Top Down Image Segmentation using Congealing and Graph-Cut

Douglas Moore, John Stevens, Scott Lundberg and Bruce A. Draper
Computer Science Dept., Colorado State University, Fort Collins, CO 80523 USA
$\{$*tropgeek,stevens,lundberg,draper*$\}$*@cs.colostate.edu*

Abstract

This paper develops a weakly supervised algorithm that learns to segment rigid multi-colored objects from a set of training images and key points. The approach uses congealing to learn a probabilistic spatial model of the multi-colored object class and graph-cut to separate the foreground from the background. The result is a novel approach which can segment heterogeneous objects, in contrast to other recent approaches which are better at segmenting uniform but possibly flexible objects.

1 Introduction

This paper presents a novel top down algorithm that learns to segment objects from a small number of image presentations. Object segmentation is the process of labeling image pixels as either object (foreground) or non-object (background), and has been a difficult problem in computer vision for decades [15]. Segmentation techniques fall into two broad categories: bottom up approaches, which only use data from a single image, and top down approaches, which incorporate external information such as category prototypes. Bottom up techniques are limited in their ability to segment a broad range of heterogeneous objects. Our approach assumes rigid objects seen from similar viewpoints; i.e. the appearance of the object undergoes only affine transformations among images. It also assumes a single on-target key point per image of the type produced by SIFT [10], Scale-saliency [8], or any of several interest point operators [11]. Unlike related approaches, we allow the object to be multi-colored with arbitrary surface markings. These conditions match the output of many recent attention-based object recognition systems [5, 6, 7, 19]. This algorithm can therefore be used to segment objects which have been recognized based on selective attention windows.

Given a set of images and key points, our algorithm builds a probabilistic spatial model of the target object class. Once the source images are aligned, target pixels should be relatively consistent (low entropy) while background pixels should be comparatively inconsistent (high entropy). More specifically, the spatial model is created by aligning the source images through congealing [12], and then fitting a two-class (foreground/background) Gaussian Mixture Model (GMM; [2, 4]) to the resulting entropy values. The resulting probabilistic spatial model is used as the predictive component in the graph-cut algorithm [3], which segments images by balancing a-priori expectations with image-specific edge information.

As shown experimentally in section 4, the algorithm is able to segment objects when the key points are manually chosen and therefore accurate. Section 4 also presents a perturbation study relating errors in key point position to congealing error, and therefore indirectly to segmentation error.

2 Related Work

Research into top-down image segmentation has been revolutionized by the introduction of graph-cut algorithms [9]. Graph-cut provides a well-motivated and computationally efficient method for segmenting images based on (1) a probabilistic predictive model and (2) local edge data. In essence, it finds the segmentation that simultaneously optimizes both the fit to the predicted model and the available boundary information.

Graph-cut in turn spurred research, including this work, into methods of learning predictive object models for use in top-down image segmentation [13, 17, 18]. Among previous research efforts, the closest to this work is LOCUS [17]. LOCUS begins with a class of unsegmented images which it segments through inference and learning. While able to segment deformable objects, it suffers from one of the same limitations as bottom-up techniques: it assumes that the target objects are internally homogeneous. Simon and Seitz [14] generate a predictive model based on extracted attention

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

windows from a pre-segmented template image. Our approach differs in that our algorithm generates a predictive model based on many views of a template object with an accompanying point-in-object and no pre-segmented template.

At the same time that top-down image segmentation techniques were improving, research was also expanding into object recognition based on selective attention windows (e. g. [5, 6, 7, 19]). These techniques recognize objects (or more commonly, views of objects) by extracting attention windows from images, and then labeling sets of categorized attention windows. These techniques have the disadvantage of not knowing the extents or boundaries of the objects recognized. Arora and Loeff et. al. [1] present a technique for determining object extents from keypoint templates using a conditional random field model. Our approach differs in that our algorithm generates a predictive model based on many views of a template object with an accompanying point-in-object and no pre-segmented template or random field model.

This paper presents a segmentation algorithm designed to work with attention-based recognition. Attention-based recognition algorithms identify specific object classes, based on at least one key point per object. As a result, they produce the input assumed by this algorithm (images of a single object with key points). Our segmentation algorithm extends these systems by allowing them to determine the spatial extent of the objects they have recognized.

3 Method

Our algorithm is divided into two phases, as shown in Figure 1. The first phase learns a probabilistic spatial model of the target class from a set of input images and corresponding key points. The second phase applies the spatial model to segment every source image.

Let $\mathcal{I} = I_1, I_2, ...$ be the set of input images, and let $K = k_1, k_2, ...$ be the set of corresponding key points, where $\forall i, k_i \in I_i$. The key points K are assumed to mark essentially the same point on the target object, and have estimated positions, rotations, and scales. These estimates are used by the congealer as the initial affine transformations for aligning the source images; the congealer then optimizes these transformations.

3.1 Congealing

As explained in [12], congealing is an iterative method that brings a set of images into alignment by iteratively adjusting the affine transformation parameters $T = t_1, t_2, ...$ for each image so as to minimize the sum of the pixel-wise entropy. In our algorithm congealing is applied to Sobel edge images create from the source images. To quantify how well windows are aligned, a window stack $\Omega = W(T)$ is defined, and the overall entropy of each pixel column $\Omega_{x,y}$ is computed using Vasick's entropy approximator [16] to create an entropy image e. The total entropy of the window stack is the sum of the entropy of the pixel columns $\Omega_{x,y}$ as in Equation 1 ($\mathcal{P}$ is the set of all valid pixel locations in a window).

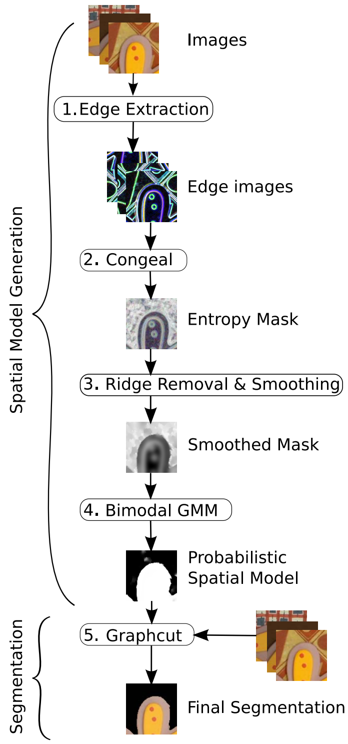

Figure 1. Outline of steps taken during segmentation.

$$H(\Omega) = \sum_{(x,y)\in\mathcal{P}} H(\Omega_{x,y}) \qquad (1)$$

Minimizing the summed pixel-wise entropy directly can lead to pathological transformations, such as scaling an image by 0. Following Miller, et. al. [12], we introduce a regularization term to avoid such transformations.

The minimization of e proceeds through a set of gradient decent searches along seven affine parameters (x-translation x_t, y-translation y_t, x-log-scale x_{ls}, y-log-scale y_{ls}, x-shear x_s, y-shear y_s, and rotation θ) using a step size ϵ for each parameter. We search over seven parameters, instead of the minimum of six parameters defined in an affine matrix, as suggested by [12].

The minimization process is repeated until convergence or a maximum number of iterations is reached. After this, the ϵ_p step sizes are decayed (a.k.a. annealed) by a factor of $\frac{1}{2}$, and the process is repeated. Once the step sizes have decayed five times the process is considered to have converged.

3.2 From Entropies to Probabilities

The output of the congealer is a set of alignment transformations and an entropy image. The assumption underlying this work is that after alignment, target pixels will look approximately the same across images and therefore have low entropy. Background pixels, on the other hand, change from image to image and consequently have higher entropy values.

We use a two-class Gaussian Mixture Model (GMM) to model the foreground and background distributions. The Gaussian process with the lower mean models target pixels, while the mixture with the higher mean models background pixels. The value of the confidence mask at each pixel is the probability that the pixel's value comes from the foreground distribution.

We observed that high entropy values can arise from two sources: they can be created by background pixels, or by the edges of markings in the interior of the target. Sub-pixel misalignments create high entropy values near high-frequency edges. To eliminate large entropy spikes caused by an object's internal edges, we first perform an edge preserving smoothing on the entropy image. Next, an edge removal process replaces each edge with the value of its nearest non-edge neighbors. A blurred version of this result is then used to create a confidence mask c. This helps eliminate the high entropy ridges within an object, enabling the result to be modeled more effectively using a two mixture GMM.

3.3 Graph Cut

The source images are segmented individually by applying graph-cut, using the spatial probability image as the source of top-down information. Graph cut is an optimization technique based on max-flow algorithms [3] which finds a global optimum to a certain class of energy functions [9]. Graph cut uses a linear weighting between local image information and a prior probability mask to optimally group similar pixels into segments while taking top down information into account. For more details, see references.

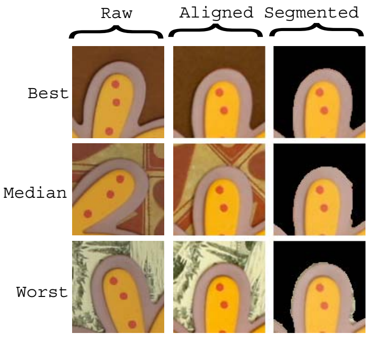

Figure 2. Results from the flower data set.

4 Experimental Results

Figure 2 demonstrates results on heterogeneous objects and backgrounds when using hand labeled key points. These results are representative of three other data sets not shown here.

The final congealed stack entropy depends on the accuracy of the key point locations. Congealing compensates for keypoint misalignments across images to a limited extent. To test how far keypoints could be misaligned, we perturbed the keypoints in an image stack from 0 to 30 pixels in random directions. Figure 3 shows the perturbations effect on the final congealed stack entropy. Small changes in keypoint location produce no change in the final congealed entropy value in the case of the flower data set, and a small change in the congealed entropy value in the case of the depth-of-field data set.

4.1 Test images

We tested the segmentation algorithm on two image sets that were selected to either maximize background variation or evaluate the performance when perspective is not constant.

Set 1: A child's toy (a simple, relatively homogeneous, object) in the shape of a flower, set on varied colorful backgrounds.

Set 2: An office chair viewed at slightly different perspectives.

The left side of Figure 4 shows the pixel-wise entropy values from Set 1 before segmentation, using the initial alignment transformations estimated from the key points. The right side of Figure 4 shows the entropy values after congealing. Together, the two sides

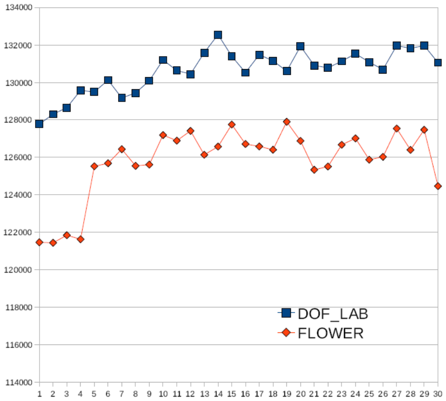

Figure 3. Stack entropy vs key point location error.

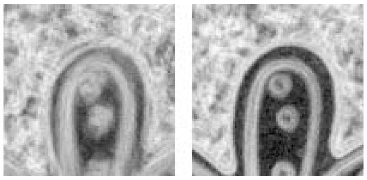

Figure 4. Example entropy image e before and after congealing.

of Figure 4 show the importance of image congealing for estimating entropy values.

5 Conclusions and Future Work

This work presents an algorithm for foreground/background segmentation, given sample images of the object class and one key point per image. None of the steps in the segmentation algorithm are novel; the congealing and graph-cut algorithms have been around for over five years, as have Gaussian Mixture Models. The contribution of this paper is the way in which these techniques have been combined to create and apply a probabilistic spatial object model. The resulting segmentation algorithm is able to segment complex heterogeneous objects based only on sample images with a single key point in each. No other top-down segmentation algorithm that we know of can do this.

References

[1] H. Arora, N. Loeff, D. Forsyth, and N. Ahuja. Unsupervised Segmentation of Objects using Efficient Learning. *CVPR*, pages 1–7, 2007.

[2] J. Bilmes. A gentle tutorial on the em algorithm and its applications to parameter estimation for gaussian mixture and hidden markov models. Technical report, International Computer Science Institute Department of EE and CS, U.C. Berkeley: Berkeley, CA., 1998.

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

[4] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(B):1–39, 1977.

[5] B. Draper, K. Back, and J. Boody. Implementing the Expert Object Recognition Pathway. *ICVS 2003, Graz, Austria, April 1-3, 2003: Proceedings*, 2003.

[6] L. Fei-Fei, R. Fergus, and P. Perona. earning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, 2007.

[7] R. Fergus, P. Perona, , and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[8] T. Kadir and M. Brady. Scale, saliency and image description. *IJCV*, 45(2):83–105, November 2001.

[9] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *PAMI*, 26(2):147–159, 2004.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[11] K. e. a. Mikolajczyk. A comparison of affine region detectors. *IJCV*, 65(1):43–72, 2005.

[12] E. Miller, N. Matsakis, and P. Viola. Learning from one example through shared densities on transforms. In *CVPR*, volume 1, pages 464–471, 2000.

[13] C. Rother, V. Kolmogorov, and A. Blake. Grabcut – interactive foreground extraction using iterated graph cuts. *SIGGRAPH*, 23(3):309–314, 2004.

[14] I. Simon and S. Seitz. A Probablistic Model for Object Recognition, Segmentation, and Non-Rigid Correspondence. *CVPR*, pages 1–7, 2007.

[15] W. Skarbek and A. Koschan. Colour image segmentation - a survey. Technical report, Institute for Technical Informatics, Technical University of Berlin, October 1994.

[16] O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(1):54–59, 1976.

[17] J. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *ICCV*, pages 756–763, 2005.

[18] S. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. *NIPS*, 1, 2002.

[19] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. *Proc. CVPR*, 2006.

We gratefully acknowledge code from Artem Sokolov, and Walter Bell (www.cs.cornell.edu/vision/wbell).