MODIFIED GRABCUT FOR UNSUPERVISED OBJECT SEGMENTATION

Mohammad Jahangiri

Imperial College London
Department of Electrical and Electronic Engineering
London, UK
m.jahangiri@imperial.ac.uk

Daniel Heesch

Pixsta Research
London, UK
daniel.heesch@pixsta.com

ABSTRACT

We propose a fully automated variation of the GrabCut technique for segmenting comparatively simple images with little variation in background colour and relatively high contrast between foreground and background. The interactive trimap generation central to the original formulation of GrabCut is replaced by a tentative approximation of the background using active contours. Instead of waiting until convergence of the iterated graph cut, we terminate as soon as the Gaussian models of foreground and background are (locally) maximally separated. We demonstrate that this results in equivalent segmentation quality at significantly lower cost. A comparison with three alternative segmentation techniques, including normalised cut, indicates that the method is eminently suitable for the chosen image domain.

Index Terms— Segmentation, GrabCut, Graph Cut, Active Contours

1. INTRODUCTION

Automated object segmentation in unconstrained images lies well beyond the capability of current technology. Much research is being invested therefore towards solving three slightly simpler problems: (i) to obtain a segmentation into regions, not necessarily objects ([1], [2]); (ii) to achieve object segmentation by incorporating additional human input (e.g. [3], [4], [5]); and (iii) to obtain a segmentation in a fully automated way on images satisfying certain constraints. Our work falls into the third class. Our goal is to segment foreground objects, that is various types of apparel often including a human model, from unstructured but non-uniform backgrounds. We assume that an image (loosely) satisfies the following constraints: (1) Images consist of possibly quite complicated foreground with high variation in the pixel colour values, and a simple background with little variation in pixel colour values; (2) The contrast between the foreground and background is high. An example is shown in Figure 1. The proposed segmentation method is based on GrabCut [4], which iteratively refines Gaussian mixture models of the background and foreground using a combination of graph cuts and EM-like parameter estimation. GrabCut takes as input a set of known background pixels specified, for example, by users drawing a bounding rectangle around the object to be segmented. These pixels remain firmly assigned to the background class during subsequent iterations. By contrast, we assume that all pixel labels need to be estimated and that pixels may change their label in the course of the algorithm. The initial Gaussian models of the foreground and background are constructed following an automated approximation of the foreground using active contours.

Graph cut techniques for segmentation abound. A staple algorithm in the graph cut family is normalised cut [2] which mitigates the tendency of other graph cut techniques towards unbalanced partitions. It requires few parameters, leads naturally to a bi-partitioning of the image and has found widespread application in the region-wise segmentation of complex images. In [6] the authors consider the problem of segmenting gray-scale images with multi-modal intensity distributions on the assumption that the number of modes is known. The distribution is modelled as a mixture of Gaussian distributions the parameters of which are iteratively refined within the standard GrabCut framework. The authors in [7] determine a possibly large set of uniform regions (not necessarily a partitioning). For each region pair a minimum cut segmentation is obtained on the assumption that the pixels of the two regions belong to different segments. The set of segmentations are combined and the resulting regions merged.

The main contributions of our work are (i) the combination of a modified version of GrabCut with automated initialisation to achieve a fully unsupervised segmentation algorithm and (ii) the proposal of a new stopping criterion for GrabCut that allows early termination whilst locally maximising the contrast between foreground and background.

The rest of the paper is organised as follows. Section 2 describes our technique for achieving an initial assignment of pixels to foreground and background. In Section 3 we present the technique for iteratively updating the labelling of the pixels. Section 4 considers a border smoothing technique based on morphological operators. Our experimental evaluation is presented in Section 5. Section 6 concludes the paper.

2. INITIALISATION OF BACKGROUND AND FOREGROUND MODELS

To initialise the Gaussian models for foreground and background, we define initial sets of pixels for each class by letting an active contour converge towards a rough approximation of the object outline. We select k equi-distant points $\mathbf{v} = \{v_i \in \mathbb{N}^2 : i = 1, \ldots, k\}$ along a rectangular contour near the borders of the image. We use the active contour model of Kass *et al.* [8] in order to evolve the initial contour towards the border of the object. In [8] the aim is to minimise the following energy function

$$E_{total}(\mathbf{v}) = \sum_{i=1}^{k}(E_{int}(\mathbf{v}, i) + E_{ext}(\mathbf{v}, i)), \qquad (1)$$

where $E_{int}(\mathbf{v}, i)$ and $E_{ext}(\mathbf{v}, i)$ are defined, respectively, as

$$E_{int}(\mathbf{v}, i) = \alpha \left\| v_i - v_{i-1} \right\|_2^2 + \beta \left\| v_{i+1} - 2v_i + v_{i-1} \right\|_2^2, \qquad (2)$$

and

$$E_{ext}(\mathbf{v}, i) = -||\nabla I(v_i)||_2^2. \qquad (3)$$

∇I corresponds to the gradient map of I. E_{int} is composed of the first order and second order term of the contour. α and β are parameters that control the local geometry of the contour. As we have no prior information about the shape of the object, we set both parameters to unity. E_{ext} forces the active contour to move towards regions with higher gradients. Whilst other terms may of course be added to this energy term, we found the gradient to be sufficient for our application. In order to minimise E_{total} we use the dynamic programming optimisation technique proposed by Amini *et al.* in [9]. An example of the initial contour and the contour obtained after iteratively optimising the energy function of (1) are shown in Figure 1. Pixels outside the estimated contour are labelled as background and are used for computing the Gaussian parameters of the background model, $\mu_B, \sigma_B \in \mathbb{R}^3$ (each component of which corresponds to one of the RGB colour channels). From the pixels which lie inside the detected contour a subset of pixels with values outside $[\mu_B - 2\sigma_B, \mu_B + 2\sigma_B]$ is selected to construct a Gaussian Mixture Model (GMM). Following [4] we choose 5 components, and use EM for parameter estimation.

3. ITERATED GRAPH CUT

Let $\mathbf{y} = \{y_i : i = 1, 2, \ldots, N\}$ denote the set of RGB colour triples of each of the N pixels. We treat the segmentation task as that of estimating a set of binary variables $\mathbf{x} = \{x_i \in \{0, 1\} : i = 1, 2, \ldots, N\}$, each indicating whether the corresponding pixel belongs to foreground or background. In conventional graph cut algorithms the optimal labelling minimises an energy function $E(\mathbf{x}, \mathbf{y})$ which takes into account the pixel data as well as the degree to which the label of the pixel differs from those of its neighbours. Let $\mathcal{V}$ denote the set of pixel indices, and $\mathcal{E}$ the set of index pairs of adjacent vertices. The standard form of the cost function is then

$$E(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} E_1(x_i, y_i) + \sum_{(i,j) \in \mathcal{E}} E_2(x_i, x_j). \qquad (4)$$

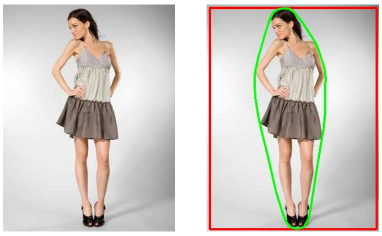

Fig. 1. Left: Original image; Right: converged state of active contours

The data term, $E_1(x_i, y_i)$, in [4] is defined as the sum over all pixels of the pixel's component likelihood,

$$E_1(x_i, y_i) = -\log p(y_i | x_i, \theta; k_i) - \log \pi(k_i, x_i), \qquad (5)$$

where $p(.)$ is the k_ith Gaussian probability distribution estimated with parameters $\theta_i = \{\mu_i, \Sigma_i\}$. In total, there are 6 Gaussians in the proposed model i.e. $k_i \in \{1, \cdots, 6\}$ from which $k_i \in \{1, 2, 3, 4, 5\}$ correspond to class $x_i = 1$ and $k_i = 6$ associates with class $x_i = 0$. $\pi(.)$ is the mixing coefficient corresponding to each Gaussian component. Note that, more conventionally, each pixel contributes with the log likelihood of its GMM, i.e. $E_1(x_i, y_i) = \sum_{i=1}^{i=5} - \log p(.) - \log \pi(.)$ not of the GMM component (5). The former would allow the more accurate, but also more expensive expectation-maximisation algorithm to be applied in order to find the parameters of the GMM.

The smoothness term $E_2(x_i, x_j)$ is defined as:

$$E_2(x_i, x_j) = \lambda \mathbf{1}(x_i \neq x_j) \exp(-\gamma |y_i - y_j|) \frac{1}{dist(i, j)}, \qquad (6)$$

where $\mathbf{1}(c) = 1$ if condition c is satisfied and 0 otherwise, $dist(\cdot, \cdot)$ is the Euclidean distance of the neighbouring pixels, and γ is the average variation in colour values in the two pixels considered (where the average is taken over the three colour channels). λ is a parameter that specifies the relative importance of the two terms making up the energy function. Choosing it is an active research area in its own right; we use cross-validation for choosing a reasonable value for λ. The value of E_2 is either zero (if the labels are the same) or a positive number that increases with the degree of similarity between the pixel values. To minimise the cost function of equation (4) we follow the algorithm proposed in [4]. The proposed segmentation methodology is summarised in the following:

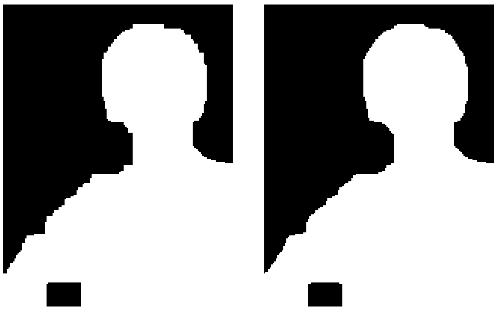

Fig. 2. Left: Binary result of automated GrabCut; Right: Result after subsequent morphological post-processing

1. Initialise pixel labels as described in section 2
2. Estimate the GMMs of the foreground and background pixels by first assigning to each pixel its most likely component and then computing parameters for each component using max-likelihood
3. Estimate new pixel labels (foreground and background) using the min cut/max flow algorithm
4. If the stopping criterion is fulfilled, terminate, else go to step 2

We note two differences to the original GrabCut formulation. Firstly, in our scheme all pixels may change their labels, whilst GrabCut assumes certain pixel labels (by defining a Trimap) to be immutable. Secondly, we introduce a new stopping criterion. While GrabCut waits until convergence of the Gibbs energy, we terminate when reaching a local maximum of the following symmetric contrast measure

$$\text{Contrast}(P,Q) = D_{KL}(P,Q) + D_{KL}(Q,P), \qquad (7)$$

where $D_{KL}(P,Q) = \sum_{i=1}^{N} p_i \log p_i/q_i$ is the Kullback Leibler divergence, and P and Q are the normalised RGB colour histograms of the foreground and background pixels.

4. CONTOUR POST PROCESSING

Discrete graph cuts lead to sharp edges between foreground and background. To soften the appearance we apply standard techniques from mathematical morphology. The object region is first opened (dilation + erosion) and subsequently closed (erosion + dilation) using a diamond-shaped smoothing element (to avoid Manhattan artifacts). The result is a smoothening of the contour as well as a reduction in the intensity gradient across the edge. The effect is visible in Figure 2.

5. EVALUATION

5.1. Experimental setup

The image collection consists of 200 images of apparel that are obtained from Pixsta Ltd's online fashion marketplace Empora.com (`www.empora.com`). A large proportion of the images are shots of models wearing different products, but also close-up views of items such as shoes and bags. The backgrounds tend to include shadows and exhibit variation in intensity. All 200 images were hand-segmented using a fine polygon approximation of the contour. The segmentation accuracy is measured in terms of the F-measure, which combines the two complementary measures of precision (fraction of hypothesised foreground that *is* foreground) and recall (fraction of actual foreground found in the hypothesised foreground). The proposed unsupervised variant of GrabCut is compared against three alternative segmentation methods, which will be described in turn.

k-means segmentation: We apply the k-means clustering algorithm with $k = 2$ and random initialisation to the array of RGB triples (not taking into account coordinate information) to obtain a segmentation into foreground and background. For evaluation purposes we choose as foreground the segment that exhibits greater overlap with the true foreground.

Maximum likelihood segmentation: To quantify the extent to which the most expensive part of the overall algorithm, namely iterated graphcut, contributes to segmentation accuracy, we compared the algorithm against a restricted version in which pixels are assigned to one of the two GMMs as estimated from the converged state of the active contour. The pixel is assigned to the model which has the greater likelihood.

Normalised cut: We use the publicly available ncut implementation (`www.cis.upenn.edu/~jshi/software`). Again we choose as foreground the segment that exhibits greater overlap with the true foreground.

5.2. Results

Segmentation performance for the three methods are summarised in Table 1. The max KL and min Gibbs correspond to the maximum contrast and converged solution in the original GrabCut, respectively. We summarise the results in four observations.

	F-measure	Precision	Recall
k-means	78.0 ± 16.4	94.6 ± 9.5	89.2 ± 19.8
n-cut	63.8 ± 14.3	53.0 ± 20.5	68.8 ± 12.0
max likelihood	93.2 ± 4.6	88.5 ± 6.8	98.8 ± 4.2
max KL	94.0 ± 5.9	90.5 ± 6.1	98.3 ± 6.8
min Gibbs	94.1 ± 5.9	90.8 ± 6.2	98.2 ± 6.8

Table 1. Results in percentage

Firstly, the maximum likelihood method achieves remarkably good performance. Yet, the small difference in F-measure between it and the graph-based methods is visually significant in many cases, as illustrated in Figure 3. The F-measure clearly is too coarse (and global) a performance measure to accurately reflect human perception of differences in segmentation quality.

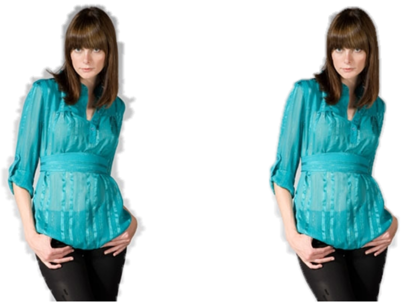

Fig. 3. Visual inspection reveals marked differences between methods of similar performance under the F-measure

Secondly, not only does the iterated graphcut perform significantly better than k-means and n-cut, performance also varies much less.

Thirdly, the point at which the contrast is maximised coincides with a performance maximum although the effect appears very weak.

Lastly, we note that the contrast criterion allows us to terminate the algorithm 60% earlier than if we waited until convergence. This criterion reduced the computational time of the original GrabCut optimisation by 16% with an average of $5s$ per image (on a standard Pentium IV machine (3.2 GHz using C++ and the OpenCV library). Note that the stopping criterion requires the computation of the KL-divergence measure which adds an extra computational burden. Figure 4 shows for one particular image how the Gibbs energy and the contrast measure change during GrabCut. The image results below illustrate how the traditional criterion may sometimes exclude protruding parts of the foreground object.

6. CONCLUSIONS

We presented an automated version of GrabCut for the purpose of reliably segmenting foreground objects against relatively simple backgrounds. We also introduced a new stopping criterion for GrabCut that allows significantly earlier termination without sacrificing segmentation quality. With 94% accuracy under the F-measure, the results are very promising and well above those of normalised cut and k-means.

7. REFERENCES

[1] D Comaniciu and P Meer, “Mean shift: a robust approach toward feature space analysis,” *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[2] J Shi and J Malik, “Normalized cuts and image segmentation,” *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

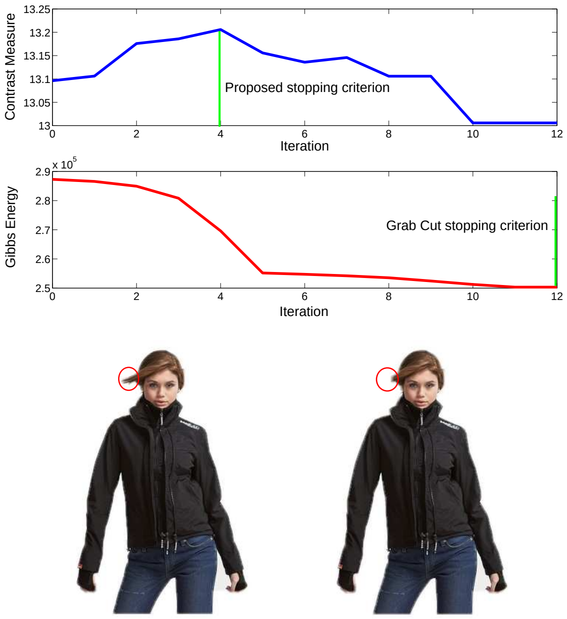

Fig. 4. Above: Changes of contrast and Gibbs energy during GrabCut; Below: Images that maximise contrast (left) and minimise Gibbs (right).

[3] A Criminisi, T Sharp, and A Blake, “GeoS: Geodesic image segmentation,” in *Proc European Conf Computer Vision*, 2008.

[4] C Rother, V Kolmogorov, and A Blake, “GrabCut: interactive foreground extraction using iterated graph cuts,” *ACM Trans SigGraph*, vol. 23, no. 3, pp. 309–314, 2004.

[5] M Unger, T Pock, and B Horst, “Interactive globally optimal image segmentation,” in *Proc British Conf Machine Vision*, 2008.

[6] A Ali and A Farag, “A novel framework for n-d multimodal image segmentation using graph cuts,” in *Proc Int’l Conf Image Processing*, 2008, pp. 729–732.

[7] F Estrada and A Jepson, “Quantitative evaluation of a novel image segmentation algorithm,” in *Proc Int’l Conf Computer Vision and Pattern Recognition*, 2005, pp. 1132–1139.

[8] M Kass A Wittkin and D Terzopolous, “Snakes: Active contour models,” in *Proc Int’l Conf Computer Vision*, 1987, pp. 259–268.

[9] A Amini, S Tehrani, and T Weymouth, “Using dynamic programming for minimizing the energy of active contours in the presence of hard constraints,” in *Proc Int’l Conf Computer Vision*, 1988, pp. 95–99.