Active 3D Segmentation through Fixation of Previously Unseen Objects

Mårten Björkman
celle@csc.kth.se
Danica Kragic
danik@csc.kth.se

CSC - CAS/CVAP
Royal Institute of Technology (KTH)
Stockholm, Sweden

We present a method for active object segmentation based on integration of several cues; image point positions, binocular disparities and pixel colours. It serves as a framework for generation of object hypotheses of previously unseen objects in natural indoor scenes. The appearance, 3D shape and size of objects are modelled in an iterative manner using an approximate Expectation-Maximisation (EM) method, that takes the dependencies between neighbouring pixel labels into consideration, unlike typical methods that assume neighbouring pixels to be independent.

To better cope with situations when an object is hard to segregate from the surface it is placed on, possibly due to ambiguities in appearence, we propose a flat surface model as a complement to the two models typically used in figure-ground segmentation. A flat surface assumption is reasonable, given that most objects in indoor scenes are placed on flat, or at least locally flat, surfaces. We will show that even if no such physical plane exists in the scene, foreground segmentation succeeds anyway, since the flat surface model will become just another background model. We further let the segmentation evolve over time, this in order to provide more information and gradually improve segmentation and to facilitate tracking.

Figure-ground segmentation is done using three different models, each described by a set of parameters; the foreground model θ_f, the background model θ_b and that of the flat surface θ_s. Each pixel has an associated label $l_i \in \{l_f, l_b, l_s\}$, depending on which component it belongs to. The model parameters $\theta = \theta_f \cup \theta_b \cup \theta_s$ and the labellings of all pixels $\mathbf{l} = \{l_i\}$ are unknown and estimated from the measurements $\mathbf{m} = \{m_i\}$ at each pixel. With EM the maximum likelihood estimate of the model parameters θ is computed iteratively in two steps. In the first step (E-step) the conditional distribution $w(\mathbf{l}) = P(\mathbf{l}|\mathbf{m}, \theta')$ is computed using the current estimate θ' and in the second step (M-step) a new estimate is found by maximising $Q(\theta|\theta') = \sum_{\mathbf{l}} w(\mathbf{l}) \log P(\mathbf{m}, \mathbf{l}|\theta)$. Unfortunately, since this summation is done over N^3 different labellings, where N is the number of pixels, it quickly becomes prohibitly expensive. To make it computationally tractable we replace $w(\mathbf{l})$ with the product of the conditional marginals for each unobserved label, $w(l_i) = P(l_i|\mathbf{m}, \theta')$. Since a measurement m_i depends only on its associated label l_i, the second step becomes a maximisation of

$$Q_1(\theta|\theta') = \sum_i \sum_{l_i \in L} w(l_i) \log P(m_i, l_i|\theta), \quad (1)$$

that is a summation over just $3N$ labels. The figure-ground segmentation is implicitly determined by the marginals $w(l_i)$, which are computed with loopy belief propagation [5] in the E-step.

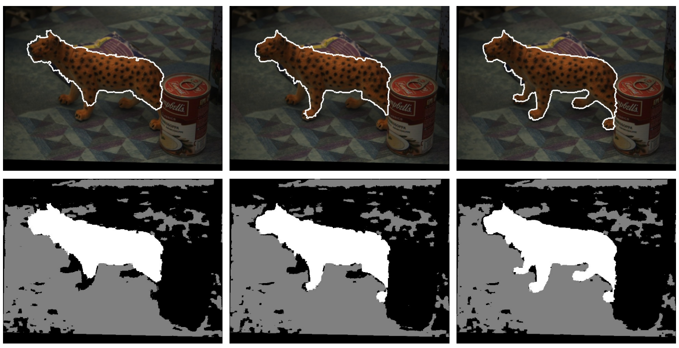

Figure 1: The 1st, 5th and 10th updates of a segmented toy cat. White areas in the bottom row show the foreground, grey the flat surface and black the remaining background points.

There are two reasons for using belief propagation, rather than maximum a posteriori (MAP) estimation with graph-cuts, which is more frequently used for segmentation [1, 4]. First, an MAP estimate might well be an extreme case, not representative of the overall distribution, since it is based on only one set of labels, the one that happens to maximise the posterior. Second, the passing of messages in belief propagation is highly deterministic and there are ways to divide the problem into steps, so that each step involves messages that are independent and can thus be easily parallalised. This is in contrast to graph-cut methods, that are less deterministic both in terms of local operations and number of required passes. Figure 1 shows some images from a real-time experiment, for which belief propagation was implemented on a GPU, resulting in an update frequency of about 8 Hz, which is fast enough for tracking.

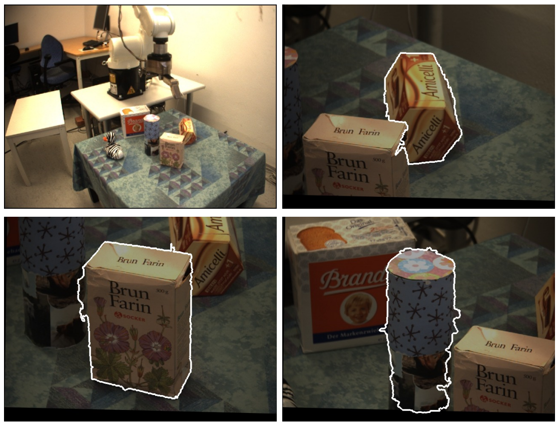

Figure 2: A wide field view in which object hypotheses are searched using attention (upper left) and segmentations of each found hypothesis in the foveated view.

Critical to any iterative segmentation system is the initialisation phase. A targeted foreground region has somehow to be pointed out, either manually or through some other mean. Unlike systems for off-line image manipulation [4], an autonomous system does not have the luxury of a human operator in the loop. Object detection has been proposed as a mean for initialisation [3], but this implies you have some model of what to detect, which is not possible when working with previously unseen objects. In this paper we instead use binocular fixation for unsupervised initialisation. While image points are densely packed, 3D points appear in clusters. These clusters may serve as bottom-up cues for object detection, regardless of appearance and shape. The only, to our knowledge, previous similar work is that of Mishra and Aloimonos [2], which however suffers from a significantly higher computational cost. The segmentation results shown in Figure 2 were autonomously produced by an attention based fixating stereo head system, that visits regions of interest and for each region segments whatever is located in the centre of view. More details on the specifics can be found in the full version of the paper.

[1] D.M. Greig, B.T. Porteous, and A.H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(2):271–279, 1989.

[2] A. Mishra and Y. Aloimonos. Active segmentation. *International Journal of Humanoid Robotics*, 6:361–386, 2009.

[3] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *European Conference on Computer Vision (ECCV02)*, pages 661–675, June 2002.

[4] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23:309–314, 2004.

[5] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12:1–41, 2000.