Active 3D Segmentation through Fixation of Previously Unseen Objects

Mårten Björkman
celle@csc.kth.se
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 appearance, 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 \( \theta_f \), the background model \( \theta_b \) and that of the flat surface \( \theta_s \). Each pixel has an associated label \( l_i \in \{\ell_f, \ell_b, \ell_s\} \), 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 \( l = \{l_i\} \) are unknown and estimated from the measurements \( m = \{m_i\} \) at each pixel. With EM the maximum likelihood estimate of the model parameters \( \theta \) is computed iteratively in two steps. In the first step (E-step) the conditional distribution \( w(l) = P(l|m, \theta^0) \) is computed using the current estimate \( \theta^0 \) and in the second step (M-step) a new estimate is found by maximising \( Q(\theta|\theta^0) = \sum l w(l) \log P(m, l|\theta) \). 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 \( w(l) \) with the product of the conditional marginals for each unobserved label, \( w(l_i) = P(l_i|m, \theta^0) \). Since a measurement \( m_i \) depends only on its associated label \( l_i \), the second step becomes a maximisation of

\[
Q_1(\theta|\theta^0) = \sum_{l \in L} \log P(m_i, l_i|\theta), \tag{1}
\]

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

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.