Learning a Model for Shape-Constrained Image Segmentation from Weakly Labeled Data

Boris Yangel and Dmitry Vetrov

Lomonosov Moscow State University

Abstract. In the paper we address a challenging problem of incorporating preferences on possible shapes of an object in a binary image segmentation framework. We extend the well-known conditional random fields model by adding new variables that are responsible for the shape of an object. We describe the shape via a flexible graph augmented with vertex positions and edge widths. We derive exact and approximate algorithms for MAP estimation of label and shape variables given an image. An original learning procedure for tuning parameters of our model based on unlabeled images with only shape descriptions given is also presented. Experiments confirm that our model improves the segmentation quality in hard-to-segment images by taking into account the knowledge about typical shapes of the object.

Keywords: MRF, image segmentation, shape priors, weakly-labeled data, part-based models

1 Introduction

Image segmentation is a well-studied problem in computer vision. It can be solved well (see, for example, [13]) when objects differ in color and texture from background significantly. However, in case of non-discriminative color models and weak object boundaries some high-level knowledge about the scene is required to make segmentation more robust. For example, one can have some clues about the shape of the object being segmented, and use those as a segmentation prior.

Taking such knowledge into account while segmenting an image is not an easy task. Introduction of high-level constraints into state-of-the-art approaches to image segmentation such as conditional random fields (CRF) [9] leads either to oversimplified models, or to complex models with high-order terms, which are hard to infer solutions from and even harder to learn. As a result, such models are tweaked manually and tend to perform much worse than they could have.

In this paper we aim to introduce a model for shape-constrained binary segmentation (i.e. segmentation to object and background) that is powerful enough to describe complex shapes, and yet has tractable inference and learning procedures. Our model is built upon a popular way (see, for example, [8, 10, 16]) of introducing global constraints into CRF, which uses unary terms of the CRF energy to constrain labeling together with an additional term as a prior for high-level clues, which in our case are shape descriptions. We describe object shape via a graph augmented with vertex positions and edge widths, a way that is similar to part-based models [4, 3] but seems to be more rich. One interesting property of our segmentation model is that it can be seen as a shape fitting model that uses pixel labels as latent variables. It allows us to come up with a training formulation that does not require pixel labels to learn from.

1.1 Related Work

The description of object shape proposed in this work can be seen as an application of part-based object modeling, technique that is well-developed in the context of the object part detection problem. Furthermore, both exact inference and learning procedures proposed in this paper are build on top of techniques developed for this kind of models. In [4] one such model, namely pictorial structure, is presented together with an efficient inference algorithm based on dynamic programming. A follow-up work [5] introduces a max-margin semi-supervised learning procedure for it. Another example is [3], which uses a trained mixture of part-based models for pose estimation and action classification.

Part-based models were used to constrain segmentation before. One example is [16], which proposes a description of object shape that is quite similar to ours. In fact, it is even more powerful since we don't allow width to vary along the edges of graph representing shape. However, this limitation makes it possible to build an algorithm for exact inference in our model. Another similar work is [12], which proposes a two-step algorithm for human segmentation. The proposed algorithm first tries to find the most plausible configuration of a part-based human model given an image using MCMC, and then uses it to constrain the segmentation. We think, however, that pixel labelings induced by shapes have lots of information about the correctness of shape fitting, and, therefore, fitting should not be decoupled from the segmentation process. Another example of a work that uses part-based models for image segmentation is [8], which represents object shape via a layered pictorial structure and finds segmentation by using Monte Carlo EM combined with loopy BP, since the model is too complicated for exact inference.

Other approaches to shape-constrained segmentation that do not involve modeling of object parts are known in literature. One notable class of such techniques includes star-shape prior [14] and tightness prior [11]. Both come with an efficient segmentation procedure that has good optimality guarantees, but impose quite weak restrictions on object shape and, therefore, can be of limited utility when segmenting certain kinds of objects. Another group of works includes those trying to describe shape via a hard mask [15, 2], representation that we find to be improper for classes of objects with high shape variability. One notable member of this group is [10], which models object shape via a huge set of different masks and also provides a framework for exact inference in CRF with high-order terms resulting from such constraints. We've used this framework as a basis for the exact inference algorithm we propose in this paper.

1.2 Contribution

Main contributions of this paper are

- a flexible model of object shape, which is invariant to translation and rotation and allows to describe classes of forms with high variability;
- two inference procedures for the segmentation constrained with the proposed shape model. The first one is fast but approximate, while the other one is slower but capable of obtaining exact solutions;
- a new formulation of learning from weakly labeled data problem for adjusting the parameters of our shape-constrained segmentation model. As a weak labeling we only use object shape descriptions for each image without pixel-wise image labeling.

2 Shape-Constrained Binary Segmentation

We state the problem of image segmentation with shape constraints as the problem of finding minimum of the energy function

$$\begin{aligned} E(L,s) = f(s) + w_c^0 \sum_i (1-L_i)\phi_i^0 + w_s^0 \sum_i (1-L_i)\psi_i^0(s) + \\ + w_c^1 \sum_i L_i \phi_i^1 + w_s^1 \sum_i L_i \psi_i^1(s) + w_p \sum_{(i,j)\in\mathcal{N}} |L_i - L_j| \phi_{ij} \end{aligned} \tag{1}$$

w.r.t. variables $L_i \in \{0,1\}$ representing pixel labels and variable s describing object shape. Variable i indexes all pixels and $\mathcal{N}$ stands for the set of pairs of indices of neighboring image pixels. Pairwise terms ϕ_{ij}, which we require to be non-negative for reasons explained later, can be used to move object boundary towards the areas with high color gradient magnitude. There are two types of unary terms: constant $\phi_i^{0,1}$ that can be used to encode known color distributions of object and background, and $\psi_i^{0,1}$ that depend on s and, thus, allow us to relate shape descriptions with labeling configurations. Energy term $f(s)$ is used to penalize improbable shape descriptions. We give more information about terms involving variable s in the following sections of the paper. Term weights $w_c^{0,1}$, $w_s^{0,1}$ and w_p act as model parameters.

2.1 The Model of Object Shape

We describe object shape via a graph augmented with vertex positions and edge width values. One example of such description, which we would call *shape graph* from now, is shown in Fig. 1(a). By varying angles, lengths, and widths in graphs we can obtain different shape variations. However incidence relation in the graph is fixed so all shape variations have similar structure.

More formally, shape graph is a tuple $(\mathcal{E}_s, \mathcal{V}_s)$, where

$$\begin{aligned} \mathcal{E}_s = \{e_k^s\}, \quad e_k^s = \left(f_k^s, t_k^s, b_k^s\right), \\ \mathcal{V}_s = \{v_l^s\}, \quad v_l^s = \left(x_l^s, y_l^s\right). \end{aligned}$$

Here $\mathcal{E}_s$ stands for the set of edges of shape graph s, in which every edge e_k^s has width b_k^s and connects vertices with indices f_k^s t_k^s. The set of graph vertices is denoted as $\mathcal{V}_s$. Vertex v_l^s has coords (x_l^s, y_l^s) on the image.

Each shape graph has an associated value of the shape energy

$$\begin{aligned} f(s) = w_r \left(\frac{\|e_r^s\|}{l_r} - 1 \right)^2 + \sum_k w_k^b \left(\frac{b_k^s}{\|e_k^s\|} - \rho_k^b \right)^2 + \\ + \sum_{(k_1,k_2)\in\mathcal{M}} w_{k_1,k_2}^l \left(\|e_{k_1}^s\| - \rho_{k_1,k_2}^l \|e_{k_2}^s\| \right)^2 + \\ + \sum_{(k_1,k_2)\in\mathcal{M}} w_{k_1,k_2}^\alpha \min \Big\{ \big(\angle(e_{k_1}^s, e_{k_2}^s) - \alpha_{k_1,k_2} + 2\pi\big)^2, \\ \big(\angle(e_{k_1}^s, e_{k_2}^s) - \alpha_{k_1,k_2} - 2\pi\big)^2, \big(\angle(e_{k_1}^s, e_{k_2}^s) - \alpha_{k_1,k_2}\big)^2 \Big\}, \end{aligned} \tag{2}$$

where $\mathcal{M}$ stands for the set of pairs of edge indices involved in pairwise constraints,

$$\|e_k^s\| = \|v_{f_k}^s - v_{t_k}^s\| \tag{3}$$

is the length of edge e_k^s, and

$$\angle(e_{k_1}^s, e_{k_2}^s) = \angle\big((v_{f_{k_1}}^s, v_{t_{k_1}}^s), (v_{f_{k_2}}^s, v_{t_{k_2}}^s)\big) \tag{4}$$

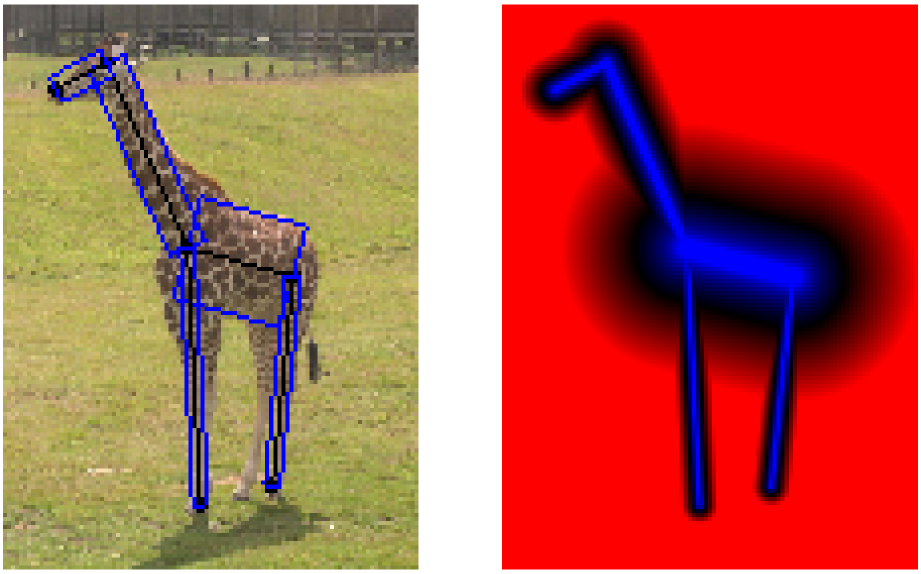
is the directed angle between edges with indices k_1 and k_2. Parameter ρ_{k_1,k_2}^l stands for the mean length ratio between edges with indices k_1 and k_2, while w_{k_1,k_2}^l measures softness of the constraint. Parameters α_{k_1,k_2}, w_{k_1,k_2}^α, ρ_k^b, w_k^b have analogous meaning for angles between edges and width-to-length ratios respectively. Parameters l_r and w_r are used to specify the scale of the so-called *root* edge e_r^s, thus enforcing the scale of the whole model through pairwise edge length constraints.

The form of angle penalty terms is justified by the fact that the difference of two angles inside the $[-\pi, \pi]$ range can lie outside this range and, thus, can have an alternative representation with lower penalty inside.

The idea behind this energy function is to penalize uncommon shapes while allowing different object parts to have different variability. For example, a shape model aiming to describe both running and standing horses can enforce quite soft constraints on the angle between horse body and legs compared to the angle between its neck and head.

Since the shape energy terms depend only on edge length, width and angles between edges, the energy is invariant to both rotation and translation. However such information, if available, can be easily incorporated into the energy function as a constraint on the root edge, as it was done for scale.

The reason this model is not invariant to scale lies in the form of pairwise terms enforcing constraints on lengths of neighboring edges. These terms could have been made scale-invariant by replacing length difference by ratio difference as it was done for terms involving edge width, but then we won't be able to apply the exact inference algorithm described in Sect. 2.4 to our problem.



(a) A photo of a giraffe with a shape graph modeling its shape.

(b) $\psi_i^1(s) - \psi_i^0(s)$ induced by the graph on the left. Positive values are colored red.

Fig. 1: An illustration of a graph-based shape model.

2.2 Relation to pixel labeling

In order to complete the description of the proposed shape model, we need to specify the way it affects image labeling. It is natural to assume that pixels located near some edge of the shape graph will probably belong to object, while pixels that are far from any edge will most likely belong to background. Based on this assumption we define shape-based unary terms for pixel i as

$$\psi_i^1(s) = \frac{1}{|I|} \min_{e_k^s \in \mathcal{E}_s} \beta(b_k^s) d^2[p_i, (v_{f_k}^s, v_{t_k}^s)], \tag{5}$$

$$\psi_i^0(s) = \frac{1}{|I|} \max_{e_k^s \in \mathcal{E}_s} -\log\left[1 - \exp\left(- \beta(b_k^s) d^2[p_i, (v_{f_k}^s, v_{t_k}^s)]\right)\right], \tag{6}$$

$$\beta(b) = \frac{4 \log 2}{b^2}, \tag{7}$$

where $d[p, (v_1, v_2)])$ is the distance from a point p to a segment (v_1, v_2), p_i represents the coordinates of the i-th pixel of an image and $|I|$ is the total number of pixels. Reasons for scaling terms by inverse image size are explained in Sect. 4.2. Form of the coefficient $\beta(b)$ is justified by the requirement that object and background labels should be equally likely at half-width distance from an edge. An example of shape terms induced by a shape graph can be seen in Fig. 1(b).

2.3 Approximate inference

Energy function (1) has a nice property: if the shape description s is fixed and $\phi_{ij} \geq 0$ for every pair of neighboring pixels, its minimum can be efficiently computed via graph cuts [1]. It allows us to cast the problem of minimizing $E(L, s)$ to the problem of finding

$$\min_s F(s) = \min_s \min_L E(L, s). \tag{8}$$

Function $F(s)$ has quite a few parameters compared to the original energy function, can be efficiently computed in each point, and, thus, can be minimized using some derivative-free optimization technique.

In this work we use simulated annealing as a minimization technique. It was selected mostly due to its natural applicability to the graph-based shape description. We use the following transition moves:

- randomly changing the length of a random edge e_k^s by sampling l^{new} from the truncated Gaussian $\mathcal{N}_{l \geq l^{min}}(l; \|e_k^s\|, \sigma_l^2 T^2)$ and shifting $v_{t_k^s}$ so that l^{new} is the new length of e_k^s (which, in turn, changes the configuration of all other edges incident to $v_{t_k^s}$);
- randomly changing the width of a random edge e_k^s by sampling b_k^s from $\mathcal{N}_{b \geq b^{min}}(b; b_k^s, \sigma_b^2 T^2)$;
- randomly changing the angle between a random edge e_k^s and a fixed direction by sampling a new angle from $\mathcal{N}(\alpha; \angle e_k^s, \sigma_\alpha^s T^2)$ and shifting $v_{t_k^s}$ accordingly (again, it will change the configuration of all neighboring edges);
- applying random amount of translation, rotation or scale to the whole shape graph. Amounts of transformation were sampled from Gaussian distributions (truncated Gaussian in the case of scale) in the same fashion as above, with standard deviation being proportional to T. Rotation and scale transforms use the mean of vertex coordinates as the origin.

The current annealing temperature, T, was set to $\frac{1}{\log_2(n+1)}$ on the n-th annealing iteration. We've used $\min\left[1, \exp\left(\frac{F(s_{best}) - F(s)}{T}\right)\right]$ as the acceptance probability for the proposed shape graph s, where s_{best} is the graph with the lowest value of $F(s)$ found so far.

We were able to obtain quite good local minima starting the optimization process from a *mean* shape graph automatically fitted into image bounds. It was produced by first building a graph with the lowest possible energy in which root edge has the specified direction, and then shifting the graph so that its center corresponds to the image center. If the expected object orientation was known, we used it as the orientation of a fitted graph. Otherwise, we tried to start the process from a number of mean shape graphs fitted with different orientations.

2.4 Exact inference

The exact inference procedure for the discussed model is built upon the branch-and-mincut framework [10]. This framework aims to minimize an energy function of form (1) via a breadth-first branch-and-bound procedure, which uses the expression

$$\begin{aligned}\min_L \Bigg[f^S + w_c^0 \sum_i (1 - L_i)\phi_i^0 + w_s^0 \sum_i (1 - L_i)\psi_i^{0,S} + \\ + w_c^1 \sum_i L_i \phi_i^1 + w_s^1 \sum_i L_i \psi_i^{1,S} + w_p \sum_{(i,j)\in\mathcal{N}} |L_i - L_j|\phi_{ij} \Bigg],\end{aligned} \tag{9}$$

where

$$f^S = \min_{s\in S} f(s), \quad \psi_i^{0,S} = \min_{s\in S} \psi_i^0(s), \quad \psi_i^{1,S} = \min_{s\in S} \psi_i^1(s), \tag{10}$$

to bound below the minimum of the energy when s is constrained to be in S. This lower bound can be efficiently computed for any set S via graph cuts if the *aggregated potentials* f^S, $\{\psi_i^{0,S}\}$ and $\{\psi_i^{1,S}\}$ are known. In order to apply this framework to our model we need to provide a way to describe a set of shape graphs S together with a subdivision scheme for it. We also need to provide an efficient algorithm for computing aggregated potentials for any given set S.

We choose to represent S as a set of axis-aligned bounding boxes (AABB) limiting possible positions of shape graph vertices, together with a set of one-dimensional ranges limiting the width of each edge. More formally,

$$\begin{gathered} s \in S \iff \forall k \; b_k^s \in B_k^S, \; \forall l \; v_l^s \in V_l^S, \\ B_k^S = [b_k^{S,min}, b_k^{S,max}], V_l^S = [x_l^{S,min}, x_l^{S,max}] \times [y_l^{S,min}, y_l^{S,max}]. \end{gathered} \tag{11}$$

A natural subdivision scheme for this representation is to either split one of the vertex constraints in four, or split some edge width constraint in two until the constraints become singletons.

In order to compute $\psi_i^{1,S}$ we first note that

$$\min_{s\in S} \psi_i^1(s) = \frac{1}{|I|} \min_k \min_{b_k^s \in B_k^S} \beta(b) \min_{v_{f_k}^s \in V_{f_k}^S} \min_{v_{t_k}^s \in V_{t_k}^S} d^2\big(p_i, (v_{f_k}^s, v_{t_k}^s)\big). \tag{12}$$

So, in order to compute this aggregated potential we need to find the closest location to the pixel i for every edge of s w.r.t. constraints on S, and then take the minimum across all edges using the maximum possible width for every edge. Similar considerations apply to $\psi_i^{0,S}$, but in this case every edge should be taken away from pixel as far as possible, while its width should be made as small as possible. When edge constraints are given by AABB, finding closest (or farthest) location of an edge from a pixel is a simple problem that can be solved in constant time.

In order to compute f^S we first note that angles $\angle(e_{k_1}^s, e_{k_2}^s)$ in (2) can be rewritten as $\angle e_{k_2}^s - \angle e_{k_1}^s$, where $\angle e$ is an angle between an edge and some fixed direction. Then all pairwise terms of (2) will be quadratic w.r.t. values $\|e_{k_1}^s\|$, $\|e_{k_2}^s\|$, $\angle e_{k_1}^s$, $\angle e_{k_2}^s$. Thus, given $\mathcal{M}$ corresponds to a tree, the energy can be efficiently minimized via dynamic programming accelerated using generalized distance transform (GDT) technique [4], considering these values together with b_k as variables. Feasible sets for edge lengths and angles are given by V_l, while widths should be constrained by B_k. Note that the shape graph itself is not forced to be a tree, only the graph of pairwise edge constraints given by $\mathcal{M}$.

3 Learning from weakly-labeled data

In this section we describe a way to estimate the parameters of the shape-constrained binary segmentation model (1). We first explain our training problem formulation and the way we build training set, and then show a way to estimate non-linear and linear parameters of the model separately.

3.1 Training set

Since the energy (1) is a function of shape graph and pixel labeling, it is natural to use a set of images augmented with both ground truth labelings and graphs as a training set. However, true labelings are much more expensive to obtain in terms of human labor amount required to label one image compared to shape graphs. We aim to solve this issue by proposing a novel formulation of the segmentation model training problem. We note that the segmentation energy minimum, if expressed in a form (8), can be seen as the most plausible configuration of a shape graph on a given image, with pixel labels acting as latent variables of the model. We therefore state the problem of training our model as the problem of training shape graph fitting model given shape graph ground truth only. The hope is that the best labeling associated with the fitted graph corresponds to a meaningful segmentation. Experimental results shown later confirm that this is actually the case.

We denote the training set for our training problem as

$$(I^1, s^1), \ldots, (I^m, s^m), \tag{13}$$

where I^m is m-th training image, s_m is a ground truth shape graph for it and M is the total number of images in the training set.

3.2 Learning non-linear parameters

Our model has a number of non-linear parameters, r, l_r, $\{\rho^l_{k_1,k_2}\}$, $\{\alpha_{k_1,k_2}\}$, $\{\rho^b_k\}$, which we will together denote as θ from now. We propose to train these parameters before the rest of the model in order to simplify the procedure. Training objective is to choose θ that minimizes the total energy of the training set

$$\sum_{m=1}^{M} f(s^m; \theta), \tag{14}$$

where $f(s; \theta)$ is given by (2). While the shape energy also depends on the values of term weights, the optimal parameter values do not and are given by the following expressions:

$$l_k = \frac{\sum_{m=1}^{M} |e_k^{s^m}|^2}{\sum_{m=1}^{M} |e_k^{s^m}|}, \tag{15}$$

$$r = \arg\min_k \sum_{m=1}^{M} \Big(\frac{|e_k^{s_m}|}{l_k} - 1 \Big)^2, \tag{16}$$

$$\rho_k^b = \frac{1}{M} \sum_{m=1}^{M} \frac{b_k^{s_m}}{|e_k^{s_m}|}, \tag{17}$$

$$\rho_{k_1,k_2}^l = \sum_{m=1}^{M} \frac{|e_{k_1}^{s_m}||e_{k_2}^{s_m}|}{|e_{k_2}^{s_m}|^2}. \tag{18}$$

The estimation procedure for α_{k_1,k_2} is also trivial, but hard to write in a closed form since multiple angle representations should be considered when averaging angles.

3.3 Learning linear parameters

Discussed model has w_r, $\{w_{k_1,k_2}^l\}$, $\{w_{k_1,k_2}^\alpha\}$, $\{w_k^b\}$, $w_c^{0,1}$, $w_s^{0,1}$ and w_p as its linear parameters. Let us denote the vector containing all these parameters as w and introduce the vector $\Phi(I, L, s)$ containing negated energy terms corresponding to weights in w for a given image I with an associated labeling L and a shape graph s. Now the problem of segmenting image I can be stated as the problem of finding

$$\arg\min_{L,s} E(L, s) = \arg\max_{L,s} w^T \Phi(I, L, s). \tag{19}$$

Such a reformulation allows us to use the latent structural SVM formulation [17] for weight training. Thus, we want to find

$$\begin{aligned} \arg\min_w \Bigg[\frac{1}{2} w^T w + \frac{C}{M} \sum_{m=1}^{M} \Big\{ \max_{L,s} [w^T \Phi(I^m, L, s) + \Delta(s, s^m)] - \\ - \max_L w^T \Phi(I^m, L, s^m) \Big\} \Bigg]. \end{aligned} \tag{20}$$

The idea behind this objective is to enforce a ground truth shape graph together with its best possible labeling have lesser energy value than any other shape graph and labeling. Loss function $\Delta(s_1, s_2)$ is used to rescale the energy margin depending on how close the two graphs are: graphs similar to the ground truth should also have low energy, while those significantly distinct from the ground truth should have much higher energy values. The loss function we use is of form

$$\Delta(s_1, s_2) = \sum_l \min\{\|v_l^{s_1} - v_l^{s_2}\|, t_v\} + \lambda \sum_k \min\{|b_k^{s_1} - b_k^{s_2}|, t_e\}. \tag{21}$$

The motivation behind truncating the loss is that when some graph is too much apart from the ground truth, one should not really care about the value of margin as soon as it's large.

Convex-concave procedure (CCCP) [18] can be used to minimize (20), which would result in an iterative procedure with two-step iterations. On each CCCP iteration the expression $w^T \Phi(I^m, L, s^m)$ is first maximized w.r.t. L for each m in order to obtain $\tilde{L}^m$, new optimal values of the latent variables. For the model considered in this paper it can be done via a graph cut. Obtained $\tilde{L}^m$ are then substituted into (20), which results in a regular structural SVM problem

$$\begin{gathered}\frac{1}{2} w^T w + \frac{C}{M} \sum_{m=1}^{M} \Big\{ \max_{L,s} [w^T \Phi(I^m, L, s) + \Delta(s, s^m)] - \\ - w^T \Phi(I^m, \tilde{L}^m, s^m) \Big\} \to \min_{w}\end{gathered} \tag{22}$$

that should be solved in order to update weights.

In order to solve (22) we employ the cutting plane algorithm [7] in which most violated constraints $\arg\max_{L,s} [w^T \Phi(I^m, L, s) + \Delta(s, s^m)]$ are found by the inference algorithm described in Sect. 2.3. Loss function can be easily incorporated into it, since the annealing only requires the ability to evaluate the objective function at a point. We've found it useful to start the annealing from both a mean shape graph fitted into image and a ground truth shape graph and then choose the best solution among two. While we are unable to obtain exact solutions to the SSVM problem this way due to the usage of an approximate procedure for finding most violated constraints, weights we've found on each iteration were close enough to global optima for CCCP to converge.

4 Experiments

4.1 Datasets

In order to validate the model proposed in this paper, we've trained it on two different datasets. The first dataset was build from a specific subset of ETHZ shape classes dataset [6] that contained images of giraffes. It featured images that are hard to segment using low-level cues only due to weak boundaries and significantly overlapping color distributions of object and background pixels. Meanwhile, giraffe shape has a simple structure and can be described with a graph-based model well. Another dataset was build from synthetic images of capital "E" letter that feature totally non-discriminative object and background color distributions, and, so, the only information segmentation model can rely on is object shape and boundaries. All the images were downscaled to size of about 140×140 pixels in order to speedup the training process.

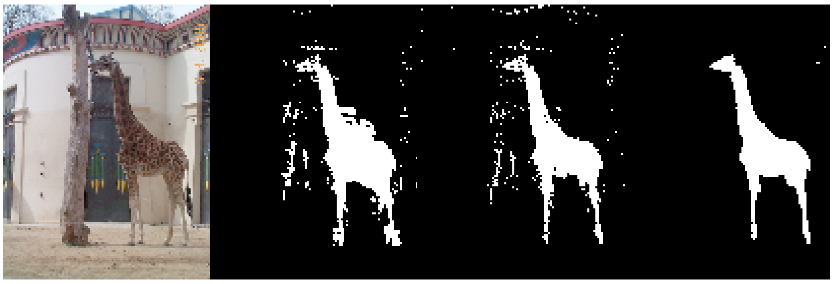


Fig. 2: Convergence of latent variables during training.

4.2 Color-based terms

In our experiments we used unary color terms of the form

$$\phi_i^l = -\frac{1}{|I|} \log P(I_i \mid L_i = l), \ l \in \{0, 1\}. \tag{23}$$

Color distributions for object and background were represented by 3-component GMMs in RGB space, which were learned from seeds placed on a few training images. Pairwise terms we used were of the form

$$\phi_{ij} = \frac{1}{|I|^{\frac{1}{2}}} e^{-\alpha \frac{(\|I_i\| - \|I_j\|)^2}{D^2}}, \tag{24}$$

where $\|I_i\|$ is brightness value of i-th pixel of image I and D is the mean brightness difference value for all the pixel pairs from $\mathcal{N}$ in that image. Value of α was set to 0.2. Color-based terms were scaled according to image size in the same way as (5) and (6) to make model features independent of image size.

4.3 Model training

Weight vector with $w_c^0 = w_c^1 = 1$, $w_s^0 = w_s^1 = 0.3$, $w_p = 0.001$ and all other components set to 0 was used to initialize the training process. Constant C in latent SSVM objective (20) was set to 300. Parameters of the loss function (21) were set as follows: $\lambda = 10$, $t_v = 0.25|I|^{\frac{1}{2}}$, $t_e = 0.1|I|^{\frac{1}{2}}$. Training process usually converged in 5-7 iterations, each iteration requiring about 70-100 cutting planes.

The most interesting observation about the training process is that the latent variables $\tilde{L}^k$ found on each iteration of CCCP tend to converge to ground truth labelings (see Fig. 2 for an example). Thus, true pixel labels are actually not required for learning since they can be closely approximated during the training process.

4.4 Segmentation

We've then applied trained models to images not involved in the training process using the proposed approximate inference algorithm. Some images together

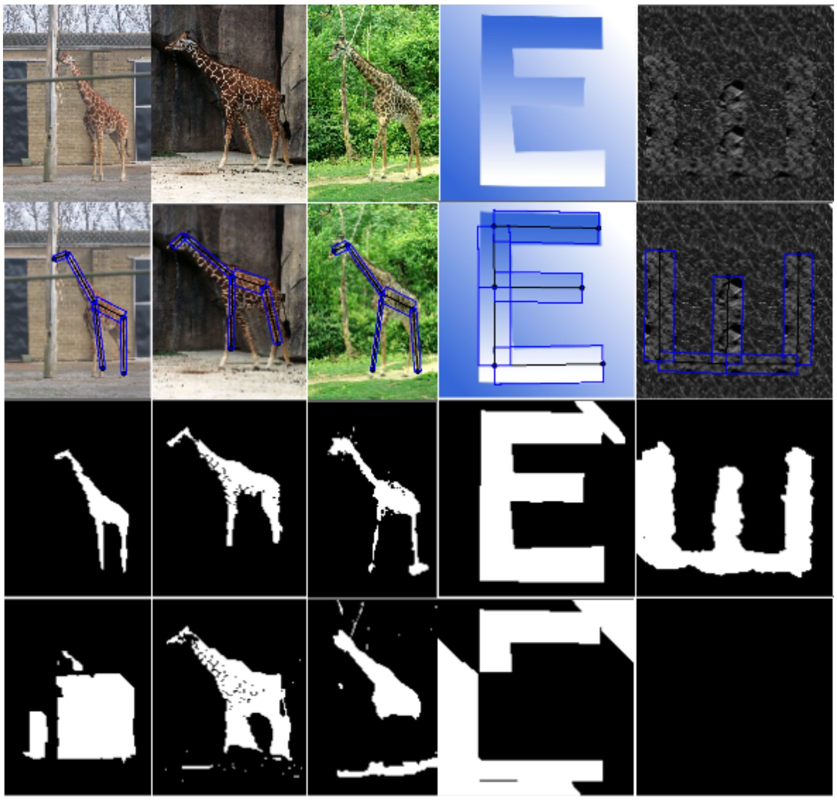


Fig. 3: Top to bottom: image, found shape graph, shape-constrained segmentation, color-based segmentation.

with the best found shape graph and labeling configuration can be seen in Fig. 3. Segmentations based solely on color-based terms (both unary and pairwise) are provided for comparison. It can be seen that our algorithm is able to fit shape graph into an image correctly. Fitted graph induces quite reasonable segmentation of an image, much better than the one obtained from color-based terms only.

In order to compare the approximate inference algorithm with the exact one, we've applied the latter to same images. As a result, we found out that in many cases approximate algorithm is able to find solutions very close to global optima. However, it's not always the case. One example where approximate inference fails while exact succeeds is shown in Fig. 4. Unfortunately, our exact algorithm may took a lot of time (hours sometimes) to converge despite its very efficient implementation (utilization of GPU for calculation of aggregated potentials, extensive caching). The reason is that the bound (9) is not tight, and, thus, lots of lower bound computations are needed before a good subset of solution space would be discovered. Many of images in the test set require a few

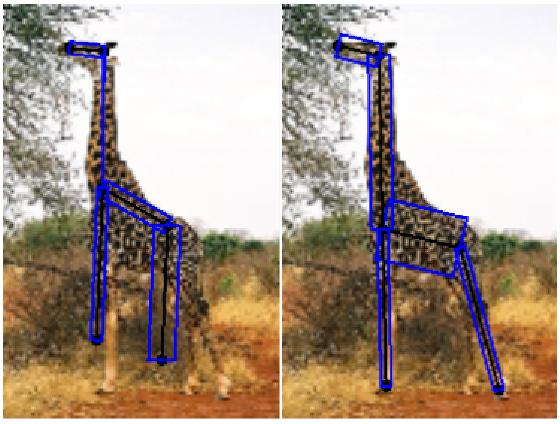


Fig. 4: Solution found by simulated annealing (left) vs branch-and-bound.

millions of lower bound computations. There are also some images that require too many lower bound computations to discover the solution, so we were not able to segment those using our exact inference algorithm. At the same time, the proposed approximate algorithm is quite fast. It usually takes less than a minute per image and can be further accelerated by reducing the number of annealing iterations or reannealing attempts.

5 Conclusion

In this paper we've proposed a model for shape-constrained binary segmentation of an image. The model emerges from combining regular CRF for binary segmentation with high-order terms based on a specific form of object shape description, namely graph-based shape representation. This representation is invariant to object rotation and translation and can describe classes of objects with complex shapes and high in-class variability. We present two inference algorithms for the proposed model. One, which is based on simulated annealing, is fast but approximate, while the other can obtain exact solutions via the branch-and-bound procedure.

We've also proposed a novel training formulation for our model, which requires only ground truth shape descriptions but no pixel-wise labelings to learn its parameters. Training procedures are provided for both linear and non-linear parameters of the model. Experiments on artificial as well as real-world images confirm that ground truth image labelings are indeed not required to learn a well-performing model.

One interesting direction of future work would be to combine the model proposed in this paper with state-of-the-art object part detectors, which output can be used as additional clues about possible positions of shape graph vertices. Another question is whether the exact inference algorithm we've presented can be significantly accelerated by, for example, tightening the lower bound (9).

References

1. Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Computer Vision, 2001. Proceedings. 8 IEEE International Conference on*, volume 1, pages 105–112. IEEE, 2001.
2. D. Cremers, F. Schmidt, and F. Barthel. Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–6. IEEE, 2008.
3. C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. *Computer Vision–ECCV 2012*, pages 158–172, 2012.
4. P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.
5. P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
6. V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *International Journal of Computer Vision*, 87(3):284–303, 2010.
7. T. Joachims, T. Finley, and C. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.
8. M. Kumar, P. Torr, and A. Zisserman. Obj cut. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 18–25. IEEE, 2005.
9. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.
10. V. Lempitsky, A. Blake, and C. Rother. Image segmentation by branch-and-mincut. *Computer Vision–ECCV 2008*, pages 15–29, 2008.
11. V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 277–284. IEEE, 2009.
12. I. Rauschert and R. Collins. A generative model for simultaneous estimation of human body shape and pixel-level segmentation. *Computer Vision–ECCV 2012*, pages 704–717, 2012.
13. C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.
14. O. Veksler. Star shape prior for graph-cut image segmentation. *Computer Vision–ECCV 2008*, pages 454–467, 2008.
15. N. Vu and B. Manjunath. Shape prior segmentation of multiple objects with graph cuts. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
16. B. Yangel and D. Vetrov. Image segmentation with a shape prior based on simplified skeleton. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 247–260. Springer, 2011.
17. C. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM, 2009.
18. A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.