

# A Unified Probabilistic Framework for Object Segmentation and Recognition

Huei-Ju Chen, Kuang-Chih Lee, Erik Murphy-Chutorian, and Jochen Triesch

Traditionally, image segmentation and object recognition have been viewed as two sequential stages in image analysis [1]. Images are first segmented into regions using a bottom-up approach, and then segmented regions are recognized individually. However, purely bottom-up segmentation approaches have difficulty to handle cluttered natural scenes. Consider a simple scene shown in Figure 1 (a). The Normalized Cut [3], a state-of-the-art segmentation technique, fails to segment the three objects on the table from the background. One way to overcome this problem is to simultaneously tackle segmentation and recognition. However, only quite recently have there been attempts to integrate or unify the segmentation and recognition problems [4, 5]. Tu *et al.* proposed a Bayesian framework to unify segmentation and recognition using a Data Driven Markov Chain Monte Carlo method [4]. Their model successfully recognizes and segments two object classes, text and faces, by utilizing two specific detection engines. However, it is not clear how this approach can be scaled up to make recognition of hundreds or thousands of objects tractable. Moreover, it appears that their system may have severe problems when dealing with objects that are mostly occluded.

We present a novel and effective probabilistic approach to integrate object segmentation and recognition. The probabilistic framework is modelled by a Bayesian belief network which consists of three latent variables representing object segmentation, recognition hypotheses, and shared features extracted at interest points. In this structure, segmentation and recognition occur simultaneously, and the integrated task of both is formulated as a *maximum a posteriori* estimation problem. Figure 1 (f) illustrates the structure of our model. Latent variables  $S$ ,  $F$ , and  $H$  respectively represent **segmentation**, **shared features** which include number of active features ( $n^f$ ), feature identities and their locations, and **object hypothesis** which includes object identity, location of presence, and associated contours.  $G$ ,  $E$  are two evidences representing the edge and Gabor-jets information from the input image  $I$ . According to the structure of the belief network,

the following three important posterior probability distribution can be decomposed to

$$P(F|G, H, S) = \alpha_1 P(F|H)P(G|F), \quad (1)$$

$$P(H|S, F) = \alpha_2 P(S|H)P(F|H)P(H), \quad (2)$$

$$P(S|E, F, H) = \alpha_3 P(S|H)P(E|S), \quad (3)$$

where  $\alpha_1, \alpha_2, \alpha_3$  are normalization constants. The interpretation of each posterior probability in Equations 1, 2, and 3 can be associated with the problem of feature activation, object recognition, and segmentation.  $P(G|F)$  represents the likelihood of the activation pattern of shared features given the observed Gabor-jets map.  $P(F|H)$  represents the degree of association between object hypotheses and active shared features.  $P(H)$  denotes prior probabilities of object hypotheses.  $P(S|H)$  computes how well hypothesized object contours match a sampled segmentation instance.  $P(E|S)$  computes the degree of consistency between a segmentation instance and the observed edge map. Initialization of the system is performed by generating a belief on the feature node first by computing  $P(G|F)$  and then propagating the beliefs to the object hypothesis node by computing  $P(F|H)$ , and then to the segmentation node by computing  $P(S|H)P(E|S)$ . This path of the initialization drastically reduces computational complexity of our system since the instances in the F and H nodes can be picked almost deterministically. From this initialization, we maximize the posterior probabilities by stochastic sampling (e.g. Gibbs sampling), which iteratively propagates belief messages between these three nodes. Finally, we resolve partial occlusions by checking the edge consistency in the boundary of the overlapping areas between each pair of objects.

Our model scales well to large object databases because the complexity of our belief network does not increase with the number of the objects we want to recognize. This is achieved by representing all objects with the same type(s) of features (i.e. constellations of Gabor jets). The scene is analyzed in an active vision fashion on one object at a time. Furthermore, these features are *shared*, as proposed by Murphy-Chutorian *et al.* [2],



(a) Test Scene 1



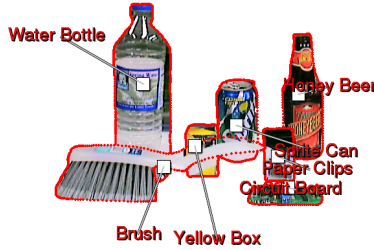
(b)



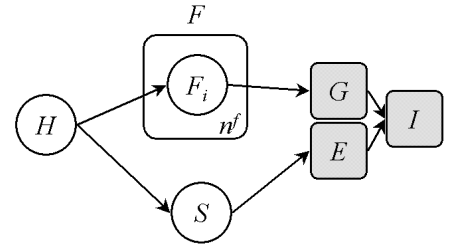
(c)



(d) Test Scene 2



(e)



(f)

Figure 1: **(a)** Test scene 1 **(b)** The segmentation and recognition results from our proposed algorithm. **(c)** The segmentation result from the Normalized Cut algorithm of the test scene 1. **(d)** Test scene 2 **(e)** Results of the test scene 2. The red lines depict the contour of each object, and show that the partial occlusion has been resolved correctly for each object. **(f)** Graphical representation of the proposed Bayesian belief network. The shaded box nodes denote the evidences. The circles denote the hidden variables. The big plate around hidden variables comprises  $n^f$  number of shared features  $F_i$ .

such that adding a new object model benefits from re-using features learned for the recognition of other objects. The experimental results shown in Figure 1 (b) and (e) demonstrate our system can successfully recognize and segment objects despite partial occlusions.

In our model, we start from the shared feature node, instead of the segmentation node, to initiate beliefs of the network. It would be impractical to initialize from the segmentation node since it is much slower and more difficult to estimate its belief given an edge map. We hypothesize that human vision should also work in a similar fashion in order to efficiently achieve multiple object recognition and segmentation in complex natural scenes with significant occlusions. Initially, the activations of a limited set of patterns best associated with the current visual scene are generated in a fast feedforward neural network. Then recurrent connections are used to validate or refine each of these hypotheses.

## References

- [1] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W H Freeman, 1983.
- [2] E. Murphy-Chutorian and J. Triesch. Shared features for scalable appearance-based object recognition. In *Proc. of IEEE Workshop on Applications of Computer Vision (WACV 2005)*, Breckenridge, Colorado, USA, 2005.
- [3] J. Shi and J. Malik. Normalized cuts and image segmentation. *Computer Vision and Pattern Recognition (CVPR)*, pages 731–737, 1997.
- [4] Z. Tu, X. Chen, A. Yuille, and a. Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and object recognition. *International Journal of Computer Vision*, 2004.
- [5] S. Yu, R. Gross, and a. J. Shi. Concurrent object segmentation and recognition with graph partitioning. *Proceedings of Neural Information Processing Systems (NIPS)*, 2002.