Saliency-driven GrabCut in Realtime

Anonymous CVPR submission

Paper ID 648

Abstract

Figure-ground segmentation from bounding box input, provided either automatically or manually, has been extremely popular in the last decade and influenced various applications. A lot of research has focused on high-quality segmentation, using complex formulations which often lead to slow techniques, and often hamper practical usage. In this paper we demonstrate a very fast segmentation technique which still achieves very high quality results. We show that by combing two simple ideas we are able to achieve an order of magnitude (10x) speed-up with respect to the closest competitor, and at the same time achieve a considerably higher accuracy. Our first idea is to include a simple salient object prior which is motivated by recent work in the field of salient object detection. Our second idea is to replace the iterative refinement of global colour models by a densely connected CRF. *We motivate this decision by showing that a dense* CRF *implicitly models an unnormalised global colour model.*

1. Introduction

Figure-ground image segmentation from bounding box input, provided either automatically [10, 13] or manually [25], has been extremely popular in the last decade and influenced various computer vision and computer graphics applications, including image editing [19], object detection [27], image classification [31], photo composition [11], scene understanding [16], automatic object class discovery [32], and fine-grained categorization [10]. In order to achieve high quality results, recent methods have focused on complex formulations [21, 28, 29], which typically leads to slow techniques.

In this works we aim to design a very fast figure-ground image segmentation technique which still achieves high quality results. Our method relies on two insights about the problem. First, given a user provided bounding box constraint, the target object region often stands out as a non-ambiguous salient object [22]. This salient object prior assumption enables a good initial estimation for the target object region by borrowing ideas from state-of-the-art salient object detection techniques [1, 12, 13, 22, 24]. We reformulate the salient object detection problem, aiming to maximize the amount of information generated within the specified bounding box while still maintaining real-time efficiency (Sec. 3.1). Second, we observe that a dense CRF implicitly models an unnormalized global colour model, the explicit estimation of which often leads to slow optimization (Sec. 3.3). This enables us to replace the iterative refinement of global colour models [25] with a densely connected CRF, for which very efficient inference techniques have been recently developed [18].

Following recent advances in GrabCut segmentation [28], we extensively evaluate our method on two standard benchmarks, the GRABCUT [25] and the MSRA1000 [1, 22] datasets, containing 50 and 1000 images, respectively, with corresponding binary segmentation masks. Our formulation achieves $F_\beta = 92.8\%$ and $F_\beta = 95.5\%$ on the GRABCUT [25] and the MSRA1000 [1, 22] dataset respectively, where the F_β represents the harmonic mean of precision and recall. Along with generating better segmentations, our method enables real-time CPU processing which is ($10\times$ on average) faster than its closest competitor [28].

2. Related work

Here we review related work that performs interactive figure-ground segmentation [8,26]. Among the many different approaches proposed over the years, the most successful technique incorporates a per-pixel appearance model and pairwise consistency constraints [4], and uses graph cut for efficient energy minimization [7].

Rother *et al*. [25] proposed the first bounding box based segmentation system that optimised both the appearance model and the segments, using initial appearance models computed from a given bounding box. It was shown by Vicente *et al*. [29] that it is possible to reformulate the GrabCut energy functional [25] in closed form as a higher order MRF, by maximizing over global appearance parameters. This was possible by switching from a GMM to a histogram representation for the appearance model. However, the optimization of the higher-order MRF is unfortunately NP-hard.

Nevertheless, the proposed dual decomposition technique is able to achieve globally optimality in about 60% of cases. Recently, OneCut [28] by Tang *et al*. has derived a similar formulation. They argue, however, that the part of the higher-order MRF that make the problem NP-hard, *i.e*. the volume regularization term, is not relevant in practical applications. Hence, they drop this term and can guarantee a globally optimal solution. It is important to note, that on an abstract level our paper has exactly the same reasoning. We show that the GrabCut functional and a densely connected CRF formulation are the same under some approximations. We then argue and show that these approximations are not too critical for practical scenarios.

Our method is inspired by the active research field of visual attention modelling. We refer the reader to a recent survey paper [5] for details about the three major branches of this area: fixation prediction [15], salient object detection [6], and objectness estimation [14]. Here we discuss recent advancements in the most related areas of salient object detection and segmentation. Liu *et al*. [22] proposed a CRF framework that combines multiple saliency measures for an effective salient object detection. Achanta *et al*. [1] proposed a frequency tuned approach to detect salient object regions. Cheng *et al*. [13] proposed a salient object detection method by modelling the global contrast of a region to all other regions in the image. A filtering based framework [24] and an efficient data representation [12] have also been used for the salient object detection task. Although these methods have successfully achieved accurate salient object detection and segmentation results on images with a *single* non-ambiguous object, dealing with more complicated images with *multiple* objects remains a challenging issue. Moreover, such methods are designed for unsupervised salient object segmentation, so they do not explicitly explore the background prior available in interactive segmentation applications.

3. Methodology

We formulate the figure-ground segmentation problem as a binary label Conditional Random Field (CRF) problem. The CRF is defined over the random variables $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, where each $X_i \in \{0, 1\}$, 0 for background and 1 for foreground, represents a binary label of the pixel $i \in \mathcal{N} = \{1, 2, ..., n\}$ such that each random variable corresponds to an image pixel. We denote with $\mathbf{x}$ a joint configuration of these random variables, and $\mathbf{I}$ the observed image data. Based on the general formulation of [18], a fully connected binary label CRF can be defined as:

$$E(\mathbf{x}) = \sum\nolimits_{i \in \mathcal{N}} \psi_i(x_i) + \sum\nolimits_{i<j} \psi_{ij}(x_i, x_j), \qquad (1)$$

where i and j are pixel indices, ψ_i and ψ_{ij} are unary (see Sec. 3.1) and pairwise (see Sec. 3.2) potentials respectively.

3.1. Saliency aware unary term estimation

The unary term ψ_i measures the cost of assigning a binary label x_i to the pixel i, defined as,

$$\psi_i = -\log(S(x_i)), \qquad (2)$$

where $S(x_i) \in (0, 1)$ will be estimated in a saliency aware fashion. Different from other state-of-the-art figure-ground segmentation approaches, which pay more attention to the iterative optimization [25] or to finding the globally minimal solution for the energy function [20, 28, 29], we focus on the complementary task of efficiently obtaining better unary potentials. We argue that although pairwise terms are very informative and important, previous approaches have primarily tried to get the best results according to the noisy unary. Using an incorrect unary term, even if we manage to find the globally optimal solution for the entire energy function, we might still obtain incorrect results. We therefore believe that the unary terms should receive at least the same amount of attention as the pairwise ones.

For object segmentation in complicated images with multiple objects, the target typically stands out as a non-ambiguous salient region [22] within the specified bounding box. This saliency assumption however has yet to be explored in the current interactive image segmentation literature. We initially considered using the state-of-the-art salient object detection method to find the saliency map from the image region cropped according to the selected bounding box. However, directly using traditional saliency detection methods in the bounding box scenario does not explicitly use the background prior, *i.e*. the fact that pixels outside the selection window belong to the background. This leads to a suboptimal appearance model estimation and missing parts for regions with similar appearance to background. We show examples of such images in the top two rows of Fig. 1, and statistical results in Tab. 1.

For an effective use of both background and saliency priors for segmentation, we need to carefully consider what are the major underlying hints contained by two priors, and use them in an efficient and unified way. The background prior not only tells us that the pixels outside the bounding box could be excluded from possible object regions, but it also supplies robust samples for understanding the background appearance. One of the most effective feature for salient object region detection is that the region should have high contrast with respect to the entire image, *i.e*. global contrast. Such cues have been extensively explored in the state-of-the-art salient object detection methods [1, 12, 13, 24].

For foreground labeling, we propose the following equations to capture both background and saliency priors,

$$S(x_i = 1) = \frac{P(\Theta_B, I_i)}{P(\Theta_B, I_i) + w_s P(\Theta_F, I_i)}, \qquad (3)$$

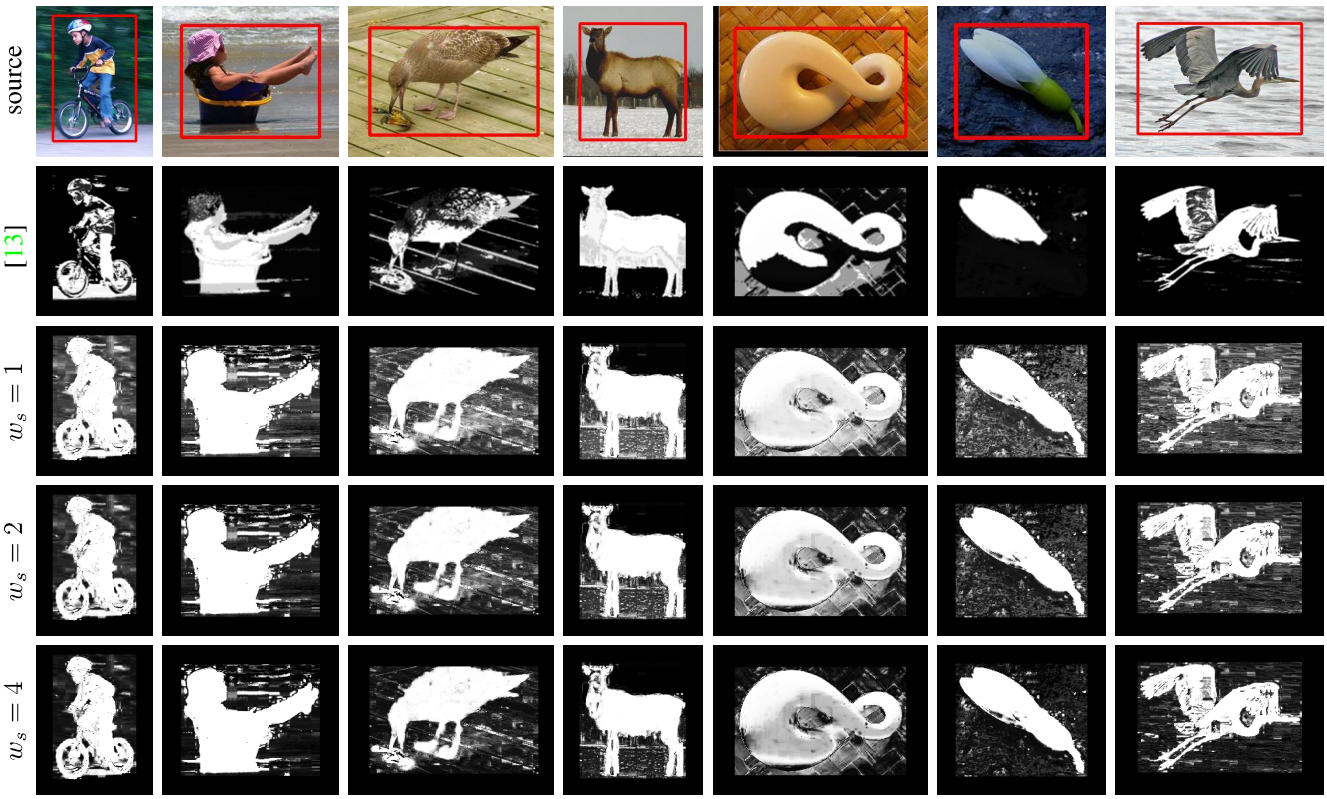


Figure 1. Sample results for global contrast saliency estimation [12] and saliency aware unary maps with different w_s values (see also (3)).

where $P(\Theta_B, I_i)$ and $P(\Theta_F, I_i)$ represent the probability of a pixel color I_i belonging to the background color model Θ_B and the foreground color model Θ_F, respectively. The parameter w_s is a scalar value that indicates how salient the target object should be. The larger w_s is the more salient an object should be in order to get a high value of $S(x_i = 1)$. For background labeling, we use $S(x_i = 0) = 1 - S(x_i = 1)$. Notice that due to the nature of the bounding box interaction, Θ_B is trained from confident background samples. Other pixels (containing both foreground and background pixels) are used to train a less confident foreground color model Θ_F .

In Fig. 1, we show results obtained using different values of w_s. When $w_s = 1$, (3) represents the value used in a standard GrabCut implementation [9, 23, 25]. When $w_s > 1$, the value $S(x_i) \in (0, 1)$ in (3) measures the saliency of an unknown pixel x_i with respect to all other pixels, while at the same time giving a larger weight to nearby background pixels, which is the most effective feature in state-of-the-art salient object detection methods [1, 12, 13, 24]. In our experiments we use $w_s = 2$, obtained using 5-fold cross validation.

3.2. Fully connected pairwise term

The pairwise term ψ_{ij} encourages similar and nearby pixels to take consistent labels. We use a contrast sensitive two kernel potential:

$$\psi_{ij} = [x_i \neq x_j] g(i, j), \tag{4}$$

$$g(i, j) = w_1 \cdot g_1(i, j) + w_2 \cdot g_2(i, j) \tag{5}$$

where the Iverson bracket $[\cdot]$ is 1 for a true condition and 0 otherwise, and the similarity function (5) is defined in terms of color vectors I_i, I_j and position values p_i, p_j:

$$g_1(i, j) = \exp\left(-|p_i - p_j|^2/\theta_\alpha^2 - |I_i - I_j|^2/\theta_\beta^2\right), \tag{6}$$

$$g_2(i, j) = \exp\left(-|p_i - p_j|^2/\theta_\gamma^2\right). \tag{7}$$

Here, (6) models the appearance similarity and encourages nearby pixels with similar color to have the same binary label. (7) encourages smoothness and helps remove small isolated regions. The degree of nearness, similarity, and smoothness are controlled by $\theta_\alpha, \theta_\beta$ and θ_γ, respectively. Intuitively, $\theta_\alpha \gg \theta_\gamma$ should be satisfied if the first term manages the long range connections and the second term measures the local smoothness. These parameters are learned via cross validation.

3.3. Implementations

Color modelling: GMMs vs. Histogram. Effective color modelling is very important for good segmentation results. Among many different models suggested in the literature, two of the most popular ones are histograms [8] and

Gaussian Mixture Models (GMMs) [4,25]. Some important recent works use histogram [28,29] representations.

In [29], the authors suggest that the MAP estimation with the GMM model is in effect an ill-posed problem, since fitting a Gaussian to the color of a single pixel may result in an infinite likelihood (see [3]). As explained in [26], this can be avoided by adding a small constant to the covariance matrix. Compared to histograms, GMMs can better adapt to the colours of the image, while still being effective at capturing small appearance differences between foreground and background. Furthermore, the histogram representation will treat different colours equally differently, ignoring the color values of the histogram bins, *e.g.* two pixels of a banana might have slightly different color and be quantised to different bins, even if they are different from the background, with typically a much larger color difference. We experimentally verify the above discussion via extensive evaluations in Sec. 5.1.

Efficient GMM estimation. As in both the OpenCV [9] and Nvidia CUDA implementation [23], typical GMM estimation can be very computationally expensive, due to the large amount of data samples (pixels) used to train the GMMs. In the salient object detection community, more efficient GMM estimation methods have recently been developed [12]. The estimation is made more efficient using an intermediate histogram based representation. Since natural images typically cover a very small portion of all possible colours, uniformly quantizing the image colours (with each channel divided into 12 parts) and then choosing the most frequent color bins until 95% of image pixels are covered, typically results in a small histogram (an average of 85 histogram bins has been reported [13] for the MSRA1000 [1,22] benchmark). Instead of using hundreds of thousands of image pixels to train the GMM, we can use this small number of histogram bins as weighted samples to train the color GMM, enabling efficient GMM estimation.

Efficient CRF inference. Our CRF formulation satisfies the general form of the fully connected pairwise CRF with Gaussian edge potentials [18]. This enables to use highly efficient Gaussian filtering [2] to perform message passing in the mean field framework. Instead of computing the exact Gibbs distribution:

$$P(\mathbf{X}) \propto \exp\left(-E(\mathbf{x})\right) \tag{8}$$

of the CRF, we can find a mean field approximation $Q(X)$ of the true distribution $P(\mathbf{X})$, that maximizes the KL-divergence $\mathbf{D}(Q||P)$ among all distributions Q that can be expressed as a product of the independent marginal, $Q(\mathbf{X}) = \Pi_i Q_i(X_i)$ [17]. Minimizing the KL-divergence, while constraining $Q(\mathbf{X})$ and $Q(X_i)$ to be valid distributions, yields the following iterative update equation:

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp(\sum_{j \neq i} g(i,j) Q_j(l') - \psi_i(x_i)), \tag{9}$$

where $l, l' \in \{0, 1\}$ and $l' = 1 - l$ are binary variables. Naive estimation of the above equation for all image pixels have a high computational complexity, which is quadratic in the number of pixels. We can rewrite the last term of (9) by adding and then subtracting $Q_i(l')$ so that

$$\sum_{j \neq i} g(i,j) Q_j(l') = \sum_{j \in \mathcal{N}} g(i,j) Q_j(l') - Q_i(l') \tag{10}$$

where $\sum_{j \in \mathcal{N}} g(i,j) Q_j(l')$ is essentially a Gaussian filter, whose value for all image pixels can be calculated efficiently using fast filtering techniques (*e.g.* [17, 18]). This reduces the complexity of the mean field inference, enabling it to be linear to the number of pixels.

4. Relationship between fully connected CRF and global color model.

In many figure-ground segmentation methods, *e.g.* GrabCut [25], two (foreground and background) global colour models are explicitly used. Each colour model is derived from its respective region label. This coupling between the pixel labelling and the global colour model leads to a very challenging optimisation, since both parts need to be inferred jointly. In GrabCut this is done in an iterative fashion, while [29] uses dual decomposition. However, both the iterative and dual decomposition optimisations are slow, with the latter taking up to minutes per frame.

In this work we replace the global colour model with a single optimization of fully connected CRF. This is based on the insight that a fully connected CRF and a standard low-connected (*e.g.* 8-connected) CRF with associated foreground and background global colour models are very closely related. This observation suggested that we can avoid the computational expensive process of global color model estimation, and use the efficient inference for fully connected CRF to enable very fast computation.

Let us consider a specific form of our fully connected CRF, where $\theta_\alpha \to \infty$, *i.e.*

$$\hat{g}_1(i,j) = \exp\left(-|I_i - I_j|^2 / \theta_\beta^2\right). \tag{11}$$

This gives our full energy as

$$E(x) = E_1(x) + w_1 \sum_{i<j} \hat{g}_1(i,j)[x_i \neq x_j], \tag{12}$$

$$E_1(x) = \sum_{i \in \mathcal{N}} \psi_i(x_i) + w_2 \sum_{i<j} g_2(i,j)[x_i \neq x_j]. \tag{13}$$

Note that this is only a minor change to the energy (1) since the spatial smoothness term g_2 is still present, but only once

and not twice. Let us now define a specific version of the GrabCut functional as follows:

$$E(x,\Theta_B,\Theta_F) = E_1(x) + w_1 \sum_{i\in\mathcal{N}} (P_B(I_i;\Theta_B)[x_i=0] + P_F(I_i;\Theta_B)[x_i=1]). \quad (14)$$

Here $\Theta_{F/B}$ are the foreground and background Gaussian mixture models respectively, and $P_{F/B}(I_i;\Theta_{F/B})$ is the negative log probability of the color I_i under the respective Gaussian mixture model. Furthermore, if we choose a small θ_γ then the spatial smoothing term g_2 approximates well the traditional 8-connect Ising prior of Grabcut.

The only difference between the GrabCut function and the fully connected CRF is the term $\hat{g}_1$ in (12) and the sum over the negative log probability in (14).

Let us define the following Parsen-Density estimator:

$$P'_B(I_i) = \frac{1}{|\mathcal{N}_B|} \sum\nolimits_{j\in\mathcal{N}_B} K(I_i,I_j) \quad (15)$$

$$\text{with kernel: } K(I_i,I_j) = -\frac{1}{2}\exp(\frac{-|I_i-I_j|^2}{2\theta_\beta^2}) \quad (16)$$

Here $\mathcal{N}_B$ is the set of background pixels, *i.e.* $x_i = 0$. Note that $P'_B(I_i)$ is the average kernel-distance of the color I_i at pixel i with all colors that are assigned to background. The Parsen-Density estimator for foreground is defined similar: $P'_F(I_i) = \frac{1}{|\mathcal{N}_F|}\sum_{j\in\mathcal{N}_F} K(I_i,I_j)$. We can now state the following theorem (see supplemental material for the corresponding proof) that relates (12) and (14).

Theorem. The minimizer x of (12) and (14) is the same if we replace the global color-model functions $P_F(I_i;\Theta_F)$ and $P_B(I_i;\Theta_B)$ in (14) by unnormalised Parsen-Density estimators $|N_F|P'_F(I_i)$ and $|N_B|P'_B(I_i)$, respectively.

It is very important to note that this discussion is very similar to the discussion of the OneCut [28], in which they re-write the GrabCut functional and remove the "balancing term" from the functional in order to guarantee global optimality. This balancing term enforces that the segmentations with a ratio $\frac{|\mathcal{N}_F|}{|\mathcal{N}_B|} = 1$ are preferred, *i.e.* it penalizes segmentations with extreme ratios. They observe empirically that removing this term does not affect results. We argue that ignoring the balancing term and having scaled density estimators is very similar, since both approximations are negligible if the ratio $\frac{|\mathcal{N}_F|}{|\mathcal{N}_B|} = 1$. Furthermore, we also get empirically good results.

5. Experiments

We extensively evaluate our method on two well known benchmarks (MSRA1000 [1, 22] and GRABCUT [25]), and compare our results with the state-of-the-art alternatives [25, 28], in terms of segmentation quality and efficiency.

5.1. Segmentation Quality Comparison

We evaluate the binary segmentation performance of each method given a user bounding box around the object of interest. The GRABCUT [25] benchmark contains 50 images with bounding box and binary mask annotations. For MSRA1000 [1,22] benchmark, we export the bounding box annotation from its binary mask ground truth, and use this bounding box as input to each method.

To objectively evaluate our method, we compare our results with the two other state-of-the-art methods for bounding box-based figure-ground segmentation *i.e.* GrabCut [25] and OneCut [28]. For GrabCut, we use the CPU implementation from OpenCV [9] and two highly optimised commercial GPU implementations from Nvidia [23] (one uses a GMM color model and another one uses a histogram color model). Average precision, recall, and F-Measure are compared against the entire ground truth datasets, with F-Measure defined as harmonic mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2)\text{Precision}\times\text{Recall}}{\beta^2\times\text{Precision}+\text{Recall}}. \quad (17)$$

Tab. 1 shows the average precision, recall, and F_β values (we use $\beta^2 = 0.3$ as in [1, 13, 28]). Visual examples of input bounding boxes and segmentation results are shown in Fig. 2. Among the baseline methods, the commercial GPU GrabCut implementation from Nvidia [23] achieves the best segmentation results. Although faster computationally, the histogram representation has limited ability to precisely capture appearance differences, resulting in significantly worse segmentation results than the GMM based representation. The comparison between the two versions of Nvidia's commercial implementation clearly verifies our discussion in Sec. 3.3. In both the benchmarks, our method consistently produces better segmentation results than all other alternatives.

The adoption of the fully connected pairwise term [18], enables our methods to capture correlation between pixels with similar appearance even if they are spatially further away from each other. Modelling such long distance consistency constraints, rather than only modelling 4 or 8 neigh-

		MSRA1000 [1,22]		GRABCUT [25]	
		F_β	Time	F_β	Time
CPU	GrabCut [25]	0.945	1.22	0.909	2.02
	OneCut [28]	0.949	0.664	0.900	1.70
	Ours	0.955	0.075	0.928	0.143
GrabCut CUDA	[23](GMM)	0.949	0.072	0.927	0.130
	[23](Hist.)	0.889	0.065	0.714	0.116

Table 1. Average precision, recall, F_β, and processing time (measured in seconds) on two well known benchmarks (see Fig. 2 for sample results). Tested on a desktop computer with Intel Xeon E5645 2.40GHz CUP, GeForce GTX 770 GPU and 4 GB RAM.

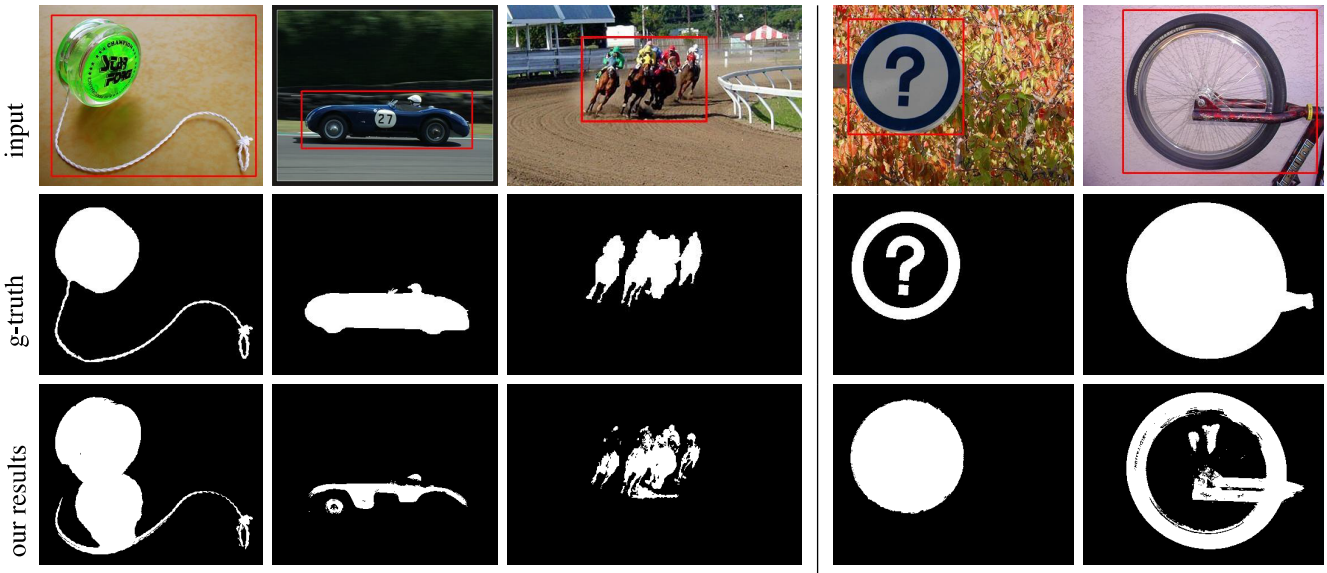
Figure 2. Sample results for images from MSRA1000 [1, 22] (a-g) and GRABCUT [25] (h-k) benchmarks, using different methods: (i) GrabCut [23]GMM), (ii) GrabCut [23]Hist.), (iii) GrabCut [25], (iv) OneCut [28], and (v) Ours.

bours for each pixel, enables our method to produce fine detailed results which could not easily be achieved by state-of-the-art interactive figure-ground segmentation tools such as GrabCut [25] or OneCut [28]. Notice that the fine details of the target object regions are successfully segmented in Fig. 2(b) and Fig. 2(d).

Due to explicitly enforcing color separation between foreground and background, similar in spirit to enforcing the foreground to become salient, only OneCut provides results similar to our own. Both methods recover more accurate fine object boundaries than the other methods, *e.g.* Fig. 2(a-c). However, on average, our method produces better results than OneCut, possibly due to the more powerful color model representation. Extending the OneCut method to incorporate GMMs for representing colours is non-trivial and known to be a NP-hard problem [28, 29].

5.2. Computational time

As shown in Tab. 1 our method is $10\times$ faster than any other current CPU based implementation. Implementing a



(a) g-truth is of high quality (b) g-truth is of low quality

Figure 3. Examples for top 50 'failing examples' shows that our results are very often comparable to ground truth annotations: (a) ground truth mask in MSRA1000 benchmark [1] is preferred, (b) our segmentation results is preferred.

GPU version to fully explore the parallel nature of the algorithm is a promising direction for future work.

Due to the use of the very efficient GMM representation of [12], the most computationally expensive part of our algorithm is the mean field based inference [18], which could be efficiently solved using advanced bilateral filtering techniques [2]. It is worth mentioning that the mean field based inference is an intrinsically parallel algorithm, and thus can be made further efficient using graphics hardware (GPU) or multi-core CPUs. In our current implementation we use OPENMP instructions to parallelize across multiple CPU cores.

5.3. Limitations

The high accuracy of our method ($F_\beta = 95.5\%$ for the MSRA1000 [1, 22] benchmark and $F_\beta = 92.8\%$ for the GRABCUT [25] benchmark), indicates that most results of our methods are very similar to the ground truth. This make it feasible to visualise and study all the clearly failing examples even for a large benchmark such as MSRA1000 [1, 22]. We do this by studying the top 50 'failing examples', which are automatically selected as the results with lowest F_β values according to ground truth. We found that the MSRA1000 [1, 22] benchmark, although used as standard benchmark for figure-ground segmentation (having currently 700+ citations), contains some clear ground truth errors as shown in Fig. 4 (where ground truth masks appear shifted due to unknown reasons). Note that, besides these errors (less than 1%), which we could easily detect from top 6% 'failing cases', most of the other ground truth annotations are of very high quality.

Fig. 3(a) shows typical examples of top 'failing cases'. In the first example, the shadow part occurs only inside the bounding box and its appearance is quite different compared with pixels outside the bounding boxes, forcing the algorithm consider it as an object region. In the other two failure cases, some foreground regions have a large portion of similar appearance regions outside the bounding box, which confuses the algorithm and leads to missing regions for the target object. We went through top 50 'failing cases' and found 12 cases with low quality (see also Fig. 3) and 8 cases with wrong mistakes (see also Fig. 4).

Figure 4. We found ground truth errors in the MSRA1000 benchmark [1] as shown above (the red lines on top of each image illustrate the contour of the ground truth mask). After a manual check, we found 9 such errors from all the annotations of 1000 images, all such ground truth errors are found in the top 6% 'failing cases'.

6. Conclusions

We have presented an efficient figure-ground image segmentation method, which simultaneously models the salient object prior assumption and uses fully connected CRF for effective label consistency modelling. Formally, we show that: a fully connected CRF, as used in this work, and a standard low-connected, *e.g.* 8-connected, CRF with associated foreground and background global colour models are very related. This motivated us to replace the global colour model in the traditional GrabCut framework with a single optimization of a fully connected CRF. Extensive evaluation on two well known benchmarks, MSRA1000 [1,22] and GRABCUT [25], demonstrates that our methods is able to get more accurate segmentation results compared to other state-of-the-art alternative methods, while achieving an order of magnitude speed-up with respect to the closest competitor.

Further introducing a bounding box prior [21], or other CPU high order terms [30] could be useful future additions to our framework. Currently, the weight w_s in (3) is set to a fixed value for the entire dataset, so another interesting area of future work is to learn to compute image specific weights.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009.

[2] A. Adams, J. Baek, and M. A. Davis. Fast high-dimensional filtering using the permutohedral lattice. In *CGF*, 2010.

[3] C. M. Bishop et al. *Pattern recognition and machine learning*. 2006.

[4] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*, pages 428–441. 2004.

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 35(1):185–207, 2013.

[6] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, pages 414–429. Springer, 2012.

[7] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE TPAMI*, 26(9):1124–1137, 2004.

[8] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *IEEE ICCV*, volume 1, pages 105–112, 2001.

[9] G. Bradski and A. Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.

[10] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *IEEE ICCV*, 2013.

[11] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: Internet image montage. *ACM TOG*, 28(5):124:1–10, 2009.

[12] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *IEEE ICCV*, pages 1529–1536, 2013.

[13] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *IEEE CVPR*, pages 409–416, 2011.

[14] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE CVPR*, 2014.

[15] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. Technical report, MIT, 2012.

[16] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009.

[17] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[18] P. Krähenbühl and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*, 2011.

[19] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. In *ACM TOG*, page 3, 2007.

[20] V. Lempitsky, A. Blake, and C. Rother. Image segmentation by branch-and-mincut. In *ECCV*, pages 15–29. 2008.

[21] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *IEEE ICCV*, 2009.

[22] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, T. X., and S. H.Y. Learning to detect a salient object. *IEEE TPAMI*, 2011.

[23] NVIDIA Corporation. CUDA Samples :: CUDA Toolkit Documentation, 2014. http://docs.nvidia.com/cuda/.

[24] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR*, pages 733–740, 2012.

[25] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut"– Interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23(3):309–314, 2004.

[26] C. Rother, V. Kolmogorov, Y. Boykov, and A. Blake. Interactive foreground extraction using graph cut. *Advances in MRF for Vision and Image Processing*, 2011.

[27] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *IEEE ICCV*, 2005.

[28] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov. Grabcut in one cut. In *IEEE ICCV*, 2013.

[29] S. Vicente, V. Kolmogorov, and C. Rother. Joint optimization of segmentation and appearance models. In *IEEE ICCV*, pages 755–762, 2009.

[30] V. Vineet, J. Warrell, and P. H. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. In *ECCV*, pages 31–44. 2012.

[31] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *IEEE ICCV*, 2005.

[32] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu. Unsupervised object class discovery via saliency-guided multiple class learning. In *IEEE CVPR*, pages 3218–3225, 2012.