Pacific Graphics 2015
N. J. Mitra, J. Stam, and K. Xu
(Guest Editors)

Volume 34 (2015), Number 7

DenseCut: Densely Connected CRFs for Realtime GrabCut

Ming-Ming Cheng[1,2] Victor Adrian Prisacariu[2] Shuai Zheng[2] Philip H. S. Torr[†2] Carsten Rother[‡3]

[1]CCCE&CS, Nankai University, China
[2]Department of Engineering, Oxford University, UK
[3]Dresden University of Technology, Germany

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 propose to replace the time consuming iterative refinement of global colour models in traditional GrabCut formulation by a densely connected CRF. *To motivate this decision, we show that a dense* CRF *implicitly models unnormalized global colour models for foreground and background. Such relationship provides insightful analysis to bridge between dense* CRF *and GrabCut functional. We extensively evaluate our algorithm using two famous benchmarks. Our experimental results demonstrated that the proposed algorithm achieves an order of magnitude (10×) speed-up with respect to the closest competitor, and at the same time achieves a considerably higher accuracy.*

Categories and Subject Descriptors (according to ACM CCS): I.4.6 [IMAGE PROCESSING AND COMPUTER VISION]: Segmentation—Region growing, partitioning

1. Introduction

Figure-ground image segmentation from bounding box input, provided either automatically [CLZ13, CZM*11, CMHH14] or manually [RKB04], has been extremely popular in the last decade and influenced various computer vision and computer graphics applications, including image editing [LHE*07, CZM*10], object detection [SBC05], image classification [WCM05], photo composition [CCT*09, CTM*13], scene understanding [KT*09], automatic object class discovery [ZWW*12], and fine-grained categorization [CLZ13]. In order to achieve high quality results, recent methods have focused on complex formulations [VKR09, LKRS09, TGVB13], which typically lead to slow techniques.

In this work, we aim to design a very fast figure-ground image segmentation technique which still achieves high quality results. We observe that a dense CRF implicitly models an unnormalized global colour model, which is similar to the ones used in the well-known GrabCut functional [RKB04]. We show empirically that the "un-normalization" is not critical in practice. Moreover, we are, to the best of our knowledge, the first to draw a close relationship between dense CRFs and the GrabCut functional. This has surprisingly gone unnoticed by the computer vision community, and yet we believe it to be an interesting result unifying two

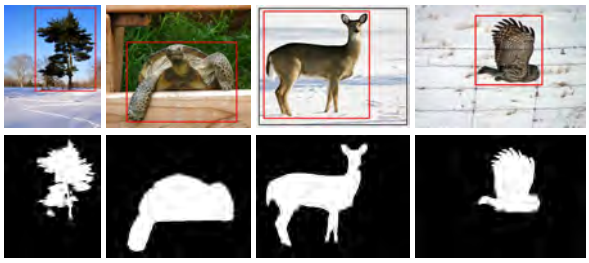


Figure 1: *Given an input image and a bounding box input (first row), our DenseCut algorithm can be used to produce high quality segmentation results (second row) at real time.*

† Leader of Torr Vision Group Oxford
‡ Head of Computer Vision Lab Dresden

strands of research on segmentation that provides a deeper insight into the success of the mean field based approach. Given this relationship, we can optimize a densely connected CRF, for which very efficient inference techniques have been recently developed [KK11], instead of running a slow, iterative refinement of global colour models as in [RKB04], or even slower techniques from [VKR09].

As demonstrated in Fig. 1, our algorithm is able to produce high quality figure-ground segmentation results at realtime. To quantitatively evaluate our method against other alternative approaches, we follow recent advances in GrabCut segmentation [TGVB13], and extensively evaluate our method on two standard benchmarks, the GRABCUT dataset [RKB04] and the MSRA1K dataset [AHES09] datasets, containing 50 and 1000 images, respectively, with corresponding binary segmentation masks. Our formulation achieves $F_\beta = 93.2\%$ and $F_\beta = 95.9\%$ on the GRABCUT dataset [RKB04] and the MSRA1K dataset [AHES09] 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 about $10\times$ faster than its closest competitor [TGVB13].

2. Related work

Here we review related work that performs interactive figure-ground segmentation [BJ01, RKBB11]. Among the many different approaches proposed over the years, the most successful technique incorporates a per-pixel appearance model and pairwise consistency constraints [BRB*04], and uses graph cut for efficient energy minimization [BK04].

Rother *et al.* [RKB04] 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.* [VKR09] that it is possible to reformulate the GrabCut energy functional [RKB04] 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, One Cut [TGVB13] 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 replace this term with a simply unary term, which prefers foreground over background, and can guarantee a globally optimal solution. It is interesting to note that on an abstract level our paper has the same line of reasoning. We show that the GrabCut functional and a densely connected CRF formulation are the same under some approximation. We then argue, and demonstrate experimentally, that this approximation is not critical in practice. Training based segmentation methods, *e.g.* "Boxsup" [DHS15] and "CRFasRCNN" [ZJRP*15], have becoming quite popular recently. These methods leverage a carefully trained deep neural networks [JSD*14,SZ15,LSD14] for high quality sematic segmentation. While these methods are suitable for offline segmentation, the heavy computational overhead makes them unsuitable for realtime interactive applications.

3. Methodology

We formulate the figure-ground segmentation problem as a binary label Conditional Random Field (CRF) problem. A CRF is a form of Markov Random Field (MRF) that defines directly the posterior probability, i.e. the probability of the output variables given the input data [BKR11]. 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 a 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 in [KK11], a fully connected binary label CRF can be defined as:

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

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. Unary term estimation

The unary term $\psi_i(x_i)$ measures the cost of assigning a binary label x_i to the pixel i, defined as,

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

which can be computed independently for each pixel by a classier that produces a distributing over the label assignment x_i. Following [LSTS04, PMC10], we use

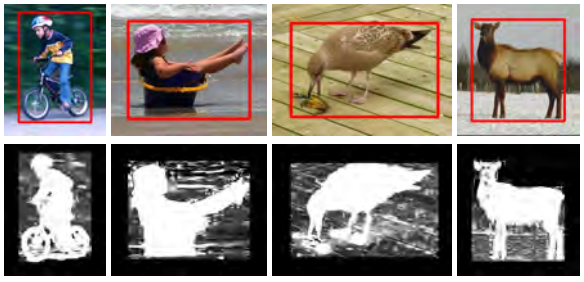


Figure 2: *Illustration of the probability of each pixel belonging to foreground color models: sample images and their corresponding $P(x_i = 1)$ are shown in the first and second row respectively.*

the foreground/background term of the form $P(x_i) = \frac{P(\Theta_{x_i}, I_i)}{P(\Theta_0, I_i)+P(\Theta_1, I_i)}$, where $P(\Theta_0, I_i), P(\Theta_1, I_i) \in (0, \infty)$ represent the probability density value of a pixel color I_i belonging to the background color model Θ_0 and the foreground color model Θ_1, respectively. We use Gaussian Mixture Models (GMMs) and follow the implementation details of [TX06] to estimate the probability density values $P(x_i)$ according to the user selection. Examples of $P(x_i = 1)$ could be found in Fig. 2.

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 three kernel potential:

$$\psi_{ij} = g(i,j)[x_i \neq x_j], \tag{3}$$

$$g(i,j) = w_1 g_1(i,j) + w_2 g_2(i,j) + w_3 g_3(i,j) \tag{4}$$

where the Iverson bracket $[\cdot]$ is 1 for a true condition and 0 otherwise, and the similarity function Equ. (4) is defined in terms of color vectors I_i, I_j and position values p_i, p_j:

$$g_1(i,j) = \exp\left(-\frac{|p_i - p_j|^2}{\theta_\alpha^2} - \frac{|I_i - I_j|^2}{\theta_\beta^2}\right), \tag{5}$$

$$g_2(i,j) = \exp\left(-\frac{|p_i - p_j|^2}{\theta_\gamma^2}\right), \tag{6}$$

$$g_3(i,j) = \exp\left(-\frac{|I_i - I_j|^2}{\theta_\mu^2}\right). \tag{7}$$

Here, Equ. (5) models the appearance similarity and encourages nearby pixels with similar color to have the same binary label. Equ. (6) encourages smoothness and helps to remove small isolated regions. The degree of nearness, similarity, and smoothness are controlled by $\theta_\alpha, \theta_\beta$, θ_γ and θ_μ. Intuitively, $\theta_\alpha \gg \theta_\gamma$ should be satisfied if the term Equ. (5) manages the long range connections and the term Equ. (6) measures the local smoothness. We use empirical values of $w_1 = 6, w_2 = 10, w_3 = 2, \theta_\alpha = 20, \theta_\beta = 33$, $\theta_\gamma = 3$ and $\theta_\mu = 43$ in all the experiments of this paper.

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 [BJ01] and Gaussian Mixture Models (GMMs) [BRB*04, RKB04]. Some important recent works use histogram [TGVB13, VKR09] representations.

In [VKR09], 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 [Bis06]). As explained in [RKBB11], 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 [B*00, BK08] and Nvidia CUDA implementation [NVI14], typical GMM estimation can be 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 [CWL*13]. 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 (*e.g.*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 (*e.g.*an average of 85 histogram bins has been reported [CZM*11, CMH*15] for the MSRA1K dataset [AHES09] 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 [KK11]. This enables us to use highly efficient Gaussian filtering [ABD10] to perform message passing in the mean field framework. Instead of computing the exact Gibbs distribution:

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

of the CRF, we can find a mean field approximation $Q(X)$ of the true distribution $P(\mathbf{X})$, that minimize the KL-divergence $\mathbf{D}(Q||P)$ among all distributions Q that can be expressed as a product of the independent marginal, $Q(\mathbf{X}) = \Pi_i Q_i(X_i)$ [KF09]. Minimizing the KL-divergence, while constraining $Q(\mathbf{X})$ and $Q(X_i)$ to be valid distributions, yields the following iterative update equation:

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

where $l, l' \in \{0, 1\}$, $l' = 1 - l$ are binary variables, and $\frac{1}{Z_i}$ is a normalization factor to constrain $Q(x_i)$ be valid distribution. Each $Q(x_i)$ can be initialized using $Q(x_i) \leftarrow \frac{1}{Z_i}\exp(-\psi_i(x_i))$ and then updated using Equ. (9) un-

til convergence [KK11]. The final label of each pixel is $\arg\max_{l\in\{0,1\}} Q(x_i = l)$, *i.e.* $Q(x_i = 1) > Q(x_i = 0)$ implies x_i is a foreground pixel.

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 Equ. (9) by adding and then subtracting $Q_i(l')$ so that

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

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.* [KF09, KK11]). 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 GrabCut functional

In many figure-ground segmentation methods, *e.g.*GrabCut [RKB04], 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 [VKR09] 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, in the sense that the former is an approximation of the latter. This approximation is basically exact when the area of the fore- and background region is the same in the final segmentation. In the following we also draw a relationship to the One Cut [TGVB13] work, since the approximations in their work and ours are 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 $w_2 = 0$. Note that this is only a minor change to the energy Equ. (1) since the spatial smoothness term is still present in g_1. The energy is then given as

$$E(\mathbf{x}) = E_1(\mathbf{x}) + w_3 \sum_{i<j} g_3(i,j)[x_i \neq x_j], \qquad (11)$$

$$E_1(\mathbf{x}) = \sum_{i\in\mathcal{N}} \psi_i(x_i) + w_1 \sum_{i<j} g_1(i,j)[x_i \neq x_j]. \qquad (12)$$

Let us now write the Grabcut functional as given in [RKB04]

$$E(\mathbf{x},\Theta_B,\Theta_F) = \sum_{i\in\mathcal{N}} (P_B(I_i;\Theta_B)[x_i = 0] + P_F(I_i;\Theta_F)[x_i = 1]) + \sum_{(i,j)\in\mathcal{N}_8} \frac{1}{|p_i - p_j|^2} \exp(-\beta|I_i - I_j|^2)[x_i \neq x_j]. \qquad (13)$$

Here Θ_F and Θ_B are the foreground and background Gaussian mixture models respectively, $P_F(I_i;\Theta_F)$ and $P_B(I_i;\Theta_B)$ are the negative log probability of the color I_i under the respective Gaussian mixture model. The second summand represents the popular edge-preserving smoothing term, here over an 8-Neighborhood grid, and β is a constant defined in [RKB04]. Note, we are interested in the minimizer $\mathbf{x}^* = \arg\min_{\mathbf{x}} \min_{\Theta_F,\Theta_B} E(\mathbf{x},\Theta_B,\Theta_F)$.

One difference between Equ. (11) and Equ. (13) is that the unary term is missing, i.e. $\sum_{i\in\mathcal{N}} \psi_i(x_i)$, in Equ. (13). Furthermore, let us show that the edge-preserving smoothing term in Equ. (13) is very similar to g_1. This can be seen by re-writing the second summand as:

$$\sum_{(i,j)\in\mathcal{N}_8} \frac{1}{|p_i - p_j|^2} \exp(-\beta|I_i - I_j|^2)[x_i \neq x_j] = \sum_{(i,j)\in\mathcal{N}_8} \exp(-\log|p_i - p_j|^2 - \beta|I_i - I_j|^2)[x_i \neq x_j]. \qquad (14)$$

If you compare this equation with Equ. (5) then the first difference is the "log" operator for the pixel distance. The second difference is that we have an 8-neighborhood system instead of a fully connected system. However, by choosing θ_α and θ_β accordingly this can approximated.

Let us now define a version of GrabCut, with a slightly modified edge-preserving smoothing as

$$E(\mathbf{x},\Theta_B,\Theta_F) = E_1(\mathbf{x}) + \sum_{i\in\mathcal{N}} (P_B(I_i,\Theta_B)[x_i = 0] + P_F(I_i;\Theta_F)[x_i = 1]). \qquad (15)$$

The only difference between the GrabCut function and the fully connected CRF is the term g_3 in Equ. (11) and the sum over the negative log probability in Equ. (15).

Let us define the following function that computes a distance between a color, here I_i, and distribution of colors, here all colors of the background region:

$$P'_B(I_i) = \frac{1}{|\mathcal{N}_B|} \sum_{j\in\mathcal{N}_B} K(I_i,I_j) \qquad (16)$$

$$\text{with kernel: } K(I_i,I_j) = -\frac{1}{2}\exp\left(\frac{-|I_i - I_j|^2}{2\theta_\mu^2}\right), \qquad (17)$$

where $\mathcal{N}_B$ is the set of background pixels, *i.e.*$x_i = 0$. Note that this can be seen as a Parzen-Density estimator with an infinity support region. In essence, $P'_B(I_i)$ is the average kernel-distance of the color I_i at pixel i with all colors that are assigned to background. The equivalent distance estimator for foreground is defined as: $P'_F(I_i) = \frac{1}{|\mathcal{N}_F|}\sum_{j\in\mathcal{N}_F} K(I_i,I_j)$.

We can now state the following theorem that relates the GrabCut function in Equ. (15) with our fully connected CRF in Equ. (11).

Theorem 4.1 The minimizer $\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x})$ of Equ. (11) and $\mathbf{x}^* = \arg\min_{\mathbf{x}} \min_{\Theta_F,\Theta_B} E(\mathbf{x},\Theta_F,\Theta_B)$ of Equ. (15) 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 Equ. (15) by weighted functions $|\mathcal{N}_F|P'_F(I_i)$ and $|\mathcal{N}_B|P'_B(I_i)$, respectively.

Proof Let us look at the function $\sum_{i<j} g_3(i,j)[x_i \neq x_j]$, which is part of Equ. (11) but not Equ. (15). The minimizer for the function can be re-written as follows:

$$\arg\min_{\mathbf{x}} \sum_{i<j} g_3(i,j)[x_i \neq x_j] \tag{18}$$

$$= \arg\min_{\mathbf{x}} \sum_{i<j} g_3(i,j)[x_i \neq x_j] - \sum_{i<j} g_3(i,j)$$

$$= \arg\min_{\mathbf{x}} \sum_{i<j} -g_3(i,j)[x_i = x_j]$$

$$= \arg\min_{\mathbf{x}} \sum_{i\in\mathcal{N}} \Big(\sum_{j\in\mathcal{N}} -\frac{1}{2} g_3(i,j)[x_i = x_j] \Big)$$

$$= \arg\min_{\mathbf{x}} \sum_{i\in\mathcal{N}} \Big(\sum_{j\in\mathcal{N}_B} K(I_i,I_j)[x_i = 0] + \sum_{j\in\mathcal{N}_F} K(I_i,I_j)[x_i = 1] \Big) \tag{19}$$

$$= \arg\min_{\mathbf{x}} \sum_{i\in\mathcal{N}} \big(|\mathcal{N}_B|P'_B(I_i)[x_i = 0] + |\mathcal{N}_F|P'_F(I_i)[x_i = 1] \big). \tag{20}$$

Comparing Equ. (20) and Equ. (15) shows the required relationship. □

The remaining question is: What is the effect of the "weighting" of the functions $P'_F(I_i)$ and $P'_B(I_i)$? First of all, observe that we would ideally like to get rid of the weights $|\mathcal{N}_F|$ and $|\mathcal{N}_B|$, since this would give us a proper (infinite) Parzen-window estimator. However, intuitively this is not possible since [VKR09] has shown that solving the GrabCut function is NP-hard. We call this approximation, *i.e.* $|\mathcal{N}_F|P'_F(I_i)$ instead of $P'_F(I_i)$ the "unnormalized global color model". It can be seen that if the ratio $\frac{|\mathcal{N}_F|}{|\mathcal{N}_B|} = 1$ then we actually have a proper density estimator, since all weights can be globally re-scaled. This means that we can compute the global minimizer $\mathbf{x}$ for Equ. (11) and analyze its ratio. If the ratio is close to 1, it means that it is closer to a proper density estimation. By choosing a rectangle image region outside the bounding box input as a working region to build CRF, we can roughly control this ratio. In our experiments, we select a $w_b = 5$ pixel wider region than the bounding box input as working region, which generates an average ratio of 1.5 and 1.2 for MSRA1000 and GRABCUT benchmarks, respectively. We experimentally find that changing w_b in a large range, *e.g.* $[2,10]$, has an ignorable influence to the algorithm performance.

It is interesting to note that this discussion is related to the main line of argumentation in the One Cut [TGVB13] work. In One Cut [TGVB13], the authors re-write the GrabCut functional by replacing the "volume regularization term" with a simple ballooning force (unary term) that prefers to have all pixels being foreground. This change makes it possible to optimize the new GrabCut functional globally optimal. The "volume regularization term" enforces that 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 regularization term does not affect results. In the above discussion we also derived a theoretically sound method for the case that $\frac{|\mathcal{N}_F|}{|\mathcal{N}_B|} = 1$. However, as in [TGVB13], ignoring this ratio constraint gives us good results in practice.

5. Experiments

We extensively evaluate our method on two well known benchmarks (MSRA1K dataset [AHES09] and GRABCUT dataset [RKB04]), and compare our results with the state-of-the-art alternatives [RKB04, TGVB13], 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 dataset [RKB04] benchmark contains 50 images with bounding box and binary mask annotations. For MSRA1K dataset [AHES09] 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 [RKB04] and One Cut [TGVB13]. For GrabCut, we use the CPU implementation from OpenCV [BK08] and two highly optimised commercial GPU implementations from Nvidia [NVI14] (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:

$$F_\beta = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}. \tag{21}$$

Tab. 1 shows the average precision, recall, and F_β values (we use $\beta^2 = 0.3$ as in [AHES09, CZM*11, TGVB13]). Visual examples of input bounding boxes and segmentation results are shown in Fig. 3. Among the baseline methods, the commercial GPU GrabCut implementation from Nvidia [NVI14] 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

		MSRA1K dataset [AHES09]		GRABCUT dataset [RKB04]	
		F_β measure	Time (s)	F_β measure	Time (s)
CPU	GrabCut [RKB04]	0.945	1.22	0.909	2.02
	One Cut [TGVB13]	0.949	0.664	0.900	1.70
	Ours	0.959	0.075	0.932	0.143
CUDA	GrabCut(GMM) [NVI14]	0.949	0.074	0.918	0.149
	GrabCut (Histogram) [NVI14]	0.889	0.059	0.714	0.135

Table 1: *Average precision, recall, F_β, and processing time (measured in seconds) on two well known benchmarks (see Fig. 3 for sample results). Tested on a computer with Intel Xeon E5645 2.40GHz CUP, 4GB RAM, Nvidia Tesla K40 GPU and CUDA 7.0 SDK.*

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.

While we have shown theoretically that GrabCut, One Cut and our Dense CRF are very related, we believe that these differences in performance stem from the fact that we have more parameters to adjust. Hence the weighting between the kernels that relate to spatial smoothing, contrast based smoothing, and global color models, are more finely tuned. This is noticeable visually - see for instance the fine details of the target object regions that are successfully segmented in Fig. 3(c) and Fig. 3(f) by our method.

Comparing One Cut with our method, we notice that, on average, our method produces better results than One Cut, possibly due to the more powerful color model representation. Extending the One Cut method to incorporate GMMs for representing colours is non-trivial and known to be a NP-hard problem [TGVB13, VKR09].

Due to explicitly enforcing color separation between foreground and background, only One Cut provides results similar to our own. Both methods recover more accurate fine object boundaries than the other methods, *e.g.* Fig. 3(c)(d)(f).

5.2. Computational time

As shown in Tab. 1 our method is about 10× faster than any other current CPU based implementation. Implementing a GPU version to fully explore the parallel nature of the algorithm is a promising direction for future work.

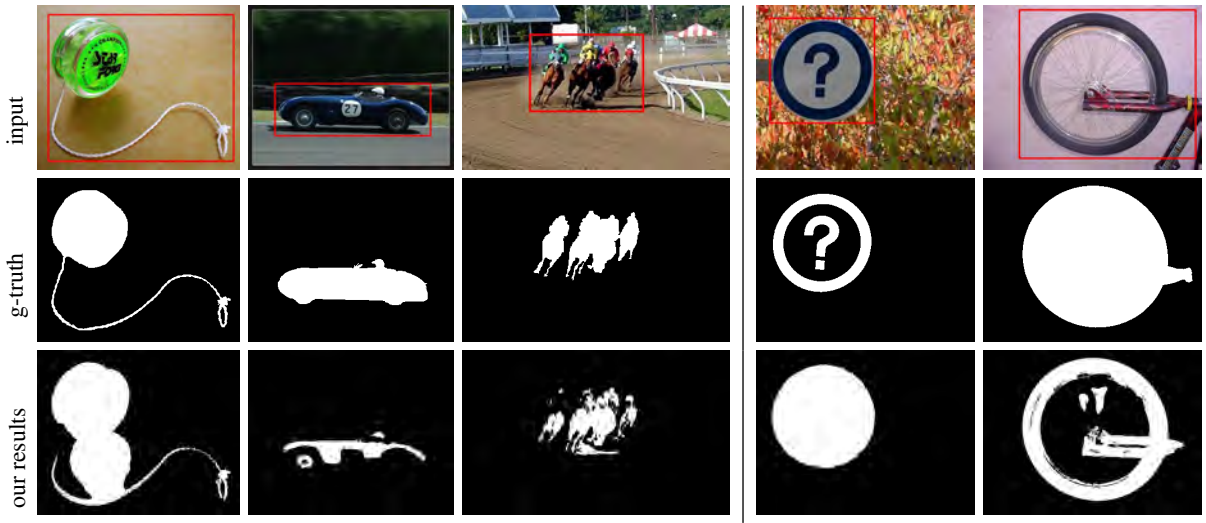
Due to the use of the very efficient GMM representation of [CWL*13], the most computationally expensive part of our algorithm is the mean field based inference [KK11], which could be efficiently solved using advanced bilateral filtering techniques [ABD10]. It is worth mentioning that the mean field based 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.

Figure 5: *We found ground truth errors in the MSRA1000 benchmark [AHES09] 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'.*

5.3. Limitations

The high accuracy of our method ($F_\beta = 95.9\%$ for the MSRA1K dataset [AHES09] benchmark and $F_\beta = 93.2\%$ for the GRABCUT dataset [RKB04] 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 MSRA1K dataset [AHES09]. 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 MSRA1K dataset [AHES09] benchmark, although used as standard benchmark for figure-ground segmentation (having currently 1100+ citations), contains some

Figure 3: *Sample results for images from* MSRA1K *dataset [AHES09] (a-g) and* GRABCUT *dataset [RKB04] (h-j) benchmarks, using different methods: (i) GrabCut [NVI14]GMM, (ii) GrabCut [NVI14]Hist., (iii) GrabCut [RKB04], (iv) One Cut [TGVB13], and (v) Ours.*



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

Figure 4: *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 [AHES09] is preferred, (b) our segmentation results is preferred.*

clear ground truth errors as shown in Fig. 5 (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. 4(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 ground truth segmentation (see also Fig. 4) and 8 cases with incorrect segmentation (see also Fig. 5).

6. Conclusions

We have presented an efficient figure-ground image segmentation method, which uses fully connected CRF for effective label consistency modelling. Formally, we show that a fully connected CRF, as used in this work, and the well-known GrabCut functional, with a low-connected, *e.g.* 8-connected, CRF with associated foreground and background global colour models are closely 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, MSRA1K dataset [AHES09] and GRABCUT dataset [RKB04], 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 [LKRS09], or other CPU high order terms [VWT12] could be useful future additions to our framework.

To encourage future works, we make the source code, links to related methods, and live discussions available in the project page: `http://mmcheng.net/densecut/`.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedbacks. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. This research was supported by Natural Science Foundation of China, Youth Leader Program of Nankai University, EPSRC (EP/I001107/1), ERC-2012-AdG 321162-HELIOS, and EPSRC (EP/J014990). Shuai is receiving a CSC scholarship, and partially fund from EPSRC (EP/I001107/2).

References

[ABD10] ADAMS A., BAEK J., DAVIS M. A.: Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum* (2010). 3, 6

[AHES09] ACHANTA R., HEMAMI S., ESTRADA F., SÜSSTRUNK S.: Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 1597–1604. 2, 3, 5, 6, 7, 8

[B*00] BRADSKI G., ET AL.: The opencv library. *Doctor Dobbs Journal 25*, 11 (2000), 120–126. 3

[Bis06] BISHOP C. M.: *Pattern recognition and machine learning*. Springer, 2006. 3

[BJ01] BOYKOV Y. Y., JOLLY M.-P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *IEEE International Conference on Computer Vision* (2001), vol. 1, pp. 105–112. 2, 3

[BK04] BOYKOV Y., KOLMOGOROV V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Recognition and Machine Intelligence 26*, 9 (2004), 1124–1137. 2

[BK08] BRADSKI G., KAEHLER A.: *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008. 3, 5

[BKR11] BLAKE A., KOHLI P., ROTHER C.: *Markov random fields for vision and image processing*. Mit Press, 2011. 2

[BRB*04] BLAKE A., ROTHER C., BROWN M., PEREZ P., TORR P.: Interactive image segmentation using an adaptive gmmrf model. In *European Conference on Computer Vision*. 2004, pp. 428–441. 2, 3

[CCT*09] CHEN T., CHENG M.-M., TAN P., SHAMIR A., HU S.-M.: Sketch2photo: Internet image montage. *ACM TOG 28*, 5 (2009), 124:1–10. 1

[CLZ13] CHAI Y., LEMPITSKY V., ZISSERMAN A.: Symbiotic segmentation and part localization for fine-grained categorization. In *IEEE International Conference on Computer Vision* (2013). 1

[CMH*15] CHENG M.-M., MITRA N. J., HUANG X., TORR P. H. S., HU S.-M.: Global contrast based salient region detection. *IEEE Transactions on Pattern Recognition and Machine Intelligence 37*, 3 (2015), 569–582. 3

[CMHH14] CHENG M.-M., MITRA N., HUANG X., HU S.-M.: Salientshape: group saliency in image collections. *The Visual Computer 30*, 4 (2014), 443–453. 1

[CTM*13] CHEN T., TAN P., MA L.-Q., CHENG M.-M., SHAMIR A., HU S.-M.: Poseshop: Human image database construction and personalized content synthesis. *Visualization and Computer Graphics, IEEE Transactions on 19*, 5 (2013), 824–837. 1

[CWL*13] CHENG M.-M., WARRELL J., LIN W.-Y., ZHENG S., VINEET V., CROOK N.: Efficient salient region detection with soft image abstraction. In *IEEE International Conference on Computer Vision* (2013), pp. 1529–1536. 3, 6

[CZM*10] CHENG M.-M., ZHANG F.-L., MITRA N. J., HUANG X., HU S.-M.: Repfinder: Finding approximately repeated scene elements for image editing. *ACM Transactions on Graphics 29*, 4 (2010), 83:1–8. 1

[CZM*11] CHENG M.-M., ZHANG G.-X., MITRA N. J., HUANG X., HU S.-M.: Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2011), pp. 409–416. 1, 3, 5

[DHS15] DAI J., HE K., SUN J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *arXiv preprint arXiv:1503.01640* (2015). 2

[JSD*14] JIA Y., SHELHAMER E., DONAHUE J., KARAYEV S., LONG J., GIRSHICK R., GUADARRAMA S., DARRELL T.: Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014). 2

[KF09] KOLLER D., FRIEDMAN N.: *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 3, 4

[KK11] KRÄHENBÜHL P., KOLTUN V.: Efficient inference in fully connected CRFs with gaussian edge potentials. In *Neural Information Processing Systems* (2011). 2, 3, 4, 6

[KT*09] KOHLI P., TORR P. H., ET AL.: Robust higher order potentials for enforcing label consistency. *International Journal on Computer Vision 82*, 3 (2009), 302–324. 1

[LHE*07] LALONDE J.-F., HOIEM D., EFROS A. A., ROTHER C., WINN J., CRIMINISI A.: Photo clip art. In *ACM Transactions on Graphics* (2007), p. 3. 1

[LKRS09] LEMPITSKY V., KOHLI P., ROTHER C., SHARP T.: Image segmentation with a bounding box prior. In *IEEE International Conference on Computer Vision* (2009). 1, 8

[LSD14] LONG J., SHELHAMER E., DARRELL T.: Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038* (2014). 2

[LSTS04] LI Y., SUN J., TANG C.-K., SHUM H.-Y.: Lazy snapping. *ACM Transactions on Graphics 23*, 3 (2004), 303–308. 2

[NVI14] NVIDIA CORPORATION: CUDA Samples :: CUDA Toolkit Documentation, 2014. URL: http://docs.nvidia.com/cuda/cuda-samples/. 3, 5, 6, 7

[PMC10] PRICE B. L., MORSE B., COHEN S.: Geodesic graph cut for interactive image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 3161–3168. 2

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A.: "GrabCut"– Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics 23*, 3 (2004), 309–314. 1, 2, 3, 4, 5, 6, 7, 8

[RKBB11] ROTHER C., KOLMOGOROV V., BOYKOV Y., BLAKE A.: Interactive foreground extraction using graph cut. *Advances in MRF for Vision and Image Processing* (2011). 2, 3

[SBC05] SHOTTON J., BLAKE A., CIPOLLA R.: Contour-based learning for object detection. In *IEEE International Conference on Computer Vision* (2005). 1

[SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *ICLR* (2015). 2

[TGVB13] TANG M., GORELICK L., VEKSLER O., BOYKOV Y.: Grabcut in one cut. In *IEEE International Conference on Computer Vision* (2013). 1, 2, 3, 4, 5, 6, 7

[TX06] TALBOT J. F., XU X.: Implementing grabcut. *Brigham Young University* (2006). 3

[VKR09] VICENTE S., KOLMOGOROV V., ROTHER C.: Joint optimization of segmentation and appearance models. In *IEEE International Conference on Computer Vision* (2009), pp. 755–762. 1, 2, 3, 4, 5, 6

[VWT12] VINEET V., WARRELL J., TORR P. H.: Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. In *European Conference on Computer Vision*. 2012, pp. 31–44. 8

[WCM05] WINN J., CRIMINISI A., MINKA T.: Object categorization by learned universal visual dictionary. In *IEEE International Conference on Computer Vision* (2005). 1

[ZJRP*15] ZHENG S., JAYASUMANA S., ROMERA-PAREDES B., VINEET V., SU Z., DU D., HUANG C., TORR P.: Conditional random fields as recurrent neural networks. *arXiv preprint arXiv:1502.03240* (2015). 2

[ZWW*12] ZHU J.-Y., WU J., WEI Y., CHANG E., TU Z.: Unsupervised object class discovery via saliency-guided multiple class learning. In *IEEE CVPR* (2012), pp. 3218–3225. 1