Proceedings of Machine Learning Research vol 72, 249-259, 2018 PGM 2018

A Novel Approach to Handle Inference in Discrete Markov Networks with Large Label Sets

Alexander Oliver Mader ALEXANDER.O.MADER@FH-KIEL.DE
Institute of Computer Science, Kiel University of Applied Sciences, Kiel, Germany
Department of Computer Science, Faculty of Engineering, Kiel University, Kiel, Germany
Department of Digital Imaging, Philips Research Hamburg, Hamburg, Germany

Jens von Berg JENS.VON.BERG@PHILIPS.COM
Cristian Lorenz CRISTIAN.LORENZ@PHILIPS.COM
Department of Digital Imaging, Philips Research Hamburg, Germany

Carsten Meyer CARSTEN.MEYER@FH-KIEL.DE
Institute of Computer Science, Kiel University of Applied Sciences, Kiel, Germany
Department of Computer Science, Faculty of Engineering, Kiel University, Kiel, Germany
Department of Digital Imaging, Philips Research Hamburg, Hamburg, Germany

Abstract

MAP inference over discrete Markov networks with large label sets is often applied, e.g., in localizing multiple key points in the image domain. Often, approximate or domain specific methods are used to make the problem feasible. An alternative method is to preselect a limited (much smaller) set of suitable labels, which bears the risk to exclude the correct solution. To solve the latter problem, we propose a two-step approach: First, the reduced label sets are extended by a novel "refine" label, which — when chosen during inference — marks nodes where the label set is insufficient. The energies for this additional label are learned in conjunction with the network's potential weights. Second, for all nodes marked with the "refine" label, additional local inference steps over the full label set are performed. This greedy refinement becomes feasible by extracting small subgraphs around the marked nodes and fixing all other nodes. We thoroughly evaluate and analyze our approach by solving the problem of localizing and identifying 16 posterior ribs in 2D chest radiographs.

Keywords: Markov networks; inference; feasibility; object localization; posterior ribs.

1. Introduction

The optimization of discrete Markov random fields (MRFs) is of great importance to many fields. For instance in computer vision, it is used – inter alia – for image denoising (Shao et al., 2014), depth estimation (Zhu et al., 2010), spatial regularization (Koch et al., 2015), etc. However, the large label space in combination with large number of nodes quickly render the problem infeasible. Thus, either approximate inference algorithms (see, e.g., Wang et al. (2013)) are applied or domain specific solutions are developed.

One example for such a domain specific formulation is the artificial reduction of the label space. This is quite common in the localization of spatially correlated key points. Often, a MRF is used to incorporate geometric information between key points (see, e.g., Donner et al. (2013), Štern et al. (2016), and Mader et al. (2017)). The comparably large label space (i.e., all possible positions in an image) is normally too large to be used directly. Thus, a heuristic is used to preselect a limited

number of reasonable labels, drastically reducing the label space and enabling MRF inference. However, the MRF solution is only as good as the quality of the preselected labels.

In this paper, we propose an approach to overcome the imposed limitation of artificially reduced label spaces in MRFs. First, we introduce a novel "refine" label that can be chosen among the preselected labels during inference. To this end, we assign an energy to this "refine" label that is automatically learned from data. Second, after the first inference a second stage of potentially multiple inference steps is performed. For each "refine"-labeled node, a subgraph is extracted to optimize all marked nodes independently, each over the full label space. In addition to the missing energies for the "refine" label, we also automatically learn the weights of the MRF potentials, generally improving the accuracy of the MRF inference output (Komodakis et al., 2015).

We evaluate the proposed approach on the unsolved problem of localizing and labeling posterior ribs in chest radiographs. To this end, multiple key point localizers – which are used to preselect key point localization hypotheses – are followed by a MRF to capture spatial relations between key points. On a public dataset we illustrate how the proposed refinement step is able to overcome the implicit upper bound caused by an imperfect reduction of the label space to the preselected localization hypotheses. Preliminary results on this task have been published in Mader et al. (2018). In this work, we further analyze in detail the choice of the preselection method and the influence of the size of the reduced label set. In particular, we compare two commonly applied methods to preselect key point hypotheses, namely a deep, fully convolutional neural network and ensembles of random regression forests. Additionally, we analyze the inference time of our novel method compared to the baseline, single inference method.

The remainder of this paper is structured as follows. Section 2 describes the general problem and our novel solution approach. Section 3 introduces a particular application, namely localizing multiple, spatially correlated key points in medical images, which serves to evaluate and analyze our approach. Section 4 summarizes and discusses the results of our experiments while Section 5 draws our final conclusions.

2. Method

We start by formulating the original inference problem in Section 2.1, followed by the introduction of our novel "refine" label in the first inference step in Section 2.2 and our refinement as second inference step in Section 2.3. How missing energies and weights are learned is outlined in Section 2.4.

2.1 Original Problem

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ defined by the N nodes $\mathcal{V} = \{v_1, \ldots, v_N\}$ and the edges $\mathcal{E}$, the goal is to assign a label l_{v_i} from the discrete label set $\mathcal{L}_{v_i}$ to each node $v_i \in \mathcal{V}$, such that the MRF energy

$$E(\mathbf{x}) = \sum_{c \in C} \lambda_c \cdot \phi_c(\mathbf{x}_c) \tag{1}$$

for one particular state $\mathbf{x} \in \mathcal{X} = \mathcal{L}_{v_1} \times \ldots \times \mathcal{L}_{v_N}$, parameterized by a set of potential functions $\Phi = \{\phi_c \mid c \in C\}$ for cliques C, is minimized:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \,. \tag{2}$$

Note that we explicitly introduced the weights $\mathbf{\Lambda} = \{\lambda_c\}$ scaling the potential functions (Komodakis et al., 2015), in order to automatically learn them later on (Section 2.4).

In the general case, this exact inference is a #P-complete problem, so people mostly resort to approximate inference algorithms (e.g., the "lazy flipper" by Andres et al. (2012) as a move making strategy or dual decomposition approaches (Komodakis et al., 2011)). Alternatively, for particular classes of MRFs polynomial-time inference algorithms exist (Wang et al., 2013), e.g., belief propagation for trees (forests). The number of nodes $|\mathcal{V}|$ and the number of labels to assign a node $|\mathcal{L}_{v_i}|$ have a direct influence on the practical feasibility, i.e., they form the search space $\mathcal{X}$. While most often the number of nodes is dictated by the problem domain, the reduction of the label set seems to be a viable way for practical performance improvements in terms of runtime and memory demand.

Let us assume the label sets $\mathcal{L}_{v_i}$ have been reduced in a preselection step to feasible subsets $\mathcal{L}'_{v_i} \subset \mathcal{L}_{v_i}$ of labels such that $|\mathcal{L}'_{v_i}| \ll |\mathcal{L}_{v_i}|$. In real world applications this drastic reduction of the label space allows for acceptable inference times, given that the reduced search space $\mathcal{X}' = \mathcal{L}'_{v_1} \times \ldots \times \mathcal{L}'_{v_N}$ is much smaller than the original search space $\mathcal{X}$. Our new search problem becomes now

$$\hat{\mathbf{x}}' = \arg\min_{\mathbf{x} \in \mathcal{X}'} E(\mathbf{x}) \,, \tag{3}$$

while bearing the risk that the correct solution might not be found at all because it has been removed from the search space.

2.2 First Inference Step: MRF Formulation with "Refine" Label

To overcome this problem, we introduce an additional "refine" label to be included in our reduced label set $\mathcal{L}^{\mathrm{R}}_{v_i} := \mathcal{L}'_{v_i} \cup \{\text{"refine"}\}$. This label can be chosen during inference instead of the preselected labels, in case none of them fit the MRF model well (i.e., due to large potential energies).

Since it is often not possible to compute the potential function for the newly introduced "refine" label, we introduce the energies $\boldsymbol{\beta} = \{\beta_c\}$ to be used instead. The values β_c are — as well as the potential weights λ_c — automatically learned from training data, as explained in Section 2.4. Then, our new energy formulation is defined as

$$E^{\mathrm{R}}(\mathbf{x}) = \sum_{c \in C} \lambda_c \cdot \begin{cases} \beta_c & \text{if } x_{v_i} = \text{"refine" for any } v_i \in \mathrm{Scope}(\phi_c) \\ \phi_c(\mathbf{x}_c) & \text{else} \end{cases} \,. \tag{4}$$

Note that these newly introduced energies often have a very intuitive interpretation. For instance, assuming there is a potential function assessing the distance between key points, the corresponding energy β_c resembles a threshold value beyond which the distance between two correct key point positions is considered unlikely. With our new energy formulation from Eq. (4) the general search problem can be stated similar to Eq. (3), just using a slightly modified search space $\mathcal{X}^{\mathrm{R}} = \mathcal{L}^{\mathrm{R}}_{v_1} \times \ldots \times \mathcal{L}^{\mathrm{R}}_{v_N}$:

$$\hat{\mathbf{x}}^{\mathrm{R}} = \arg\min_{\mathbf{x} \in \mathcal{X}^{\mathrm{R}}} E^{\mathrm{R}}(\mathbf{x}) \,. \tag{5}$$

2.3 Second Inference Step: Refinement over Subgraphs

After the first inference step, the optimal configuration $\hat{\mathbf{x}}^{\mathrm{R}}$ may contain zero or more "refine"-labeled nodes $\mathcal{V}^{\mathrm{R}} = \{v_i \in \mathcal{V} \mid \hat{x}^{\mathrm{R}}_{v_i} = \text{"refine"}\}$. In order to assign those nodes a valid label as well, we optimize each node independently over the full label set $\mathcal{L}_{v_i}$. This is done by considering subgraphs around each "refine"-labeled node, which contain all nodes which are in the scope of any potential

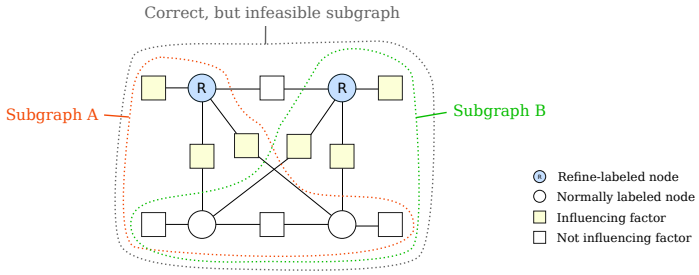


Figure 1: A fully connected MRF shown as factor graph with unary and binary potentials. Depicted are the two subgraphs (in red and green) which are consistent with the requirement to contain exactly one "refine"-labeled node.

function involving the "refine"-labeled node. To ensure that this optimization is feasible, we extract only subgraphs with exactly one "refine"-labeled node $v_i \in \mathcal{V}^{\mathrm{R}}$ (i.e., we neglect all potentials involving a second "refine"-labeled node). Thus, the extracted subgraphs are parameterized by the potential functions $\Phi^{\mathrm{R}}_{v_i} = \{\phi_c \mid v_i \in \mathrm{Scope}(\phi_c) \wedge \exists! v_j \in \mathrm{Scope}(\phi_c)\, \hat{x}^{\mathrm{R}}_{v_j} = \text{"refine"}\}$. This extraction process is illustrated for one example in a factor graph representation in Fig. 1. Then, for each subgraph, we perform inference with respect to the single "refine" node $v_i \in \mathcal{V}^{\mathrm{R}}$ only, using the full label set $\mathcal{L}_{v_i}$ (instead of the restricted label set $\mathcal{L}^{\mathrm{R}}_{v_i}$), while the labels of the other nodes in the subgraph are fixed to the labels assigned in the first inference step:

$$\hat{x}^{\mathrm{R}}_{v_i} = \underset{l \in \mathcal{L}_{v_i}}{\arg\min} \sum_{\phi_c \in \Phi^{\mathrm{R}}_{v_i}} \lambda_c \cdot \phi_c(l)\,. \tag{6}$$

Note, we used some hand-wavy notation ($\phi_c(l)$) to indicate that only one label is altered. This drastically reduces the complexity $O(|\mathcal{L}|^{|\mathcal{V}|})$ of the original inference problem from Eq. (2) (assuming the label set size is equal for each node for simplicity) to $O(|\mathcal{L}^{\mathrm{R}}|^{|\mathcal{V}|} + s \cdot |\mathcal{L}|) \ll O(|\mathcal{L}|^{|\mathcal{V}|})$, with s being the number of extracted subgraphs. Still, the "refine"-labeled nodes can assume any possible label (instead of only the preselected ones), significantly reducing a potentially detrimental effect of the label space reduction from $\mathcal{X}$ to $\mathcal{X}'$.

This refinement is optimal if there is no potential function that has more than one "refine"-labeled node in its scope, in which case the refinement step is guaranteed to produce an equal or better solution. For other cases (see Fig. 1 for an example) we need a heuristic to determine the order in which the subgraphs are processed, since we only consider subgraphs with exactly one "refine"-labeled node to maintain feasibility. We process the subgraphs in decreasing order of the number of potentials $|\Phi^{\mathrm{R}}_{v_i}|$ parameterizing the subgraph, i.e., start with the graph with most knowledge sources. If there are two or more subgraphs for which the number of potentials is identical, a subgraph is chosen randomly. After one subgraph has been optimized and the corresponding node has been assigned a valid label from the full label set $\mathcal{L}_{v_i}$ as formulated in Eq. (6), we continue with the next "refine"-labeled node and the corresponding subgraph. Note that we use the label of a refined node in successive optimizations. Using this strategy, we can, for instance, fully resolve a chain structure where the last three nodes have been assigned the "refine" label.

2.4 Parameter Estimation

To estimate the missing energies $\boldsymbol{\beta}$ and the potential weights $\boldsymbol{\Lambda}$, we minimize a max-margin hinge loss L over data $\mathcal{D}$ (LeCun et al., 2006) using a gradient descent scheme. The basic idea is to increase the energy gap between the "correct" state $\mathbf{x}^+$ as defined by the restricted state space $\mathcal{X}^{\mathrm{R}}$ and the best (lowest energy) "incorrect" state $\mathbf{x}^-$ as defined by a MRF inference step until a certain margin m is satisfied. Let our loss function be defined as

$$L(\boldsymbol{\Lambda}, \boldsymbol{\beta}) = \frac{1}{K} \sum_{k=1}^{K} \max\left(0, m + E(\mathbf{x}_k^+) - E(\mathbf{x}_k^-)\right) \cdot \xi_k \tag{7}$$

subject to $\lambda_c \geq 0$ for all $c \in C$. The factor ξ_k is set to the number of errors of the "incorrect" state $\mathbf{x}_k^-$, since a sample with more errors should have a higher influence to steer the optimization. Remember, this is a non-convex problem which requires inference after each update step to find the currently best "incorrect" state $\mathbf{x}_k^-$, which might change after the parameter update. We use a variant of stochastic gradient descent in form of the Adam algorithm by Kingma and Ba (2014) to perform the optimization.

By optimizing the potential weights λ_c, we we can simultaneously optimize the topology: Starting from a fully connected graph, after convergence of the parameter optimization algorithm, we drop all potential functions with a zero weight $\lambda_c < 1\mathrm{e}{-10}$. If it is not possible to start with a fully connected graph since the number of nodes is too large, simpler initial topologies (e.g., trees) can be used to apply, e.g., polynomial-time inference algorithms.

3. Empirical Evaluation

We evaluate our proposed method on the problem of localizing spatially correlated key points in medical images, outlined in Section 3.1. The algorithm for this application is described in detail in Section 3.2 and evaluated on a public dataset as explained in Section 3.3.

3.1 Localization of Spatially Correlated Key Points in Medical Images

A common problem in medical imaging is the localization of spatially correlated key points. It is necessary for clinical tasks like diagnosis, surgical planning, and post-operative assessment. We apply the approach described in Section 2 to the still unsolved problem of localizing and labeling 16 posterior ribs in 2D chest radiographs (see Fig. 2a). One common way to solve such a problem is to use multiple localizers (one for each key point), e.g., a random forest (Donner et al., 2013) or a deep convolutional network (Payer et al., 2016), to preselect likely key point positions, i.e., to define the restricted label sets $\mathcal{L}'_{v_i}$. Then, a MRF is used as spatial regularizer (Donner et al., 2013). As discussed in Section 2.3, the restriction of the label sets is necessary, since the number of possible combinations of key point locations w.r.t. the original label space — the image domain — is infeasible to search.

3.2 Detailed Algorithm for Localization of Spatially Correlated Key Points

In our approach, we either use ensembles of random regression forests (RF; one ensemble for each key point) or a deep, fully convolutional neural network (FCN) to produce one "heatmap" (a single value assigned to each image pixel which can be interpreted as pseudo-probability that the key point

is located at that pixel) for each key point[1]. The heatmaps are post-processed by non-maximum suppression to generate n localization hypotheses for each key point. The outputs of the localizers, i.e., the heatmap values, are used as unary potentials $\phi^{\text{loc}}(\cdot)$ in our MRF comprised of $N = 16$ nodes, one for each key point. To compensate for potentially incorrect first best hypotheses, we add binary potentials assessing the spatial configuration between two key point hypotheses for each key point pair (fully connected). $\phi^{\text{ang}}(\cdot)$ uses a von Mises distribution to model the angle formed by the vector between two key points and the x-axis, $\phi^{\text{dist}}(\cdot)$ uses a Gaussian distribution to model the distance between key points and $\phi^{\text{vec}}(\cdot)$ uses a multivariate Gaussian distribution to model the vector from one key point to the other. In total, our MRF is parameterized by 376 unary and binary potential functions. We apply exact inference in form of the A* search by Bergtholdt et al. (2010), which is also used during training to find the "incorrect" configuration $\mathbf{x}_k^-$. A more detailed description of the potential functions can be found in Mader et al. (2018).

3.3 Dataset and Evaluation Metrics

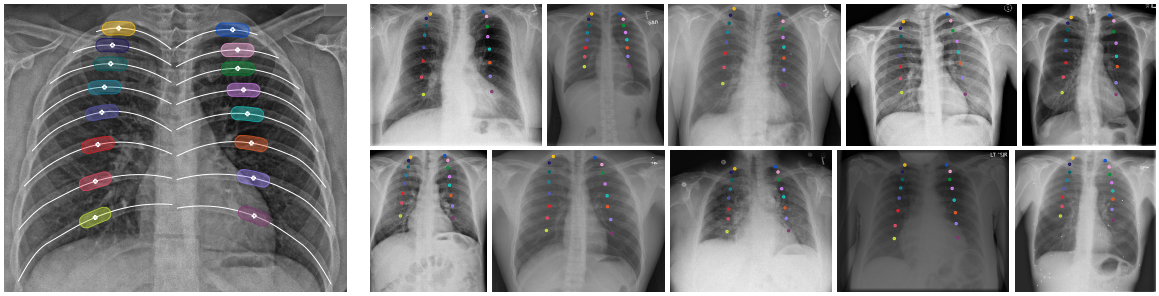
For evaluation, we use 642 images of the publicly available Indiana chest X-ray collection from the U.S. National Library of Medicine (2017). The images were rescaled to an isotropic resolution of 1 x 1 mm/px resulting in an average image size of 410 x 388 px. A semi-automatic approach has been used to derive annotations in form of labeled rib centerlines and labeled key point positions for each of the 16 posterior ribs L2, . . . , L9, R2, . . . , R9. For each key point, a corresponding localization prediction is treated correct (localization and labeling criterion) if it is close to the annotated key point (distance $\leq$ 15 mm) and very close to the annotated centerline (distance $\leq$ 7.5 mm). This criterion ensures that the localization hypotheses lays on the correct rib while allowing for some translation along the rib centerline. An annotated example case and the resulting localization and labeling criterion is depicted in Fig. 2a. All experiments are run in a 3-fold cross-validation setup, which provided us with 428 training images in each fold. 50 % of the images were used to train the localizers and estimate the distributions, 40 % were used to learn the missing energies $\boldsymbol{\beta}$ and potential weights $\boldsymbol{\Lambda}$. The last 10 % was used as validation corpus.

We use two evaluation measures: "% cases" is the percentage of images for which the solution of the MRF inference problem is considered correct for *all* $N = 16$ key points of the input image, according to the above-mentioned criterion. "% key points" is the percentage of correctly predicted key point positions relative to the total number of key points in all images (16×642), independent of how many key points are correct in the individual images.

4. Results and Discussion

First, we compare our novel two-step MRF inference approach including the "refine" labels, Eqs. (5) and (6), against the baseline, i.e., a single MRF inference step without "refine" labels, Eq. (3). This is done for the two methods to preselect key point hypotheses described in Section 3.2, namely the FCN and the RF, using $n = |\mathcal{L}_{v_i}| = 20$ localization hypotheses per key point; see Table 1. As can be seen, our novel approach outperforms the baseline for both preselection methods and for both evaluation measures (percentage of correct cases and percentage of correct key points). In terms of correct cases, the RF preselection method performs better than the FCN (while the contrary holds in

1. While each ensemble of regression forests produces one heatmap for a single key point, one FCN can output the heatmaps of the N key points simultaneously.



(a) Annotations (b) Fully correct cases

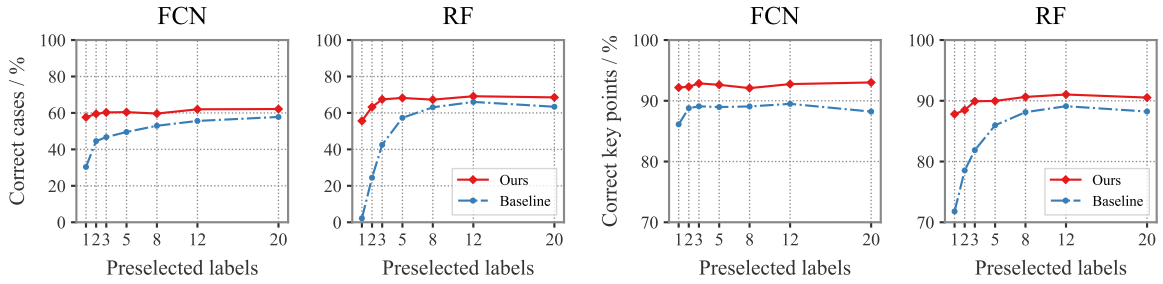
Figure 2: (a) Illustration of the centerline and key point annotations and the resulting localization criterion, i.e., the area where a localization prediction is considered correct. (b) Randomly selected examples from the database for which our two-step inference approach resulted in a configuration where the localization predictions for all key points are considered correct.

Table 1: Evaluation of our novel two-step MRF inference approach with "refine" labels (Eqs. (5) and (6)), to the baseline, i.e., a single MRF inference with restricted label set without "refine" label (Eq. (3)), for two different preselection methods FCN and RF, for $n = |\mathcal{L}_{v_i}| = 20$ preselected hypotheses per key point.

Approach	Correct cases / %		Correct key points / %	
	FCN	RF	FCN	RF
Baseline	57.8	63.4	88.2	88.2
Our two-step inference	**62.1**	**68.5**	**93.0**	**90.5**
Our two-step inference, modified	75.5	68.4	93.9	90.4

terms of correct key points). Analyzing the restricted label sets $\mathcal{L}'_{v_i}$ produced by the FCN in more detail, there are a number of cases where the correct location was not contained in the restricted label set $\mathcal{L}'_{v_i}$ even for large $n = 20$ in contrast to the RF. This is also evident in the amount of correctly chosen "refine" labels per case as illustrated in Fig. 4b, i.e., no correct label was contained in the restricted label set $\mathcal{L}'_{v_i}$, thus the "refine" label was correctly chosen instead. Additionally it can be seen by calculating the theoretical maximal performance of the baseline, single inference method, which amounts to "a posteriori" selecting the optimal hypothesis out of the $n = 20$ candidates per key point in form of an oracle which knows the correct solution: This theoretical maximal performance for the baseline is just 60.1 % correct cases for the FCN, but 87.5 % for the RF.

The RF is not as accurate as the FCN if only considering the top-ranked candidate; on the other hand, if the top-ranked candidate is wrong, the true position was contained in the restricted label set more often for the RF than for the FCN (see Fig. 4b). Therefore, the improvement by our new refinement step in terms of "% key points" is larger for the FCN than for the RF. Inspecting the FCN outputs, we observed that there is sometimes a large gap between the pseudo-probability (and thus the unary potentials) for the top-ranked and the second ranked candidate. This is much more pronounced for the FCN than for the RF. This means that in the refinement step, the binary potentials would have to compensate for this gap. It is thus interesting to see whether the refinement performs better in such a case when excluding the unary potentials in the refinement step. This is listed in



(a) Cases (b) Key points

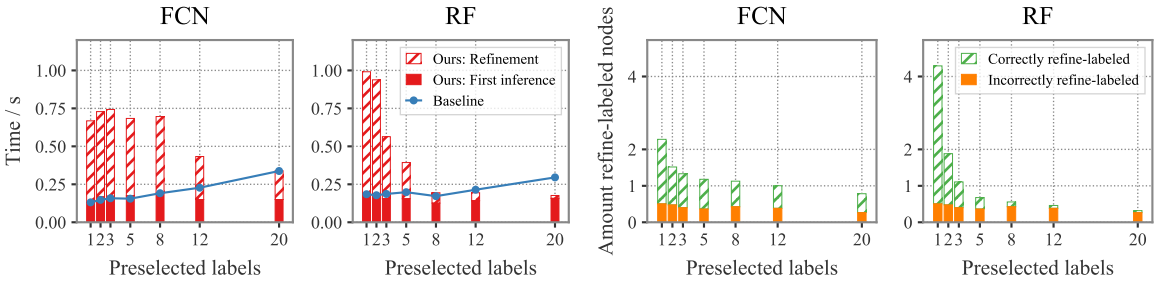
Figure 3: Evaluation of our approach, i.e., % cases (a) and % key points (b), as a function of the number $n = |\mathcal{L}_{v_i}|$ of preselected labels, for both preselection methods FCN and RF.

Table 1 as "Our two-step inference, modified" and indeed, this improves especially the percentage of cases for the FCN (with a smaller improvement of the percentage of key points and only a minor effect on the results of the RF). This is reasonable, given that we assume the preselection method failed in case a node was assigned the "refine" label and our unary potential makes use of the preselection method's output as well. [2]

One of the key parameters is the number $n = |\mathcal{L}_{v_i}|$ of labels to preselect. It is directly linked to the performance in terms of accuracy as well as runtime in training and test. Therefore, in Fig. 3 we present evaluation results (% cases and % key points) as a function of the number $n = |\mathcal{L}_{v_i}|$ of preselected labels for both preselection methods FCN and RF. For the baseline approach, the optimal performance (w.r.t. % cases and % key points) is obtained for about $n = 12$ key point hypotheses. If the restricted label set contains more hypotheses — as selected by the FCN or RF — performance may degrade slightly. Using our two-step inference approach with the "refine" label, performance is optimal already for $n = 3$ key point hypotheses. Our approach outperforms the baseline for all n (compare Table 1 for $n = 20$), and we observe only a slight performance degradation with increasing n (if any). In the extreme case, we may restrict the label set to even a single hypothesis; while the performance of the baseline single step MRF inference is unacceptably low, our novel MRF inference approach leads to acceptable performance (at least for the % key points evaluation measure).

The corresponding test times w.r.t. different numbers of preselected labels are illustrated in Fig. 4a. We compare the test time of our novel two-step MRF inference approach (runtime of our refinement step stacked on top of the runtime of our first inference step including the "refine" label) to the baseline, single MRF inference without the "refine" label. Unsurprisingly, the runtime of the baseline increases with the number of preselected labels. However, our very similar first inference step, shows a contrary more stable behavior. This is probably related to the energy thresholds introduced by β_c that are beneficial for the heuristic of the A* search by Bergtholdt et al. (2010). The runtime of the refinement step decreases with an increasing number of preselected labels, which is in turn related to the general decrease of selected "refine" labels in the first step, which is illustrated in Fig. 4b. Generally, the runtime of our approach starts out worse than the runtime of the baseline approach but gets better with an increasing number of labels until it is even faster ($n = 20$) than

2. This method was used in Mader et al. (2018) with an additional L1 regularization of λ_c in the loss formulation Eq. (7) (to accelerate the sparsification of potential functions) and a more elaborate cross-validated parameter selection starting from a simpler (not fully connected) topology; these modifications resulted in slightly increased results.



(a) Test time (b) Avg. count of "refine"-labeled nodes per case

Figure 4: (a) Runtime of the baseline method (blue line) compared to the runtime of our novel refinement step illustrated as bar stacked on top of our first inference step including the "refine" label for different numbers of preselected labels n for FCN and RF. We used the output of one fold only to create comparable results. (b) Average amount of assigned "refine" labels per case (image) splitted into correct assignments (i.e., none of the preselected labels was correct) and incorrect assignments (i.e., at least one preselected label was correct) for different numbers of preselected labels n for FCN and RF.

the baseline. This might also be related to the fact that more terms are dropped after the parameter optimization (Section 2.4) for increasing numbers of preselected labels. E.g., to produce the results from Table 1 with $n = 20$, graphs with, on average, 225 remaining of 376 initial potential functions were used.

Note that approximate approaches like, e.g., the move making strategy "lazy flipper" by Andres et al. (2012) or loopy belief propagation (Wang et al., 2013), are not able to solve the full problem from Eq. (2) in an acceptable period of time (< 1 minute) in order to compare them to our method.

5. Conclusions

In this paper, we presented an approach to overcome the limitation of artificially reduced label sets (to enable acceptable MRF inference times) and the inherent risk of excluding correct solutions. First, we extended the reduced label set by a novel "refine" label. The corresponding missing energies were jointly learned together with the potential weights using a max-margin hinge loss and gradient descent. Second, after a first inference, we greedily extracted subgraphs around the "refine"-labeled nodes and performed (potentially multiple) inference steps over the full label set instead of the reduced one.

We demonstrated how our approach can be applied to the common task of multi-object localization, where a preselection method is used to reduce the full label set (all possible pixel positions) to a much smaller set of viable candidates only. We compared two different preselection methods, namely a fully convolutional network and ensembles of random forests, followed by a MRF as a spatial regularizer. This setup was applied to the unsolved problem of posterior rib detection in 2D chest radiographs. We analyzed the accuracy of our approach in terms of correct cases and correct key points as well as the inference runtime compared to the baseline as a function of the number of preselected hypotheses. In particular, we showed that our approach is superior to the baseline, single inference approach in terms of accuracy or runtime or both.

Apart from using the novel "refine" label to mark nodes the location of which is suboptimal and needs to be refined in a second inference step, there are also other ways of interpreting the additional "refine" label: First, it can be interpreted as a label indicating uncertainty (w.r.t. key point location), necessitating reasoning methods, e.g., for quality control ensuring certain levels of certainty. Second, the semantic meaning of this additional label can be changed to a "missing" label, indicating nodes which are not contained in the field of view of the current image. This is especially useful in applications where only a subset of the N key points may be contained in an image (Mader et al., 2017).

In future work, it is interesting to apply our proposed approach to other, non-image related inference problems.

Acknowledgments

This work has been financially supported by the Federal Ministry of Education and Research under the grant 03FH013IX5. The liability for the content of this work lies with the authors.

References

B. Andres, J. H. Kappes, T. Beier, U. Köthe, and F. A. Hamprecht. The lazy flipper: Efficient depth-limited exhaustive search in discrete graphical models. In *European Conference on Computer Vision*, pages 154–166. Springer, 2012.

M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *International journal of computer vision*, 87(1):93–117, 2010.

R. Donner, B. H. Menze, H. Bischof, and G. Langs. Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization. *Medical image analysis*, 17(8):1304–1314, 2013.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ICLR*, 2014.

L. M. Koch, M. Rajchl, T. Tong, J. Passerat-Palmbach, P. Aljabar, and D. Rueckert. Multi-atlas segmentation as a graph labelling problem: Application to partially annotated atlas data. In *International Conference on Information Processing in Medical Imaging*, pages 221–232. Springer, 2015.

N. Komodakis, N. Paragios, and G. Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):531–552, 2011.

N. Komodakis, B. Xiang, and N. Paragios. A framework for efficient structured max-margin learning of high-order mrf models. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1425–1441, 2015.

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

A. O. Mader, C. Lorenz, M. Bergtholdt, J. von Berg, H. Schramm, J. Modersitzki, and C. Meyer. Detection and Localization of Landmarks in the Lower Extremities Using an Automatically Learned Conditional Random Field. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 64–75. Springer, 2017.

A. O. Mader, J. von Berg, A. Fabritz, C. Lorenz, and C. Meyer. Localization and Labeling of Posterior Ribs in Chest Radiographs using a CRF-regularized FCN with Local Refinement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018.

C. Payer, D. Štern, H. Bischof, and M. Urschler. Regressing heatmaps for multiple landmark localization using cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–238. Springer, 2016.

L. Shao, R. Yan, X. Li, and Y. Liu. From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms. *IEEE Transactions on Cybernetics*, 44(7):1001–1013, 2014.

D. Štern, T. Ebner, and M. Urschler. From local to global random regression forests: exploring anatomical landmark localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 221–229. Springer, 2016.

U.S. National Library of Medicine. Open-i Open Access Biomedical Image Search Engine. `https://openi.nlm.nih.gov`, 2017. Accessed February 14, 2018.

C. Wang, N. Komodakis, and N. Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627, 2013.

J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):899–909, 2010.