Least biased target selection in probabilistic atlas construction. In *Proceedings of the 8th international conference on Medical image computing and computer-assisted intervention - Volume Part II*, MICCAI'05, pages 419–426, Berlin, Heidelberg, 2005.

[55] Olivier Pauly, Ben Glocker, Antonio Criminisi, Diana Mateus, Axel Martinez Möller, Stephan Nekolla, and Nassir Navab. Fast Multiple Organ Detection and Localization in Whole-Body MR Dixon Sequences. In *Proc. MICCAI*, pages 239–247, 2011.

[56] P Peloschek, G Langs, M Weber, J Sailer, M Reisegger, H Imhof, H Bischof, and F Kainberger. An automatic model-based system for joint space measurements on hand radiographs: Initial experience. *Radiology*, 245(3):855–862, 2007.

[57] Martin Roberts, Timothy F. Cootes, and Judith E. Adams. Vertebral morphometry - semiautomatic determination of detailed shape from dual-energy x-ray absorptiometry images using active appearance models. *Investigative Radiology*, 41(12):849–459, 2006.

[58] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Non-rigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712 –721, August 1999.

[59] Stefan Schmidt, Jörg Kappes, Martin Bergtholdt, Vladimir Pekar, Sebastian Dries, Daniel Bystrov, and Christoph Schnörr. Spine Detection and Labeling Using a Parts-Based Graphical Model. In *Proc. IPMI*, pages 122–133, 2007.

[60] G.L. Scott and H.C. Longuet-Higgins. An algorithm for associating the features of two images. *Proceedings: Biological Sciences*, 244(1309):21–26, 1991.

[61] S. Seifert, A. Barbu, S. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu. Hierarchical Parsing and Semantic Navigation of Full Body CT Data. In *SPIE Proceedings*, volume 7259, 2009.

[62] Toby Sharp. Implementing Decision Trees and Forests on a GPU. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Proc. ECCV*, pages 595–608, 2008.

[63] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moorea, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from a Single Depth Image. In *Proc. CVPR*, pages 1297–1304, 2011.

[64] Leo Breiman Statistics and Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[65] C. Studholme, D.L.G. Hill, and D.J. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71 – 86, 1999.

[66] Zhuowen Tu and Xiang Bai. Auto-context and Its Application to High-level Vision Tasks and 3D Brain Image Segmentation. *TPAMI*, 32(10):1744–1757, 2010.

[67] Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[68] Yefeng Zheng, Bogdan Georgescu, and Dorin Comaniciu. Marginal Space Learning for Efficient Detection of 2D/3D Anatomical Structures in Medical Images. In *Proc. IPMI*, pages 411–422, 2009.

[69] Xiangrong Zhou, Song Wang, Huayue Chen, Takeshi Hara, Ryujiro Yokoyama, Masayuki Kanematsu, and Hiroshi Fujita. Automatic localization of solid organs on 3D CT images by a collaborative majority voting decision based on ensemble learning. *Computerized Medical Imaging and Graphics, doi:10.1016/j.compmedimag.2011.12.004*, 2012.