GeoF: Geodesic Forests for Learning Coupled Predictors

†P. Kontschieder †P. Kohli
‡Graz University of Technology, Austria

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 T exton F orest, Auto-Context, and Entangled F orest models. A second contribution is to show the connection between the typical Conditional Random F ield (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,
medical 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 textron 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 \(J : \Omega \subset \mathbb{N}^2 \to \mathbb{R}\), and a 2D pixel position is denoted \(p \in \Omega\). We cast the semantic segmentation task as that of associating each pixel \(p\) with its corresponding discrete class label \(c \in C\). 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 \(p\) is denoted \(v(p) = (v_1, \ldots, v_l) \in \mathbb{R}^l\). A set of training data points (and associated labels) is denoted \(S = \{z_i\}\), with each training point-class pair being \(z = (v, c)\). Let \(c = \{c_p | p \in \Omega\}\) denote the vector of class variables predicted by our classifier on the entire image. We use \(c_p\) 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:

\[
c^* = \arg \max_c p(c|J) = \arg \max_c p(J|c)p(c) \tag{1}
\]

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

\[
p(c|J) = \prod_p \psi(c_p, v(p)) \prod_{(p,q) \in \mathcal{N}} \phi(c_p, c_q, v(p), v(q)) \tag{2}
\]

\(^1\)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 \( p \) and \( q \). From (3) we can see that the MAP labels \( c_p^* \) and \( c_q^* \) are functions of the input features responses \( \nu(p) \) and \( \nu(q) \) i.e. \( c_p^* = f(\nu(p)) \) and \( c_q^* = f(\nu(q)) \). 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 \( p_3 \) and \( p_2 \) 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 \( p_2 \) and \( p_2 \), despite being far from each other in Euclidean terms, 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.

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

\[
Q(p; M, \nabla J) = \min_{p' \in \Omega} (\delta(p, p') + \nu M(p'))
\]

with the geodesic distance between two points \( p \) and \( q \):

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

where \( \Gamma \) is a path connecting the two points and \( \mathcal{P}_{p,q} \) 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-
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: $\sum_c g_{si}(c|v) = 1$. $Q(\cdot)$ 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 $x$ 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.

$$v_i^d(x) = F_k^d(p_1) + F_k^d(p_2),$$

$$v_i^d(x) = F_k^d(p_1) - F_k^d(p_2),$$

$$v_i^d(x) = |F_k^d(p_1) - F_k^d(p_2)|$$

where $k \in \{0, 1, 2\}$ denotes the channel where features are computed, and: (i) $F_0^d(p) = J(p)$, i.e. the raw image intensities, (ii) $F_1^d(p) = p_{s(d)}(c|(v))$, i.e. the intermediate class posteriors computed in the section $s(d)$ defined by the depth $d$, and (iii) $F_2^d(p) = g_{s(d)}(c|(p))$, i.e. the geodesic-filtered posteriors, capturing connectivity of point $p$ 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 $\theta_j = \arg\min_{\theta_j} E(S_j, \theta_j)$. The traditional choice for the

3Here 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 \( E_{IT} \), which after some algebraic manipulation reduces to

\[
E_{IT}(S, \theta) = - \sum_{i \in \{1, \ldots, L\}} \sum_{c \in C} n(c, S_i) \log \frac{n(c, S_i)}{|S_i|}
\]

with \( n(c, S) \) denoting the number of training pixels of class \( c \) in the training subset \( S \) (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 \( E = E_{gb} \), which is typically defined as

\[
E_{gb}(S, \theta) = \sum_{i \in \{1, \ldots, L\}} \sum_{z \in Z_i} \psi(z; S_i) + \lambda \sum_{z \in Z_i, r \in N(z_i)} \phi(z, r)
\]

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

\[
E_{gb}(S, \theta) = - \sum_{i \in \{1, \ldots, L\}} \sum_{c \in C} n(c, S_i) \log \frac{n(c, S_i)}{|S_i|}. \tag{8}
\]

So, we discover that under the above assumptions, (8) and (7) are identical. Thus, conventional entangled 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 \( S_0 \) we have \( n(c_1, S_0) \neq n(c_2, S_0) \), \( c_1, c_2 \in C \). 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: \( \omega_c = \frac{\sum_{k \in C} n(k, S_0)}{n(c_0, S_0)} \) and the corresponding node-based normalization factor \( Z(S_j) = \sum_{k \in C} \omega_k n(k, S_j) \).

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

\[
E_{IT}(S_j, \theta) = - \sum_{i \in \{1, \ldots, L\}} \sum_{c \in C} w_c \ n(c, S_i) \log \frac{w_c n(c, S_i)}{Z(S_i)}. \tag{9}
\]

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

\[
E_{gb}(S_j, \theta) = - \sum_{i \in \{1, \ldots, L\}} Z(S_i) \sum_{c \in C} w_c n(c, S_i) \log \frac{w_c n(c, S_i)}{Z(S_i)}. \tag{10}
\]

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 negate 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.

**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 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-sensitive Potts model. Thus, 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.
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.

In terms of runtime, the standard forest + CRF (02) takes $\sim$ 0.71s (per frame) vs. $\sim$ 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 ~1.2s per frame while geodesic forests (12) need ~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 ~1.39s/frame). The CRF approach (02) takes ~1.35s per frame while entangled geodesic forests are much faster (~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: ~1.07s/frame for (02) ~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).

**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**

<table>
<thead>
<tr>
<th>Number of training / testing images</th>
<th>LFW</th>
<th>Image datasets</th>
<th>CT</th>
<th>KinBG</th>
<th>CamVid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td>512</td>
<td>1250</td>
<td>2500</td>
</tr>
</tbody>
</table>

**Accuracy of existing algorithms**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3 - Classification forest (pixel-wise classification only, no spatial context, no geodesic features)</td>
<td>38.1</td>
<td>53.2</td>
<td>57.1</td>
<td>33.3</td>
<td></td>
</tr>
<tr>
<td>0.2 - Classification forest followed by conditional random field (CRF)</td>
<td>45.2</td>
<td>68.3</td>
<td>60.0</td>
<td>41.7</td>
<td></td>
</tr>
<tr>
<td>0.3 - Auto-context classification forest (energy: $E_{IT}$; no geo. features)</td>
<td>48.1</td>
<td>65.9</td>
<td>61.9</td>
<td>35.2</td>
<td></td>
</tr>
<tr>
<td>0.4 - Entangled classification forest (energy: $E_{IT}$; no geo. features; 1 entgl. section at depth=10)</td>
<td>43.2</td>
<td>58.3</td>
<td>55.7</td>
<td>35.5</td>
<td></td>
</tr>
<tr>
<td>0.5 - Structured class-labels in random forests [13]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>36.2</td>
</tr>
<tr>
<td>0.6 - Local label descriptor [29]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>29.6</td>
</tr>
</tbody>
</table>

**Accuracy of variants of proposed geodesic forests (GeoF)**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Two auto-context forests)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7 - Auto-context geodesic forests (energy $E_{IT}$; eval. using raw class posterior $p_D$ as output)</td>
<td>49.8</td>
<td>65.7</td>
<td>62.4</td>
<td>35.2</td>
<td></td>
</tr>
<tr>
<td>0.8 - Auto-context geodesic forests (energy $E_{IT}$; eval. using smoothed class posterior $g_D$ as output)</td>
<td>50.4</td>
<td>69.2</td>
<td>63.9</td>
<td>36.6</td>
<td></td>
</tr>
<tr>
<td>(One entanglement section at depth=10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9 - Entangled geodesic forests (energy $E_{IT}$; evaluation using $p_D$)</td>
<td>46.8</td>
<td>58.6</td>
<td>55.9</td>
<td>36.8</td>
<td></td>
</tr>
<tr>
<td>1.0 - Entangled geodesic forests (energy $E_{IT}$; evaluation using $g_D$)</td>
<td>46.2</td>
<td>60.2</td>
<td>55.4</td>
<td>35.1</td>
<td></td>
</tr>
<tr>
<td>1.1 - Entangled geodesic forests (energy $E_{RF}$; evaluation using $p_D$)</td>
<td>54.3</td>
<td>69.1</td>
<td>59.8</td>
<td>34.9</td>
<td></td>
</tr>
<tr>
<td>1.2 - Entangled geodesic forests (energy $E_{RF}$; evaluation using $g_D$)</td>
<td>54.6</td>
<td>72.3</td>
<td>60.0</td>
<td>37.7</td>
<td></td>
</tr>
<tr>
<td>(Two entanglement sections, at depth=10 and depth=15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3 - Entangled geodesic forests (energy $E_{IT}$; evaluation using $p_D$)</td>
<td>49.5</td>
<td>60.3</td>
<td>56.6</td>
<td>37.9</td>
<td></td>
</tr>
<tr>
<td>1.4 - Entangled geodesic forests (energy $E_{IT}$; evaluation using $g_D$)</td>
<td>50.1</td>
<td>61.1</td>
<td>56.8</td>
<td>38.0</td>
<td></td>
</tr>
<tr>
<td>1.5 - Entangled geodesic forests (energy $E_{RF}$; evaluation using $p_D$)</td>
<td>56.6</td>
<td>69.9</td>
<td>59.8</td>
<td>36.6</td>
<td></td>
</tr>
<tr>
<td>1.6 - Entangled geodesic forests (energy $E_{RF}$; evaluation using $g_D$)</td>
<td>56.8</td>
<td>72.2</td>
<td>60.3</td>
<td>38.3</td>
<td></td>
</tr>
</tbody>
</table>

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.