2013 IEEE Conference on Computer Vision and Pattern Recognition

GeoF: Geodesic Forests for Learning Coupled Predictors

[†]P. Kontschieder [‡]P. Kohli [‡]J. Shotton [‡]A. Criminisi
[†]Graz University of Technology, Austria [‡]Microsoft Research Ltd, Cambridge, United Kingdom

Abstract

Conventional decision forest based methods for image labelling tasks like object segmentation make predictions for each variable (pixel) independently [3, 5, 8]. This prevents them from enforcing dependencies between variables and translates into locally inconsistent pixel labellings. Random field models, instead, encourage spatial consistency of labels at increased computational expense.

This paper presents a new and efficient forest based model that achieves spatially consistent semantic image segmentation by encoding variable dependencies directly in the feature space the forests operate on. Such correlations are captured via new long-range, soft connectivity features, computed via generalized geodesic distance transforms. Our model can be thought of as a generalization of the successful Semantic Texton Forest, Auto-Context, and Entangled Forest models. A second contribution is to show the connection between the typical Conditional Random Field (CRF) energy and the forest training objective. This analysis yields a new objective for training decision forests that encourages more accurate structured prediction.

Our GeoF model is validated quantitatively on the task of semantic image segmentation, on four challenging and very diverse image datasets. GeoF outperforms both state-of-the-art forest models and the conventional pairwise CRF.

1. Introduction

Many problems in computer vision can be formulated in terms of structured output prediction. Here, the term 'structured' relates to the presence of dependencies between output variables. For instance, in image labelling problems such as object segmentation or image denoising, the variables associated with neighboring pixels are more likely to take the same labels. In recent years, decision forests [3, 5, 8] have become very popular for the solution of a wide variety of image labelling problems - from anatomy delineation in 3D medical images [17] and semantic segmentation in natural images [24, 25] to human pose estimation for the Microsoft Kinect sensor [23].

The success of forest models is largely due to: their scalability to large amount of data, ability to learn long-range dependencies between features and output variables, relative robustness to overfitting, and finally, efficient predictions. The last of these qualities is derived from the independence assumption made by these methods. In fact, conventional decision forests ignore the structure in output spaces and make predictions for each output variable independently. This assumption prevents them from enforcing dependencies between variables, and for semantic segmentation tasks, translates into pixel labellings that do not follow object boundaries and are inconsistent with context.

To overcome these problems, Markov or Conditional random fields (MRF/CRF) [4] are used as a post-processing step [19, 25]. For instance, in [10, 25] image segmentation is achieved by first computing pixel-wise unaries via supervised classification, and then smoothing the labels with a CRF. The more recent works in [11, 19] essentially present a CRF model, where the pairwise potentials (and not just the unaries) are conditioned on the data and predicted via a *single* tree. Another way of mixing trees and fields is presented in [20], where again, the underlying model is a CRF. In the forest approach in [13], spatial smoothness is achieved by combining structured class-labels that are learned by incorporating joint statistics in a small neighborhood. Although all the above approaches lead to improved results, this comes at the cost of increased computation at test time.

The main contribution of this paper is a new and efficient forest-based model for structured output prediction. Our framework overcomes the above-mentioned problem by incorporating learned spatial context directly *within* the forest itself. This leads to smooth, pixel-wise labellings without the need for field-based post-processing. Long-range correlations between pixel labels are captured via new *soft connectivity features* which can be computed efficiently using generalized geodesic distance transforms. Another contribution is to analyse the relationship between a typical CRF-like energy and the forest training objective. This analysis leads to a new objective for training decision forests that produces more accurate semantic segmentation.

We validate our model on the task of segmenting four challenging and very diverse image datasets: face images,

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.16

medical scans, depth images and driving videos. Quantitative results demonstrate the superiority of our model both in terms of accuracy and efficiency, with respect to state-of-the-art forest models and grid-based pairwise CRFs.

In the literature. Our work is related to methods based on sequential classification. The recent work on auto-context [24, 26], stacking [18, 28], deep learning [14, 15] and entanglement [17] has shown how a sequence of classifiers using the output of the previous classifier as input to the next can both effectively capture spatial context (*e.g.* learning that the heart is between the lungs) and improve accuracy. In [9], the relationship between anytime classification and intermediate predictions within decision trees is shown. In [21] the authors reinterpret conventional message-passing inference on graphical models as a sequential probabilistic inference algorithm.

Our geodesic forest model (GeoF) can be seen as a generalization of semantic texton forests [24], auto-context [24, 26], and entanglement forests [17]. In fact, GeoF builds upon these models by using: (i) new, long-range *soft connectivity features*, and (ii) a new *field-inspired objective* for forest training. The combination of those novel features and objective function encourage GeoF to produce context-consistent, spatially-smooth semantic image segmentation.

2. Background and Problem Formulation

In this work, an image[1] is denoted $J : \Omega \subset \mathbb{N}^2 \rightarrow \mathbb{R}$, and a 2D pixel position is denoted $\mathbf{p} \in \Omega$. We cast the semantic segmentation task as that of associating each pixel $\mathbf{p}$ with its corresponding discrete class label $c \in \mathcal{C}$. It is a typical, supervised classification task, where we assume provided a set of labelled training images $\{J\}$ and their segmentations. A vector of feature responses at position $\mathbf{p}$ is denoted $\mathbf{v}(\mathbf{p}) = (v_1, \dots, v_i, \dots v_m) \in \mathbb{R}^m$. A set of training data points (and associated labels) is denoted $\mathcal{S} = \{\mathbf{z}_i\}$, with each training point-class pair being $\mathbf{z} = (\mathbf{v}, c)$. Let $\mathbf{c} = \{c_\mathbf{p} | \mathbf{p} \in \Omega\}$ denote the vector of class variables predicted by our classifier on the entire image. We use $\mathbf{c}_d$ to denote predictions obtained at depth d in the tree. D denotes the maximum tree depth and T the number of trees.

Random field models. Given an image J, its most probable labelling can be inferred by maximizing the posterior:

$$\mathbf{c}^* = \arg\max_\mathbf{c} p(\mathbf{c}|J) = \arg\max_\mathbf{c} p(J|\mathbf{c})p(\mathbf{c}) \qquad (1)$$

The conventional pairwise random field models assume that the posterior distribution factorizes into a product of unary and pairwise potential functions as:

$$p(\mathbf{c}|J) = \prod_\mathbf{p} \psi\left(c_\mathbf{p}, \mathbf{v}(\mathbf{p})\right) \prod_{(\mathbf{p},\mathbf{q}) \in \mathcal{N}} \phi\left(c_\mathbf{p}, c_\mathbf{q}, \mathbf{v}(\mathbf{p}), \mathbf{v}(\mathbf{q})\right) \quad (2)$$

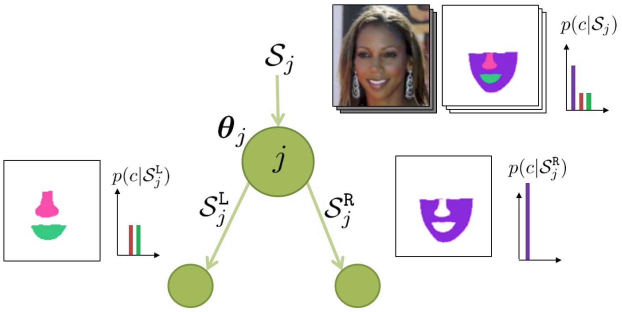

Figure 1. **Training a split node j of a tree.** Our model training seeks parameters $\boldsymbol{\theta}_j$ which aim to maximize both the class purity *and* spatial compactnes of pixel clusters in child nodes.

where the set $\mathcal{N}$ of pixel pairs describes a pre-defined neighbourhood system. Although this factorization assumption makes inference of the Maximum a Posteriori (MAP) solution for many models tractable, it severely limits the expressive power of the model. Furthermore, inference and learning are computationally expensive.

Decision forest models. Decision forests [3, 5, 8] further assume that the posterior decomposes over individual variables as: $p(\mathbf{c}|J) = \prod_\mathbf{p} \psi\left(c_\mathbf{p}, \mathbf{v}(\mathbf{p})\right)$. Ignoring the dependency between output variables makes predictions independent and efficient.

Forest prediction. To make predictions, a series of feature tests starting at the root node are applied to each pixel independently. At each node a test is computed on the feature response $\mathbf{v}(\mathbf{p})$ and, depending on the results, the pixel is sent to the left or right child. The procedure is repeated until the pixel reaches a leaf node. At this point the empirical class distribution $\psi(c_\mathbf{p}, \mathbf{v}(\mathbf{p}))$ associated with the leaf is read off. The MAP label for the pixel is obtained as:

$$c^*_\mathbf{p} = \arg\max_{c_\mathbf{p}} \psi(c_\mathbf{p}, \mathbf{v}(\mathbf{p})). \qquad (3)$$

Forest training. Training involves: (i) selecting the feature tests at each split node of each tree, and (ii) estimating the distribution $\psi(c_\mathbf{p}, \mathbf{v}(\mathbf{p}))$ associated with each leaf. Typically, a decision tree is trained greedily, where for each split node j the parameters $\boldsymbol{\theta}_j$ associated with a low energy (*e.g.* low class entropy) in the child nodes are selected. Figure 1 illustrates this point and suggests that ideally we would like training to maximize class purity as well as encouraging spatial compactness of the resulting pixel clusters.

Coupling forest predictions to reveal hidden correlations. Although the independence assumption enables efficient training and rapid predictions with random forests, it prevents the model from enforcing dependencies between variables, and for image segmentation problems, translates into pixel labellings that do not follow object boundaries and are not consistent with local or global context. In this paper, we overcome this problem and encourage forests

[1] For simplicity the notation refers to single-banded images. The extension to multi-channel images is straightforward.

to produce spatially compact/coherent pixel labellings. In what follows, we will show how a learned model of spatial context can be encoded *within* a decision forest directly. This leads to smooth, pixel-wise image labellings without the need for additional post-processing.

One of the key theoretical insights of our work is the observation that although forests make predictions for each variable independently, these predictions are related due to correlations at the feature level. For instance, in the semantic image segmentation task consider the class predictions of two pixels $\mathbf{p}$ and $\mathbf{q}$. From (3) we can see that the MAP labels $c^*_{\mathbf{p}}$ and $c^*_{\mathbf{q}}$ are functions of the input features responses $\mathbf{v}(\mathbf{p})$ and $\mathbf{v}(\mathbf{q})$ *i.e.* $c^*_{\mathbf{p}} = f(\mathbf{v}(\mathbf{p}))$ and $c^*_{\mathbf{q}} = f(\mathbf{v}(\mathbf{q}))$. Therefore, output-variable dependencies can be encoded in the features that the forest operates on. We exploit this insight to couple forest predictions in two ways: (i) we enable *long-range geodesic features* for soft connectivity between image regions; (ii) we train *entangled* classification forests, where geodesically smoothed, intermediate class posteriors estimated at higher levels in each tree are used as features in the training of the tree lower levels. We describe details of these two contributions in the next two sections.

3. Long-range, soft connectivity features

The need for long-range connectivity features. In [16, 23, 27] the authors have shown how simple pixel comparison features can be effective in classification tasks when used within a decision forest. Such features are extremely fast to compute (they involve just pixel-wise read-outs), but not very expressive. This is illustrated in Fig. 2 where we compare pair-wise intensity difference features with an alternative feature response based on the cost of the shortest path connecting the two points. Intuitively, path-based features should better capture connectivity between points. In turn, this could be used within a supervised segmentation algorithm to decide whether two points should be assigned the same class label or not. For example, the points r_3 and p_3 have identical intensity values. However, one is in the lungs and the other in the air outside the body. Since the shortest path connecting them has a high geodesic length (it cuts through high image gradients, see definition in (4)), this provides a hint that the two points may not be part of the same object/class. Similarly, the points r_2 and p_2, despite being far from each other in Euclidean terms, they are close in geodesic terms. This provides evidence that they may belong to the same object (the aorta in this case).

The problem. In theory using pixel-pair geodesic path lengths within a supervised classifier could enable edge-aware label smoothing, similar to CRFs. However, these features need to be available at test time, for *any* pair of pixels. But computing any-pair shortest paths within an image on the fly is infeasible. We circumvent this problem by proposing a novel set of visual features which are computationally efficient and yet manage to capture the degree of connectivity between probabilistically defined image *regions*. They are based on the use of *generalized* geodesic distances, as introduced in [7] and summarized next.

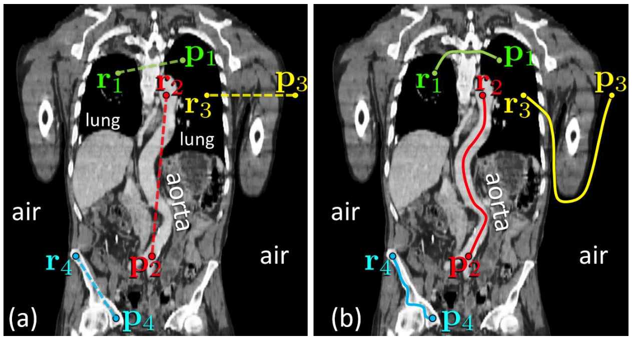

Figure 2. **Connectivity features.** A 2D frontal slice through a 3D computed tomography scan. **(a)** Given a pixel pair (a reference and a probe pixel) popular features only look at the intensities at the two pixel positions, and ignore what happens in between. **(b)** In contrast, the length of the shortest path connecting the pixel pair carries richer information. The geodesic length of the shortest path connecting two points provides hints about the points belonging (or not) to the same object class (*e.g.* the aorta in the figure).

Generalized geodesic distances. Given a grey-valued image J, and a real-valued object "soft mask" (that encodes pixel likelihood) $M(\mathbf{p}) : \Omega \in \mathbb{N}^d \to [0, 1]$ the generalized geodesic distance Q is defined as follows:

$$Q(\mathbf{p}; M, \nabla J) = \min_{\mathbf{p}' \in \Omega} (\delta(\mathbf{p}, \mathbf{p}') + \nu M(\mathbf{p}')) \qquad (4)$$

with the geodesic distance between two points $\mathbf{p}$ and $\mathbf{q}$:

$$\delta(\mathbf{p}, \mathbf{q}) = \inf_{\mathbf{\Gamma} \in \mathcal{P}_{\mathbf{p},\mathbf{q}}} \int_0^{l(\mathbf{\Gamma})} \sqrt{1 + \gamma^2 (\nabla J(s) \cdot \mathbf{\Gamma}'(s))^2} ds. \qquad (5)$$

where $\mathbf{\Gamma}$ is a path connecting the two points and $\mathcal{P}_{\mathbf{p},\mathbf{q}}$ is the set of all possible paths. Thus (4) defines the distance of any point in the image from a region in the image defined via the "soft belief" M.

Soft connectivity between a pixel and a class region. Let us assume that we have an image J and also the belief matrix M associated with a chosen class. Now we can compute the distance of every point in the image from the given class region. Note that the class region is defined in a probabilistic way and we do not need to select hard seed positions. Also, the belief M could be the output of any given probabilistic classifier[2]. We can think of having C such masks and thus C such distances associated with each input image.

Figure 3 shows an illustration. Given a depth image (*e.g.* acquired with Kinect), we assume we have a classi-

[2] For example, the map M could be defined as $1 - p$, with p the probability of a pixel belonging to a given object class (In [7] $M \sim 0$ when $p \sim 1$).

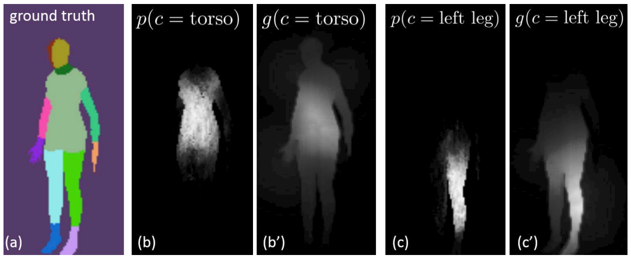

Figure 3. **Generalized geodesic distances from probabilistic class regions. (a)** Ground truth body part labels for a depth image. **(b, c)** Approximate class probability maps $p(c|\mathbf{v})$; assumed given here. **(b', c')** Geodesic-filtered probability maps $g(c|\mathbf{v})$. Notice how g may be interpreted as an edge-aware, diffused version of the noisier probabilities p (see definition in (6)). The visual features used in GeoF are pixel read-outs of the g maps. They efficiently capture long-range connectivity (of a pixel to a class region).

fier which when evaluated on an image produces the class probabilities $p(c = \text{torso})$ and $p(c = \text{left leg})$. We can use those probabilities to construct the soft masks M needed for the generalized geodesic distance transform, and the corresponding filtered probabilities will be $g(c = \text{torso})$ and $g(c = \text{left leg})$. The g maps (definition in (6)) are an edge-aware, smoothed version of the class probabilities p. Contrast sensitivity is modulated by the geodesic strength parameter $\gamma \geq 0$ in (5). Next we incorporate geodesic distances as connectivity features within a classification forest.

4. Entangled geodesic forests

Here we are interested in extremely efficient semantic segmentation. Thus, we build upon decision forests [3, 5, 8], because of their speed and flexibility. Next we describe our extension to enable coherent segmentation.

4.1. Entangled soft connectivity features

As illustrated in Fig. 4 in the spirit of entangled forests [17] we train all trees: (i) in parallel, (ii) in breadth-first order, and (iii) in sections. When training the first section (section 0) only *appearance-based features* (*e.g*. raw intensities) are available. However, when training the next section more *derived* features become available. In fact, the class posteriors $p(c|\mathbf{v})$ of the previous section may be used as input features to the next [17]. In this paper we further augment such features by using the geodesically filtered versions of those posteriors, $g(c|\mathbf{v})$.

More formally, we are given an ordered set of sections $(s_0, s_1, \ldots, D)$, where s_i indicates the maximum depth of the i^{th} section and D is the maximum tree depth. Given a class posterior $p_{s_i}(c|\mathbf{v})$ computed at the i^{th} section (with $i > 0$), its geodesically smoothed version is defined as

$$g_{s_i}(c|\mathbf{v}(\mathbf{p})) = \frac{1}{W}\, p_{s_i}(c|\mathbf{v}(\mathbf{p}))\, e^{-\frac{Q(\mathbf{p};p_{s_i}(c|\mathbf{v}(\Omega)),\nabla J)^2}{\sigma^2}} \quad (6)$$

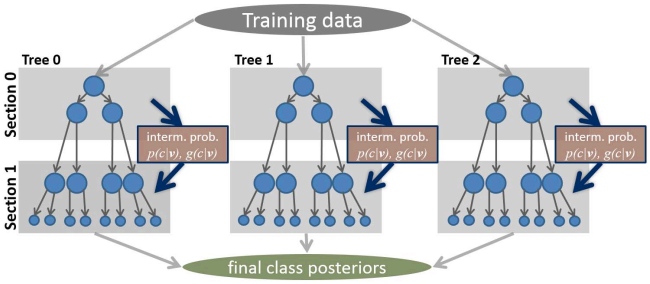

Figure 4. **An entangled geodesic forest.** A forest with three entangled trees. The trees are entangled because intermediate predictions of their top section are used (together with raw intensity features) as features for training of the lower sections. Only one entanglement section is shown here, for clarity.

where W is a normalization factor to ensure probabilistic normalization: $\sum_c g_{s_i}(c|\mathbf{v}) = 1$. $Q(\cdot)$ is defined in (4). As shown in Fig. 3, this operation has the effect of diffusing the class probabilities spatially, while preserving strong edges.

Feature responses for a reference pixel $\mathbf{r}$ are defined as a function of tree depth d, and as sum, differences or absolute differences between two pixel probe values in different *feature channels*[3], *i.e*.

$$\begin{aligned} v_i^d(\mathbf{r}) &= F_k^d(\mathbf{p}_1) + F_k^d(\mathbf{p}_2), \\ v_i^d(\mathbf{r}) &= F_k^d(\mathbf{p}_1) - F_k^d(\mathbf{p}_2), \text{or} \\ v_i^d(\mathbf{r}) &= |F_k^d(\mathbf{p}_1) - F_k^d(\mathbf{p}_2)| \end{aligned}$$

where $k \in \{0, 1, 2\}$ denotes the channel where features are computed, and: (i) $F_0^d(\mathbf{p}) = J(\mathbf{p})$, *i.e*. the raw image intensities, (ii) $F_1^d(\mathbf{p}) = p_{s(d)}(c|(\mathbf{p}))$, *i.e*. the intermediate class posteriors computed in the section $s(d)$ defined by the depth d, and (iii) $F_2^d(\mathbf{p}) = g_{s(d)}(c|(\mathbf{p}))$, *i.e*. the geodesic-filtered posteriors, capturing connectivity of point $\mathbf{p}$ to the region of class c. The entangled feature channels ($k = 1, 2$) are available only for section s_1 and greater, and are computed very efficiently as table look-ups.

4.2. Field-inspired forest training objective

This section describes our second contribution: the use of a new objective for the forest training procedure. In what follows we depart from the traditional information-theoretic training objective, typically used in classification forests, and derive a random-field inspired objective function.

Information-theory based objective (I). Most algorithms for training classification forests are greedy and find the optimal parameters for a split node j as $\boldsymbol{\theta}_j = \arg\min_{\boldsymbol{\theta}} E(\mathcal{S}_j, \boldsymbol{\theta})$ (Fig. 1). The traditional choice for the

[3] Here the term "feature channel" indicates both the original image bands (*e.g*. three bands for color images) as well as derived bands where features are computed (*e.g*. gradients or intermediate class probabilities).

objective function E is the Shannon entropy $E_{\texttt{IT}}$, which after some algebraic manipulation reduces to

$$E_{\texttt{IT}}(\mathcal{S}_j, \boldsymbol{\theta}) = - \sum_{i \in \{\texttt{L},\texttt{R}\}} \sum_{c \in \mathcal{C}} n(c, \mathcal{S}_j^i) \; \log \frac{n(c, \mathcal{S}_j^i)}{|\mathcal{S}_j^i|} \qquad (7)$$

with $n(c, \mathcal{S})$ denoting the number of training pixels of class c in the training subset $\mathcal{S}$ (please refer to Fig. 1 for notation).

Field-inspired objective (I). Similarly, we can think of training each tree split node by using an MRF energy $E = E_{\texttt{RF}}$, which is typically defined as

$$E_{\texttt{RF}}(\mathcal{S}_j, \boldsymbol{\theta}) = \sum_{i \in \{\texttt{L},\texttt{R}\}} \left(\sum_{\mathbf{z}_k \in \mathcal{S}_j^i} \psi(\mathbf{z}_k; \mathcal{S}_j^i) + \lambda \sum_{\mathbf{z}_k \in \mathcal{S}_j^i, \mathbf{r} \in \mathcal{N}(\mathbf{z}_k)} \phi(\mathbf{z}_k, \mathbf{r}) \right)$$

with $\mathcal{N}(\mathbf{z}_k)$ denoting a local neighborhood of the point $\mathbf{z}_k$. As unary potentials we choose the commonly used log-loss $\psi(\mathbf{z}; \mathcal{S}) = -\log p(c = c(\mathbf{z})|\mathcal{S})$. If we ignore the pairwise term (by setting $\lambda = 0$) we get

$$E_{\texttt{RF}}(\mathcal{S}_j, \boldsymbol{\theta}) = - \sum_{i \in \{\texttt{L},\texttt{R}\}} \sum_{c \in \mathcal{C}} n(c, \mathcal{S}_j^i) \; \log \frac{n(c, \mathcal{S}_j^i)}{|\mathcal{S}_j^i|}. \qquad (8)$$

So, we discover that under the above assumptions, (8) and (7) are identical. Thus, conventional entropy-based tree training corresponds *exactly* to minimizing an MRF-like energy which uses the log-loss as unary and no pairwise term[4]. Further interesting findings arise when we consider the effect of having unbalanced classes in the training set.

4.2.1 Correcting class imbalance

In general, in the original training set $\mathcal{S}_0$ we have $n(c_1, \mathcal{S}_0) \neq n(c_2, \mathcal{S}_0), \; c_1, c_2 \in \mathcal{C}$. So it is often beneficial to re-balance the effect of different classes (as shown *e.g.* in [24]). This is particularly important in the context of semantic segmentation, where often the pixels in the background class are much more numerous than those in other classes. Thus we define the following global re-balancing factors: $\omega_c = \frac{\sum_{k \in \mathcal{C}} n(k, \mathcal{S}_0)}{n(c, \mathcal{S}_0)}$ and the corresponding node-based normalization factor $Z(\mathcal{S}_j) = \sum_{k \in \mathcal{C}} \omega_k \; n(k, \mathcal{S}_j)$.

Information-based objective (II). Now, after some algebraic manipulation the energy in (7) becomes

$$E_{\texttt{IT}}(\mathcal{S}_j, \boldsymbol{\theta}) = - \sum_{i \in \{\texttt{L},\texttt{R}\}} \sum_{c \in \mathcal{C}} w_c \; n(c, \mathcal{S}_j^i) \; \log \frac{w_c n(c, \mathcal{S}_j^i)}{Z(\mathcal{S}_j^i)}. \qquad (9)$$

Field-inspired objective (II). Analogously, the class-rebalanced field unary in (8) becomes

$$E_{\texttt{RF}}(\mathcal{S}_j, \boldsymbol{\theta}) = - \sum_{i \in \{\texttt{L},\texttt{R}\}} Z(\mathcal{S}_j^i) \sum_{c \in \mathcal{C}} n(c, \mathcal{S}_j^i) \; \log \frac{w_c n(c, \mathcal{S}_j^i)}{Z(\mathcal{S}_j^i)}. \qquad (10)$$

Thus, after class re-balancing, the entropy-based energy in (9) and the field unary in (10) are no longer the same. Quantitative comparisons in the next section will show which training objective produces the most accurate results. Also, as discussed later, the use of connectivity features negates the need for a pair-wise term rich energy.

5. Results and Comparisons

We validate our semantic segmentation approach on four, very diverse labelled image datasets.

LFW: Labelled Faces in the Wild. This is an augmented version of the public dataset in [1], where we have manually segmented a subset of 1250 images into the following 8 categories: background, nose, mouth, L/R eye, L/R eyebrow and lower face. The contained faces exhibit strong variations in pose and appearance. Furthermore, the mouth and eyes show considerable articulation.

CT: Computed Tomography. We tested our algorithm also against a new dataset of medical images. It comprises 2D coronal slices taken at random positions within labelled, 3D CT scans. As ground truth, different anatomical entities have been segmented in 3D, using an interactive segmentation tool. We have the following 9 classes: background (BG), heart (HR), liver (LI), spleen (SP), left/right lung (LL/RL), left/right kidney (LK/RK) and aorta (AO).

KinBG: depth images. This is a new dataset, similar to the body-part Kinect dataset in [23], with the difference that the retargeted mocap characters have been inserted within a Kinect acquired, real background scene. We have 12 body parts (L/R head side, neck, torso, L/R arm, L/R hand, L/R leg, L/R foot) and 3 background classes. In fact, in contrast to [23], we do not assume a given FG/BG separation, and the background is subdivided into: floor, back wall and everything else. This yields a total of 15 classes.

CamVid: video dataset. This road scene video dataset was initially introduced in [6]. A subset of 711 image frames are almost entirely segmented into 32 classes. In our setup, we followed the training/test protocol as in recent work [6, 13, 29] and used the following 11 object classes: road, building, sky, tree, sidewalk, car, column-pole, sign-symbol, fence, pedestrian and bicyclist.

Comparisons with related methods. We provide comparisons with various state-of-the-art forest-based approaches [13, 17, 29]. We also compare against approaches using forest-based unaries followed by CRF smoothing [12]. In the latter, as energy model, we used a log-loss as unary term and a contrast-sensitive Potts model as pairwise term. Additionally, we also implemented an auto-context [26] version of classification forests where: A first forest is trained using raw intensity features; Then, a second forest is trained using both raw intensities and the probabilities from the first forest as features. Both entangled geodesic features and un-entangled class posteriors are

[4] We will discuss the effect of removing the pair-wise interactions later.

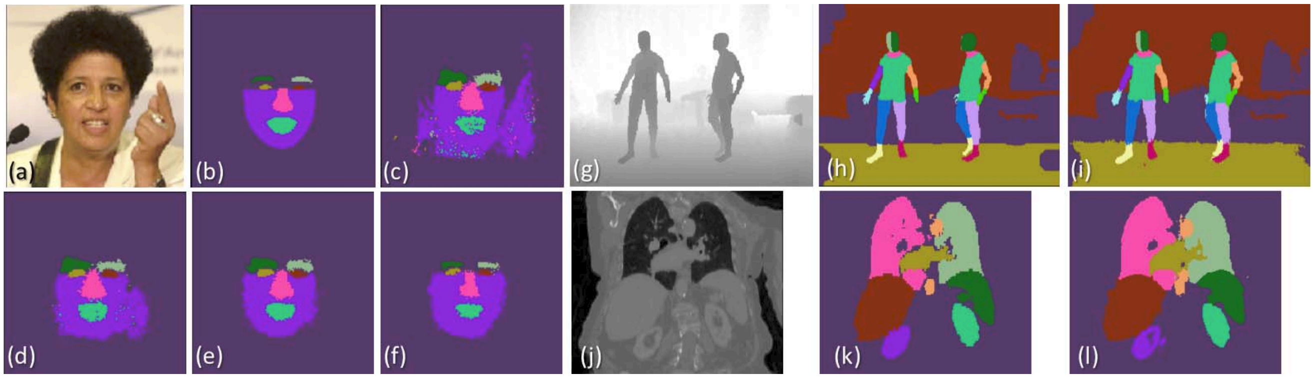

Figure 5. **The effect of geodesic entanglement on spatial coherence of the output semantic segmentation.** **(a, g, j)** Input test images, from the LFW, KinBG and CT datasets, respectively. **(b, h, k)** Ground truth labels (different colors for different classes). **(c)** Segmentation results from conventional pixel-wise classification forest. The lack of spatial smoothing produces noisy labeling. Notice also the overly large eye/eyebrow segments. **(d)** Results from forest with probability entanglement. Entangling the p feature channels only helps spatial coherence of the output. **(e)** Results from forest with geodesic entanglement. Enabling the long-range geodesic feature channels g helps spatial coherence further. The spurious hand region is gone. **(f, i, l)** Results from forest with geodesic entanglement and field-inspired energy term. Using our field-inspired energy term helps further still. *e.g.* notice the better recovered eyebrow shape in (f).

considered here. A fair comparison is ensured by training all forest-based algorithms to the same number of nodes. All baseline algorithms have been individually optimized so as to yield the highest Jaccard scores.

Qualitative results. Fig. 5 shows qualitative results on three datasets. The combination of entangled geodesic features and log-loss training produces coherent segmentations without the need for field-based post-processing.

Quantitative results are summarized in Table 5 where we compare the accuracy of various segmentation algorithms in terms of their Jaccard score (as adopted also in [2]). For all forest based algorithms we fix $T = 10$ and $D = 20$, except for the CamVid dataset where we use a maximum depth $D = 17$ since the number of training samples is considerably smaller. We also report runtimes for similarly non-optimized C# implementations. However, decision forests are well-suited for GPU implementations [22].

Labelled Faces in the Wild. The baseline forest (01) yields a mean Jaccard score of only 38.1% as it produces noisy segmentations and overly bold segments for the smaller objects such as the eyebrows (see Fig. 5 (c)). CRF-based post-processing (02) boosts the score to 45.2%, still lower than what our implemented auto-context forest (03) and our proposed geodesic forests achieve (07-16). Both the use of entangled geodesic features and the field-inspired energy help achieve the highest accuracy in this dataset. As shown in fig. 5f, GeoF better delineates small structures.

Figure 6 plots the testing accuracy of algorithms (01,08,14 and 16) as a function of the tree depth. Entangled geodesic forests using either of the two energy models (14,16) work better than the conventional forest (01). Using the field-inspired energy (16) works better than the conventional information gain (14). Using two entanglement sections works better than a single one on this data. Our auto-context geodesic forest (08) does well, but the second forest does not seem to yield much additional improvement.

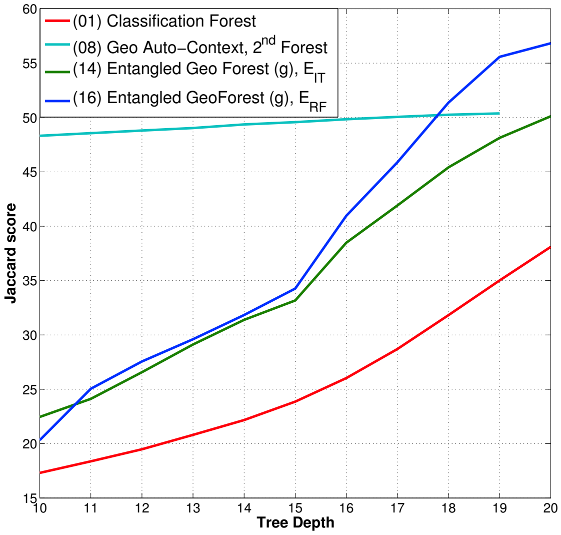

Figure 6. **Accuracy as a function of tree depth** D, for different forest variants, evaluated on the LFW face dataset.

In terms of runtime, the standard forest + CRF (02) takes ∼ 0.71s (per frame) *vs.* ∼ 0.42s for a single-section entangled geodesic forest. Also, forest-based inference is simpler and more easily parallelizable than using graph-cut algorithms for inference on CRFs.

CT scans. Starting with baseline scores of 53.2% (01) and 68.3% (02) we find again that providing entangled geodesic features improves on all our compared methods. The auto-context forest performs well here too, even without these additional features. However, the best results are achieved with one or two sections of entanglement in geodesic forests (`12, 16`). The CRF approach (`02`) takes $\sim$ 1.2s per frame while geodesic forests (`12`) need $\sim$ 0.72s.

KinBG depth images. In this dataset the best results are achieved by our auto-context geodesic forests (`07, 08`) which yield strong improvements over the baseline (+ 6.8% over (`01`), + 3.9% over (`02`)). However, using auto-context forest variants (*e.g.* `03, 07, 08`) results in higher runtimes as two forests need to be evaluated (resulting in $\sim$ 1.39s/frame). The CRF approach (`02`) takes $\sim$ 1.35s per frame while entangled geodesic forests are much faster ($\sim$ 0.64s/frame). In contrast to [23], here we achieve simultaneous body parts *and* background labeling without the need for a preliminary background removal stage.

CamVid videos. For this dataset we have followed the experimental setup described in [13], providing Lab raw channel intensities, first and second order image gradients and HOG-like features. The baseline result for (`01`) is 33.3% which we are able to considerably outperform with all our geodesic forest variants. The best performing geodesic forest (`16`) improves over the recent work in [13] (+2.1%) and [29] (+8.7%). The highest score is obtained by the CRF (`02`) (41.7%), but at the expense of twice the runtime: $\sim$ 1.07s/frame for (`02`) $\sim$ 0.56s/frame for geodesic forests.

Smoother energy models? In further experiments we have tried training forests by adding pairwise terms or other global smoothness terms in the energy (10), but without consistently improving the accuracy further. These results suggest that perhaps our long-range connectivity features already do a sufficient job at capturing spatial smoothness.

Capturing semantic context via entangled geodesic features. Figure 7 illustrates how GeoF captures long-range *semantic* context on the CT dataset. For a reference pixel of a given class (*e.g.* liver) the elements of each matrix indicate the frequency of classes in the two automatically selected probes (probe 1 in the rows and probe 2 in the columns). For example, in Fig. 7a we see that at depth 10 (after one level of entanglement) when the reference pixel is in the liver, the two probes tend to be selected (during training) to also be in the liver. This encourages local context and label smoothing; and can be thought of as a generalization of MRFs where the discriminative cliques are learned automatically, rather than being manually predefined. For deeper trees we start to see the effect of longer-range semantic context. For example, in Fig. 7b the probes tend to be selected frequently also in the heart and right lung regions. This indeed makes sense when the goal is to identify liver pixels. Similar reasoning applies to other classes (*e.g.* see Fig. 7a',b',c' for pixels in the left kidney).

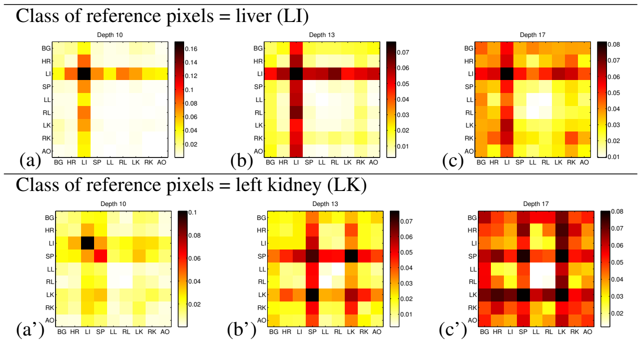

Figure 7. **Class co-occurrence matrices for the feature probe pixels. (a, b, c)** The reference point is in the liver. **(a', b', c')** The reference point is in the left kidney. Co-occurrence matrices are shown for three different tree depths: $D = 10$, $D = 13$, $D = 17$. In this dataset (CT) classes are: background (BG), heart (HR), liver (LI), spleen (SP), l./r. lung(LL/RL), l./r. kidney (LK/RK) and aorta (AO). This figure demonstrates capturing semantic context. *e.g.* in b' when trying to identify the left kidney it helps to use probes either in the spleen region (just above the left kidney) or in the left kidney itself (encouraging local smoothness).

6. Conclusion

This paper has presented a new forest-based model for structured-output learning, applied to the task of semantic image segmentation. Our model encourages spatial smoothness and long-range, semantic context within the forest itself, via the use of new, soft connectivity features which build upon entangled, generalized geodesic distances. In addition, the paper shows how training forests by minimizing a new random field-inspired energy yields higher accuracy than entropy based approaches. Quantitative validation on four diverse image datasets shows at par or better accuracy than state-of-the-art approaches, with faster runtimes.

Acknowledgements. We would like to thank F. Wen for making available the labelled face images to us. We are also very grateful to S. Nowozin, C. Rother, A. Fitzgibbon and J. Jancsary for inspiring and heated conversations.

References

[1] http://vis-www.cs.umass.edu/lfw/.

[2] http://www.pascal-network.org/challenges/voc/.

[3] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7), 1997.

[4] A. Blake, P. Kohli, and C. Rother. *Markov Random Fields for Vision and Image Processing*. The MIT Press, 2011.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1), 2001.

[6] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *Proc. ECCV*. Springer, 2008.

[7] A. Criminisi, T. Sharp, and A. Blake. GeoS: Geodesic image segmentation. In *Proc. ECCV*. Springer, 2008.

[8] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013.

	Image datasets			
	LFW	**CT**	**KinBG**	**CamVid**
Number of training / testing images	1000 / 250	512 / 250	2500 / 250	367 / 233
Accuracy of existing algorithms				
01- Classification forest (pixel-wise classification only, no spatial context, no geodesic features)	38.1	53.2	57.1	33.3
02- Classification forest followed by conditional random field (CRF)	45.2	68.3	60.0	**41.7**
03- Auto-context classification forest (energy: E_{IT}; no geo. features)	48.1	65.9	61.9	35.2
04- Entangled classification forest (energy: E_{IT}; no geo. features; 1 entgl. section at depth=10)	43.2	58.3	55.7	35.5
05- Structured class-labels in random forests [13]	–	–	–	36.2
06- Local label descriptor [29]	–	–	–	29.6
Accuracy of variants of proposed geodesic forests (GeoF)				
(*Two autocontext forests*)				
07- Auto-context geodesic forests (energy: E_{IT}; eval. using raw class posterior p_D as output)	49.8	65.7	**62.4**	35.2
08- Auto-context geodesic forests (energy: E_{IT}; eval. using smoothed class posterior g_D as output)	50.4	69.2	**63.9**	36.6
(*One entanglement section at depth=10*)				
09- Entangled geodesic forests (energy E_{IT}; evaluation using p_D)	46.8	58.6	55.9	36.8
10- Entangled geodesic forests (energy E_{IT}; evaluation using g_D)	46.2	60.2	55.4	35.1
11- Entangled geodesic forests (energy E_{RF}; evaluation using p_D)	54.3	69.1	59.8	34.9
12- Entangled geodesic forests (energy E_{RF}; evaluation using g_D)	54.6	**72.3**	60.0	37.7
(*Two entanglement sections, at depth=10 and depth=15*)				
13- Entangled geodesic forests (energy E_{IT}; evaluation using p_D)	49.5	60.3	56.6	37.9
14- Entangled geodesic forests (energy E_{IT}; evaluation using g_D)	50.1	61.1	56.8	38.0
15- Entangled geodesic forests (energy E_{RF}; evaluation using p_D)	**56.6**	69.9	59.8	36.6
16- Entangled geodesic forests (energy E_{RF}; evaluation using g_D)	**56.8**	**72.2**	60.3	**38.3**

Table 1. **Quantitative validation and comparison.** Average Jaccard accuracy measures (in %, larger values are better) across all classes, for our geodesic forest algorithm as compared to existing techniques (*e.g*. random classification forest, and forest + CRF), for four different labelled image databases. Bold-face numbers indicate the top two algorithms for each image dataset.

[9] B. Fröhlich, E. Rodner, and J. Denzler. As time goes by - anytime semantic segmentation with iterative context forests. In *Proc. DAGM*, 2012.

[10] X. He, R. S. Zemel, and M. A. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Proc. IEEE CVPR*, 2004.

[11] J. Jancsary, S. Nowozin, T. Sharp, and C. Rother. Regression tree fields - an efficient, non-parametric approach to image labeling problems. In *Proc. IEEE CVPR*, 2012.

[12] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. PAMI*, 28(10), 2006.

[13] P. Kontschieder, S. Rota Buló, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. In *Proc. IEEE ICCV*, 2011.

[14] P. Kontschieder, S. Rota Bulò, A. Criminisi, P. Kohli, M. Pelillo, and H. Bischof. Context-sensitive decision forests for object detection. In *Proc. NIPS*, 2012.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[16] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Trans. PAMI*, 2006.

[17] A. Montillo, J. Shotton, J. Winn, J. Iglesias, D. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of CT images. In *Proc. IPMI*, 2011.

[18] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *Proc. ECCV*. Springer, 2010.

[19] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *Proc. IEEE ICCV*, 2011.

[20] N. Payet and S. Todorovic. Hough forest random field for object recognition and segmentation. *IEEE Trans. PAMI*, 2012.

[21] S. Ross, D. Munoz, M. Hebert, and J. A. Bagnell. Learning message-passing inference machines for structured prediction. In *Proc. IEEE CVPR*, 2011.

[22] T. Sharp. Implementing decision trees and forests on a GPU. In *Proc. ECCV*. Springer, 2008.

[23] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Trans. PAMI*, 2013.

[24] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proc. IEEE CVPR*, 2008.

[25] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. *TextonBoost* for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1), 2009.

[26] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Trans. PAMI*, 32(10), 2010.

[27] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *Proc. IEEE CVPR*, 2006.

[28] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5, 1992.

[29] Y. Yang, Z. Li, L. Zhang, C. Murphy, J. Hoeve, and H. Jiang. Local label descriptor for example based semantic image labeling. In *Proc. ECCV*. Springer, 2012.