Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization

Martin Urschler a,b,c, Thomas Ebner d, Darko Štern a,b,*

a Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
b Institute for Computer Graphics and Vision, Graz University of Technology, Austria
c BioTechMed-Graz, Austria

A R T I C L E   I N F O
Article history:
Received 29 March 2017
Revised 27 July 2017
Accepted 11 September 2017
Available online 21 September 2017

Keywords:
Anatomical landmarks
Localization
Random regression forest
Coordinate descent

A B S T R A C T
In approaches for automatic localization of multiple anatomical landmarks, disambiguation of locally similar structures as obtained by locally accurate candidate generation is often performed by solely including high level knowledge about geometric landmark configuration. In our novel localization approach, we propose to combine both image appearance information and geometric landmark configuration into a unified random forest framework integrated into an optimization procedure that iteratively refines joint landmark predictions by using the coordinate descent algorithm. Depending on how strong multiple landmarks are correlated in a specific localization task, this integration has the benefit that it remains flexible in deciding whether appearance information or the geometric configuration of multiple landmarks is the stronger cue for solving a localization problem both accurately and robustly. Furthermore, no preliminary choice on how to encode a graphical model describing landmark configuration has to be made. In an extensive evaluation on five challenging datasets involving different 2D and 3D imaging modalities, we show that our proposed method is widely applicable and delivers state-of-the-art results when compared to various other related methods.

© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.

1. Introduction

A large number of medical image analysis applications rely on automatic anatomical landmark localization algorithms as a preliminary step for, e.g. segmentation based on deformable and statistical shape models (Heimann and Meinzer, 2009; Zhang et al., 2012; Lay et al., 2013), registration of images using rigid (Hajnal et al., 2001) and deformable transformations (Johnson and Christensen, 2002; Urschler et al., 2006), construction of anatomical atlases for population studies (Toews et al., 2010), or to focus on anatomical structures of interest for computer-aided diagnosis (Doi, 2007) and regression tasks like skeletal age estimation (Thodberg et al., 2009; Stern et al., 2014). However, due to anatomical variation and especially in the presence of potentially ambiguous (i.e. locally similar) landmarks, the task of both accurate and robust anatomical landmark localization becomes challenging.

The majority of recent machine learning approaches for multiple landmark localization either solely makes use of appearance features from the whole input image, or combines appearance features derived from the local vicinities of landmarks with a parametric or graphical model fitting step on top. Criminisi et al. (2013) have shown that global anatomical configuration can be captured with a regression-based prediction strategy, when the scale of exclusively used appearance features is allowed to vary up to the image size. By learning from the appearance of all anatomical structures present in the training dataset, their regression-based random forest (RF) for organ bounding box localization demonstrates robustness in the presence of ambiguities, however, their approach lacks in accuracy. On the other hand, restricting appearance features to a local neighborhood during training enables accurate prediction of landmark positions, however, on its own it inevitably leads to false positives due to locally similar anatomical structures not being distinguishable. Therefore, to select a likely global landmark configuration, sophisticated geometric configuration models like graphical (Donner et al., 2013) or constrained local models (Lindner et al., 2015) are required on top for disambiguation.

* Corresponding author at: Institute for Computer Graphics and Vision, Graz University of Technology, Austria.
E-mail address: stern@icg.tugraz.at (D. Štern).

https://doi.org/10.1016/j.media.2017.09.003
1361-8415/© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)
In this work we propose to integrate image appearance information and geometric landmark configuration into a unified framework. Thus, we aim for structured output prediction, where the landmark configuration is implicitly modeled by simultaneously taking into account both appearance features derived from the whole image and geometric constraints that landmarks impose on each other. We pose our novel approach as an optimization problem that maximizes the joint probability for locating all landmarks simultaneously given an input image. This problem is solved by the coordinate descent algorithm (Wright, 2015), which decouples the NP hard problem of simultaneously searching for all landmarks by successively optimizing for individual ones. Thus, as illustrated in Fig. 1, by passing information to other landmarks and receiving information required to update its own position, each landmark location iteratively achieves a self-improvement according to both image appearance features and the current best beliefs of where the other landmarks are located.

1.1. Related work

While localization of multiple anatomical landmarks from medical images is in principle possible by solely relying on heuristic intensity-based image processing techniques (Wörz and Rohr, 2006; Donner et al., 2007), nowadays machine learning approaches capturing prior knowledge on appearance and shape from training data are predominantly used due to better generalizability in the presence of anatomical variation. These learning methods allow landmark localization either by using predictions for each landmark based on appearance information extracted from the whole image, or by combining potentially ambiguous predictions based on locally restricted appearance information with an explicit regularization model encoding geometric landmark configuration.

Using a regression-based approach implemented in an RF framework, Pauly et al. (2011) and Criminisi et al. (2013) proposed to learn the relative positions between the organs of interest and all the anatomy available in the training data solely with arbitrary scale Haar-like appearance features, i.e. the size of the appearance features and their distance from the training voxels are arbitrary. In a test image, the position of organs of interest is then obtained from the recognized anatomy, implemented by accumulating the relative positions of each voxel of the image to the organ of interest. Mainly aiming for image retrieval applications, the focus of Criminisi et al. (2013) lay on fast and robust but approximate and inaccurate bounding box localization of larger organ structures. The stratified decision forest method (Oktay et al., 2017) extended the work of Criminisi et al. (2013) for cases when there is significant variation of pose and size in a dataset. Gauriau et al. (2015) combined cascaded regression with a simple statistical shape prior derived from segmentation masks. For a multi-object segmentation task, Glocker et al. (2012b) developed an RF based method combining a classification and regression objective function. Ebner et al. (2014) adapted the multi-output regression approach of Criminisi et al. (2013) to the landmark localization task. They improved localization accuracy by putting more trust into the surrounding anatomy of the predicted landmark, since the relative position of closer anatomy shows less variation regarding the landmark location. This was achieved by introducing a distance weight at testing time that reduces influence of regions farther away from the anatomical landmark. Additionally, they used a two-stage regression RF cascade, with the scale range of Haar-like appearance features in the second stage restricted on capturing only the appearance of the structures locally surrounding a landmark. Applied to a segmentation task, Peter et al. (2015) developed this idea further by automatically selecting the most informative scale range of the appearance features defining surrounding

**Fig. 1.** Our proposed method for multiple landmark localization integrates appearance information and geometric configuration by first accurately generating landmark candidates (top), followed by iteratively refining landmark configuration based on coordinate descent optimization to achieve robustness in the presence of locally similar structures (bottom).
anatomy. All these regression-based approaches showed that by learning just from appearance information extracted from training data, they are robust in the presence of locally similar structures. However, due to variations in relative positions of the anatomy used for robust prediction of landmarks, a tradeoff between accuracy and robustness is involved. Recently, a new direction in regression-based landmark localization started to appear, where CNNs, i.e., convolutional neural networks (LeCun et al., 1998), are used to model geometric configuration solely from appearance features (Payer et al., 2016). Despite their promising initial results, accuracy and robustness of this method depends on the availability of large training datasets and is limited by the need to restrict resolution of training images due to computational complexity and memory requirements.

To achieve high accuracy in landmark localization, most state-of-the-art methods derive predictions solely from local appearance features describing close anatomy, with feature extraction performed in the RF framework (Breiman, 2001). As a consequence, an additional regularization step involving an explicit model of geometric configuration is required for jointly treating all landmarks to improve robustness in the presence of locally similar structures. Here, a widely used regularization strategy fits a learned point distribution model to local appearance-based continuous predictions containing potentially ambiguous responses. This strategy goes back to seminal work from Cootes et al. (1995), where Active Shape Model fitting was implemented as iteratively updating a point configuration by optimizing landmark locations locally using appearance-based predictions and regularizing the landmark configuration with the point distribution model. Variations of this shape prior constraint (Heimann and Meinzer, 2009) are still widely used in medical image analysis applications, however, these methods have shown limited flexibility in the presence of anatomical abnormalities or pathologies (Beichel et al., 2005) due to their strong reliance on the constructed parametric model and the dependency on robust initialization to avoid local minima during fitting. Despite achieving state-of-the-art localization results with their accurate local RF regression voting step, the constrained local model of Lindner et al. (2015) still suffers from some of these drawbacks due to their complex, parameter-intensive multiscale landmark fitting.

An alternative strategy for regularizing potentially ambiguous responses discretizes the prediction derived from local appearance features by extracting a set of local maximas as candidates for each landmark. This enables the use of learned spatial priors by tree-structured (Felzenszwalb and Huttenlocher, 2005) or Markov Random Field (MRF)-based (Besbes et al., 2009) graphical models, which were successfully used for lung segmentation (Ibragimov et al., 2012), robust detection of the spinal column in human (Glocker et al., 2012a; 2013; Kelm et al., 2013) or zebrafish (Richmond et al., 2015), landmark localization in brain images (Toews and Arbel, 2007), and for labeling landmarks in hand (Donner et al., 2013) or whole-body CT images (Potesil et al., 2015). Representative for these algorithms, Donner et al. (2013) performs landmark localization using local appearance based RFs in a Hough voting scheme (Gall et al., 2011), which leads to landmark candidates that are regularized by solving an MRF encoding the spatial constraints for the most likely landmark configuration.

A common property of all methods using an explicit model of geometric configuration is that after generating local predictions, appearance information is never further used in the regularization stage. Štern et al. (2016) have recently shown that the regression-based approach using the same arbitrary scale appearance features as in (Criminisi et al., 2013) can also be used for regularization without any explicitly modeled shape constraint. However, since each landmark is trained independently of each other, a simple MRF model is required on top for joint prediction to achieve state-of-the-art results on a challenging hand X-ray dataset.

1.2. Contributions

By integrating image appearance information and geometric landmark configuration into a single regression-based landmark prediction framework, we overcome the limitations of related localization approaches. Depending on how strong multiple landmarks are correlated in a specific localization task, this integration has the benefit that it remains flexible in deciding whether appearance information or the geometric configuration of multiple landmarks is the stronger cue for solving a localization problem. Furthermore, no preliminary choice on how to encode a graphical model describing landmark configuration has to be made, instead geometric constraints are learned directly from training data and encoded implicitly in the regression framework. For this integration, RFs are perfectly suited to implement the decision between geometric and appearance features randomly at the node level. Finally, using a single framework for both accurate and robust landmark localization removes the need for heuristics to initialize landmark fitting and simplifies parameter tuning, which are limitations of point distribution and constrained local models.

Crucial for the performance of our novel method is the iterative refinement of joint landmark predictions, implemented as coordinate descent optimization. Inspiration for this idea comes from the auto context work of Tu and Bai (2010) used for image segmentation. In auto context, initial predictions of a labeling are refined in a cascade of RFs, where prediction results of a previous cascade stage are used as additional features in the next stage. Since each cascade stage has to be trained on a separate dataset, this method is dependent on a large amount of available training data. Differently to auto context, in our approach refinement is not performed with a cascade, but is solely based on a single RF trained on the ground-truth landmark annotations, which is iteratively applied during testing as explained in Section 2.

Overall, our novel integrated landmark localization approach is applicable in a very generic way and can be used for localization of multiple anatomical landmarks in different 2D and 3D imaging modalities (X-ray, CT, MRI) depicting distinct anatomical structures. We consider our experimental evaluation (see Section 3) as another important contribution, since, in addition to the diversity of the investigated datasets, we also extensively compare our proposed method to a variety of approaches, both regression-based and approaches that incorporate geometric constraints using explicitly defined models. As shown in Section 4, with these experiments we quantitatively demonstrate state-of-the-art performance on all evaluated benchmark datasets, as well as superior performance when no strong geometric correlation between landmarks is available.

2. Method

To jointly localize multiple anatomical landmarks from a given input image, we propose a coordinate descent optimization strategy (Wright, 2015), which is an iterative scheme decoupling the simultaneous search for all landmark locations by successively optimizing for individual landmarks. During each iteration, this scheme maximizes the conditional probability of each landmark position given appearance information from the input image and current best beliefs of where all other landmarks are located (see Fig. 1, bottom). To estimate this conditional probability, we use a classifier trained on ground-truth landmark locations that encodes geometric constraints. However, during the iterative optimization anatomically infeasible landmark configurations may occur. We reduce the dependency of the estimated conditional probability on
feasible landmark configurations, by training the classifier as an ensemble with a bagging strategy (Breiman, 1996), where each classifier instance predicts using solely a subset of landmarks (see Fig. 2). As shown in Fig. 1 (top), our coordinate descent strategy is initialized with landmark candidates predicted from independently trained landmark regressors based on local appearance features. Since this initialization represents highly accurate localizations of multiple landmarks, but a potentially erroneous global landmark configuration due to ambiguities, our proposed iterative coordinate descent strategy effectively leads to both accurate and robust landmark predictions. In Section 2.1 we describe our optimization strategy in more detail, followed by a description of the RF framework that we use for implementing this strategy in Section 2.2.

2.1. Coordinate descent optimization

Selecting the most likely landmark configuration \( \hat{x} = (x_1, \ldots, x_t)^T \) from a set of candidate positions for each landmark can be defined by the following optimization problem:

\[
\hat{x} = \arg \max_{x} p(x \mid \Theta, I),
\]

where \( \Theta \) represents a model derived from annotated training data that jointly estimates landmark positions and \( I \) is an input image at testing time. This general discrete optimization problem is NP-hard, a widely used approach to solve it approximately is by using an MRF of discretized landmark candidates as proposed e.g. in Donner et al. (2013). The discrete MRF solver requires to specify a graph structure since connecting all landmarks with each other quickly makes solving the MRF intractable when the number of landmarks \( L \) exceeds small numbers. Instead of simultaneously optimizing over a large set of variables, we consider a simpler problem that iteratively optimizes each landmark location independently, while keeping the other locations fixed. With \( t \in \{1, \ldots, T\} \) resembling multiple iterations, this strategy is given by

\[
x_t^{i+1} = \arg \max_{x_t \in \mathcal{X}_t} p(x_t^{i+1} \mid x_1^{i+1}, \ldots, x_{t-1}^{i+1}, x_t^0, \ldots, x_t^i \mid \Theta_t, I).
\]

Defining \( x_t^{i+1} \) as the set \( \{x_t^{i+1}, x_1^{i+1}, \ldots, x_T^{i+1}\} \), the notation is simplified:

\[
x_t^{i+1} = \arg \max_{x_t \in \mathcal{X}_t} p(x_t \mid x_t^0, \Theta_t, I).
\]

where \( \mathcal{X}_t \) represents the set of candidate positions for landmark \( i \in \{1, \ldots, L\} \) and \( \Theta_t \) is a model independently estimating the position of landmark \( i \). By alternating between each of the \( L \) landmarks, the strategy in (2) is known as the coordinate descent optimization algorithm (Wright, 2015), which can be shown to converge to the global optimum in case of convex and smooth cost functions. In our case of a non-convex cost function, a local optimum is achieved which depends on the initialization, however, most importantly, the solution takes into account geometrical constraints that landmarks impose onto each other. When optimizing for a single landmark, the joint probability distribution in (3) can be reformulated using the product rule of probability:

\[
x_t^{i+1} = \arg \max_{x_t \in \mathcal{X}_t} p(x_t \mid x_t^0, \Theta_t, I) \cdot p(x_t^0 \mid \Theta_t, I).
\]

By omitting the prior probabilities of landmarks other than \( i \) as constant for the optimization, the proposed iterative optimization can be written as maximizing the following conditional probabilities for all \( i \in \{1, \ldots, L\} \):

\[
x_t^{i+1} = \arg \max_{x_t \in \mathcal{X}_t} p(x_t \mid x_t^0, \Theta_t, I).
\]

We distinguish two kinds of models \( \Theta_t \), a model \( \Theta_t^0 \) that is based on local appearance features providing accurate but ambiguous predictions, and our novel global context model \( \Theta_t^f \), which combines geometrical configuration of landmarks with arbitrarily scaled appearance features to improve robustness of predictions in the presence of locally similar structures. Thus, (5) can be rewritten as:

\[
x_t^{i+1} = \arg \max_{x_t \in \mathcal{X}_t} p(x_t \mid x_t^0, \Theta_t^0, \Theta_t^f, I).
\]

By using Bayesian rule we can reformulate the conditional probability in (6) under the assumption that the two models \( \Theta_t^0, \Theta_t^f \) are conditionally independent regarding \( x \) and image \( I \):

\[
x_t^{i+1} = \arg \max_{x_t \in \mathcal{X}_t} p(x_t \mid \Theta_t^f, I) \cdot p(x_t^0 \mid \Theta_t^0, \Theta_t^f, I).
\]

where the probabilities being constant for the optimization are omitted. Without giving any prior preferences to landmark candidates, the probability \( p(x_t^0) \) is constant for the maximization. Thus, as illustrated in Figs. 1 and 2, we propose our coordinate descent optimization scheme as follows:

\[
x_t^{i+1} = \arg \max_{x_t \in \mathcal{X}_t} p(x_t \mid x_t^0, \Theta_t^0, I) \cdot p(x_t \mid x_t^0, \Theta_t^f, I).
\]

Since \( p(x_t \mid \Theta_t^f, I) \) does not take into account the iterative improvement of the geometrical landmark configuration, the prediction based on local appearance features has a fixed contribution to each landmark candidate during the iteration. While the multiplication in (8) can be performed in a fully probabilistic manner, in our implementation the \( N_M \) strongest local maxima of \( p(x_t \mid \Theta_t^0, I) \) are used as the set of candidate positions \( \mathcal{X}_t \) for each landmark \( i \) during iterations (see Table 3 in Appendix A for more details). Additionally, for each landmark \( i \) the global maximum of the local appearance based prediction

\[
x_t^i = \arg \max_{x_t \in \mathcal{X}_t} p(x_t \mid \Theta_t^0, I)
\]

provides the initial landmark configuration \( x_t^0 = \{x_t^0, \ldots, x_t^L\} \) for optimization (see Fig. 1, top).
Overall, in our method the robustness of the accurate, but ambiguous prediction from local appearance features is iteratively improved with a prediction based on a model \( \Theta^*_i \) combining arbitrarily scaled appearance features and the geometrical configuration of the current best beliefs of where other landmarks are located (\( x^*_i \)), as shown in an example in Fig. 3(b).

2.1.1. Estimation of conditional probabilities

To obtain the most probable structured output prediction \( \hat{x} \) using (8), we require an estimation of conditional probabilities modeling local appearance \( \Theta^*_i \) and global context \( \Theta^*_g \), respectively. The estimation of \( p(x \mid \Theta^*_i, I) \) is performed using an RF approach, modeling \( \Theta^*_i \) based on local appearance features extracted from a set of training images \( D \) for each landmark \( i \) independently, which will be explained in Section 2.2.1.

Estimating for each landmark \( i \) the conditional probability describing global context requires modeling \( \Theta^*_g \), from arbitrarily scaled appearance features extracted from a set of training images \( D \) and from the geometric configuration of the landmarks \( x_i \) other than landmark \( i \). While learning a model for all combinations of geometric configurations of landmarks \( x_i \) that may be seen during testing is computationally intractable, learning from the configuration that is expected to be seen during testing as in auto context (Tu and Bai, 2010) has its own drawbacks. Adaptation of the auto context strategy for localization requires larger amounts of training data and may still lead to infeasible landmark configurations during testing due to overfitting to predictions from the training dataset.

We suggest a different strategy to model parameters \( \hat{\Theta}^g_i \) of the conditional probability by describing global context solely from ground truth annotations of the landmark locations \( x^*_i \) together with arbitrarily scaled appearance features extracted from training images \( D \) as follows:

\[
\hat{\Theta}^g_i = \arg\max_{\Theta} p(\Theta \mid x^*_i, D). \tag{10}
\]

Since a single model per landmark \( \Theta^*_g \) is used for all iterations of our coordinate descent optimization, this strategy does not require additional validation sets for training as in auto context. However, in our approach there is a discrepancy between training on ground truth landmark annotations \( x^*_i \) and testing on the best beliefs of where other landmarks are located resulting from the coordinate descent iteration \( x^*_i \). To overcome inevitable overfitting to geometrical configurations \( x^*_i \), we treat possible anatomically infeasible landmark configurations \( x^*_i \) as noisy observations. Thus, the discrepancy between training a model \( \hat{\Theta}^g_i \) on ground truth landmark annotations \( x^*_i \) and testing on configurations unseen during training \( x^*_i \) can be interpreted as finding a suitable bias-variance tradeoff involving different levels of model complexity. When using all \( L - 1 \) available landmarks for modeling \( \hat{\Theta}^g_i \) during training, we achieve a model with low bias but high variance, which will overfit, i.e. it will not generalize well to testing data. On the other hand, a model with very low complexity (\( K < L - 1 \)) will have better generalization capabilities but it shows low overall localization performance. To avoid overfitting, we therefore use an ensemble of lower complexity models trained on randomly chosen subsets of \( K \) landmarks, where the \( T_k \) predictions of the ensemble are averaged:

\[
\hat{\Theta}^g_i = \arg\max_{\Theta} \frac{1}{T_k} \sum_{e=1}^{T_k} p(\Theta \mid x^*_k(i), D). \tag{11}
\]

Here, for predictor \( e \in \{1, ..., T_k\} \), \( R_k(i) \) is a random subset of \( K \) landmarks other than the landmark of interest \( i \). Depending on the number of landmarks located on anatomically infeasible positions during initialization, an optimal tradeoff between bias and variance can be found by varying model complexity parameter \( K \).

For estimating conditional probability (11) combining global context from arbitrarily scaled appearance features and geometric configuration of landmarks, we use the RF framework due to its simplicity in integrating different feature types as explained in the next section.

2.2. Random forest framework

We use RFs (Breiman, 2001) in (8) to model our local appearance based regressor \( \hat{\Theta}^g_i \) generating locally accurate landmark predictions as well as our ensemble of classifiers \( \hat{\Theta}^g_i \) that is used to remove ambiguities due to locally similar structures by taking global landmark configuration into account.

During RF training, for each landmark a separate forest consisting of \( T_k \) trees is constructed by using selected voxels \( v_j \) from training images \( D \). Starting at the root node with the set \( S_n \), the voxels of set \( S_n \), reaching node \( n \), are recursively split into two subsets \( S_{n+1} \), \( S_{n+2} \) according to a splitting decision associated with the node. The splitting decision is made by thresholding a feature response evaluated on the voxels in \( S_n \). We use two different types of features:

- Haar-like appearance feature responses are calculated as differences between mean image intensity of two rectangles with maximal size \( s \) and maximal offset \( o \) relative to a voxel position \( v_j \in S_n \).
- Geometric feature (GF) responses are calculated as the difference in a random dimension \( x, y \) or \( z \) between the voxel \( v_j \in S_n \) and a randomly selected reference landmark from the set of landmarks \( R_k(i) \) as defined in (11).
Until maximum depth $N_0$ of the tree is reached, each node stores a feature and threshold selected from a pool of $N_T$ randomly generated Haar-like appearance or GFs and $N_T$ thresholds, maximizing a regression or classification objective function.

- **Regression objective function $J_k$** is given with:
\[
J_k = \sum_{j \in S_k} \| d_j - \overline{d}(S_k) \|_2^2 - \sum_{c \in \{0,1\text{-rgt}\}} \sum_{j \in S_k} \| d_j - \overline{d}(S_k) \|_2^2, 
\]
where $d_j$ is the $j$th distance vector between the voxel $v_j$ and the landmark $i$ for which the forest is trained, while $\overline{d}(S_k)$ is the mean voting vector of voxels in $S_k$. For later testing, we store at each leaf node $\xi$ the mean value of relative voting vectors $\overline{d}_\xi$ of all voxels reaching $\xi$.

- **Classification objective function $J_c$** is defined using information gain:
\[
J_c = H(S_n) - \sum_{c \in \{0,1\text{-rgt}\} \mid |S_c| \neq |S_n|} \frac{|S_c|}{|S_n|} H(S_c),
\]
\[
H(S_n) = -\sum_{b \in \text{pos}, \text{neg}} p(b) \log (p(b)),
\]
where $|S|$ denotes the cardinality of the respective set $S$ and $p(b)$ is calculated as the normalized empirical histogram of positive pos or negative neg class corresponding to the training voxels in $S_n$. For later testing, the probability $p(b)$ of being the landmark is stored at each leaf node $\xi$.

2.2.1. **LocalRFF - accurate candidate generation**

As the first stage of our method, different approaches for feature extraction could be used, delivering accurate but locally ambiguous predictions, e.g. using convolutional neural networks as in Aubert et al. (2016). However, we implemented the LocalRFF stage using Hough regression RFs (Gall et al., 2011) to benefit from having a unified framework for all stages of our method.

To capture local appearance in a model $\Theta^L_i$ for each landmark $i$ as required in (8), we select training voxels $S_0$ locally around the landmark with radius $r_{\text{rgt}}$ and use only Haar-like appearance features with small maximum offset $o_{\text{rgt}}$ and size $s_{\text{rgt}}$ (see Fig. 1, top). LocalRFF is trained as a regression RF using objective function (12).

During testing, all voxels of a previously unseen image $I$ are pushed through all the $T_n$ trees of LocalRFF until they reach the leaf nodes. Based on the voxel’s position and the relative voting vector $\overline{d}_\xi$, stored in the leaf node, each node contributes to an accumulator image $a_i$ defining the most likely location of landmark $i \in \{1, \ldots, L\}$ in the coordinate system of the image $I$. We extract the $N_M$ strongest maxima $F_i$ from the accumulator $a_i$ of each landmark $i$, defining possible candidate locations, which are accurately located on locally similar structures, see Fig. 1 (top). Finally, the strongest candidate $X^i$ for each landmark $i$ defines the landmark configuration $X^i = [X^1_i, \ldots, X^L_i]$ that is used to initialize coordinate descent optimization.

2.2.2. **Global context random forest (GCRF)**

A crucial part of our proposed method is the integration of appearance information and geometrical configuration of landmarks in our global context RF model $\Theta^G_i$ as required in (8). Global appearance information is captured using Haar-like features with a large maximum offset $o_{\text{rgt}}$ and a large maximum size $s_{\text{rgt}}$. The geometrical configuration is captured using GFs, where during training for each landmark $i$, information about the ground-truth location of all other landmarks $X^i_j$ is used. Since during testing the ground truth position of other landmarks is not available, they are replaced by the best beliefs of where other landmarks are located $\{X^i_j\}$, as explained in Section 2.1.1. To overcome the overfitting to ground-truth landmark positions $X^i_j$, we train the GCRF as an ensemble with a bagging strategy by generating for each tree a random pool of $K$ reference landmarks $R_k(i)$, which are used as input for computing GFs (see Fig. 2).

Optimizing classification objective (13), GCRF training starts by selecting $N_{\text{pos}}$ positive and $N_{\text{neg}}$ negative voxels from each training image. Inspired by the voxel selection in Stern et al. (2016), training voxels for landmark $i$ are randomly selected within a maximal radius $r_{\text{gcrf}}$ around candidate locations $R_k(i)$ as obtained for the training images with localRFF. The voxels around ground-truth landmark positions are defined as positive, remaining training voxels are used as negatives. Thus, negative voxels are concentrated around locally similar anatomical structures, forcing GCRF to learn how to distinguish potentially ambiguous landmark candidates.

During testing, to calculate the conditional probability $p(x \mid X^i_j, \Theta^G_i, I)$ for landmark $i$ in (8), we push candidate locations $R_k(i)$ through all trees and average the probabilities $p(b)$ of being landmark $i$ as stored in the reached leaf nodes for all trees. Embedded in the coordinate descent scheme, this leads to an iterative improvement of all landmark locations $X^i$ according to information on both appearance and geometrical configuration. Thus, an implicit connection between the independent landmarks modeled in separate RFs is established.

3. **Experimental setup**

To show that our proposed landmark localization algorithm is applicable to a large variety of medical data from different 2D and 3D imaging modalities, we evaluate performance on five challenging datasets (three publicly available, two in-house) and compare our results extensively to state-of-the-art landmark localization methods. As evaluation measure, we use Euclidean distance between estimated and ground truth landmark positions. We perform cross-validation experiments and report accuracy by median deviations on testing images, while robustness is assessed using the number of outliers defined as localization errors larger than a dataset specific threshold. Additionally, mean and standard deviation of localization errors are computed.

3.1. **Data sets**

**2D hand dataset (2DHand)** consists of 895 2D X-ray hand images of subjects older than 10 years from the publicly available Digital Hand Atlas Database 1. Due to their lacking physical pixel resolution, we normalized image resolution according to wrist widths as suggested in Lindner et al., 2015; Stern et al., 2016. We used a histogram matching step to normalize intensities regarding contrast and brightness. This normalization was solely performed inside an Otsu threshold-based segmentation of the outline of the hand by using histogram matching to find a nonlinear intensity mapping transforming each image according to the histogram of a reference image (Gonzalez and Woods, 1992). For evaluation, $L = 37$ landmarks were manually annotated by three experts (see Fig. 4). These annotations are available at our website2. The dataset was evaluated using a three-fold cross-validation, splitting images into 67% training and 33% testing data for each fold.

**Cephalogram dataset (2DCep)** contains 2D X-ray images of the skull from 400 subjects with an image size of 1935 × 2400 and a pixel spacing of 0.1 × 0.1 mm² from the Automatic Cephalometric X-Ray Landmark Detection Challenge (Wang et al., 2016). Annotation consists of $L = 19$ anatomical landmarks (see Fig. 4). The dataset was evaluated by training on 150 images and testing on

two test-sets, Test1 and Test2 (total of 250 images) as proposed in Wang et al. (2016). We compare to the challenge results in terms of different success detection thresholds between 2 and 4 mm.

Whole body dataset (3DBody) consists of 20 whole body CTs from the Whole Body Morphometry Project with an average size of 512 × 512 × 1900 voxels and a voxel size of 1.3 × 1.3 × 1 mm$^3$. Data and annotation, consisting of $L = 57$ anatomical landmarks, were provided by Donner et al. (2013) (compare Fig. 4 for annotation). The dataset was evaluated using a two-fold cross-validation with 10 training and 10 testing images.

3D hand dataset (3DHand) consists of left hand T1-weighted 3D gradient echo MR images of scans from 60 male Caucasian subjects between 13 and 23 years from an in-house forensic age estimation study (Urschler et al., 2015). The average dimension of the volumes is 294 × 512 × 72 with a voxel size of 0.45 × 0.45 × 0.9 mm$^3$. In each volume, $L = 28$ landmarks were manually annotated by a scientist similar to the annotation of 2DHand. The dataset was evaluated using the same setup as in Ebner et al. (2014), a cross-validation with five rounds by splitting randomly into 43 training and 17 testing images.

3D teeth dataset (3DTeeth) consists of 280 3D proton density weighted MR images of left or right side of the head from an in-house forensic age estimation study (Stern et al., 2017). Left side images were mirrored to create a consistent dataset of images with 208 × 256 × 30 voxels and a physical resolution of 0.59 × 0.59 × 1 mm$^3$ per voxel. Specifying their center locations, two wisdom teeth per dataset were annotated by a dentist. The dataset was evaluated using a three-fold cross-validation, splitting the data into 67% training and 33% testing data for each fold.

3.2. Implementation details

All trees of the RFs in our proposed method (localRFF and GCRF) were trained independently for each landmark up to a maximum depth of $N_D = 25$. When experimenting with different numbers of randomly generated features $N_F$ and thresholds $N_T$, we found that $N_F = 20$ and $N_T = 10$ were sufficient to keep localization performance while minimizing training time. Haar-like appearance features were calculated using integral image data structures (Viola and Jones, 2004), with a memory efficient integral volume implementation (Urschler et al., 2013) being used for 3D datasets. For localRFF the forests consisted of $T_F = 8$ trees, while for GCRF we used $T_F = 128$ trees. LocalRFF was trained by selecting training voxels inside a radius of $r_{trf} = 10$ mm around each landmark. Haar-like features were restricted to a maximum offset $a_{trf} = 3$ mm and a maximum size $s_{trf} = 1$ mm for 2DHand and 3DHand datasets, for 2Dcep we used $a_{trf} = 5$ mm; $s_{trf} = 3$ mm and for 3DBody $a_{trf} = 10$ mm; $s_{trf} = 4$ mm. To generate candidate locations $P_l$, the $N_D = 100$ strongest maxima with a minimum distance of 10 mm to each other were extracted using non-maximum suppression. GCRF was trained by selecting for all datasets $N_{pos} = N_{neg} = 10,000$ voxels, except for 3DBody where 100,000 voxels per
The radius \( r_{\text{gof}} \) for voxel selection in GCRF was set to 10 mm, while Haar-like appearance feature offset \( r_{\text{gof}} \) and size \( s_{\text{gof}} \) were restricted to 50 mm, respectively. We implemented the coordinate descent optimization by simultaneously updating all landmarks during each iteration to benefit from parallelized execution. As it will be explained in Section 4.4, we found that \( K = 4 \) is a parameter value that can be applied generically to all investigated datasets, except for 3DTeeth, where only two landmarks were available. Therefore, in accordance with the resulting bias-variance tradeoff behavior as shown in Fig. 8, we chose to train 10% of GCRF trees with both geometric and appearance features on this dataset, while the other trees solely used appearance information. For localRRF+MRF, solutions to the MRF models were computed using loopy belief propagation (Wainwright and Jordan, 2008). All experiments were performed on an Intel CoreDuo i7 920 workstation with 24 GB RAM.

4. Experimental results and discussion

In the following sections we provide detailed discussion of the quantitative and qualitative results from the experimental evaluation, with Section 4.1 describing our MainExperiment and Section 4.2 describing the experiment when no strong geometrical model is available (ReducedExperiment). Further, we investigate in more detail the benefit of combining geometrical and appearance information in Section 4.3 (IntegrationExperiment), evaluate model complexity in Section 4.4 (ModelComplexityExperiment), and comment on the runtime of our proposed method in Section 4.5. We finish the discussion with an outlook in Section 4.6.

4.1. MainExperiment

The main evaluation of our novel approach was performed on a variety of challenging datasets, including three publicly available and one in-house dataset, that show different anatomical structures and come from three distinct imaging modalities (2D X-ray, 3D CT and 3D MRI). Furthermore, we extensively compared our approach to a large number of state-of-the-art methods resembling different strategies (direct regression, model based regularization and forest based regularization) for localization of multiple anatomical landmarks.

As a baseline method resembling a direct regression strategy, we used the accurate but ambiguous localRRF, where for each landmark independent RFs were trained. The implementations of methods Criminisi,2013 and Ebner,2014 were taken from the work of Ebner et al. (2014) and applied to 2DHand, while results for 3DHand were taken from their paper. Both methods were trained as multi-output regression RFs, where each tree votes for all landmark locations simultaneously. Payer et al. (2016) provided implementations of two convolutional neural network based landmark localization algorithms, involving their SpatialConfigurationNet (Payer,2016) and a U-Net (Ronneberger et al., 2015) that was adapted for the localization task. Both algorithms were applied to all four datasets. Lindner et al. (2015) applied their model based regularization method to our 2DHand dataset using our preprocessed images and for 2DCEP we looked up their results from Wang et al. (2016). Similar to Schnorr et al. (2016), we implemented an MRF based approach that adds robustness to the localRRF using an explicit graphical model (localRRF+MRF), and applied it to all available datasets. We experimented with different hand-crafted loosely connected MRF configurations and chose the graphical models depicted in Fig. 4. The results of Donner et al. (2013) on 3DBody and Ibragimov et al. (2012) on 2DCEP were taken from the respective publications. Finally, we compared our method with forest based regularization approaches. For localRRF+Criminisi,2013 we used the multi-output forest implementation of Criminisi,2013 from Ebner et al. (2014) and multiplied it with the localRRF response to achieve a probabilistic prediction that takes into account the geometrical configuration among landmarks. For localRRF+Stern,2016 we used the classification based backprojection RF trained independently for each landmark as proposed in Stern et al. (2016), which regularizes locally accurate but ambiguous predictions of localRRF solely based on appearance information.

Regarding accuracy and robustness, the overall quantitative results of our MainExperiment are shown in Table 1. Fig. 5 compares the better performing methods graphically by cumulative localization error distributions. For our proposed method, Fig. 6(a) illustrates qualitatively the landmark localization results of all five datasets of MainExperiment, drawn relative to ground truth positions of the depicted image based on the error vectors from cross-validation. To enable this visualization, for 3D datasets we project volumetric data to sagittal and coronal views, respectively. Additionally, Fig. 6(b) shows a few characteristic cases where outliers occur. In the following subsections, we compared the performance of the methods resembling direct regression, model based regularization and forest based regularization strategy in more details.

4.1.1. Direct regression methods

Local appearance based regression methods (localRRF) are capable of accurately localizing landmarks as can be seen from Table 1 and Fig. 5a–d, however, in the presence of locally similar structures their robustness is low leading to approximately 10% outliers for all investigated datasets. Extending the scale of appearance features to the whole image as proposed in Criminisi et al. (2013) may improve robustness by giving localization results near to the mean landmark locations that were seen in the training data, but at the same time reduces accuracy of predictions (see results of 2DHand experiment in Table 1). Using the same large scale appearance features, the work of Ebner et al. (2014) showed how to improve on accuracy by introducing a distance weighting and a refinement stage, but still the amount of outliers is large, especially for the 2DHand dataset (see Table 1 and Fig. 5b). A novel direction in machine learning based medical image analysis is to use deep neural networks to automatically learn the scale and type of appearance features instead of pre-defined Haar-like features. If a large enough training dataset is available and image resolution is small enough to be stored in memory, the regression method of Payer et al. (2016) and the adaptation of the U-Net (Ronneberger et al., 2015) architecture for landmark localization show state-of-the-art performance in robustness, while due to working with downsampled images accuracy is worse than the state-of-the-art methods on 2DHand and 2DCEP (see Fig. 5a,c). Still for the 3DBody dataset, which has only 10 available training images that additionally are too large for downsampling without merging individual landmarks, both CNN methods failed to converge during training.

4.1.2. Model based regularization methods

Starting from accurate local appearance based regression methods like localRRF, robustness can be improved by introducing an explicit model of geometrical configuration for regularization. Using a constrained local model for regularization as in Lindner et al. (2015) shows best results on the 2DCEP dataset as well as close to the best results for 2DHand, when using their implementation with parameters tuned for the respective datasets (see Table 1 and Fig. 5a and c). Drawbacks of their method are its dependency on a good initialization of their multi-scale landmark fitting algorithm and that extension to 3D data is not trivial. After discretizing predictions from a localRRF, finding the best geometrical configuration from a set of landmark candidate locations can efficiently be performed by pre-defining a graphical model and
solving for the optimal configuration. As illustrated in Fig. 5a–d, in our experiments such an approach (localRFF+MRF) shows one of the best performances on all evaluated datasets, being robust in terms of outliers while keeping high accuracy. To obtain these results we had to test different hand-crafted graphical models for each dataset. For 2DHand, 2DCep and 3DHand we found loosely connected, cyclic graphs, for which belief propagation could be used to efficiently compute solutions. However, for 3DBody we had to further reduce the number of connections for belief propagation, which was achieved by modeling the connections in a tree-structure reflecting anatomy at different scales (see Fig. 4). Donner et al. (2013) proposed to find an optimal graph structure for the MRF automatically, which could be solved using standard belief propagation as well. This approach is closely related to the localRFF+MRF model, but our hand-crafted MRF model performs better than the automatically generated one on the 3DBody dataset (see Table 1). Using a different strategy, Ibragimov et al. (2012) investigated a fully connected graph for representing landmark connections. However, due to computational intractability when using belief propagation in this case, to compute a solution they used a game-theoretic optimization approach to find a Nash equilibrium. Comparison on the 2Dcep dataset indicates slightly worse performance of this approach in both accuracy and robustness compared to the state-of-the-art methods (see Fig. 5c).

4.1.3. Forest based regularization methods

Random forest based approaches can also be used to disambiguate locally accurate but ambiguous predictions of local appearance based regression methods. In our experiments we have investigated two such models for regularization that are solely trained on Haar-like appearance features. Using localRFF+Criminisi,2013 the arrangement between landmarks encoded in the leaf nodes of the multi-output regression random forest (Criminisi et al., 2013) is taken into account for regularization. As shown in Table 1, an improvement in accuracy and robustness over Criminisi,2013 can be observed for 2DHand and 3DHand, however, all evaluated methods with a strong geometrical model outperform it. Although trained independently for each landmark as a classification RF, localRFF+Stern,2016 does not have any embedded geometrical model, but due to its backprojection step that focuses on hard to distinguish locally similar structures it shows state-of-the-art results on 2Dcep (see Fig. 5c), however on 2DHand and 3DBody it lacks in robustness (see Table 1). These comparisons show that training an RF to discriminate locally similar structures based on
Fig. 6. Qualitative results on all five datasets, presented on 2D projections for 3D datasets. Images in (a) show the results of all cross-validation folds with error vectors drawn relative to the ground truth landmark positions of the respective image, while the images (b) depict some specific outlier cases of predicted landmarks compared to their ground-truth locations.

appearance features alone is not sufficient and a stronger model of the anatomical configuration of landmarks is needed.

In our proposed approach, we used our novel geometric features to integrate a strong regularization model into the RF framework. This allows geometrical information of relative distances between landmarks to interact with appearance information during coordinate descent optimization, thus iteratively improving initial landmark locations from an accurate but ambiguous local prediction stage. On the 2DHand dataset, performance of our integrated approach is in-line with the accuracy of the best performing methods using explicit geometric models (localRRF+MRF, Lindner et al. (2015)), and the robustness of upcoming approaches based on CNNs (see Table 1 and Fig. 5a). Regarding results of the 2D cep dataset, the approaches Lindner et al. (2015), localRRF+MRF,
Localization results of ReducedExperiment, given as number of outliers when using a subset of 2DHand landmarks and 3DTeeth.

<table>
<thead>
<tr>
<th>Method</th>
<th>2DHand</th>
<th>3DTeeth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt;10mm</td>
<td>&gt;7mm</td>
</tr>
<tr>
<td>localRRF+MRF</td>
<td>495</td>
<td>(9.22%)</td>
</tr>
<tr>
<td>localRRF+localRRF+MRF+</td>
<td>11</td>
<td>(0.20%)</td>
</tr>
<tr>
<td>proposed</td>
<td>4</td>
<td>(0.07%)</td>
</tr>
</tbody>
</table>

Localization results with different number of candidates on the 2DHand dataset. We use the same error measures as in Table 1.

<table>
<thead>
<tr>
<th>Number of candidates</th>
<th>2DHand</th>
<th>3DTeeth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt;10mm</td>
<td>&gt;7mm</td>
</tr>
<tr>
<td>5 candidates</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>25 candidates</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>50 candidates</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>100 candidates</td>
<td>0.52</td>
<td>0.52</td>
</tr>
</tbody>
</table>

4.2. ReducedExperiment

While the strategies of combining local appearance based regression with a graphical model show state-of-the-art results in localization (see Section 4.1.2), they are limited when there is no strong geometric model available. Such limitations range from the case of weak correlation between landmarks, where the contribution of a geometric model is negligible, up to the extreme case of only a single annotated landmark being available, where the use of a graphical model is not meaningful. Furthermore, the need for costly and tedious expert annotation is a regular problem in medical image analysis, with localization of age-relevant anatomical structures for automatic age estimation (Stern et al., 2014) being one specific example. In multi-factorial age estimation (Schmeling et al., 2011), several anatomical structures like hand bones and teeth have to be localized for studying closing of epiphyseal gaps and mineralization of tooth roots and crowns. This could be achieved by the costly expert annotation of all teeth as well as hand bones, as in our 2DHand datasets, or by solely annotating wisdom teeth (Demirjian et al., 1973) as well as radius and ulna bones (Tanner et al., 1983) in an age estimation application. In our ReducedExperiment, we therefore reduced the number of investigated landmarks to six radius/ulna landmarks in our 2DHand dataset, and we utilized the 3DTeeth dataset where only two landmarks representing the wisdom teeth are available. This experiment was performed for localRRF+MRF and localRRF+localRRF+MRF in addition to our proposed method. As can be seen from Table 2, the localization performance of the localRRF+MRF method explicitly representing geometrical configuration is poor in terms of outliers compared to our proposed approach. Thus, if solely a low number of weakly correlated annotated landmarks is available, our method outperforms state-of-the-art approaches based on a strong graphi-
4.3. IntegrationExperiment

So, solely using appearance information to select the best among a set of local appearance based landmark candidates, localRRF+MRF from [2016] outperforms the model-based localRRF+MRF method when only a low number of weakly correlated annotated landmarks is available (see Table 2). On the other hand, when there is strong correlation between landmarks, we show in MainExperiment that methods utilizing geometrical configuration are superior (see Table 1 and Fig. 5). In our proposed approach, we therefore exploit the strengths of integrating appearance information with geometrical landmark configuration for regularization in the localization task. To investigate this integration in more detail, we compare the cumulative error distribution of our proposed approach with the two variants, where solely Haar-like appearance features, very similar to localRRF+MRF from [2016], or solely GFs were used in the nodes of the GCRF regularizer. This experiment was performed on a single fold of the 2DHand cross-validation that had the largest number of outliers in MainExperiment. Fig. 7 shows that both types of features alone demonstrate reasonable localization results (43 and 96 outliers, respectively), but are outperformed significantly regarding robustness (13 outliers) by our proposed combination that integrates both kinds of features. Therefore, both kinds of information were needed to achieve the state-of-the-art results in our MainExperiment, while it outperforms both, methods that are using solely appearance information or solely a graphical model, when there is weak correlation among landmarks as in our ReducedExperiment. Thus, of the compared approaches our proposed method is the most flexible regarding representation of different landmark configurations.

4.4. ModelComplexityExperiment

In our proposed approach landmark locations are iteratively improved from previous predictions, which has similarities to the auto context strategy [2010] that is based on a cascade of stages. The drawback of auto context is the need for a large amount of data separated into individual sets for training a model in each cascade stage. Contrary to that, our single prediction model is trained on the full amount of data, while predictions are iteratively improving at testing time. On the 2DHand dataset, we exemplarily show in Fig. 3(a) how during the coordinate descent iterations the number of outliers is decreased, reaching an optimum after four iterations. This behavior is visualized qualitatively in Fig. 3(b), where, starting from the localRRF initialization, two depicted landmarks are iteratively refined.

Since we use the annotated ground-truth locations of landmarks (compare Fig. 4) for our GFs, training a model with high model complexity by using all other $L - 1$ landmarks leads to overweighting to the ground-truth landmark configuration. We show this with our ModelComplexityExperiment performed on a reduced training and test set consisting of 50% images of one fold of the 2DHand cross-validation, due to the large number of training steps required for the evaluation. Overfitting can be seen as the difference between the red and blue lines for $K = L - 1 = 36$ in Fig. 8, where the red line shows the error in terms of number of outliers larger than 10 mm on the test set when the coordinate descent optimization is initialized with the localRRF predictions $x^0$. When the localRRF prediction is artificially replaced with the known ground-truth locations $x^*$ for initializing coordinate descent, the blue error line resembles the best-case scenario for testing, since it simulates the case that the local regression stage produces a perfect initialization. At the same time, the blue line represents the worst-case scenario of the training error, allowing to observe the tradeoff between bias and variance when varying the model complexity parameter $K$. To overcome overfitting to the ground-truth landmark configuration, we use a bagging strategy and ensemble learning by reducing the number of randomly selected landmarks $K$ when training the GFs of each tree of the RF. Effectively, by lowering $K$ in our bagging strategy, we reduce the probability of a tree to select landmarks being located on anatomically infeasible positions, thus increasing the probability of the tree voting for the correct position and leading to more robust predictions of the RF ensemble.

While a model with large $K$ leads to a low bias–high variance model that does not generalize well, reducing $K$ to too small numbers generates a high bias model with low prediction performance in terms of robustness. When varying the model complexity parameter $K$, we observed for all of our evaluated datasets a similar
behavior regarding the tradeoff between bias and variance, with the sweet spot located in between 10 – 20% of total landmarks, as exemplarily shown in Fig. 8 for 2DHand. By choosing $K = 4$ we obtained a generically applicable parameter value located inside this range for all investigated datasets, except for the special case of the 3DTeeth dataset, where only one other landmark can be used for computing GFs. Nevertheless, by choosing 10% of trees trained on both geometric and appearance features, while the other trees use appearance information only, we reproduce the same behavior regarding the tradeoff between bias and variance as for the other evaluated datasets.

4.5. Runtime

Exemplarily shown for the 2DHand dataset, on our workstation localRRF requires an execution time of 56 s per image during testing due to the need for pushing all pixels/voxels of an image through the RF. Runtime of this method scales with the number of input pixels/voxels. All methods that use localRRF as a first step require additional time for regularization. For Štern,2016 an additional 56 s are needed, MRF requires three extra seconds, and Criminisi,2013 as well as our proposed method need an additional four seconds. The total runtime of the approach from Ebner et al. (2014) is 60 s. Without being able to directly compare runtimes due to the implementations on their workstations, the runtimes of the CNN-based methods (Payer et al., 2016) were ten seconds and the fastest reported timings are those from Lindner et al. (2015) with five seconds. To produce results of our proposed method, we did not use any GPU acceleration for forest evaluation as proposed e.g. in Sharp (2008), nor did we investigate a hierarchical localization approach as in Lindner et al. (2015). While runtimes of the regularization steps for our proposed method and comparable state-of-the-art methods are essentially the same, we did not focus on execution speed in our C++ implementation of the localRRF, thus improvements in runtime are possible.

4.6. Outlook

From our quantitative results we still see a few outliers that are caused by occlusions in the datasets during testing, as illustrated in Fig. 6(b), where A shows a missing tooth in the lower jaw, and B an occlusion due to a label on the X-ray image. We intend to further study this type of limitation of our method in future work, by investigating datasets with stronger occlusions or varying fields of view, where not all the landmarks are always visible. This problem is also of interest in computer vision tasks like human pose estimation (Girshick et al., 2011; Chen and Yuille, 2014), which could be a potential future application of our method. Another source of outliers in our results are landmarks where annotation is ambiguous due to hard to define anatomical structures, an example is shown in Fig. 6(b) with the cyan landmark in C and D. However, our method shares this problem with all other state-of-the-art approaches. Finally, we have not yet explored how models containing many densely sampled landmarks, like they are often used in point distribution models for segmentation of anatomical structures, will behave in our framework, an issue that may be considered as another interesting direction for future work.

5. Conclusions

State-of-the-art methods for multiple anatomical landmark localization use locally accurate but ambiguous predictions regularized by explicit models of geometric configuration. While in these methods appearance and geometric information do not interact during the landmark prediction stage, our proposed novel approach follows this regularization strategy differently by combining geometrical landmark configuration with appearance information in a unified random forest based framework. Thus, our method is capable of learning from both kinds of information simultaneously, internally deciding on the combination of geometric configuration and appearance information that is most informative for a specific localization task. Without requiring an explicit hand-crafted geometrical model and with only a few iterations of coordinate descent, our model is capable of regularizing initial predictions of landmark positions according to the best beliefs of where other landmarks are located.

From the extensive comparison of our proposed approach to various state-of-the-art methods on a total of five challenging datasets, we observe that there is currently no single method that dominates localization performance, despite compared methods being tuned for best parameters in author-provided or re-implemented versions. However, our approach is able to achieve results that are best or similar to the individual best performing methods for all datasets simultaneously, without dataset specific parameter tuning or the need for task-specific hand-crafted models, thus showing its generic applicability and flexibility. Moreover, when there is no strong correlation among landmarks available and therefore the appearance information might be most informative for the localization task, our novel combination of otherwise separate knowledge about appearance information and geometrical configuration outperformed the compared state-of-the-art methods.

Due to its simplicity and its lack of requiring an explicitly defined model of geometric configuration, we believe that our novel flexible approach will generalize well to other application scenarios.

Acknowledgements

We thank the anonymous reviewers for their thorough work and excellent comments that helped us to improve our manuscript. We acknowledge Franz Zehentner for helping with our quantitative evaluation and Alexander Shekhovtsov for fruitful discussions. We thank Claudia Lindner for applying their algorithm on our 2DHand dataset, Christian Payer for results from the CNN based methods, and Rene Donner for providing 3DBody landmark annotations. This work was partly supported by the KIRAS program (no 850183) under supervision of the Austrian Research Promotion Agency (FGF) as well as by the Austrian Science Fund (FWF): P28078-N33.

Appendix A. Number of candidates from localRRF predictions

We performed an additional experiment on one cross-validation fold of the 2DHand dataset, where we compare the results of varying the number of candidates per landmark (5,10,25,50,100) with the fully probabilistic treatment when all voxels in the image are used as candidates (dense). Results of this experiment are shown in Table 3. We conclude that even with 50 candidates, accurate and robust landmark localization can be achieved, thus our choice of 100 selected candidates is a conservative one. Further increasing the number of candidates to achieve a fully probabilistic treatment does not bring any benefit in localization performance, however it would lead to an unnecessary increase in runtime.

References
