Learning object segmentation from motion

Michael G. Ross and Leslie Pack Kaelbling

Introduction There are many image segmentation algorithms, but integrating them into larger systems is difficult. The segmentation of an image into regions implies the optimization of a region criterion appropriate to the system's task and environment. A higher-level system does not require generic image segmentation, but needs object segmentation, the division of an image into regions that correspond to the objects the system manipulates or monitors.

The appropriate definition of objects is dependent on the system and its environment. For mobile robots, manipulators, or activity monitoring systems, a useful definition of an object is "a collection of elements which undergo common motions in the world." For example, if an image contains a static desk or a moving person, each is segmented as an individual object; each is a set of elements that move, or would move, together. The common motion definition corresponds well to our intuitive definition of objects, and in humans the ability to segment moving objects develops before the ability to segment objects by color, texture, or shape properties [7].

Given these inspirations, this work attempts to create a system for learning a model for segmenting objects in static images by observing moving objects in videos. The goal is to create an object segmentation algorithm that discovers likely common-motion boundaries that are useful to a higher-level intelligent system, and that can adapt to new environments using self-supervised learning methods.

Learning a segmentation model Our training data (Figure 1) consists of a video stream, which is fed into a background subtraction algorithm [8] that outputs a series of images and a binary image of the motion boundaries present. This provides us with a large, cheaply acquired database of sample object segmentations. Some recent work in learning segmentation or edge-detection models [3, 4] has required smaller, expensive, human-labeled databases which may contain many non-object boundaries.

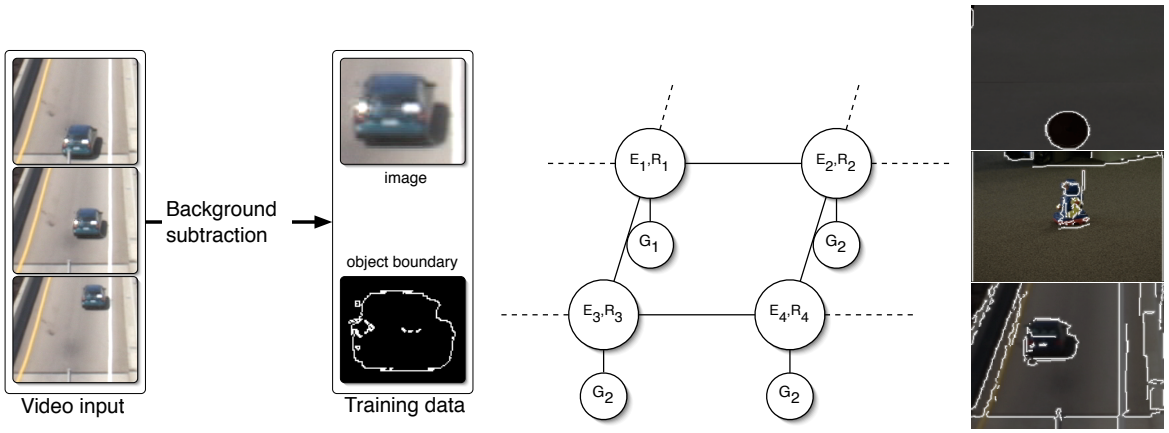

Figure 1: Left: A moving car is captured and motion-segmented using background subtraction, and then added to the segmentation training set. Center: A piece of an MRF model for image segmentation. Right: Segmentation results on a black disc, a robot, and a traffic image.

The goal of this work is to infer the object boundaries present in single static images using the shape and image statistics of such boundaries in the training data. Markov random fields (MRFs) have been used in image segmentation and low-level vision problems for nearly twenty years [2, 1] and they are the basis for our model because they allow us to capture shape properties via the noncausal interactions of neighboring boundary patches. The algorithm tiles an input image into 5 by 5 pixel, non-overlapping patches. Local brightness gradient information for each patch i is stored in a visible data node, G_i, which is independent of the rest of the model given the value of its associated boundary node.

The boundary nodes (Figure 1) represent the edges, shapes, and regional properties of the objects in the image. Each boundary node i consists of a pair of variables (E_i, R_i), where E_i specifies a type

of boundary (no boundary, straight line, corner, etc.) and R_i contains local image color and texture information. All variables are visible in training, but the E variables, representing the object boundary, are hidden in inference. Each E_i can have one of 2713 possible boundary fragment values (for details, see [6]) and interacts with its neighbors to infer the object's shape. The R data enforces observed region properties. For example, the model may learn that the boundary of a red region should not include edge fragments that border a green region.

Completed and future work Currently, we have completed an algorithm that learns information about object shape and brightness gradients from video streams and is able to discover object boundaries, sometimes in very difficult data (Figure 1)[6]. The inference and parameter learning problems of MRFs are often difficult, but approximate belief-propagation inference algorithms [5, 10] can alleviate both problems. Wainwright et al. [9] have shown that the use of belief-propagation leads to closed-form estimates of MRF compatibilities from observed marginal distributions through approximate maximum-likelihood estimation. Our Java implementation produces boundary estimates in a 150 by 150 pixel image in approximately 30 seconds on modern hardware.

In the near future, we plan to incorporate the local color and texture variables, R, into the learning and inference process. This will allow the model to capture the internal texture and color relationships of the training set objects and further improve the results. Multi-resolution models or the use of larger MRF neighborhoods can capture more shape information from the training data. We also intend to continue gathering larger data sets and to compare our inferred common-motion boundaries on the test data to those boundaries that can be detected by background subtraction.

The ultimate goal of this work is to produce an object segmentation system that adapts to new environments without supervision and produces output relevant to the larger system.

Research Support: This research is supported in part by the Office of Naval Research under contract #N00014-00-1-0298, in part by the Singapore-MIT Alliance under the November 6, 1998 agreement, and in part by a National Science Foundation Graduate Student Fellowship.

References:

[1] W.T. Freeman, E.C. Pasztor, and O.T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1), 2000.

[2] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), November 1984.

[3] S.M. Konishi, A.L. Yuille, J.M. Coughlan, and Song Chun Zhu. Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1), 2003.

[4] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, In press, 2003.

[5] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[6] M.G. Ross and L.P. Kaelbling. Learning object segmentation from video data. Technical Report AIM-2003-022, MIT Artificial Intelligence Laboratory, 2003.

[7] E.S. Spelke, P. Vishton, and C. von Hofsten. Object perception, object-directed action, and physical knowledge in infancy. In *The Cognitive Neurosciences*, pages 165–179. The MIT Press, 1994.

[8] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition*, 1999.

[9] M.J. Wainwrigh, T. Jaakkola, and A.S. Willsky. Tree-reweighted belief propagation and approximate ML estimation by pseudo-moment matching. In *Workshop on Artificial Intelligence and Statistics*, 2003.

[10] Y. Weiss. Belief propagation and revision in networks with loops. Technical Report 1616, MIT Artificial Intelligence Laboratory, November 1997.