SUPERPIXEL CUT FOR FIGURE-GROUND IMAGE SEGMENTATION

Michael Ying Yang[a], Bodo Rosenhahn[b]

[a] University of Twente, ITC Faculty, EOS department, Enschede, The Netherlands - yang@tnt.uni-hannover.de
[b] Leibniz University Hannover, Institute of Information Processing, Germany

Commission III,WG III/4

KEY WORDS: Computer Vision, Superpixel Cut, Min-Cut, Image Segmentation

ABSTRACT:

Figure-ground image segmentation has been a challenging problem in computer vision. Apart from the difficulties in establishing an effective framework to divide the image pixels into meaningful groups, the notions of figure and ground often need to be properly defined by providing either user inputs or object models. In this paper, we propose a novel graph-based segmentation framework, called superpixel cut. The key idea is to formulate foreground segmentation as finding a subset of superpixels that partitions a graph over superpixels. The problem is formulated as Min-Cut. Therefore, we propose a novel cost function that simultaneously minimizes the inter-class similarity while maximizing the intra-class similarity. This cost function is optimized using parametric programming. After a small learning step, our approach is fully automatic and fully bottom-up, which requires no high-level knowledge such as shape priors and scene content. It recovers coherent components of images, providing a set of multiscale hypotheses for high-level reasoning. We evaluate our proposed framework by comparing it to other generic figure-ground segmentation approaches. Our method achieves improved performance on state-of-the-art benchmark databases.

1. INTRODUCTION

Despite a variety of segmentation techniques have been proposed, figure-ground image segmentation remains challenging for any single method to do segmentation successfully due to the diversity and ambiguity in an image. The task is to produce a binary segmentation of the image, separating foreground objects from their background (Rother et al., 2004). In image segmentation, one has to consider a prohibitive number of possible pixel groupings that separate the figure from the background. Apart from the difficulties in establishing an effective framework to divide the image pixels into meaningful groups, the notions of figure and ground often need to be properly defined by providing either user inputs (Rother et al., 2004, Vicente et al., 2008) or object models. Prior knowledge about object appearance, or other scene content could significantly simplify the problem. For instance, many segmentation techniques are formulated as Markov random field based energy minimization problems that could be solved using Min-Cut in an efficient manner. However, the corresponding energy functions typically include terms that require prior object knowledge in terms of user interaction (Rother et al., 2004, Vicente et al., 2008) or knowledge about object appearance. A good figure-ground segmentation is a valuable input for many higher-level tasks. For example, object recognition (Belongie et al., 2002) benefits from segmentation as shape descriptors can be derived from segmentation results. One can consider segmentation as a necessary bottom-up preprocessing step for recognition or indexing, providing substantial reduction in the computational complexity of these tasks. It is therefore unclear how segmentation methods that use strong prior knowledge are applicable for object recognition from large databases.

In recent years an increasingly popular way to solve various image labeling problems like object segmentation, stereo and single view reconstruction is to formulate them using superpixels obtained from unsupervised segmentation algorithms (Li et al., 2004, Levinshtein et al., 2010, Brendel and Todorovic, 2010). For instance, they may belong to the same object or may have the same surface orientation. These methods are inspired from the observation that pixels constituting a particular superpixel often have the same label. This approach has the benefit that higher order features based on all the pixels constituting the superpixel can be computed (Yang et al., 2010). Further, it is also much faster for segmentation as inference now only needs to be performed over a small number of superpixels rather than all the pixels in the image (Yang and Förstner, 2011).

In this paper we address the figure-ground segmentation as a superpixel selection problem. Segmenting foreground is formulated as finding a subset of superpixels that partitions a graph over superpixels, with graph edges encoding superpixel similarity. Our approach has the following two important characters, which distinguish our work from most of the others: First, it is fully automatic, efficient, and requires no user input; Second, it is fully bottom-up, which requires no high-level knowledge such as shape priors and scene content.

1.1 Contributions

The main contributions of this paper are:

- We propose a novel graph-based segmentation framework, called superpixel cut, which formulates foreground segmentation as finding a subset of superpixels that partitions a graph over superpixels. Mathematically, we formulate it as Min-Cut, with a novel cost function that simultaneously minimizes the inter-class similarity while maximizing the intra-class similarity.
- We give proof for the proposed cost function with estimation of a lower bound and a upper bound. Then, this cost function is optimized by parametric programming.
- Finally, we achieve highly competitive results on the Weizmann Horse Database and the Berkeley Segmentation Data Set.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-III-3-387-2016

The illustration in Fig. 1 shows an overview of our approach for figure-ground segmentation. Given an image (Fig. 1(a)), we first extract image contours (Fig. 1(b)). Meanwhile, we compute the superpixel segmentation of the image (Fig. 1(c)), in which the superpixel resolution is chosen to ensure that object boundaries are reasonably well approximated by superpixel boundaries. Based on the contour image and superpixel image, figure-ground segmentation is performed as a superpixel selection problem (Fig. 1(d)).

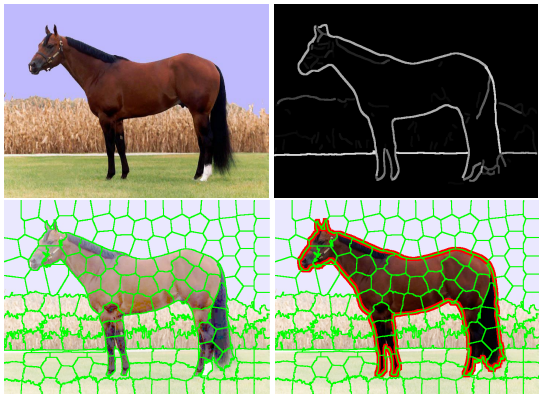

Figure 1: Overview of our approach for figure-ground segmentation: (a) original image; (b) contour image (Pb-detector); (c) superpixel segmentation (SLIC), in which superpixel resolution is chosen to ensure that target boundaries are reasonably well approximated by superpixel boundaries; (d) figure-ground segmentation as a superpixel selection problem (the red boundary overlays with the superpixel segmentation image for visualization).

The following sections are organized as follows. The related works are discussed in Section 2.. In Section 3., the model for the segmentation problem is formulated. In Section 4., experimental results are presented. Finally, this work is concluded and future work is discussed in Section 5..

2. RELATED WORK

A number of figure-ground segmentation methods have been recently pursued. Interactive segmentation (Rother et al., 2004, Vicente et al., 2008) has been thoroughly researched since the very popular GrabCut work (Rother et al., 2004). Most of these approaches minimize a binary pairwise energy function (Boykov and Jolly, 2001) whose unary potentials are determined by appearance models estimated based on user input on the test image. Bagon et al. (Bagon et al., 2008) proposed an algorithm that generates figure-ground segmentations by maximizing a self-similarity criterion around a user selected image point. While the method of (Küttel and Ferrari, 2012) is also based on minimizing an energy function of the same form as interactive segmentation, it is fully automatic. The unary potentials of the energy function of the test image is derived from the transferred segmentation masks. Similarly, Carreira and Sminchisescu (Carreira and Sminchisescu, 2012) extracted multiple figure-ground hypotheses based on energy minimization using parametric Min-Cut and learned to score them using region and Gestalt-based features. Bertelli et al. (Bertelli et al., 2011) presented a supervised learning approach for segmentation using kernelized structural SVM. By designing non-linear kernel functions, high-level object similarity information is integrated with multiple low-level segmentation cues. In contrast, we only exploit the bottom-up approach without any high-level knowledge in this paper. Joulin et al. (Joulin et al., 2010) presented a discriminative clustering framework for image cosegmentation. Foreground (background) labels are assigned jointly to all images, so that a supervised classifier trained with these labels leads to maximal separation of the two classes. Our method only uses features derived from a single image. The second category of approaches uses superpixels, which have been exploited to aid segmentation recently. In most cases, they are used to initialize segmentation. Malisiewicz and Efros (Malisiewicz and Efros, 2007) showed that superpixels with good object overlap could be obtained by merging pairs and triplets of superpixels from multiregion segmentations, but at the expense of generating also a large quantity of implausible ones. Endres and Hoiem (Endres and Hoiem, 2010) generated multiple proposals by varying the parameters of a conditional random field built over a superpixel graph.

Segmenting figure in an image has been addressed by many researchers in different ways. Contour grouping methods naturally lead to figure-ground segmentation (Kennedy et al., 2011). Computing closure can be achieved by using only weak shape priors, such as compactness, continuity and proximity. The most basic closure based cost function uses a notion of boundary gap, which is a measure of missing image edges along the closed contour. Wang et al. (Wang et al., 2005) optimized a measure of average gap using the ratio cut approach (Wang and Siskind, 2003). However, a measure based purely on the total boundary gap is insufficient for perceptual closure. Ren et al. (Ren et al., 2005) presented a model of curvilinear grouping using piecewise linear representations of contours and a conditional random field to capture continuity and the frequency of different junction types. Stahl and Wang (Stahl and Wang, 2007) gave a grouping cost function in a ratio form, where the numerator measures the boundary proximity of the resulting structure and the denominator measures the area.

The previous work most related to ours is Levinshtein et al. (Levinshtein et al., 2010), which transformed the problem of finding contour closure to finding subsets of superpixels. They defined the cost function as a ratio of a boundary gap measure to area, which promotes spatially coherent sets of superpixels. Image contour closure is extended to include spatiotemporal closure in (Levinshtein et al., 2012). Inspired by their approach, we define our similarity measure as boundary gap. However, as we will argue in Section 3., our proposed cost function, which minimizes the inter-class similarity and maximizing the intra-class similarity, is a more reasonable cost function than the closure cost proposed in (Levinshtein et al., 2010, Levinshtein et al., 2012).

3. PROBLEM FORMULATION

In this section, we propose a novel graph-based segmentation framework. We formulate figure-ground segmentation as a superpixel selection problem. Segmenting foreground is formulated as finding a subset of superpixels that partitions a graph over superpixels, with graph edges encoding superpixel similarity. Mathematically, we formulate it as Min-Cut, then we propose a more reasonable cost function that simultaneously minimizes the inter-class similarity while maximizing the intra-class similarity. Our framework reduces grouping complexity from an exponential number of superpixel subsets by restricting foreground boundary to lie along superpixel boundaries. We derive a lower bound and a upper bound for the proposed cost function. To solve the optimization problem, this cost function is converted to a parametric programming and solved approximately by parametric Max-Flow algorithm.

3.1 Cost function and graph construction

We construct a graph over superpixels of an image I, as shown in Figure 2. Superpixels are generated by some unsupervised segmentation algorithms, such as NCut (Shi and Malik, 2000), gPb-OWT-UCM (Arbelaez et al., 2011), SLIC (Achanta et al., 2012), etc. Formally, let $G = (V, E)$ be a graph with node set V

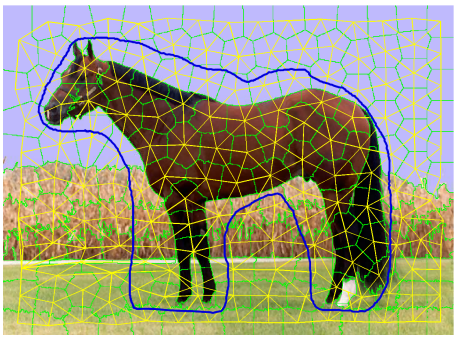

Figure 2: The graph model with superpixel segmentation of an image. Every superpixel is represented by a node, with the green curves as superpixel boundaries, the yellow lines as the edges connecting adjacency nodes. The blue curve represents a cut of the graph.

corresponding to the superpixels and E corresponding to graph edges, where $V = \{v_i\}_{i=1}^{n}$. We further define an edge weight w_{ij} to encode the similarity between two superpixels v_i and v_j in the image that are connected by an edge. $w_{ij} = 0$ if superpixels v_i and v_j have no edge between them. The weight matrix $W = (w_{ij})_{n\times n}$ is an affinity matrix between superpixels, which is symmetric.

Given the above graph $G = (V, E)$, the task is to partition it into 2 groups, namely figure and ground [1]. Various techniques can be employed for such a task, such as graph cut and spectral clustering (Shi and Malik, 2000). We are interested in determining a partition of the graph such that the sum of the edge weights from E that have one endpoint in each set is minimal, also called the Min-Cut. If we let $x_i \in \{0, 1\}$ be a binary indicator variable for the superpixel v_i, the vector $\mathbf{x}$ yields a full labeling of superpixels of I as foreground (1) or background (0). Min-Cut can be naturally posed as the minimization of the integer quadratic program

$$\begin{aligned} \min \quad & \textstyle\sum_{(i,j)\in E} w_{ij}x_i(1-x_j), \\ \text{s.t.} \quad & x_i \in \{0,1\}, i = 1, \cdots, n. \end{aligned} \tag{1}$$

However, minimizing the above cut alone unfairly penalizes larger superpixel selections. Previously, (Shi and Malik, 2000, Stahl and Wang, 2007) show that normalization of the cut by a measure that is proportional to the size of the selection yields better results. (Levinshtein et al., 2010, Levinshtein et al., 2012) also adopt this measure and try to minimize a cost that is a ratio of a cut to a selection size. However, foreground may not be chosen optimally in complex scenes due to the background clutter or similarity of foreground and background. It turns out that even using these cuts, one cannot simultaneously minimize the inter-class similarity while maximizing the similarity within the groups. Therefore, a better cost function would seek to minimize the inter-class similarity and at the same time it seeks to maximize the intra-class similarity. Optimizing this cost over superpixels enables us to efficiently recover coherent foreground segments out of an exponential number of superpixel subsets. Our cost is defined as follows, which we call the superpixel cut [2],

$$\begin{aligned} C(\mathbf{x}) &= \frac{P(\mathbf{x})}{Q(\mathbf{x})} \\ &= \frac{\sum_{(i,j)\in E} w_{ij}x_i(1-x_j)}{\sum_{(i,j)\in E} w_{ij}x_ix_j + \sum_{(i,j)\in E} w_{ij}(1-x_i)(1-x_j)} \end{aligned} \tag{2}$$

where $P(\mathbf{x})$ is the sum of the affinities of all the graph edges between selected ($x_i = 1$) and unselected ($x_i = 0$) superpixels, and $Q(\mathbf{x})$ is the sum of the affinities of all the edges except the cut edges. The first term of $Q(\mathbf{x})$ is the sum of the affinities of all the edges within selected ($x_i = 1$) superpixels, and the second term of $Q(\mathbf{x})$ is the sum of the affinities of all the edges within unselected ($x_i = 0$) superpixels.

Minimizing the ratio $C(\mathbf{x})$ is equivalent to minimizing the numerator $P(\mathbf{x})$ (Min-Cut) while maximizing the denominator $Q(\mathbf{x})$. Note that the numerator $P(\mathbf{x})$ and the denominator $Q(\mathbf{x})$ does not sum up to a constant. So minimizing the ratio $C(\mathbf{x})$ is not equivalent to the minimization of $P(\mathbf{x})$ (Min-Cut). Replacing $Q(\mathbf{x})$ by $P'(\mathbf{x})$ in Eq. (2) results $C'(\mathbf{x})$, which is the ratio-cut (Wang and Siskind, 2003). Here $P'(\mathbf{x})$ is the Min-Cut with different weight matrix. Our cost function in Eq. (2) tries to obtain a partition where the weight of the partition is directly proportional to the sum of the weights on the edges between the two partite sets and the sum of the reciprocals of the weights on the edges inside the partite sets. Naturally, the cut between foreground and background superpixels is small when foreground superpixels are strongly dissimilar from the background.

Let $D_i = \sum_{j=1}^{n} w_{ij}$, due to the symmetry of W, Eq. (2) is reformulated as

$$C(\mathbf{x}) = \frac{\sum_{i=1}^{n} D_ix_i - 2\sum_{i<j,(i,j)\in E} w_{ij}x_ix_j}{A - 2\sum_{i=1}^{n} D_ix_i + 4\sum_{i<j,(i,j)\in E} w_{ij}x_ix_j} \tag{3}$$

where $A = \sum_{(i,j)\in E} w_{ij}$ is a constant as W is given. Section 3.3 will provide details on the settings for W that we used for figure-ground segmentation. Note that in the current form Eq. (2) has a trivial solution by setting $\mathbf{x}$ to be vector $\mathbf{1}$ (all foreground). This issue can be resolved by penalizing some superpixels, as proposed in (Levinshtein et al., 2010). Specific details on this penalty will be given in Section 3.3.

Theoretical analysis This paragraph presents the proof that our proposed cost function has a upper bound as half of Normalized Cut (Shi and Malik, 2000). We have the following proposition.

Proposition 1 *The cost function in Eq. (2) has a lower bound as 0 and a upper bound as half of Normalized Cut (Shi and Malik, 2000).*

$$0 \le C(\boldsymbol{x}) \le \frac{1}{2}Ncut \tag{4}$$

[1] In this paper, the notions of figure (ground) and foreground (background) are identically.

[2] Intuitively, this cost function could also be used for (hierarchical) clustering, which is out of the scope of this paper.

Lower bound For $x_i \in \{0, 1\}$, it is trivial that both the numerator $P(\mathbf{x})$ and the denominator $Q(\mathbf{x})$ in Eq. (2) are greater than 0. When $\mathbf{x} = \mathbf{1}$ or $\mathbf{x} = \mathbf{0}$, the lower bound is *tight*, $C(\mathbf{x}) = 0$.
Upper bound Let $a \in \Re^+$ and $b \in \Re^+$, we have $2ab \leq a^2 + b^2$. So, $\frac{2}{a+b} \leq \frac{a+b}{ab} = \frac{1}{a} + \frac{1}{b}$. Let $a = \sum_{(i,j)\in E} w_{ij} x_i x_j$, and $b = \sum_{(i,j)\in E} w_{ij}(1 - x_i)(1 - x_j)$, it then follows that $C(\mathbf{x}) = \frac{P(\mathbf{X})}{a+b} \leq \frac{1}{2}(\frac{P(\mathbf{X})}{a} + \frac{P(\mathbf{X})}{b})$. Recall that Normalized Cut in (Shi and Malik, 2000) has the following cost function $\frac{P(\mathbf{X})}{a} + \frac{P(\mathbf{X})}{b}$. Therefore, $C(\mathbf{x}) \leq \frac{1}{2}$Ncut holds. When $a = b$, the upper bound is *tight*, $C(\mathbf{x}) = \frac{1}{2}$Ncut.

As **consequence** of this proof, our proposed cost is at least as good as Normalized Cut.

3.2 Optimization using parametric Max-Flow

To solve the optimization problem of Eq. (2), a common approach in fractional optimization is to minimize a parametrized difference $E(\mathbf{x}, \lambda) = P(\mathbf{x}) - \lambda Q(\mathbf{x})$, instead of minimizing the ratio $C(\mathbf{x}) = \frac{P(\mathbf{X})}{Q(\mathbf{X})}$ directly. It is shown in Appendix the optimal λ corresponds to the optimal ratio $\frac{P(\mathbf{X})}{Q(\mathbf{X})}$. The optimal λ can be efficiently recovered using a binary search or Newton's method for fractional optimization (Kolmogorov et al., 2007). The constraints on the ratio guarantee that the resulting difference is concave and thus can be minimized globally.

In the case of binary variables, ratio minimization can be reduced to solving a parametric Max-Flow problem. Kolmogorov et al. (Kolmogorov et al., 2007) showed that under certain constraints on the ratio $C(\mathbf{x})$, the energy $E(\mathbf{x}, \lambda)$ is submodular and can thus be minimized globally in polynomial time using Min-Cut. Converting our cost $C(\mathbf{x})$ in Eq. (2) to a parametric programming results in

$$E(\mathbf{x}, \lambda) = P(\mathbf{x}) - \lambda Q(\mathbf{x})$$
$$= -A\lambda + (1 + 2\lambda) \sum_{i=1}^{n} D_i x_i - (2 + 4\lambda) \sum_{i<j,(i,j)\in E} w_{ij} x_i x_j \quad (5)$$
$$\approx -A\lambda + (1 + 2\lambda) \sum_{i=1}^{n} D_i x_i - 2 \sum_{i<j,(i,j)\in E} w_{ij} x_i x_j \quad (6)$$

Because of λ involving in the quadratic term in Eq. (5), the method of parametric Max-Flow in (Kolmogorov et al., 2007) is not directly applicable for minimizing $E(\mathbf{x}, \lambda)$. In this paper, we omit the quadratic term involving λ in Eq. (5), and apply a parametric Max-Flow algorithm (Kolmogorov et al., 2007) to solve the optimization in Eq. (6). The parametric Max-Flow algorithm in (Kolmogorov et al., 2007) does not only optimize the ratio, but also finds all intervals of λ (also the corresponding $\mathbf{x}$) for which $\mathbf{x}$ remains constant. The parametric Max-Flow can optimize the above parametric programming in Eq. (6), and efficiently find all the different breakpoints (interval boundaries) of with stationary optimal solution $\mathbf{x}$, resulting in a monotonically increasing sequence of breakpoints, also yielding a set of K solutions. We refer the reader to (Kolmogorov et al., 2007) for more details on parametric Max-Flow algorithm.

3.3 Choice of weight matrix

The weight w_{ij} encodes the similarity between two superpixels that are connected by an edge. Following the work of (Levinshtein et al., 2010, Levinshtein et al., 2012), we define the boundary gap between two superpixels as the similarity measure in this paper. In principle, any similarity measures could be applied, which are often derived from the superpixel features.

The boundary gap is a measure of the disagreement between the boundary of an image and is defined as $w_{ij} = g_{ij} - h_{ij}$, where g_{ij} is the boundary length and h_{ij} is the *edginess* of the boundary. Specifically, if s_{ij} is the set of pixels on the boundary between superpixel node v_i and v_j, then $g_{ij} = |s_{ij}|$, and $h_{ij} = \sum_{p\in s_{ij}} h^p_{ij}$, where $h^p_{ij} = [Pr(\mathbf{f}^p) > T_e]$ is an edge indicator for pixel p. $Pr(\cdot)$ is a logistic regressor and $\mathbf{f}^p$ is a feature vector for pixel p. T_e is the parameter of the edginess measure, which controls the contribution of weak edges from the contour image (Levinshtein et al., 2010). Note that if two superpixels do not have a shared boundary, then both g_{ij} and h_{ij} will be 0. Then w_{ij} will also be 0, indicating that superpixel nodes v_i and v_j have no edge in the graph. In addition, superpixels that touch the image boundary incur a natural penalty because all image boundary pixels have 0 edginess (Levinshtein et al., 2010). As a result, cuts are discouraged from touching the image boundary thereby avoiding the trivial solution discussed in Section 3.1.

Given a pixel p on the superpixel boundary, the feature vector $\mathbf{f}^p$ is a function of both the local geometry of the superpixel boundary and the detected image edge response in its neighborhood. The feature vector consists of three components: (a) distance to the nearest image edge; (b) strength of the nearest image edge; (c) alignment between the tangent to the superpixel boundary point and the tangent to the nearest image edge. Given a dataset of images with manually labeled figure-ground masks, we map the ground-truth onto superpixels. Our training data consists of all pixels falling on superpixel boundaries where positive training data consists of pixels that fall on figure-ground boundaries and negative training data consists of all the other pixels on superpixel boundaries. It is used to train a logistic classifier over the feature vector $\mathbf{f}^p$ to predict the likelihood of p being on an object boundary. This training process is identical to (Levinshtein et al., 2010).

4. EXPERIMENTAL RESULTS

In this section, we compare our proposed superpixel cut, to three other methods: superpixel closure (SC) (Levinshtein et al., 2010, Levinshtein et al., 2012), cosegmentation (Joulin et al., 2010), and a multiscale version of normalized cut from (Cour et al., 2005). We follow the evaluation protocol of (Levinshtein et al., 2010, Levinshtein et al., 2012) and use the last 50 images from Weizmann Horse Database (WHD) (Borenstein et al., 2004) for learning the weight matrix. For testing, we use the rest of images from WHD and some images from Berkeley Segmentation Data Set (BSDS500) (Arbelaez et al., 2011).

4.1 Datasets and implementation details

We present experiments on two datasets: WHD and BSDS500.

WHD This popular dataset contains 328 horse images, with different poses and backgrounds. The dataset is annotated with ground-truth segmentation masks for all images. We use images from this dataset for quantitative evaluation.

BSDS500 This new dataset is an extension of the BSDS300, where the original 300 images are used for training/validation and 200 fresh images, together with human annotations, are added for

testing. Each image is manually segmented by a number of different human subjects, and on average, five ground truths are available per image. This is one of the most challenging datasets for benchmarking segmentation algorithms. Many images have multiple foreground objects appearing at a variety of scales and locations. We use some images from this dataset for qualitative evaluation.

Implementation details Our framework builds a graph on superpixel nodes, which are generated by SLIC (Achanta et al., 2012), though other choices are also possible. The main reason of choosing SLIC is that it is currently state-of-the-art superpixel segmentation algorithm and practically efficient. The SLIC parameters are the region size and the regularizer. For our experiments, we set region size proportional to the image size to make around 200 superpixels for every image. The regularizer is set as 0.15 for all the images. The contour image is used to extract features for learning the gap measure. We use Pb detector (Martin et al., 2004) in the cost of relatively worse detected contours, instead of globalPb (Arbelaez et al., 2011) in (Levinshtein et al., 2012) which takes long time to compute. We follow the parameter setting in (Levinshtein et al., 2012) to set edge threshold $T_e = 0.05$ and maximal number of solutions $K = 10$.

4.2 Results

4.2.1 Qualitative results We provide a qualitative evaluation of our approach by testing it on images from the two datasets.

Fig. 3 shows the top 3 segmentation results for an example image, corresponding to different λ in Eq. (5) of optimization. The red figure boundary overlays with the superpixel segmentation image for visualization. By comparing with ground-truth, the second result fully recovered the horse boundary. Notice that the rest 2 results are also reasonable in the sense that they represent the object and surrounding context.

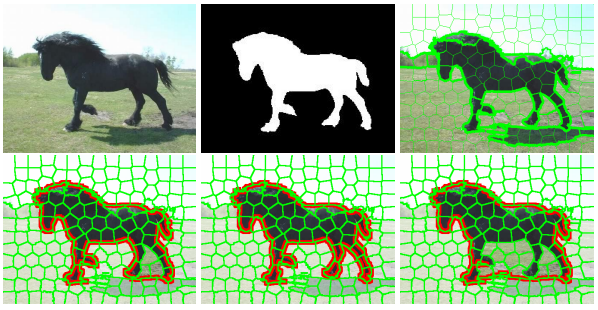

Figure 3: Segmentation example of a horse image. Top row: original image, ground-truth, and superpixel boundary edginess; Bottom row: top 3 segmentation results (the red figure boundary overlays with the superpixel segmentation image for visualization). The thickness of the boundary between superpixels corresponds to the average edge probability of its superpixel boundary pixels, which shows the thick edges are more likely to be the object boundaries. By comparing with ground-truth, the second result fully recovered the horse boundary.

Fig. 4 illustrates the performance of our method compared to the competing superpixel closure approach (SC) (Levinshtein et al., 2012). Top 3 rows are from WHD, and bottom 2 rows are from BSDS5000 Train&Val. We manually select the best result for each method. Notice that both SC and our method produce reasonable results for all 5 images, our results are more compact. We observe that the superpixels between the horse's legs are selected as foreground for SC approach in the first row (Fig. 4). This is because if there is a more compact contour that results in lower energy, it will be preferred. For our method, the superpixels between the horse's legs are not selected, because our cost function reflects the maximization of the intra-class similarity. The optimization process finds that setting these nodes as 0 resulting in lowest energy. Also note that the area between the ears of the horse is selected as foreground for SC approach since the cost function of SC tries to maximize the area. In addition, a significant number of images in the horse dataset have a picture frame boundary around the image, eg. the second row of Fig. 4. These boundaries provide the largest and most compact solutions for SC cost function, and are therefore found by SC instead of finding the horse. Our method still performs well on these images due to the novel cost function. Although we use an approximation of the original cost function for optimization, our method usually results in better quality compared to SC w.r.t. coverage (Fig. 4 Flower image) and compactness (Fig. 4 Birds image).

Table 1: Average pixelwise accuracy of five methods on WHD.

Algorithm	Accuracy (%)	Propetry
MNcut	50.1%	low-level
Cosegmentation	80.1%	low-level
SC	82.0%	low-level
Bertelli et al.	94.6%	high-level
Our method	82.6%	low-level

Some more segmentation examples of BSDS500 Test images are visualized in Fig. 5. The top 5 rows are perceptually satisfactory results, and the bottom 2 rows show the failure cases of our method. Our method often segments single foreground object successfully, despite of relatively large illumination changes and complex layouts of distinct textures, e.g. Fig. 5 (top 3 rows). For the Swan image, although there is reflection on the water surface, our method is able to recover almost all of the boundary in the image. It is usually very difficult for many segmentation algorithms, even the ones incorporating high-level shape priors, to segment a highly textured object from textured background. Our method provides perceptually satisfactory results in the tortoise and fish images. Our method can also cope with multiple foreground objects to a certain degree, e.g. Fig. 5 the Rhino image. For many images of BSDS500, it is difficult for human to decide which is foreground. We notice that although some test images containing multiple foreground objects are well segmented, our method prefers single foreground. The main reason is that single figure gives better implication of possible foreground object.

4.2.2 Quantitative results We quantify performance as pixelwise accuracy, as suggested in (Joulin et al., 2010, Bertelli et al., 2011). It measures the percentage of pixels classified correctly into foreground or background. We compare to the other work in Table 1.

We first compare our method to MNcut (Cour et al., 2005), which is a multiscale version of Normalized Cut (Shi and Malik, 2000). Our method has around 30% accuracy gain. Both SC and our method outperform the Cosegmentation framework, which accesses all image information. Our method also outperforms SC by a margin. As a reference we also include the performance of (Bertelli et al., 2011). While their score is much higher, their method utilizes high-level object similarity information and employs a sliding-window horse detector. In contrast, we only exploit the bottom-up information, making our method simpler and more generic. Therefore, we outperform the competing approaches on WHD dataset, which we attribute to the more reasonable cost function in our framework.

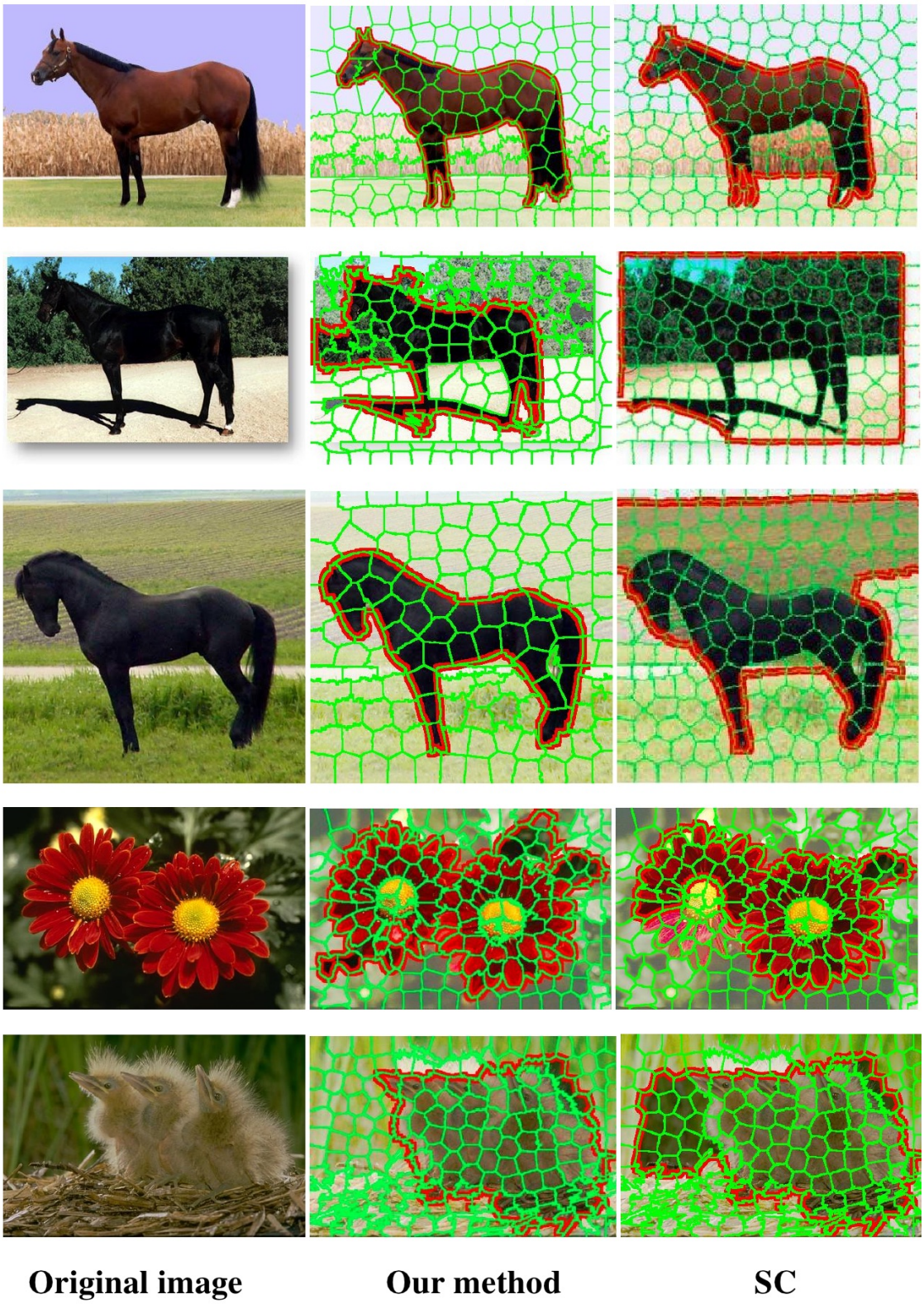

Figure 4: Qualitative results. We compare our results (middle) to superpixel closure algorithm (SC) (right). Left column is the original images. Top 3 rows are from WHD, and bottom 2 rows are from BSDS5000 Train&Val.

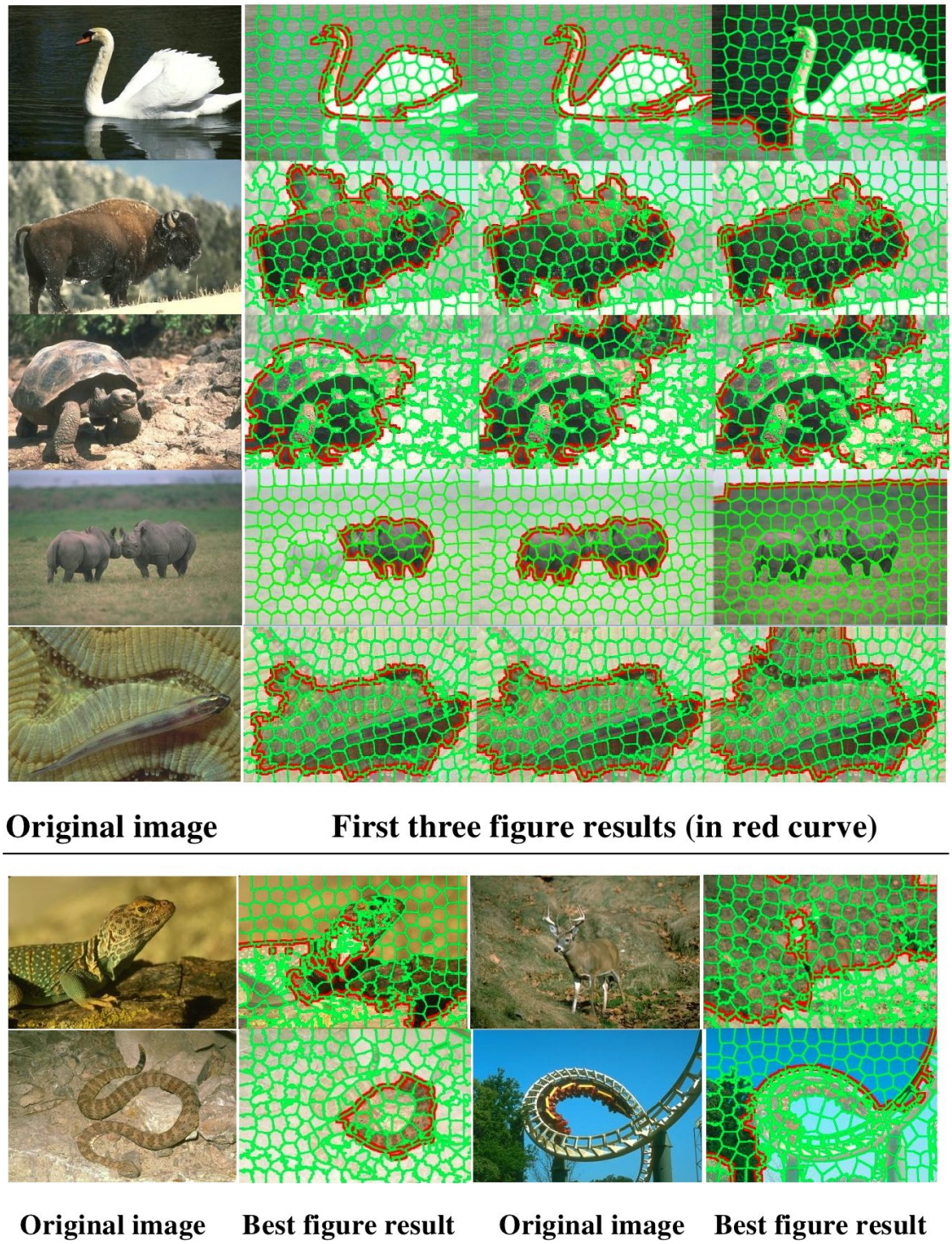

Figure 5: Segmentation examples of BSDS500 Test images. Top 5 rows: original image and first 3 figure segmentation results (the red figure boundary overlays with the superpixel segmentation image for visualization). Bottom 2 rows: typical failure cases.

5. CONCLUSION

We have presented a novel graph-based framework for figure-ground segmentation based on the observation that object boundaries are often reasonably well approximated by superpixel boundaries. We propose a new cost function that simultaneously minimizes the inter-class similarity while maximizing the intra-class similarity. No parameter needs to be tuned within this cost function. The scheme is fully automatic, efficient, and fully bottom-up. It recovers coherent components of images, corresponding to objects, object parts, and objects with surrounding context, providing a set of multiscale hypotheses for high-level reasoning. The experiments demonstrate the high performance of our approach on challenging datasets. For future work, we plan to use multiscale superpixel information to build a cost function that respects scene hierarchy. We will also evaluate our proposed cost function for hierarchical clustering.

ACKNOWLEDGEMENTS

The work is funded by DFG (German Research Foundation) YA 351/2-1. The authors gratefully acknowledge the support.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34(11), pp. 2274–2282.

Arbelaez, P., Maire, M., Fowlkes, C. and Malik, J., 2011. Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 33(5), pp. 898–916.

Bagon, S., Boiman, O. and Irani, M., 2008. What is a good image segment? a unified approach to segment extraction. In: Proc. European Conf. Computer Vision (4), pp. 30–44.

Belongie, S., Malik, J. and Puzicha, J., 2002. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24(4), pp. 509–522.

Bertelli, L., Yu, T., Vu, D. and Gokturk, B., 2011. Kernelized structural svm learning for supervised object segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2153–2160.

Borenstein, E., Sharon, E. and Ullman, S., 2004. Combining top-down and bottom-up segmentation. In: Proc. IEEE CVPR Workshop Perceptual Organization in Computer Vision, pp. 46–53.

Boykov, Y. and Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: Proc. IEEE Int. Conf. Computer Vision, pp. 105–112.

Brendel, W. and Todorovic, S., 2010. Segmentation as maximum-weight independent set. In: Neural Info. Process. Sys. (NIPS), pp. 307–315.

Carreira, J. and Sminchisescu, C., 2012. Cpmc: Automatic object segmentation using constrained parametric min-cuts. IEEE Trans. Pattern Anal. Mach. Intell. 34(7), pp. 1312–1328.

Cour, T., Bénézit, F. and Shi, J., 2005. Spectral segmentation with multiscale graph decomposition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2), pp. 1124–1131.

Endres, I. and Hoiem, D., 2010. Category independent object proposals. In: Proc. European Conf. Computer Vision (5), pp. 575–588.

Joulin, A., Bach, F. R. and Ponce, J., 2010. Discriminative clustering for image co-segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1943–1950.

Kennedy, R., Gallier, J. H. and Shi, J., 2011. Contour cut: Identifying salient contours in images by solving a hermitian eigenvalue problem. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2065–2072.

Kolmogorov, V., Boykov, Y. and Rother, C., 2007. Applications of parametric maxflow in computer vision. In: Proc. IEEE Int. Conf. Computer Vision, pp. 1–8.

Küttel, D. and Ferrari, V., 2012. Figure-ground segmentation by transferring window masks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 558–565.

Levinshtein, A., Sminchisescu, C. and Dickinson, S. J., 2010. Optimal contour closure by superpixel grouping. In: Proc. European Conf. Computer Vision (2), pp. 480–493.

Levinshtein, A., Sminchisescu, C. and Dickinson, S. J., 2012. Optimal image and video closure by superpixel grouping. Int. J. Comput. Vis. 100(1), pp. 99–119.

Li, Y., Sun, J., Tang, C.-K. and Shum, H.-Y., 2004. Lazy snapping. ACM Trans. Graph. 23(3), pp. 303–308.

Malisiewicz, T. and Efros, A. A., 2007. Improving spatial support for objects via multiple segmentations. In: Proc. British Mach. Vision Conf., pp. 1–10.

Martin, D. R., Fowlkes, C. and Malik, J., 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pattern Anal. Mach. Intell. 26(5), pp. 530–549.

Ren, X., Fowlkes, C. and Malik, J., 2005. Scale-invariant contour completion using conditional random fields. In: Proc. IEEE Int. Conf. Computer Vision, pp. 1214–1221.

Rother, C., Kolmogorov, V. and Blake, A., 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23, pp. 309–314.

Shi, J. and Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), pp. 888–905.

Stahl, J. S. and Wang, S., 2007. Edge grouping combining boundary and region information. IEEE Trans. Image Process. 16(10), pp. 2590–2606.

Vicente, S., Kolmogorov, V. and Rother, C., 2008. Graph cut based image segmentation with connectivity priors. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8.

Wang, S. and Siskind, J. M., 2003. Image segmentation with ratio cut. IEEE Trans. Pattern Anal. Mach. Intell. 25(6), pp. 675–690.

Wang, S., Kubota, T., Siskind, J. M. and Wang, J., 2005. Salient closed boundary extraction with ratio contour. IEEE Trans. Pattern Anal. Mach. Intell. 27(4), pp. 546–561.

Yang, M. Y. and Förstner, W., 2011. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In: ICCV Workshop, pp. 196–203.

Yang, M. Y., Förstner, W. and Drauschke, M., 2010. Hierarchical conditional random field for multi-class image classification. In: Int. Conf. Computer Vision Theory and Appl., pp. 464–469.