Learning to approximate global shape priors for figure-ground segmentation

Daniel Kuettel
dkuettel@vision.ee.ethz.ch
ETH Zurich

Vittorio Ferrari
vferrari@staffmail.ed.ac.uk
University of Edinburgh

Abstract

We present a technique for approximate minimization of two-label energy functions with higher-order or global potentials. Our method treats the energy function as a black-box: it does not exploit knowledge of its form nor its order, as opposed to optimization schemes specialized to a particular form. The key idea is to automatically learn a lower-order approximation of the energy function, which can then be minimized used existing efficient algorithms. We experimentally demonstrate our method for binary image segmentation, where it enables to incorporate a global shape prior into traditional models based on pairwise conditional random fields.

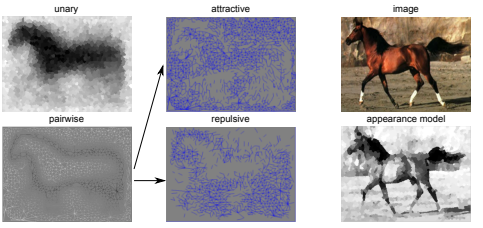
1 Introduction

Many problems in computer vision are formulated as minimizing the energy of a discrete graphical model (e.g. a conditional random field [6, 20, 24, 25, 33]). This involves designing an energy function to model the problem, and then minimizing it to find the lowest energy labelling. A good function should give lower energy to labellings that solve the problem more accurately. In figure-ground segmentation, the energy favours spatially smooth labellings that cover regions matching a certain appearance model [6, 18, 24, 25].

The energy minimization framework allows to cleanly separate modelling and inference. It has enjoyed tremendous success in computer vision, and has been applied to various problems such as segmentation [25], stereo [27] and denoising [31]. However, it involves a trade-off between expressiveness and optimizability. On the one hand we want sophisticated energy functions which model the problem accurately. On the other hand, we want to find the global optimum of the energy over all possible labellings of its variables. Unfortunately, general efficient minimization algorithms only exist for restricted classes of energy functions. A very popular such class are *pairwise* functions, involving only terms depending either on one or two variables [6, 15, 16]. The optimization of higher-order functions is still an open, highly challenging problem [2, 17, 19, 23, 26, 34].

It takes considerable effort to design an energy function that balances modelling accuracy versus minimization feasibility. A recent trend is to design energies containing higher-order potentials of a particular form, along with an optimizer specialized for that form [2, 11, 13, 17, 23, 26, 34]. The downside of this approach is that different optimizers needs to be invented for each form of higher-order potential, which requires great skill and knowledge.

In this paper, we propose an alternative approach which does not require inventing specialized minimization algorithms. The designer only provides an *arbitrary* energy function



(a) (b)

Input: regularizer $(\mathbf{R}, \mathbf{r})$
Initialization: $\mathbf{P} = 0$, $\mathbf{q} = 0$
for $k = 1 \rightarrow K$ **do**
  $\mathbf{x_k} \leftarrow$ new training sample
  $\bar{\mathbf{x}}_\mathbf{k} \leftarrow$ linearize $\mathbf{x_k}$ as in eq. 14
  $\mathbf{P} \leftarrow \mathbf{P} + \bar{\mathbf{x}}_\mathbf{k} \bar{\mathbf{x}}_\mathbf{k}^\top$
  $\mathbf{q} \leftarrow \mathbf{q} + \bar{\mathbf{x}}_\mathbf{k} E_k$
end for
output: $\Phi \leftarrow (\mathbf{R} + 2\mathbf{P})^{-1}(2\mathbf{q} - \mathbf{r})$

(c)

Figure 1: *(a) Learned proxy for the full min-Chamfer prior. Top left: the unary potentials show the mean shape of horses (darker = more foreground). Bottom left: Pairwise potentials (white = attractive, black = repulsive). Inside and outside the horse body it clearly prefers that neighboring pixels take the same label (attractive). Along the horse's boundaries the potentials are repulsive. At the location of the legs there is considerable uncertainty, as they are very variable in the training images. This leaves room for the appearance model to precisely determine their location. (b) Appearance model for a particular test image (darker = more foreground). (c) Algorithm to learn Φ.*

E that suits her problem. We make no assumption about the form of E nor about its order. It can be arbitrarily complex and, in the extreme case, contain a global potential defined over all variables. Our key idea is to *automatically* learn a pairwise function $\tilde{E}$, which tries to approximate E as well as possible. After learning the parameters of this *proxy* function, it can be efficiently minimized using standard algorithms for pairwise models [6, 15, 16].

We demonstrate our idea on binary image segmentation, where variables can take two labels: foreground or background. Standard pairwise CRF models incorporate only a preference for smooth segmentations which align well with image gradients [6, 13, 24, 25]. Including a shape prior would require higher-order potentials to capture the complex statistical dependencies between pixels that characterize a shape class (e.g. horses). A desirable higher-order prior is the smallest Chamfer distance to a set of exemplars of the shape class [8]. This min-Chamfer potential is *global*, as it can only be expressed exactly using a function involving *all* variables (pixels).

When is our approach useful? On the one hand, it is general, as it can learn to approximate *any* binary energy function. On the other hand, as the modelling capacity of our proxy is limited, the quality of the approximation degrades as the input function becomes more and more complex. In the limit, a function of N binary variables can have up to 2^N completely unrelated outputs. Although our method would not be effective for such extremely irregular functions, in practice many desirable higher-order energy functions are much more regular. While these cannot be written exactly as pairwise functions, they could be approximated well by our method. In this intermediate complexity regime our method is most useful. It enables to employ desirable higher-order functions for which no exact efficient optimizer exists. We demonstrate this advantage in our experiments by learning high quality proxies for two variants of the popular min-Chamfer global shape prior [8, 29, 30] (sec. 4).

1.1 Related Work

Pairwise functions. Minimizing pairwise discrete energy functions has been thoroughly researched and many efficient algorithms exist. Particularly relevant to our work are the QPBO [16] and TRW-S [15] algorithms, which we use to minimize our proxy (sec. 3).

Higher-order functions. In the general case, minimizing energies containing higher-order potentials defined over many variables is prohibitively expensive. However, there is a growing body of works tackling interesting special cases which can be optimized efficiently with

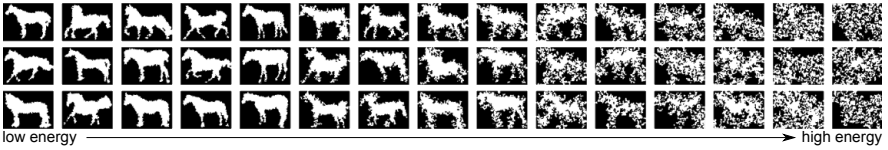


Figure 2: **min-Chamfer prior.** *Example shapes with increasing energy $E_c(\mathbf{x})$ for horses.*

dedicated algorithms.

The primal-dual technique of [2] can exactly minimize any binary (two-label) function with n variables and k-th order submodular potentials in $O(2^k n^3)$. While this is better than the most general case, it is still exponential in k. Ramalingam et al. [23] present a technique to transform certain submodular multi-label functions into binary submodular pairwise functions which can then be minimized exactly using graph-cuts. However, the transformation can be applied only to functions of order up to $k = 4$.

Several works appeared to deal with *pattern potentials* [12, 17, 26, 34]. These potentials can act on rather large sets of variables, but they give the same (high) energy value to most labellings. Only a small number of so-called pattern labellings can have their own (lower) energy values. Typically these methods involve constructions growing roughly linearly with the number of patterns. This is useful when there are much fewer patterns than the total number of labellings (which grows exponentially with k). A famous example is the P^N-Potts potential of [12], which encourages local groups of pixels to take the same label. Particularly related to our work are the potentials proposed by [17, 26], where example labellings of an image patch are represented as patterns. However, as any non-pattern labelling gets the same energy, it would need a prohibitively large number of patterns to encode a global shape prior. In practice, the experiments in [26] are limited to 25 patterns on 10×10 patches.

Efficient algorithms have also been proposed for some specific forms of global potentials. Co-occurrence potentials [9, 19, 34] encourage labellings containing certain desirable combinations of labels (e.g. horse and grass). Connectivity priors [10, 33] encourage binary segmentations where the foreground forms a single connected component. Recently [22] proposed a construction to enhance a pairwise model to incorporate a preference for the foreground area to be near a predefined size.

Figure-ground segmentation. Pairwise discrete energy functions are probably the most popular way to model figure-ground segmentation in computer vision [18, 24, 25]. Most approaches use only submodular potentials to encourage smoothness, while a few include also non-submodular terms enforcing simple local shape models [1] and then minimize the energy with QPBO [16]. In general, global shape priors are very difficult to plug into a discrete energy formulation, which is more suited for local interactions. However, there are works which model segmentation with global shape priors within continuous energy minimization frameworks, achieving some success for certain kinds of priors [28].

Structured SVMs. As our method involves learning a pairwise discrete energy function (proxy), it is related to structured SVMs [8, 32]. SSVMs find parameters so that the optimal labelling has lower energy than any other labelling, by a margin proportional to their difference (loss). Importantly, SSVMs involves minimizing the pairwise energy augmented with the loss function, which is intractable for a complex loss such as min-Chamfer. More generally, we want to tackle higher-order energies which cannot be minimized exactly with existing methods. Trying to learn a proxy for such energies with SSVMs inevitably leads to intractable loss-augmented proxies. As another difference, our method tries to approximate the higher-order function over its entire domain, not just near the optimum. In principle, this

enables to sample from the proxy, which is useful in importance sampling schemes, or as part of a more complex learning algorithm.

2 Overview

Discrete higher-order energies. We consider a binary energy function $E(\mathbf{x})$ defined over a set of variables $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$. Each variable can take a label in $\mathcal{X} = \{0,1\}$. A well designed $E(\mathbf{x})$ assigns lower energies to labelings $\mathbf{x} \in \mathcal{X}^N$ that solve the problem more accurately. Thus, the lowest energy labeling is returned as solution

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}^N} E(\mathbf{x}) \tag{1}$$

$E(\mathbf{x})$ can be decomposed into a sum of potential functions defined on cliques (subsets) of variables

$$E(\mathbf{x}) = \sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c) \tag{2}$$

where $\psi_c(\mathbf{x}_c)$ is a potential function on clique c and $\mathbf{x}_c = \{x_i, i \in c\}$. $\mathcal{C}$ is the set of cliques.

Minimizing eq. (1) is in general NP-hard. However, pairwise energies (i.e. with cliques of size up to 2) can be optimized efficiently with algorithms such as graph-cuts [6], QPBO [16] and tree-reweighted message passing [15]. Higher-order potentials can express more complex dependencies between variables. Without assumptions on the form of the higher-order potentials, the complexity of minimization increases exponentially with the size of the largest clique [4]. In the extreme case, it can have size N (global clique), i.e. the function cannot be decomposed into a sum over simpler clique potentials. Although the task is daunting in the general case, several algorithms have been recently proposed to efficiently minimize certain types of higher-order and global potentials (sec. 1.1).

Learning a pairwise approximation. In this paper we propose to *automatically* learn a lower-order approximation $\tilde{E}(\mathbf{x})$ of the original energy function $E(\mathbf{x})$. We make no assumptions about the form of E. It can be arbitrarily complex and could possibly contain a global potential defined over all variables. The approximation $\tilde{E}$ instead is limited to unary and pairwise potentials (fig. 1(a)). It is learned to hold as well as possible over the entire domain $\mathcal{X}^N$

$$\tilde{E}(\mathbf{x}) \cong E(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}^N \tag{3}$$

The key advantage of learning a pairwise approximation $\tilde{E}$ is that it can be minimized efficiently (sec. 3).

Segmentation with the min-Chamfer prior. We apply this idea to figure-ground segmentation, where variables x_i are pixels and the labels correspond to foreground ($x_i = 1$) and background ($x_i = 0$). In this context, $E(\mathbf{x})$ is a global shape prior giving lower energy to segmentations $\mathbf{x}$ that fit a shape class such as horses or mugs. In our experiments we use the smallest Chamfer distance to a set of exemplar shapes as E

$$E_c(\mathbf{x}) = \min_{\mathbf{s}\in\mathcal{S}} \left(\frac{1}{|\partial\mathbf{x}|} \sum_{x\in\partial\mathbf{x}} \min_{s\in\partial\mathbf{s}} d(x,s) + \frac{1}{|\partial\mathbf{s}|} \sum_{s\in\partial\mathbf{s}} \min_{x\in\partial\mathbf{x}} d(x,s) \right) \tag{4}$$

where $\mathcal{S}$ is the set of exemplars, d is the Euclidean distance, and ∂ denotes the outline of a segmentation (i.e. the pixels at the boundary between foreground and background).

The Chamfer distance is normalized by the length of the outlines and it is symmetric over the exemplars $\mathbf{s}$ and the input segmentation $\mathbf{x}$. It has multiple equivalent minima, one at each exemplar ($E_c(\mathbf{x}) = 0 \ \ \forall \mathbf{x} \in \mathcal{S}$) and it increases smoothly as $\mathbf{x}$ deviates from any exemplar (fig. 2). It is a truly global potential, that can only be expressed exactly using all variables.

In order to use the shape prior to segment a test image, we add a unary potential containing an appearance model preferring image regions with a certain color distribution [25] (fig. 1(b), sec. 4). Note how this does not change the difficulty of the optimization problem, and therefore in the following sections we consider E as containing the shape prior only.

3 Learning the proxy $\tilde{E}(\mathbf{x})$

Given an *arbitrary* energy function $E(\mathbf{x})$, we want to find a function $\tilde{E}(\mathbf{x})$ that approximates it well and can be can be minimized efficiently. To be as general as possible, we do not make assumptions about the form or the order of $E(\mathbf{x})$, which could even contain a global potential. We simply treat $E(\mathbf{x})$ as a *black box* mapping a configuration $\mathbf{x}$ to an energy. Instead, we constrain $\tilde{E}(\mathbf{x})$ to unary and pairwise potentials only. In the following we refer to $\tilde{E}$ as the *proxy* to the *original* energy E

$$\tilde{E}(\mathbf{x};\Psi) = \psi_\emptyset + \sum_{i=1}^{N} \psi_i(x_i) + \sum_{(i,j)\in\mathcal{C}_2} \psi_{ij}(x_i,x_j) \tag{5}$$

where $\psi_\emptyset$ is a constant offset, $\psi_i(x_i)$ is the unary potential for variable i, $\psi_{ij}(x_i,x_j)$ is the pairwise potential for variables (i,j), and $\mathcal{C}_2$ is a set of pairs defining which variables are connected. We connect all neighbouring pixels and also connect a large number of distant pixel pairs, in order to make the proxy capable of learning long-range dependencies. This class of energy functions can be efficiently minimized using QPBO [16]. As the learned $\tilde{E}$ typically contains some non-submodular pairwise potentials, QPBO might leave some variables unlabelled. We label them in a second optimization pass using TRW-S [15]. However, in all experiments in sec. 4, QPBO labelled all pixels.

The parameters Ψ of $\tilde{E}$ include 1 value for $\psi_\emptyset$, $2N$ values for $\psi_i(0)$ and $\psi_i(1)$, and $4|\mathcal{C}_2|$ values for $\psi_{ij}(0,0)$, $\psi_{ij}(0,1)$, $\psi_{ij}(1,0)$ and $\psi_{ij}(1,1)$, for a total of $1+2N+4|\mathcal{C}_2|$ parameters. However, this is an overcomplete parameterization ([35], chapter 3.2). There exist a minimal parameterization which can represent the same class of functions. In that parameterization, $\psi_i(0),\psi_{ij}(0,0),\psi_{ij}(0,1)$ and $\psi_{ij}(1,0)$ are all set to 0. This reduces eq. (5) to

$$\tilde{E}(\mathbf{x};\Phi) = \phi_\emptyset + \sum_{i=1}^{N} \phi_i x_i + \sum_{(i,j)\in\mathcal{C}_2} \phi_{ij} x_i x_j \tag{6}$$

which is known as the Ising model ([35], section 3.3). Any energy function (5), specified by $1+2N+4|\mathcal{C}_2|$ parameters, can be transformed into an equivalent energy function (6), specified by only $1+N+|\mathcal{C}_2|$ parameters ([14], section 4.1). Therefore, the vector of $\bar{N} = 1+N+|\mathcal{C}_2|$ parameters

$$\Phi = (\phi_\emptyset,\ldots,\phi_i,\ldots,\phi_{jl},\ldots)^\top \quad 1\le i\le N \quad (j,l)\in\mathcal{C}_2 \tag{7}$$

completely determines $\tilde{E}(\mathbf{x})$.

Our goal is to learn the parameters Φ so that $\tilde{E}(\mathbf{x};\Phi) \cong E(\mathbf{x})$ for $\forall\mathbf{x}\in\mathcal{X}^N$. In principle, with no assumption about the structure of $E(\mathbf{x})$, we would have to evaluate every possible labeling $\mathbf{x}\in\mathcal{X}^N$ to do this. Unfortunately this is not feasible, as there are 2^N different $\mathbf{x}$. We relax the problem by requiring $\tilde{E}$ to approximate E on a very large subset of sample labellings $\tilde{\mathcal{X}} = \{\mathbf{x}_1,\ldots,\mathbf{x}_K\} \subset \mathcal{X}^N$ (typically millions). The underlying assumption is that E varies smoothly over its domain $\mathcal{X}^N$, and therefore the approximation is good also for most other labelling. This is a weak assumption, as smoothness holds for most practical energy functions. Thus we require

$$\tilde{E}(\mathbf{x}_k;\Phi) \cong E(\mathbf{x}_k), \quad 1\le k\le K \tag{8}$$

We cast learning Φ as a least squares minimization problem

$$\min_{\Phi} \sum_{k=1}^{K} \left(\tilde{E}(\mathbf{x}_k;\Phi) - E(\mathbf{x}_k)\right)^2 \tag{9}$$

Furthermore, we add a regularization term $R(\Phi)$ on the parameters

$$\min_{\Phi} \sum_{k=1}^{K} \left(\tilde{E}(\mathbf{x}_k;\Phi) - E(\mathbf{x}_k)\right)^2 + R(\Phi) \tag{10}$$

The regularizer states a prior preference for a certain distribution of parameter values. In our experiments we use the simple L2 regularizer ($R(\Phi) = \sum_{\phi\in\Phi} \phi^2$).

To summarize, learning $\tilde{E}(\mathbf{x})$ means finding the Φ that leads to the best least-squares approximation of the original energy on a sampled subset of labellings. The following subsections will show how to solve the minimization in (10) efficiently in closed form (sec. 3.1 and 3.2), and how to best sample from $\mathcal{X}^N$ to learn a good approximation (sec. 3.3).

3.1 Learning as quadratic form minimization

This section shows how to transform the objective function in (10) into a quadratic form, which enables to solve the minimization exactly in closed form.

We assume that the regularizer $R(\Phi)$ is already in quadratic form

$$R(\Phi) = \tfrac{1}{2}\Phi^\top \mathbf{R}\Phi + \mathbf{r}^\top\Phi \tag{11}$$

This includes a wide variety of regularizers, including the L2 regularizer we use.

The sum of squared differences can be expressed as

$$\sum_{i=1}^{K} d_k^2 = \mathbf{d}^\top \mathbf{I}\mathbf{d} \tag{12}$$

with the column vector $\mathbf{d} = (d_1,\ldots,d_K)^\top$ defined as $d_k := \tilde{E}(\mathbf{x}_k;\Phi) - E(\mathbf{x}_k)$. We now show how to rewrite each d_k as a linear function of Φ, making $\mathbf{d}^\top\mathbf{I}\mathbf{d}$ a quadratic form in Φ. We first note that $E(\mathbf{x}_k)$ is simply evaluating the original energy function on a sample training configuration $\mathbf{x}_k$. As this is independent of Φ, it is a constant in the minimization problem. We denote those constants with E_k. Instead, the proxy energy $\tilde{E}(\mathbf{x}_k;\Phi)$ depends on both the parameters Φ and a sample $\mathbf{x}_k$. We construct a *linearized* $\bar{\mathbf{x}}_k$, so that

$$\tilde{E}(\mathbf{x}_k;\Phi) = \bar{\mathbf{x}}_k^\top\Phi \tag{13}$$

The linearized $\bar{\mathbf{x}}_k$ is a binary vector that selects which potentials are active among all those in Φ, according to the configuration $\mathbf{x}_k$. We show below both $\bar{\mathbf{x}}_k$ and Φ side-by-side to clarify their relation

$$\begin{aligned} \Phi &= (\ \phi_\emptyset\ ,\ \ldots\ ,\ \phi_i\ ,\ \ldots\ ,\ \ \phi_{jl}\ \ ,\ \ldots)^\top \\ \bar{\mathbf{x}}_k &= (\ \ 1\ \ ,\ \ldots\ ,\ \mathbf{x}_{ki}\ ,\ \ldots\ ,\ \mathbf{x}_{kj}\mathbf{x}_{kl}\ ,\ \ldots)^\top \end{aligned} \tag{14}$$

By inserting the linearization of eq. (13) into the sum of squared differences of eq. (12)

$$\mathbf{d}^\top\mathbf{I}\mathbf{d} = (\bar{\mathbf{X}}^\top\Phi - \mathbf{E})^\top\mathbf{I}(\bar{\mathbf{X}}^\top\Phi - \mathbf{E}) = \Phi^\top\bar{\mathbf{X}}\bar{\mathbf{X}}^\top\Phi - 2\mathbf{E}^\top\bar{\mathbf{X}}^\top\Phi + \mathbf{E}^\top\mathbf{E} \tag{15}$$

where $\bar{\mathbf{X}}$ contains $\bar{\mathbf{x}}_k$ as columns, and $\mathbf{E} = (E_1,\ldots,E_K)^\top$. With the reformulations above, the minimization problem (10) can be written as minimizing a quadratic form over Φ

$$\min_{\Phi} \quad \tfrac{1}{2}\Phi^\top(\mathbf{R} + 2\bar{\mathbf{X}}\bar{\mathbf{X}}^\top)\Phi + (\mathbf{r}^\top - 2\mathbf{E}^\top\bar{\mathbf{X}}^\top)\Phi \tag{16}$$

In general, any quadratic form $\frac{1}{2}\mathbf{y}^\top\mathbf{H}\mathbf{y}+\mathbf{f}^\top\mathbf{y}$ can be minimized by $\mathbf{y}=-2\mathbf{H}^{-1}\mathbf{f}$, if $\mathbf{H}$ is positive-definite. Therefore, we can solve our minimization problem in closed form by

$$\Phi=(\mathbf{R}+2\bar{\mathbf{X}}\bar{\mathbf{X}}^\top)^{-1}(2\bar{\mathbf{X}}\mathbf{E}-\mathbf{r}) \tag{17}$$

when $\mathbf{R}+2\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ is positive-definite. The symmetric matrix $\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ is always positive-semi-definite. Given $\geq\bar{N}$ training samples, it is also full rank and therefore positive-definite. Moreover, note how the regularizer R is positive-definite by construction. If given enough weight, it can make $\mathbf{R}+2\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ positive-definite even without enough training samples.

3.2 Efficiently setting up the quadratic form

In order to learn a good approximation $\tilde{E}(\mathbf{x};\Phi)\cong E(\mathbf{x})$, we need many training samples $\mathbf{x}_k$, typically millions. Unfortunately, the matrix $\bar{\mathbf{X}}$ is of size $\bar{N}\times K$, i.e. it grows with the number K of training samples. For a typical segmentation problem, $K\approx 10^7$ and $\bar{N}\approx 10^4$. Hence, $\bar{\mathbf{X}}$ takes about 100 GBs and does not fit in memory.

However, $\bar{\mathbf{X}}$ is not necessary on its own. It only appears in the terms $\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$, which has size $\bar{N}\times\bar{N}$, and $\bar{\mathbf{X}}\mathbf{E}$, which has size $\bar{N}\times 1$. Hence, the size of both terms involving $\bar{\mathbf{X}}$ *does not depend on the number of training samples* K.

We show here how to incrementally compute $\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ and $\bar{\mathbf{X}}\mathbf{E}$, one training sample at a time. In this fashion we never need to handle the full $\bar{\mathbf{X}}$. We rewrite $\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ as

$$\bar{\mathbf{X}}\bar{\mathbf{X}}^\top=\begin{bmatrix}\bar{\mathbf{x}}_1 & \dots & \bar{\mathbf{x}}_k & \dots & \bar{\mathbf{x}}_K\end{bmatrix}\cdot\begin{bmatrix}\bar{\mathbf{x}}_1^\top & \dots & \bar{\mathbf{x}}_k^\top & \dots & \bar{\mathbf{x}}_K^\top\end{bmatrix}^\top=\sum_{k=1}^{K}\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^\top \tag{18}$$

and analogously $\bar{\mathbf{X}}\mathbf{E}=\sum_{k=1}^{K}\bar{\mathbf{x}}_k E_k$. Therefore, both terms can be computed as a sum over individual training samples. This idea enables to learn Φ very efficiently. Fig. 1(c) shows the algorithm. To keep the notation clean, we use $\mathbf{P}:=\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$ and $\mathbf{q}:=\bar{\mathbf{X}}\mathbf{E}$, as they are incrementally built during its execution.

The largest matrix involved in this algorithm is $\bar{N}\times\bar{N}$, instead of $\bar{N}\times K$. For good results, the number K of training samples needs to be much larger than the number $\bar{N}$ of parameters. Hence, the algorithm requires much less memory than a direct implementation of learning as in eq. (17). The algorithm also offers a considerable speed up in runtime because it avoids computing the large matrix multiplication $\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$. The runtime for computing $\mathbf{P}$ and $\mathbf{q}$ is $O(K\bar{N}^2)$, compared to $O(K^2\bar{N})$ for directly computing $\bar{\mathbf{X}}\bar{\mathbf{X}}^\top$. Finally, note how inverting $\mathbf{R}+2\mathbf{P}$ takes $O(\bar{N}^{2.373})$. However, since $K\gg\bar{N}$, $O(\bar{N}^{2.373})<O(\bar{N}^3)<O(K\bar{N}^2)$. Hence, the runtime of learning is dominated by the cost of constructing $\mathbf{R}+2\mathbf{P}$ and not by its inversion.

3.3 Training samples

As $\mathcal{X}^N$ is huge, we need many training samples for accurate training. Training with only random samples is not recommended as they typically all have high energy. For the proxy to learn about regions of $\mathcal{X}^N$ with low energy, it is important to include many such samples as well. In our experiments we generate 50% purely random samples and 50% low energy samples. For the min-Chamfer prior, we generate low energy samples by first drawing a random exemplar shape $\mathbf{s}\in\mathcal{S}$, and then adding small random perturbations to it (fig. 2).

4 Experiments and conclusions

We conduct experiments on the Weizmann Horses dataset [5], which contains 327 images of horses along with ground-truth segmentations. We divide them into 228 for training the

(a) (b)

Figure 3: *(a) Compares the exact solution of the energy of the model E_n^{test} based on the unnormalized asymmetric min-Chamfer versus the solution of our the corresponding energy $\tilde{E}_n^{test}$ based on our proxy. They are very close. (b) Compares the solution when using only the appearance terms E_a versus using the complete model $\tilde{E}_c^{test}$ containing also our full min-Chamfer proxy. The latter leads to better segmentations.*

shape prior proxy and 90 for segmenting horses in new test images. The remaining 9 images are used to set a hyper-parameter λ (see sec. 4.2). Before all experiments we crop the images to a rectangle around the horse based on the ground-truth segmentation. To reduce the number of pixels, we partition each image into superpixels using [21]. These preserve object boundaries and give only about 1000 superpixel per image. In subsection 4.1 we evaluate the quality of the learned prior proxy on its own, and then in subsection 4.2 we use it in conjunction with an appearance model to segment new test images.

4.1 Quality of the learned prior proxy alone

In this experiment we learn a proxy $\tilde{E}_c$ for the min-Chamfer shape prior E_c of eq. (4) using our method of sec. 3. The 228 ground-truth segmentations in the training set form the set $\mathcal{S}$. We generate 5×10^6 training samples derived from $\mathcal{S}$ (sec. 3.3). Training takes about 350h on a 2.6GHz CPU, most of which is spent computing the energies of the training samples.

Neighbor connections only. In a first experiment we connect all neighboring superpixels in the proxy model. This results in 1469 parameters for the unary potentials and 4309 for the pairwise potentials. To evaluate how well the learned proxy $\tilde{E}_c$ approximates the min-Chamfer prior E_c, we generate a second, disjoint set of $K = 2 \times 10^6$ samples from $\mathcal{S}$ (as the prior is defined by $\mathcal{S}$, eq. 4). We quantify the approximation error as the square-root of the mean squared difference between $\tilde{E}_c$ and E_c on these samples. The approximation error is $5.6 \cdot 10^{-4}$, which is two orders of magnitudes smaller than the difference between the average energy of the good low-energy samples and the bad high energy ones ($0.1 \cdot 10^{-2}$, sec. 3.3). This shows that the learned proxy approximates well the min-Chamfer prior. Note the importance of the training samples: when training with only random samples or only good low-energy samples the error increases to $2.5 \cdot 10^{-3}$ or $1.1 \cdot 10^{-3}$, respectively.

Adding distant connections. The approximation quality can be improved by adding more pairwise connections to the proxy $\tilde{E}_c$. This increases its capacity, enabling it to better approximate complex energies. To confirm this experimentally, we added 5000 randomly sampled connections between distant superpixels. The approximation error decreased to $5.2 \cdot 10^{-4}$.

4.2 Segmenting new test images

Appearance model. After learning the proxy prior $\tilde{E}_c$, we use it to help segmenting a new test image I. We combine the prior with the appearance model of [25], i.e. a GMM over RGB space (fig. 1(b)). The appearance model is estimated from a rectangle centered in I and covering 50% of it (as done in [18, 25]). This gives an energy function $E_a(\mathbf{x};I)$ with only unary terms based on the appearance of superpixels in I. The overall energy function we use to segment I is then

$$\tilde{E}_c^{test}(\mathbf{x};I) = E_a(\mathbf{x};I) + \lambda \tilde{E}_c(\mathbf{x}) \tag{19}$$

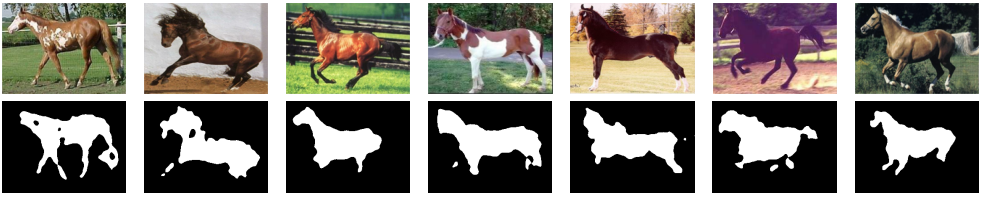


Figure 4: *Examples segmentations using the model $\tilde{E}_c^{test}$ based on our full min-Chamfer proxy.*

where λ is the weight of the shape prior. We choose λ so that minimizing eq. (19) leads to the best segmentation accuracy on a held-out validation set of 9 images (not used for the shape prior). Segmentation accuracy is measured by the percentage of pixels labeled as in the ground-truth.

Results: unnormalized asymmetric min-Chamfer. We evaluate how well the model (19) approximates the model using the true prior $E_c^{test} = E_a(\mathbf{x};I) + \lambda E_c(\mathbf{x})$. More precisely: How close is the labelling returned by minimizing $\tilde{E}_c^{test}$ to the true global optimum of E_c^{test}? We cannot answer this question exactly for the min-Chamfer prior E_c, as exact minimization is prohibitively time consuming. Indeed, this fact is the basic motivation for this paper. Instead, we first report experiments on a simplified version E_n of min-Chamfer, which is not symmetric and not normalized by outline length

$$E_n(\mathbf{x}) = \min_{\mathbf{s}\in\mathcal{S}} \left(\sum_{x\in\partial\mathbf{x}} \min_{s\in\partial\mathbf{s}} d(x,s) \right) \tag{20}$$

As shown by [7], with a single exemplar, $E_n^{test}(\mathbf{x},I) = E_a(\mathbf{x}|I) + \lambda E_n(\mathbf{x})$ can be represented using unary and (submodular) pairwise potentials. Thus it can be minimized exactly. We extend this idea to an arbitrary number of exemplars by solving the minimization for each exemplar separately, and then returning the lowest energy solution. This corresponds to explicitly looping over the first min in (20). Although exact, this procedure is extremely slow, as there are > 200 exemplars in $\mathcal{S}$. We learn a proxy prior $\tilde{E}_n$ to approximate E_n using the same procedure as in sec. 4.1. For a test image I, we very efficiently minimize the proxy test energy $\tilde{E}_n^{test}(\mathbf{x};I) = E_a(\mathbf{x};I) + \lambda\tilde{E}_n(\mathbf{x})$, obtaining a labelling $\tilde{\mathbf{x}}^*$. We also compute the globally optimal labelling $\mathbf{x}^*$ of the original test energy E_n^{test}, using the slow procedure above. We now quantify how well our procedure approximates the true global optimum by comparing the energy of $\tilde{\mathbf{x}}^*$ to that of $\mathbf{x}^*$ under the exact model E_n^{test}. On average over the 90 test images, the ratio $E_n^{test}(\tilde{\mathbf{x}}^*,I)/E_n^{test}(\mathbf{x}^*,I)$ is 1.3. Hence, our approximate solution is close to the global optimum (fig. 3(a)).

Another relevant question is whether using the shape prior improves the accuracy of the output segmentations. We evaluate segmentation accuracy averaged over the 90 test images. Using only the appearance term E_a, the accuracy is 76.2%. Incorporating the proxy prior, i.e. minimizing $\tilde{E}_n^{test}$, improves results considerably to 83.0%. Moreover, this is very similar to the accuracy produced by minimizing E_n^{test} (82.7%). This brings further evidence that our proxy-based inference scheme accurately approximates the behavior of E_n^{test}, i.e. the original segmentation energy using both the appearance term and the exact unnormalized asymmetric min-Chamfer shape prior. Note how our proxy-based inference scheme is $\sim 200\times$ faster than the exact minimization of E_n^{test} as it only requires minimizing a single pairwise energy (no loop over exemplars).

Results: full min-Chamfer. As it is not possible to exactly minimize the test energy $E_c^{test}(\mathbf{x};I) = E_a(\mathbf{x};I) + \lambda E_c(\mathbf{x})$, for comparison we implemented a slow minimization baseline, which evaluates E_c^{test} on $2.5\cdot 10^6$ random perturbations of the exemplars and selects the

one with the minimum energy. On average over the 90 test images, the ratio of the energy of the solution output by minimizing our proxy-based energy $\tilde{E}_c^{test}(\mathbf{x};I) = E_a(\mathbf{x};I) + \lambda \tilde{E}_c(\mathbf{x})$ to the baseline solution is 1.58. Importantly, the baseline is more than 1000× slower than our technique. Hence, assuming this semi-exhaustive baseline reaches an energy close to the global optimum, we conclude that our proxy-based approximate optimization scheme is very effective.

In terms of segmentation accuracy, minimizing our energy $\tilde{E}_c^{test}$ based on the full min-Chamfer proxy prior leads to 86.8%, which is a further improvement over using the unnormalized asymmetric min-Chamfer prior (82.7%). This is because the unnormalized asymmetric min-Chamfer prior has the drawback of preferring short outlines $\partial\mathbf{x}$ (a segmentation with no foreground has distance 0 to any exemplar). Hence, the normalized symmetric min-Chamfer distance (4) is more desirable [8, 29]. Fig. 1(a) shows the learned proxy and fig. 3(b) shows the improvement by our full min-Chamfer proxy over appearance alone.

4.3 Conclusion

We presented an approach to approximately minimize arbitrary binary energy functions. It bridges the gap between accurate modelling and ease of optimization in a principled way. It combines the convenience of modelling using a complex energy function, with the computational benefits of using a pairwise model.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. ClassCut for unsupervised class segmentation. In *ECCV*, 2010.

[2] C. Arora, S. Banerjee, P. Kalra, and S. Maheshwari. Generic cuts: An efficient algorithm for optimal inference in higher order mrf-map. In *ECCV*, 2012.

[3] L. Bertelli, T. Yu, D. Vu, and S. Gokturk. Kernelized structural SVM learning for supervised object segmentation. In *CVPR*, 2011.

[4] C. Bishop. Pattern recognition and machine learning. *Springer*, 2006.

[5] E. Borenstein and S. Ullman. Learning to segment. In *ECCV*, 2004.

[6] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, 2001.

[7] D. Freedman and T. Zhang. Interactive graph cut based segmentation with shape priors. In *CVPR*, 2005.

[8] D. M. Gavrila. Pedestrian detection from a moving vehicle. In *ECCV*, volume 2, pages 37–49, 2000.

[9] J. Gonfaus, X. Boix, J. Weijer, A. Bagdanov, J. Serrat, and J. Gonzalez. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010.

[10] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: coupling edges in graph cuts. In *CVPR*, 2011.

[11] T. Kim, K. Lee, and S. Lee. Nonparametric higher-order learning for interactive segmentation. In *CVPR*, 2010.

[12] P. Kohli, M. Kumar, and P. Torr. P3 & beyond: Solving energies with higher order cliques. In *CVPR*, 2007.

[13] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.

[14] V Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Trans. on PAMI*, 26(2):147–159, 2004.

[15] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on PAMI*, 28(10):1568 – 1583, 2006.

[16] Vladimir Kolmogorov and Carsten Rother. Minimizing nonsubmodular functions with graph cuts – a review. *IEEE Trans. on PAMI*, 29(7):1274–1279, 2007.

[17] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *CVPR*, 2009.

[18] Daniel Kuettel and Vittorio Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012.

[19] L. Ladicky, C. Russel, P. Kohli, and P.H.S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[20] Victor Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

[21] A. Levinshtein, A. Stere, K. Kutulakos, D. Fleed, and S. Dickinson. Turbopixels: Fast superpixels using geometric flows. In *IEEE Trans. on PAMI*, 2009.

[22] P. Pletscher and K. Pushmeet. Learning low-order models for enforcing high-order statistics. In *AISTATS*, 2012.

[23] S. Ramalingam, P. Kohli, K. Alahari, and P. Torr. Exact inference in multi-label crfs with higher order cliques. In *CVPR*, 2008.

[24] Amir Rosenfeld and Daphna Weinshall. Extracting foreground masks towards object recognition. In *ICCV*, 2011.

[25] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[26] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, 2009.

[27] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *CVPR*, 2001.

[28] T. Schoenemann and D. Cremers. Matching non-rigidly deformable shapes across images: A globally optimal solution. In *CVPR*, 2008.

[29] J. Shotton, A. Black, and R. Cipolla. Multi-scale categorical object recognition using contour fragments. In *IEEE Trans. on PAMI*, 2008.

[30] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Model-based hand tracking using a hierarchical bayesian filter. In *IEEE Trans. on PAMI*, 2006.

[31] R. Szeliski, R. Zabih, D. Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. on PAMI*, 30(6):1068–1080, 2008.

[32] Martin Szummer, Pushmeet Kohli, and Derek Hoiem. Learning CRFs using graph cuts. In *ECCV*, 2008.

[33] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *CVPR*, 2008.

[34] V. Vineet, J. Warrell, and P. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. In *ECCV*, 2012.

[35] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. In *Foundations and Trends in Machine Learning*, 2008.