Uncertainty-driven Forest Predictors for Vertebra Localization and Segmentation

David Richmond[1⋆], Dagmar Kainmueller[1⋆], Ben Glocker[2],
Carsten Rother[3⋆⋆], and Gene Myers[1⋆⋆]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Germany
[2] Biomedical Image Analysis Group, Imperial College London, UK
[3] Computer Vision Lab Dresden, Technical University Dresden, Germany
richmond@mpi-cbg.de kainmueller@mpi-cbg.de

Abstract. Accurate localization, identification and segmentation of vertebrae is an important task in medical as well as biological image analysis. The prevailing approach to solve such a task is to first generate pixel-independent features for each vertebra, e.g. via a random forest predictor, which are then fed into an MRF-based objective to infer the optimal MAP solution of a constellation model. We abandon this static, two-stage approach and mix feature generation with model-based inference in a new, more flexible, way. We evaluate our method on two data sets with different objectives. The first is semantic segmentation of a 21-part developing spine of zebrafish in microscopy images, and the second is localization and identification of vertebrae in benchmark human CT.

1 Introduction

State-of-the-art approaches for object localization or semantic segmentation typically employ pixel-wise forest predictors combined with MAP inference on a graphical constellation model [4,5,13] or a (super-)pixel graph [14,12], respectively. A recent trend in computer vision replaces single-level forest predictors by deep, *cascaded* models for feature generation, such as CNNs [8] and Auto-Context Models [15]. These models play the role of learning a complex non-linear mapping from images to features that are relevant for the task at hand.

This modeling framework is however static, as it separates feature generation from inference (i.e., "model fitting"). It has been shown that better features can be generated by *interleaving* feature generation with MAP inference [9,11,7].[4] In this work we take this idea a step further: Instead of interleaving feature generation with a pixel-level structured model or model-agnostic smoothing, we

⋆ Shared first authors.

⋆⋆ Shared last authors.

[4] Note that this is conceptually different from the classical "hierarchical" approach that, purely for the sake of pruning the search space to reduce run-time, performs feature generation and inference/model fitting multiple times on different scales.

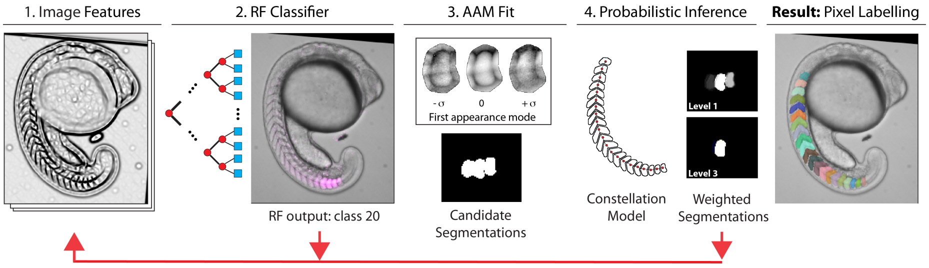


Fig. 1. Our proposed pipeline for multi-class, semantic segmentation. A stack of feature images is created by a standard filter bank, and used to train a random forest classifier. The random forest output is then used in combination with the original image to generate candidate segmentations for each class, by fitting multiple instances of appearance models. These candidate segmentations are weighted by means of probabilistic inference in a constellation model that captures relative locations of classes. The weighted and fused candidate segmentations are then fed back as additional "smoothed" features into a new random forest classifier, forming a cascade.

interleave with a global, generative constellation model. We suggest a cascaded pipeline, as illustrated in Figure 1.

We explore two applications, namely (1) light microscopic images of zebrafish embryos, where we segment developing vertebrae, called *somites*, and (2) benchmark human spine CTs [4], where we localize and identify vertebrae. Figure 1 (right) and Figure 2 (left) show exemplary zebrafish images with and without overlaid segmentation of somites, respectively.

The most important aspect of our cascaded pipeline is the question of what to infer from a constellation model at intermediate stages of the cascade. Options are the MAP solution or the marginal distributions. Interestingly, marginal distributions are a winner for the zebrafish application. The reason that *uncertainty is beneficial* here is that individual somites are highly ambiguous with respect to shape, appearance, and importantly also the appearance of surrounding tissue. Hence only the relative spatial arrangement can disambiguate them. We show that as opposed to MAP inference, the soft marginals do not commit to a certain – potentially wrong – solution "at first sight".

Closely related to our work are (1) Auto Context [15], but they do not perform any smoothing between levels of the cascade. (2) Geodesic Forests [7], but they do not use a structured model for smoothing. (3) Cascaded classifiers interleaved with MAP inference [9,11], but they do not use a (global generative) constellation model and do not explore marginals for inference. (4) Constellation models for the widely studied application of human vertebrae localization (see e.g. [16]) but none of the respective methods runs a Random Forest cascade.

To summarize, our work makes the following main **contributions**: (1) We show, for the first time, that probabilistic inference can give a boost in performance in cascaded MRF-Forest-based models. This is compared to standard

MAP inference (as in e.g. [4,13]) and model-agnostic geodesic smoothing [7]. (2) We outperform a state-of-the-art method [5] on benchmark human spine CTs of challenging pathological cases. (3) We are the first to tackle somite detection in zebrafish, where we achieve an overall average Dice score of 0.82.

2 Method

Background: Random Forests classifiers (RF) [2] are widely used in medical image analysis for organ localization, particularly for vertebrae [10,4,5]. We use RFs in a cascaded fashion [15], where the probability maps yielded by an RF are treated as features that are fed into subsequent RFs, forming a cascade. In variants of this approach, inference or smoothing operations on these probability maps are interleaved with the RF prediction, and the "smoothed" probability maps are then used as features [9,11,7]. For comparison to our proposed model-based smoothing we explore model-agnostic geodesic smoothing [7]. The rationale behind geodesic smoothing is that pixels within a small geodesic distance of each other likely belong to the same class. See [7] for details.

Generating Candidate Segmentations: We train an n-class RF, where n is the number of object parts (e.g. the number of vertebrae in the human spine), plus one background class. At test time, an RF generates one probability map per class. Given an RF-generated probability map for some foreground class, we first compute its mode via the mean shift algorithm. Second, we fit a learned, static constellation of landmarks to these centroids, yielding an optimal affine transformation w.r.t. the sum of squared landmark distances. Third, we sample a number of candidate locations around these points to get sets of initializations for the respective classes. Fourth, we fit a class specific appearance model to the image, multiple times, starting at the initial locations computed in the previous step. Depending on the application, we either use active appearance models (AAM) [3], or static average and variance images [5]. Each appearance model fit results in a binary segmentation, together with a cost for the fit. We denote the cost for the l-th fit for class v as $a(v, l)$. The cost is the sum of squared differences between the target and the template image generated by the (active) appearance model. In case of static appearance models, we weigh the squared differences by the respective pixel-wise variances stored in the variance image.

Weighting and Fusing Candidate Segmentations: The above method generates a number of candidate segmentations per class. We assign weights to these by means of a constellation model in the form of a pairwise CRF. The nodes $v \in V$ of the respective graph $G = (V, E)$ correspond to the classes. The labels $l \in L$ that each node can take correspond to the respective candidate segmentations. The edges $E \subseteq V \times V$ encode the pairs of classes for which we model relative locations. We employ either a chain model that only connects spatially neighboring classes, or a fully connected model, depending on the application.

Let Ω denote the image domain. We define unary terms $\phi(v, l)$ of the CRF as a linear combination of the cost of the respective appearance model fit $a(v, l)$ and the negative logarithm of the RF probability map $RF_v : \Omega \to [0, 1]$ accumulated over the foreground of the respective binary segmentation $S_{v,l} : \Omega \to \{0,1\}$,

$$\phi(v,l) := a(v,l) + \frac{\lambda}{|S_{v,l}^{-1}(1)|} \sum_{i \in \Omega} -log(RF_v(i)) \cdot S_{v,l}(i). \tag{1}$$

A parameter λ weighs the relative influence of the two terms. We set this parameter heuristically. We define pairwise terms to reflect the probability of relative locations of neighboring proposals. We learn the average distances $d(v,w)$ between the centroids of any two vertebrae v, w, as well as respective standard deviations $\sigma(v,w)$, and assume an according Gaussian distribution. Let $c(v,l)$ denote the centroid of the l-th candidate segmentation of class v. Our pairwise terms read

$$\psi(v,w,k,l) := \frac{\left(|c(v,k) - c(w,l)| - d(v,w)\right)^2}{\sigma(v,w)^2}. \tag{2}$$

We compute weights for each proposal and each class by means of inference in this CRF. We explore two well-known variants of inference, namely MAP inference and probabilistic inference. MAP inference finds a label l_v for each node such that the energy of the CRF,

$$E(\{l_v\}_{v \in V}) = \sum_{v \in V} \phi(v, v_l) + \sum_{(v,w) \in E} \psi(v, w, l_v, l_w), \tag{3}$$

is minimized, thus yielding binary weights $w(v,l) \in \{0,1\}$ for each proposal. Probabilistic inference computes the marginal probabilities $p_v(l)$ of the respective Gibbs distribution $p(\{l_v\}_{v \in V}) = \frac{1}{Z} \exp\left(-E(\{l_v\}_{v \in V})\right)$, yielding continuous weights $w(v,l) \in [0,1]$ for each proposal. $W_v := \sum_{l \in L} p_v(l) \cdot S_{v,l}$. We call W_v a *smoothed probability map* or *smoothed RF output* for class v.

For a chain model, both MAP and probabilistic inference can be solved optimally by means of dynamic programming. For a fully connected model, the respective optimization problem is NP hard. However, probabilistic inference by Loopy Belief Propagation, and approximate MAP inference by TRWS [6] followed by Iterated Conditional Modes (ICM) [1], yields good results in practice.

3 Experiments

Zebrafish: We applied our approach to semantic segmentation of 21 *somites* in a data set of 32 images of developing zebrafish. All images were automatically pre-aligned to a reference image by rigid registration. Experts in biology manually created ground truth segmentations of these images. This data set poses multiple challenges for automated segmentation, due to (1) the similar appearance of neighboring segments, and (2) the small amount of training data. We train three-level cascades. We compare our approach with Auto-context [15] and GeoF [7], as well as with state-of-the-art RF-predict-and-MAP. Figure 2 gives an overview of the different types of inference/smoothing that we evaluate, and an idea of how the smoothed features look. For all algorithms, we evaluate the

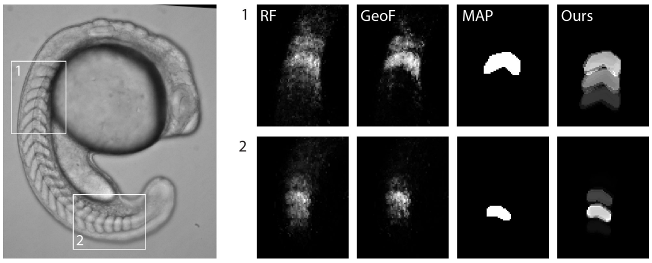


Fig. 2. Different types of inference/smoothing. Left: Zebrafish embryo. Right: Probability maps of two exemplary classes. RF probability map; geodesic smoothing (GeoF); MAP inference in our constellation model; probabilistic inference (MARG).

Dice score averaged over all 21 foreground classes, employing two-fold cross-validation to obtain scores for all 32 images. Forest parameters are as follows: 16 trees, maximum depth 12, features from a standard filter bank and local contextual features. We use a chain model as respective MRF.

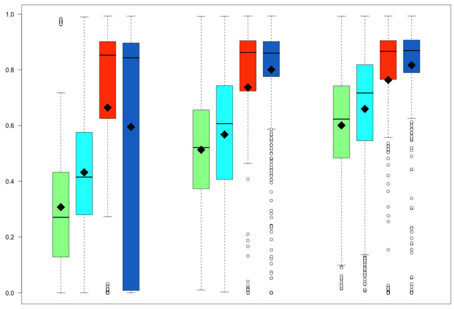
Spine CT: The data used for experimental evaluation is the publicly available database of pathological spine CT[5]. For vertebrae localization in spine CT images, we use a static appearance model constructed from a mean and variance image pair as described in [4]. We use a fully connected MRF. Forest parameters are as follows: 25 trees, maximum depth 24, features from local and contextual average intensity, as described in [5].

4 Results and Discussion

Zebrafish: Figure 3 shows box plots of the Dice scores obtained from the smoothed RF output at all three levels of the cascade, for all four methods. Figure 3a lists the average Dice scores and standard deviations of the RF output as well as the smoothed RF output for all four methods after the final level of the cascade. Auto-context returns a final average Dice score of 0.60. Compared to Auto-context, GeoF generates considerably smoother posteriors, and performs better at every level of the cascade (green vs. cyan box plots in Figure 3). The best average score obtained by GeoF is 0.66 after three levels. This increase of 6% w.r.t. Auto-context is comparable to the gains reported in [7] when applying geodesic smoothing without changing the training objective.

After the first level of the cascade MAP inference performs best among all approaches (red box plots in Figure 3), with a mean Dice Score of 0.66. This approach also improves over the cascade, reaching a final Dice score of 0.76 after three levels. However, probabilistic instead of MAP inference yields the highest overall average Dice Score of 0.82, outperforming MAP by 6%. Also, the accuracy increases considerably from level to level (blue box plots in Figure 3).

[5] http://research.microsoft.com/spine/



	RF	SRF
Auto Context	0.60 (0.20)	-
Geodesic	0.63 (0.21)	0.66 (0.22)
MAP	0.71 (0.27)	0.76 (0.27)
Marginals	**0.82 (0.16)**	**0.82** (0.18)

(a)

	MAP			Marginals		
Level	FF	RF	SRF	FF	RF	SRF
1	1	0	0	1	0	0
2	0.16	**0.60**	0.24	0.11	0.34	**0.55**
3	0.15	**0.60**	0.25	0.10	0.43	**0.47**

(b)

Fig. 3. Evaluation of four methods on 32 zebrafish datasets. Left: Segmentation accuracy as Dice scores after each level of a three-level cascade. RF (green), GeoF (cyan), MAP (red), marginals (blue). Right: (a) Average Dice Score (over 32x21 values), and standard deviation (in brackets), for segmentations obtained directly from RF-generated probability maps (RF), and from respective "smoothed" RF probability maps (SRF). (b) Variable importance of features, normalized over the three classes: Filter bank features (FF), RF output (RF) and smoothed RF output (SRF).

Observe that the accuracy of every approach increases over the levels of the cascade. Furthermore, approaches that employ any kind of smoothing between levels perform better than auto-context, confirming the power of cascading with interleaved smoothing. Model-based smoothing performs considerably better than model-agnostic geodesic smoothing, likely due to the more specific prior knowledge induced by the constellation model.

Interestingly, while MAP inference yields the best results after the first level of our cascade, probabilistic inference undergoes a much more dramatic increase in the mean Dice score and concurrent reduction in the standard deviation over the 3 levels of the cascade. We observe that this is due to failure cases that are "rescued" by our approach, but not by MAP, as shown in Figure 4.

We quantify the relative strength of features generated by probabilistic inference vs. MAP inference by means of their variable importance [2]. Figure 3b reveals that the features generated by probabilistic inference are significantly more important for forest performance than the respective MAP features.

Human Spine CT: We evaluate the results of a single-level RF and a two-level cascade on publicly available data (cf. [5]) in terms of True Positive Rates (TPR) as listed in Table 1. Note that [5] reports Precision as opposed to TPR on this data for the sake of comparability to another dataset that is guaranteed to contain all vertebrae of the spine. This measure neglects false negative detections. However, the data we evaluate shows arbitrary subsets of vertebrae. This poses an additional challenge to an automated localisation method, because it has to decide which vertebrae are present at all. Hence we decided for an error measure that accounts for false negative detections, namely TPR.

We calculated the TPR for the results of [5] obtained exactly with their method, which is 63%. Our one-level cascade without inference is a re-implemen-

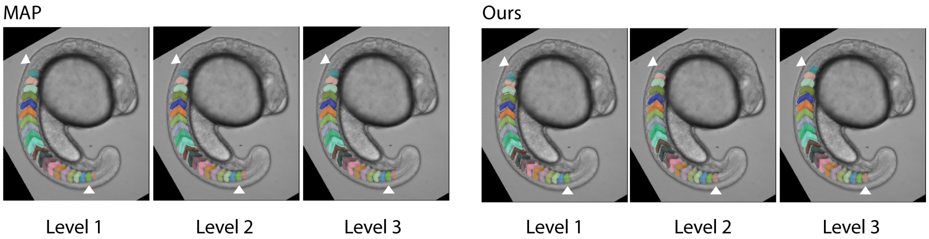


Fig. 4. Exemplary failure case that is "rescued" by our approach, but not by MAP inference. Arrows point to ground truth start and end of spine. After the first level, segmentations are off by one somite. This stays constant for MAP; however, our approach gradually recovers a correct segmentation.

	One Level	Cascade	
		Level 1	Level 2
None	8.7 / 12.7 (15.7) [0.67]	9.0 / 14.1 (17.7) [0.66]	9.0 / 12.7 (12.0) [0.68]
MAP	**7.8** / **10.1** (**8.0**) [0.73]	8.0 / 10.7 (8.6) [0.69]	**7.8** / 10.3 (8.3) [**0.74**]
Marg	-	8.0 / 10.6 (8.6) [0.70]	7.9 / 10.7 (8.5) [**0.74**]

Table 1. Evaluation of spine localization on pathological CTs. Median / mean distance, standard deviation (in brackets), True Positive Rate [in square brackets]. Distances are in mm. Rows: RF output without inference (None), MAP inference, and probabilistic inference (Marg). First column: One-level RF. Second column: Two level cascade.

tation of [5], with the slight modifications of training deeper trees and limiting image thresholds to a HU window of $[0, 1000]$. These modifications improve the TPR to 67%. Our best result, obtained by cascading and inference, has a TPR of 74%. Hence we outperform [5] by 11% in terms of TPR. For our best result we also computed the Precision, which is 79%. The Precision reported by [5] is 70%. Hence, we outperform [5] by 9% in terms of Precision.

Cascading is less powerful for the human spine CT than for the zebrafish, and MAP and marginals are en par. Potentially this is due to the extreme pathologies present in many if not most cases in this data set.

5 Conclusion

We have presented cascaded forest predictors interleaved with inference in MRF constellation models for the task of semantic segmentation and localization of vertebrae in biomedical applications. In a 21-class semantic segmentation task on biological data, probabilistic inference in the constellation model yields considerably better segmentation accuracy than the common MAP inference. Here, marginals of the constellation model allow for maintaining uncertainty in the predictions and hence help avoid sticking to a (MAP) solution too early in a cascade. These findings are of impact not only for the many types of constellation models employed in related work, but also for the recent trend of learning deep models combined with physically motivated structured models. For vertebrae localization, MAP and marginal inference are en par on a challenging pathological spine CT dataset, potentially due to the strong pathologies present in the data. However, our proposed approach of cascading interleaved with inference does improve considerably on state of the art.

Acknowledgments. The authors would like to acknowledge Daniele Soroldoni and Andrew Oates for kindly supplying the zebrafish images.

References

1. J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259–302, 1986.
2. L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
3. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
4. B. Glocker, J. Feulner, A. Criminisi, D. Haynor, and E. Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In *MICCAI*, volume 7512 of *LNCS*, pages 590–598. Springer Berlin Heidelberg, 2012.
5. B. Glocker, D. Zikic, E. Konukoglu, D. Haynor, and A. Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In *MICCAI*, volume 8150 of *LNCS*, pages 262–270. Springer, 2013.
6. V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE TPAMI*, 28(10):1568–1583, 2006.
7. P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. Geof: Geodesic forests for learning coupled predictors. In *CVPR*, pages 65–72, 2013.
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *NIPS*, pages 1097–1105. Curran Associates, Inc., 2012.
9. S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *ICCV*, pages 1668–1675, 2011.
10. M. G. Roberts, T. F. Cootes, and J. E. Adams. Automatic location of vertebrae on dxa images using random forest regression. In *MICCAI*, pages 361–368. Springer, 2012.
11. U. Schmidt, C. Rother, S. Nowozin, J. Jancsary, and S. Roth. Discriminative non-blind deblurring. In *CVPR*, pages 604–611, 2013.
12. F. Schroff, A. Criminisi, and A. Zisserman. Object class segmentation using random forests. bmvc, 2008.
13. S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu. Hierarchical parsing and semantic navigation of full body ct data. volume 7259, pages 725902–725902–8, 2009.
14. J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In A. Leonardis, H. Bischof, and A. Pinz, editors, *ECCV*, volume 3951 of *LNCS*, pages 1–15. Springer, 2006.
15. Z. Tu. Auto-context and its application to high-level vision tasks. In *CVPR 2008*, pages 1–8, June 2008.
16. J. Yao, B. Glocker, T. Klinder, and S. Li. Recent advances in computational methods and clinical applications for spine imaging, 2015.