Fully-Connected CRFs with Non-Parametric Pairwise Potentials

Neill D. F. Campbell
University College London

Kartic Subr
University College London

Jan Kautz
University College London

Abstract

Conditional Random Fields (CRFs) are used for diverse tasks, ranging from image denoising to object recognition. For images, they are commonly defined as a graph with nodes corresponding to individual pixels and pairwise links that connect nodes to their immediate neighbors. Recent work has shown that fully-connected CRFs, where each node is connected to every other node, can be solved efficiently under the restriction that the pairwise term is a Gaussian kernel over a Euclidean feature space. In this paper, we generalize the pairwise terms to a non-linear dissimilarity measure that is not required to be a distance metric. To this end, we propose a density estimation technique to derive conditional pairwise potentials in a non-parametric manner. We then use an efficient embedding technique to estimate an approximate Euclidean feature space for these potentials, in which the pairwise term can still be expressed as a Gaussian kernel. We demonstrate that the use of non-parametric models for the pairwise interactions, conditioned on the input data, greatly increases expressive power whilst maintaining efficient inference.

1. Introduction

The discrete label Markov Random Field (MRF) and Conditional Random Field (CRF) are common models used throughout Computer Vision [15], in particular for low level vision tasks: *e.g.* image denoising, optical flow, binocular stereo, segmentation, etc. These models are often solved as a discrete energy minimization task over a graph containing nodes corresponding to individual pixels. The basic model consists of the combination of a set of unary terms, defined for each node individually, and a set of pairwise terms, defined as a function of two nodes that share an edge. A first order Markov assumption is often used where each node shares an edge only with its immediate neighbors, *e.g.* the graph may be a 2D grid over the pixels with each node connected to its four (or eight) nearest neighbors. This is usually to keep the inference tractable in either computational complexity or memory requirements.

The pairwise interactions impose a smoothness cost on the final labeling. The move from MRF models to CRF models [12] allows these terms to be conditioned on the input data and thus the terms become dependent on the structure of the image. Whilst good results have been obtained using only neighboring pairwise terms, they may only be used to express a limited range of priors. The need to learn richer and more expressive prior models from training data has lead to a demand to solve models that contain higher-order cliques (potential functions of more than two nodes) or those which are able to capture the interplay between nodes that are spaced further apart — non-local pairwise interactions.

In this paper we consider the latter. We investigate the addition of *non-local pairwise potentials*. This corresponds to increasing the connectivity of the graph by adding edges between nodes that are not immediate neighbors. During inference, increasing the number of edges in the graph often leads to a dramatic scaling in computational resources, both for algorithms based on graph-cut, move making techniques and message passing methods, *e.g.* [3, 7, 8].

To overcome this limitation, recent work has produced a number of approximate inference techniques making use of cross bilateral filtering. In particular, the work of Krähenbühl and Koltun [10] proposed a method for performing inference in a fully-connected pairwise CRF (every node is connected to every other node) by taking a mean-field approximation to the original CRF. Here, the message passing is performed as a Gaussian bilateral filtering process under the limitation that the pairwise potentials be expressed as a weighted sum of Gaussian kernels over a Euclidean feature space. This allows approximate maximum posterior marginal (MPM) inference to be performed very efficiently for a multi-label CRF.

The method of [10] directly addresses the issue of increasing graph connectivity since it allows for a fully-connected CRF. However, thus far, the applications have been limited by the requirement that the pairwise terms consist of a weighted sum of Gaussian kernels over a Euclidean feature space. The work of Vineet *et al.* [16] demonstrated that the pairwise terms may be extended to include non-zero mean mixtures of Gaussians, along with an estimation procedure to fit the model parameters, at the expense of a num-

ber of extra filtering operations (one per Gaussian mixture) at each iteration that slows down the inference procedure.

In this work we generalize the pairwise potential from a simple parametric model to a conditional non-parametric model that is learnt from training data. Our learning approach is to approximate directly the conditional joint probability distributions (from the training data) in a straight forward density estimation process. This probability model may be expressed as an image specific (evaluated at test time), sparsely sampled dissimilarity measure. We then use an efficient embedding technique to estimate a Euclidean feature space that approximates this measure. The pairwise terms may then be expressed as Gaussian kernels in this new feature space and thus the inference procedure of [10] may proceed unaltered. This allows us to generalize the pairwise terms to a general, non-linear *dissimilarity measure that is not required to be a distance metric*. In particular we show that the use of non-parametric models for the pairwise interactions greatly increases the expressive power whilst maintaining the efficient inference of [10].

2. Previous Work

As discussed in § 1, our work makes use of the efficient mean-field inference method of Krähenbühl and Koltun [10] and is thus related to other inference methods based on bilateral filtering including work on image denoising [9] and other low-level vision tasks such as stereo and optical flow [5, 11] and semantic object segmentation [16]. In particular there has been some work on approximating more complex pairwise terms with [16] learning the parameters of a non-zero mean Gaussian mixture model in the bilateral space and [11] approximating a truncated penalty function as a mixture of exponentials. In this work we generalize further by approximating an arbitrary dissimilarity measure which can be non-parametric and conditioned on each specific test image, as well as training data, by finding an embedding into a Euclidean feature space that best approximates the dissimilarities and automatically minimizes the dimensionality of the embedded space to match the complexity of the provided dissimilarities.

The work of [16] also addressed the issue of initialization when performing inference on a CRF under a mean-field approximation. Whilst this is not a topic we address in this work, the insight and suggestions are equally valid for our method. This topic was also looked at in [18]. The subsequent work of [17] provides a method for extending the filter based inference algorithm for models that include potentials defined over certain types of higher-order cliques. Again, this extension is not discussed in this work but the findings are equally applicable and could be used with the feature spaces presented here.

Recent work has investigated extensions to pairwise CRFs under alternative inference methods, in particular the works of Nowozin *et al*. [13] and Jancsary *et al*. [6] are state-of-the-art decision tree based algorithms with tractable training and inference, especially efficient in the case of [6]. Our approach shares the two key desirable properties of these works. Firstly, we overcome the limitation of a fixed neighborhood structure with the fully-connected model and, secondly, we remove the requirement for the pairwise terms to have a simple parametric form by allowing arbitrary non-parametric dissimilarities to encode the pairwise potentials that may be learnt from training data and also the dissimilarities can be conditioned on the input data at test time. We demonstrate that our approach confers a competitive performance with these approaches both in terms of accuracy and computational efficiency. We would refer the reader to the references contained in [6, 13] for further details of research into parameter estimation in CRF models with parametric pairwise terms.

The work of [14] proposes a scribble-based method for selecting objects in images based on dense CRFs [10]. Two standard non-Euclidean distance metrics over patches are used (χ^2 and Earth Mover's Distance) and an embedding into a Euclidean feature space is employed to incorporate them into the dense CRF framework. In contrast, we propose to generalize away from a data-driven heuristic dissimilarity measure, rather incorporating non-parametric dissimilarities, learnt from training data.

3. Efficient Mean-Field Inference in Fully-Connected Pairwise CRFs

In the recent work of [10], Krähenbühl and Koltun described an efficient algorithm to perform inference on a fully-connected CRF in linear time (in the number of nodes) by using a mean-field approximation to the original CRF and pairwise edges with potential functions defined as Gaussian kernels in some feature space. Let us denote the set of labels as $\mathbf{x} = \{x_i\}$ with a label defined for every pixel in the set of pixels $\mathcal{P}$, such that $i \in \mathcal{P}$, in a given image $\mathcal{I}$. Each label is taken from a label space $\mathcal{L}$ such that $x_i \in \mathcal{L}$. If we denote the exact CRF distribution as $P(\mathbf{x} \,|\, \mathcal{I})$ then the mean-field approximation is given as the distribution $Q(\mathbf{x})$ that minimizes the KL-divergence $\mathrm{KL}(Q \,||\, P)$ with the constraint that the distribution Q must be decomposed as the product of a set of independent marginals $Q(\mathbf{x}) = \prod_i Q_i(x_i)$. The Gibbs energy for this model is given as

$$\mathrm{E}\left(\mathbf{x} \,|\, \mathcal{I}\right) = \sum_{i \in \mathcal{P}} \psi_i\left(x_i\right) + \sum_{\substack{i,j \in \mathcal{P} \\ i \neq j}} \phi_{ij}\left(x_i, x_j\right) \qquad (1)$$

where we have $P(\mathbf{x} \,|\, \mathcal{I}) = \frac{1}{Z(\mathcal{I})} \exp\left(-\mathrm{E}\left(\mathbf{x} \,|\, \mathcal{I}\right)\right)$.

In [10] the authors demonstrate that the distribution Q can be recovered by an iterative update equation that corresponds to a message passing algorithm on the graph. The

number of edges in a fully-connected CRF dictates that traditional message passing algorithms would be intractable in computational time and resources. However, if the pairwise terms in the Gibbs energy are expressed as

$$\phi_{ij}\left(x_i, x_j\right) = \mu(x_i, x_j) \sum_{m=1}^{M} w_m \, \mathrm{k}_m\left(\mathbf{f}_i^{(m)}, \mathbf{f}_j^{(m)}\right) \quad (2)$$

where $\mu(\cdot,\cdot)$ is a constant symmetric label compatibility function and

$$\mathrm{k}_m\left(\mathbf{f}_i^{(m)}, \mathbf{f}_j^{(m)}\right) = \exp\left(-\frac{1}{2}\left[\mathbf{f}_i^{(m)} - \mathbf{f}_j^{(m)}\right]^T \Lambda_m \left[\mathbf{f}_i^{(m)} - \mathbf{f}_j^{(m)}\right]\right) \quad (3)$$

is a Gaussian kernel with precision Λ_m in some feature space $\mathbf{f}_i^{(m)} \in \mathcal{F}^{(m)}$, for the m^{th} kernel, then the message passing step consists of a low pass filtering operation under a Gaussian kernel for which efficient approximations exist, *e.g.* [1]. This allows each update iteration to be completed in linear time with respect to the number of nodes rather than quadratic time which would be required for traditional message passing. We refer the reader to [10] for further details. Throughout this paper we use the Potts model[1] $\mu(x_i, x_j) = [x_i \neq x_j]$ for the compatibility function, M denotes the number of kernels used, and we use $\Lambda_m = I$, the identity matrix, since the feature space can always be transformed under an arbitrary covariance.

Updating all the messages in a single step removes the convergence guarantees that are normally associate with mean-field approximations. However, the authors of [10] observe good convergence properties experimentally and we found convergence would usually occur in fewer than 20 iterations. After running the algorithm for a fixed number of iterations, to get into a stable fixed state of the mean field, we extract the Maximum Posterior Marginal (MPM) solution by selecting the label that maximizes the associated factor $x_i = \arg\max_{l \in \mathcal{L}} Q_i(x_i = l)$. We also note that the fixed point of the mean field update equations is dependent on initialization and not a globally optimal solution.

4. Non-Parametric Pairwise Potentials

In order to allow for more expressive pairwise potentials we would like to relax the restriction on Gaussian parametric models, [10, 11, 16] and allow for more complex, non-parametric models that may be learnt from training data and conditioned on the input.

In this section we describe how we overcome the limitation that the pairwise potentials be expresses as a Gaussian kernel, as in (3). We do this in three stages. Firstly, we present our desired pairwise potentials as density estimates of the conditional pairwise probability (learnt from training data, conditioned on a test image). We then express these probabilities as a dissimilarity measure between nodes in the CRF. Finally, we use an efficient approximate embedding technique to find a set of feature spaces that encode the dissimilarity measure as the Euclidean distance and thus the desired pairwise potential under a Gaussian kernel in this space.

4.1. Pairwise Conditional Probabilities

The pairwise potentials in a CRF encode conditional probabilities between pairs of nodes. Our approach is to estimate these probabilities directly from a set of training data $\mathcal{T}$. We make this conditional for each node at test time by first looking at the local area (an image patch $\mathbf{s}_i$) around a particular node i in the test image $\mathcal{I}$ and then finding similar patches in the training images. For each label l in the label space $\mathcal{L}$, we want to estimate the conditional probability $P\left(x_j = l \mid x_i = l, \mathcal{I}, \mathcal{T}\right)$ for the nodes j around node i, *i.e.* the conditional local density distribution of the label l.

Density Estimation: Any density estimation or regression technique could be used to approximate these conditional probabilities; in particular, we make use of a non-parametric approach by referring to the training data directly and performing a kernel based density estimate. We take the mean of the indicator images for the label l from the training images that contain a patch similar to $\mathbf{s}_i$. By indicator image we mean a binary image equal to one for every pixel belonging to class l and zero elsewhere. In practice this corresponds to extracting much larger patches from the training indicator images for class l, that are centered on patches similar to $\mathbf{s}_i$, and finding the mean.

We place a prior that dictates the range over which we are able to infer useful information in the pairwise potential by applying a Gaussian window of size σ_w in pixel distance

$$\mathrm{g}_{\text{win}}(i,j) = \exp\left(-\| \mathbf{u}_i - \mathbf{u}_j \|_2^2 \,/\, \left(2\,\sigma_w^2\right)\right) \quad , \quad (4)$$

where $\mathbf{u}_i$ and $\mathbf{u}_j$ are the pixel coordinates of nodes i and j. This is equivalent to using a Gaussian kernel to perform the density estimation.

This procedure identifies local correlations in the training data that will then be encouraged to occur in the output by means of the pairwise potentials. For example, if a particular image patch always has label l above it in the training data then the indicator images will always be set to one above this patch. Thus, the mean of all the training indicator images, centered on the patch, will be close to one. This indicates that the pairwise term should have strong connections to the pixels above for class l.

[1] Here we use $[x_i \neq x_j]$ as an inequality indicator function.

4.2. Probabilities to Feature Spaces

We now have a method for determining the non-local pairwise potentials around node i for image $\mathcal{I}$. In order to be able to use these potentials to perform the efficient inference discussed in § 3 we must be able to express them in the form of (2); more specifically, the Gaussian kernel of (3). This corresponds to finding a set of feature vectors (*i.e.* an embedding in a feature space) where the distance between the feature vectors of each node under the Gaussian kernel is equal to the conditional probability densities. We can achieve this, in a similar fashion to [14], by creating an appropriate dissimilarity measure, based on the conditional probabilities, and finding an embedding such that the Euclidean distance in the embedded space matches this dissimilarity measure.

Dissimilarity Measure: If we denote the dissimilarity measure as $\mathrm{d}\left(i,j,\mathcal{I},\mathcal{T}\right)$ then we may express our pairwise term as having the form

$$\phi_{ij}\left(x_i,x_j\right) = \left[x_i \neq x_j\right] \exp\left(-\mathrm{d}\left(i,j,\mathcal{I},\mathcal{T}\right)\right) \ . \tag{5}$$

Let us consider the training data for a single label $l \in \mathcal{L}$. We let

$$\exp\left(-\mathrm{d}_l\left(i,j,\mathcal{I},\mathcal{T}\right)\right) = \mathrm{g}_{\text{win}}(i,j)\, P\left(x_j = l \mid x_i = l,\mathcal{I},\mathcal{T}\right) \tag{6}$$

$$\Rightarrow\ \mathrm{d}_l\left(i,j,\mathcal{I},\mathcal{T}\right) = -\log\left[\mathrm{g}_{\text{win}}(i,j)\, P\left(x_j = l \mid x_i = l,\mathcal{I},\mathcal{T}\right)\right] \tag{7}$$

where this distance between landmark location i and varying j, under label l, is the conditional distribution of the label l given the training data $\mathcal{T}$ the test image $\mathcal{I}$.

Feature Space Embedding: The dissimilarity measures obtained for each label may now be embedding into a feature space $\mathcal{F}^{(l)}$ to provide a of feature vector $\{\mathbf{f}_i^{(l)}\}$ for each node i and label l such that

$$\left\| \mathbf{f}_i^{(l)} - \mathbf{f}_j^{(l)} \right\|_2^2 \approx \mathrm{d}_l\left(i,j,\mathcal{I},\mathcal{T}\right) \ \ \forall\ i,j \ . \tag{8}$$

The set of embedded vectors may then be used as a feature space in (3) to perform inference using the filtering approach. Thus we have generalized the constraints on the pairwise potentials to the requirement that they be expressed as (5) where the functions $\mathrm{d}_l\left(i,j,\mathcal{I},\mathcal{T}\right)$ are dissimilarity measures which must satisfy $\mathrm{d}_l\left(i,i,\mathcal{I},\mathcal{T}\right) = 0$. Whilst it is not a strict requirement that $\mathrm{d}_l(\cdot,\cdot)$ be a distance function, we note that when the distance function is embedded in the form (5) the resulting approximate distance will be symmetric and therefore a symmetric distance will always be used to perform inference.

4.3. Approximate Euclidean Embedding

We make use of the Landmark version of the Multidimensional Scaling (MDS) algorithm [4] to compute the feature vectors $\{\mathbf{f}_i^{(l)}\}$ from the dissimilarity measures as an embedding in p-dimensional Euclidean space $\mathcal{R}^p$.

The landmark variant (LMDS) has the advantage over classical MDS of removing the need to store a complete pairwise dissimilarity matrix $D_{ij}^{(l)} = \mathrm{d}_l\left(i,j,\mathcal{I},\mathcal{T}\right)$ that would have a storage complexity of $O(N^2)$, where $N = |\mathcal{P}|$ is the number of pixels in the test image $\mathcal{I}$. Instead, the Nyström approximation of $D_{ij}^{(l)}$ is used and allows us, under reasonable sampling conditions, to provide only a subset of the dissimilarity matrix.

For a p-dimensional embedding, LMDS required the complete set of dissimilarities between $p+1$ points, known as the landmarks. In practice, due to potential degeneracies, the number of landmark points needs to be $c > p+1$ to ensure that they span the p-dimensional space. The remaining points have their positions triangulated from these landmark points, requiring the dissimilarities between the landmarks and the other points.

The required dimensionality of the space (p) can be determined by analysis of the the eigenvalues computed during the LMDS embedding. Further details of MDS, LMDS and the eigenvalues are provided in the supplemental material.

Random Sampling: The use of LMDS means that we don't have to estimate $\mathrm{d}_l\left(i,j,\mathcal{I},\mathcal{T}\right)$, and hence $P\left(x_j = l \mid x_i = l,\mathcal{I},\mathcal{T}\right)$, for all nodes i. Instead we pick sampling locations $\mathcal{C}$ ($c = |\mathcal{C}|$ points uniformly distributed over the test image) and estimate $P\left(x_j = l \mid x_i = l,\mathcal{I},\mathcal{T}\right)$, and the corresponding dissimilarities, for $i \in \mathcal{C}$ and all j.

Illustration: Figure 1 provides an illustrative example from the case of a binary label set $|\mathcal{L}| = 2$, with the labels as foreground text and background, used on the task of shape completion in our experiments in § 5. We are provided with a binary shape pattern and an occlusion mask and wish to infer the labels of the occluded pixels as foreground or background. In addition to the masked test image we are also provided with a training database of images containing similar statistical properties to our test data.

Consider a single sample location for the foreground label (orange); we want to compute the dissimilarity to all other pixels. We look at the local patch (conditional region) around the pixel, and find all the patches in our training database that match with a low hamming distance. The dissimilarities to the pixels in the wider region, determined by the window size σ_w, around our input patch should have the same label distribution as the regions around the training patches. Therefore, we take the mean of the set of larger

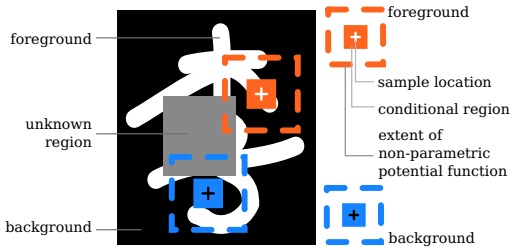


Figure 1: Random sampling to build the distance matrix. The local neighborhood of a random sample is used to condition a lookup into the training data to provide a non-parametric estimate of the potential dissimilarity measure to the wider image.

patches from the training indicator images, for the appropriate label, as our conditional probability. We then multiply by the Gaussian window function and take the negative log to obtain the required dissimilarity (7).

This process is then repeated until sufficient samples have been gathered to proceed with the embedding using LMDS. We then perform the same operation for the background label, for example taking the sample shown in blue in Fig. 1. This provides us with a set of feature spaces $\mathcal{F}^{(l)}$ which we use with a Potts model in (2), with $M = |\mathcal{L}|$, to filter each label in its own feature space. Figure 3 shows some actual examples of the raw probability estimates from one of our experiments.

Final Model: Our final model is given by

$$\mathrm{E}\left(\mathbf{x} \,|\, \mathcal{I}\right) = \sum_{i \in \mathcal{P}} \psi_i\left(x_i\right) + \beta - w \sum_{\substack{i,j \in \mathcal{P} \\ i \neq j}} \sum_{l \in \mathcal{L}} \left[\begin{array}{c} x_i = l, \\ x_j = l \end{array}\right] \exp\left(-\frac{\left\| \mathbf{f}_i^{(l)} - \mathbf{f}_j^{(l)} \right\|_2^2}{2\,\sigma_f^2}\right), \quad (9)$$

where we have used $[x_i \neq x_j] = 1 - [x_i = x_j]$ and β is a constant that may be neglected. Please see the supplementary material for further algorithmic details.

5. Experiments

Since our contribution is in the use of pairwise potentials, a direct and unbiased evaluation of our work is best obtained by removing the dependence of the results on any unary terms in the CRF. We demonstrate the effectiveness of our approach by performing in-painting on binary images. Here, we have a task where no unary is applicable in the occluded region and we must use expressive pairwise potentials to learn the wider neighborhood statistics that encode the shape distributions. The task we perform was proposed by Nowozin *et al*. [13] and used in [6]; we follow the procedure the authors describe within these papers and the supplementary materials. We used 3×3 pixel patches as the conditional region for all tests.

Method	Accuracy
Random Forest [13]	67.74 %
MRF (1 level DTF) [13]	75.18 %
Gaussian MRF (1 level RTF) [6]	74.19 %
Decision Tree Field [13]	76.01 %
Regression Tree Field [6]	77.55 %
Our Result ($\sigma_w = 7$ pixels)	82.04 %

Table 1: Quantitative comparison of test results for the KAIST Hanja2 database with small occlusions. We provide results for the accuracy (as the percentage of pixels correctly labeled) for filling in the masked regions on unseen test images after training on a separate training set. We adopt the same methodology as [13, 6] splitting the input data in a 2:1 training to test ratio with the dimensions of the masked regions drawn from $[5 \ldots 20]$ pixels.

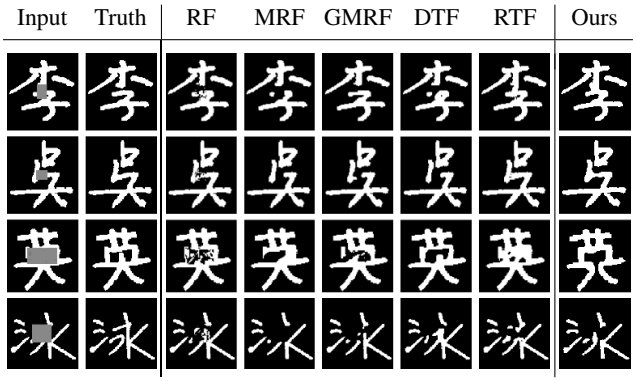


Figure 2: A qualitative comparison for the KAIST Hanja2 database with large occlusions. We fill in the grey region from the first column using the following algorithms: (RF) A baseline Random Forest. (MRF) A DTF with a depth of 1 (local neighbors). (GMRF) A Gaussian MRF, an RTF with a depth of 1. (DTF) The Decision Tree Field [13]. (RTF) Regression Tree Field [6]. (Ours) The result of our algorithm. The results in the central columns are taken from [13] and [6].

The KAIST Hanja2 Database: In this experiment we make use of the KAIST Hanja2 database: a collection of handwritten Chinese characters. The dataset displays a rich degree of shapes and variation with some characters repeated often and others with only single examples. We randomly split the database into 300 images used for training and 150 for testing. We occlude a centered rectangular region of each of the test images in two different tasks, the locations of the occlusions are obtained as detailed in [13]. The first task considers small occlusions with the mask dimensions drawn uniformly from the range $[5 \ldots 20]$ pixels, and, the second task considers large occlusions with dimensions from $[20 \ldots 40]$ pixels. The unary term is clamped to the ground truth outside the occluded region and to an uninformative uniform distribution within the occluded region.

Table 1 gives the quantitative results for the small oc-

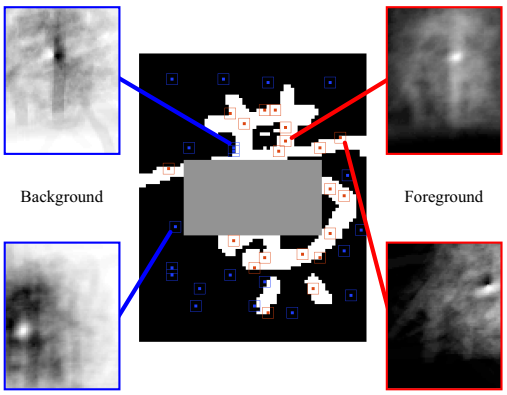


Figure 3: A sample of the dissimilarity measures used for a KAIST Hanja2 example. We show some random sample locations for foreground (red) and background (blue) 3×3 patches used as landmarks for the embedding. For each of the landmark patches we find similar patches in the training data and then estimate the density of the appropriate class (foreground or background) centered on the patch. Two samples of the density estimates are shown for each class; the colormap is black to white with increasing density. The raw probability values are shown; we apply a Gaussian window and take the negative log to obtain the dissimilarity samples for embedding.

clusion task (evaluated as the percentage of pixels correctly labeled). It provides comparisons with the Decision Tree Field [13] and Regression Tree Field [6] methods, both of which are considered state-of-the-art, and shows that our method confers a favorable performance. We also note the marked improvement of all the methods making use of increased neighborhood ranges in their potential functions; the low connectivity of the RF, MRF and and GMRF methods is indicative of this short coming.

Figure 2 provides the qualitative output for the large occlusion test cases, again showing comparisons to the state-of-the-art methods. Without any higher level inference (*i.e.* attempting to classify the characters) it is a very challenging problem to correctly recover the original character. Instead, filling in plausible structure is indicative of good performance showing that the model has captured the underlying statistics of the training data and exploited the conditional dependence on the input. We believe that our results are reasonable for the nature of the characters even though they may not accurately reconstruct the ground truth.

In Fig. 3 we provide an example of the conditional, non-parametric pairwise potentials used for the KAIST Hanja2 database. The dissimilarity measures for the foreground and background classes are observed to vary based upon the local region around the sampling locations and we can see the structure, learnt from the training data, that is promoted by the potentials.

The Weizmann Horse Database [2]: In this dataset we perform the same task but using silhouettes of horses from

Figure 4: A sample of test results for completing silhouettes from the Weizmann horse database [2]. We used 219 training images and 109 test images selected at random. We occluded with the 'large occlusion box' parameters (from the Hanja2 evaluation [13]) with the dimensions of the masked regions drawn from $[20 \ldots 40]$ pixels. Methods [10] and [14] require the color images which are not used by our method.

the Weizmann horse database [2]. We use the large occlusion parameters from the KAIST test and allow the occlusion box to move around in the test images. Figure 4 shows a random selection of results ranked in decreasing accuracy from top to bottom. Table 2 details the overall accuracy.

We compared our non-parametric model to the cross bilateral model of [10] and an input-agnostic dissimilarity used by [14]. The results obtained for [10] and [14] require the original color image in order to calculate the pairwise terms; our method makes no use of the color images during training or testing. These value is included for comparative purposes but the specific task is different since it is no longer simply binary inpainting, rather guided inpainting. The goal of [14] is robustness to inaccurate training. Consequently, the resulting images for shape completion, Fig. 4, do not consider the truth data to be reliable outside the mask region, however, we compute the accuracy only within the masked region.

Our result is shown to afford comparable accuracy to the

Method	Accuracy
Potts Model	60.17 %
Cross Bilateral (Parametric) Model [10]	*84.10 % **
χ^2 Patch Distance Model [14]	*89.87 % **
Our Result	89.78 %

Table 2: Quantitative comparison for the Weizmann Horse dataset. The Potts model provides a baseline as a generic smooth result. Both methods in italics (marked with an asterisk), the cross bilateral model (gaussian kernel in color space) of [10] and the χ^2 patch dissimilarity model of [14], needed the original color image to evaluate the pairwise terms over the occluded regions. This color image was not provided to our method during training or testing. For all methods a window size of $\sigma_w = 13$ pixels was used, the additional parameters for [10] were set to the values specified by the authors.

methods of [10] and [14] without the need for the specific color image to guide the CRF, making use of the shape training data instead. The Potts model serves as a baseline for a simple smoothness prior. All results were obtained with a window size of $\sigma_w = 13$ pixels; the increase over the $\sigma_w = 7$ pixels for the Chinese characters is indicative of the differing scales of the foreground objects.

6. Discussion

Timing: We compute the embedding using 80 landmark samples and 10 dimensions in around 1s. As in previous work, the inference is efficient, with 20 iterations in a second. Both of these timings are on the horse examples. The KAIST tests are slightly quicker. This could be improved with parallel implementations, in particular the GPU may help with filtering. The data lookup for the non-parametric, conditional potentials is less predictable. In our examples, the training data could be held in memory and accessed quickly (around 2 seconds) using a kd-tree. Our efficiency is comparable to the RTF [6], and superior to the DTF [13], and our training process is simpler and more efficient.

Short- and Long-Range Interactions: In our experiments we used the model of (9) with uninformative unary terms over the occluded region. Figure 5 shows the relationship between the window size and accuracy in performance for the small occlusion test. The graph clearly shows the boost in performance offered by increasing the neighborhood range. At a window range of 2 pixels we are approaching the performance of traditional MRF and CRF models with local neighbors. We found the best performance at $\sigma_w = 7$ pixels. Performance tails off as the range increases. This is to be expected since the local conditioning of the potentials is no longer valid over large distances; in addition, at 35 pixels we are approaching the size of the characters

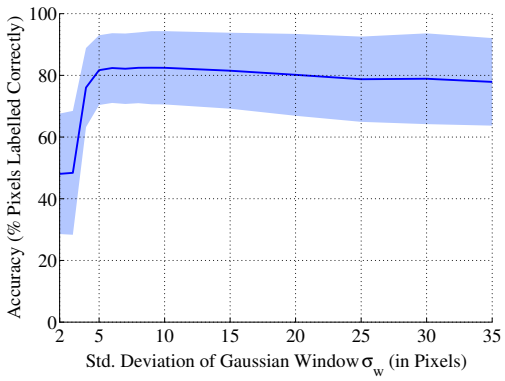


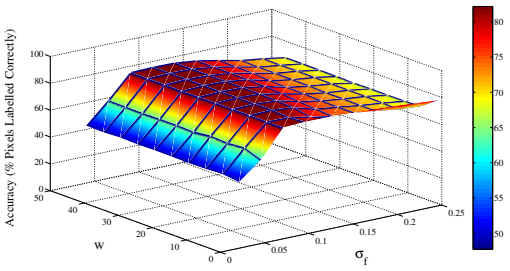
Figure 5: The variation of accuracy with the spatial window size for the KAIST Hanja2 database with small occlusions. We observe a noticeable decrease in performance for small window sizes (approaching the standard 4-connected CRF) demonstrating the advantage of having a non-local pairwise potential. There is also a drop-off in performance with large window sizes suggesting that very long range potentials act as a hinderance.

themselves.

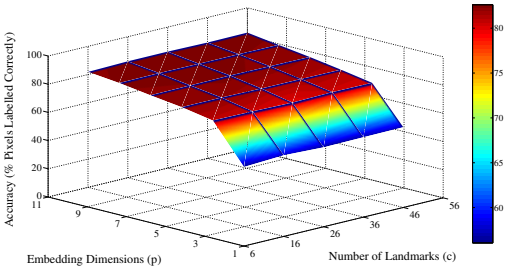
Figure 6(a) shows the variation in accuracy with the w and σ_f parameters. We observe that the weighting w of the pairwise term has relatively little impact but there is a definite optimal value for the σ_f term. We kept the weight $w = 25$ and $\sigma_f = 0.1$ terms constant for our other experiments. The weight w is of little importance since we have an uninformative unary term in the occluded region. The σ_f plays a greater role due to the windowing process applied to the pairwise potentials. Windowing the potentials shown in Fig. 3 leads to uninformative tails (at large distances) for all classes and the embedded approximation of the dissimilarity measure will be less accurate in these regions. Changing the σ_f parameter to match the window helps provide a sharper drop off in the edge potentials outside the windowed region and leads to an improved accuracy.

The Embedding: We observed several advantages of using LMDS, besides its relatively low computational complexity and memory requirements. First, the number of landmark points c is a simple parameter that may be used to trade-off error for performance (both computational as well as memory). We observed that a few samples (we used $c = 80$ for a 10 dimensional feature space) are sufficient in practice for producing embeddings with acceptable error (Fig. 6(b)). Second, the intrinsic dimensionality of the feature space may be discovered automatically as the number of positive Eigenvalues in Λ(please see supplementary material).

Limitations: Whilst our approach provides state-of-the-art performance and confers many benefits in the expressive power of the non-local and conditional potentials. Under the current model we learn a different feature space for every label. This is clearly expensive for multi label prob-



(a) Accuracy against w and σ_f.



(b) Accuracy against c and p.

Figure 6: The variation of accuracy with the parameters of (9) for the KAIST Hanja2 database with small occlusions. (a) We observe that the performance is relatively invariant to the weight parameter w but there is an optimal value of σ_f close to 0.1. (b) We observe a slight improvement in accuracy as we increase the number of landmarks c used but limiting the dimensionality of the embedded space p has a greater impact.

lems with a large label set. Also we are currently neglecting cross terms in that density estimating between different labels can also be performed and encoded into the update filtering at each iteration. The number of cross terms would scale quadratically with the size of the label set.

Further Work: This work opens a number of avenues for future investigation. In particular there are many options for estimating the conditional potential distances for a wider variety of multidimensional complex data and to improve scaling with larger training datasets. In particular non-parametric density estimators and regression techniques may prove very useful for this task.

Conclusions: We have demonstrated how to condition expressive, non-local pairwise potentials on input data. Key to our approach is the fast estimation of a feature space that is specific to the test image. This embedding of the pixels in feature space leads to an efficient mean-field inference in a fully-connected CRF model, yet with a generalized underlying dissimilarity measure. Our method confers state-of-the-art performance when compared to recent approaches that perform inference on similar models.

Acknowledgements: We thank the anonymous reviewers for their comments and suggestions. Neill Campbell and Jan Kautz received funding from EPSRC grant EP/I031170/1. Kartic Subr is supported by the Royal Society's Newton International Fellowship

References

[1] A. Adams, J. Baek, and M. A. Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum (Eurographics Proceedings)*, 2012. 3

[2] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Proc. 7^{th} Europ. Conf. on Computer Vision*, pages 109–124, 2002. 6

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, November 2001. 1

[4] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems 15.*, 2002. 4

[5] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 99, 2012. 2

[6] J. Jancsary, S. Nowozin, T. Sharp, and C. Rother. Regression tree fields - an efficient, non-parametric approach to image labeling problems. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2376–2383, 2012. 2, 5, 6, 7

[7] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1568–1583, October 2006. 1

[8] N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):531–552, 2011. 1

[9] P. Kornprobst, J. Tumblin, and F. Durand. Bilateral filtering: Theory and applications. *Foundations and Trends in Computer Graphics and Vision*, 4(1):1–74, 2009. 2

[10] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 1, 2, 3, 6, 7

[11] P. Krähenbühl and V. Koltun. Efficient nonlocal regularization for optical flow. In *Proc. 12^{th} Europ. Conf. on Computer Vision*, pages 356–369, 2012. 2, 3

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18^{th} International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 1

[13] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *Proc. 13^{th} Intl. Conf. on Computer Vision*, pages 1668–1675, 2011. 2, 5, 6, 7

[14] K. Subr, S. Paris, C. Soler, and J. Kautz. Accurate binary image selection from inaccurate user input. In *Computer Graphics Forum (Eurographics Proceedings)*, 2012. 2, 4, 6, 7

[15] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):1068–1080, June 2008. 1

[16] V. Vineet, J. Warrell, P. Sturgess, and P. Torr. Improved initialization and gaussian mixture pairwise terms for dense random fields with mean-field inference. In *Proc. 23^{rd} British Machine Vision Conference*, pages 73.1–73.11, 2012. 1, 2, 3

[17] V. Vineet, J. Warrell, and P. H. S. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. In *Proc. 12^{th} Europ. Conf. on Computer Vision*, pages 31–44, 2012. 2

[18] Y. Weiss. *Comparing the mean field method and belief propagation for approximate inference in MRFs*. MIT Press, 2001. 2