Vistas: Hierarchial boundary priors using multiscale conditional random fields.

Jonathan Warrell
j.warrell@cs.ucl.ac.uk
Alastair P. Moore
a.moore@cs.ucl.ac.uk
Simon J. D. Prince
s.prince@cs.ucl.ac.uk

Department of Computer Science
University College London
Gower Street
London
WC1E 6BT, UK

http://pvl.cs.ucl.ac.uk

Abstract

Boundary detection is a fundamental problem in computer vision. However, boundary detection is difficult as it involves integrating multiple cues (intensity, color, texture) as well as trying to incorporate object class or scene level descriptions to mitigate the ambiguity of the local signal. In this paper we investigate incorporating a priori information into boundary detection. We learn a probabilistic model that describes a prior for object boundaries over small patches of the image. We then incorporate this boundary model into a mixture of multiscale conditional random fields, where the mixture components represent different contexts formed by clustering overall spatial distributions of boundaries across images and image regions (vistas). We demonstrate this approach using challenging real-world road scenes. Importantly, we show that recent spectral methods that have been used in state-of-the-art boundary detection algorithms do not generalize well to these complex scenes. We show that our algorithm successfully learns these boundary distributions and can exploit this knowledge to improve state-of-the-art boundary detectors.

1 Introduction

Detection of natural [14] or occlusion [11] boundaries is a fundamental problem in computer vision. Unlike edge detection, boundary detection involves integrating multiple cues (intensity, color, texture) along with trying to incorporate object class or scene level descriptions to mitigate the ambiguity of the local signal. Recent work shows that boundary detection performance remains low on most real world datasets [18].

Despite these difficulties boundary detection remains an important component of many vision pipelines. For instance, boundary information has been used as the basis for feature vectors to facilitate recognition [2, 22], as a measure of region affinity for segmentation [17] and as the basis for energy terms in random field models [21].

Traditional methods for finding boundaries focused on edge detection models with white noise [5]. However, while these assumptions still prove useful under certain conditions they have been shown to generalize less reliably to natural images [14]. Therefore, recent work on boundary detection has taken three notable directions: First, improving the low-level unary

BMVC 2009 doi:10.5244/C.23.88

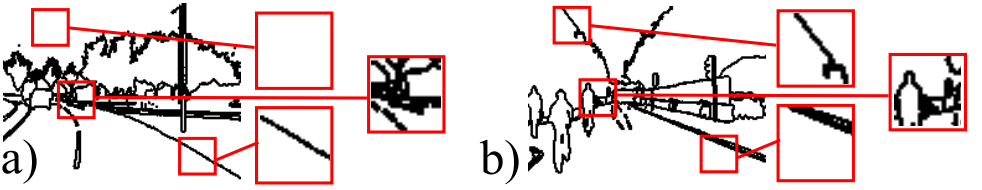

Figure 1: Non-stationary boundary density. Example patches from different regions of images of road scenes. Note the change in boundary density according to the distribution of objects in the scene. For example, the centre tends to me more cluttered. These road scene images are examples of training data adapted from publicly available road sequences [3].

detection has focused on learning statistics [14] or exemplars [6] of boundaries from large datasets using hand-labeled ground truth. Second, there has been much work on mid-level cues including multi-scale detection [18, 25], curvilinear continuity [8], contour completion [16, 19] and other grouping cues [1]. Lastly, recent work has integrated higher order scene cues [11] or the information from the eigen-spectrum of the image [13].

Until recently, the evaluation of boundary detection performance has tended to concentrated on images with low scene complexity composed of only a few object instances and recent boundary detection analysis [18] shows that performance on scenes with higher scene complexity is considerably lower. However, the scene similarity in the datasets used in [18] tends to be low.[1] Consequently, it is difficult to learn scene priors for boundaries without first learning something about the objects in the scene. Alternatively, recent work [15] has shown that for datasets with greater scene consistency [3] it is possible to learn a distribution over the density of boundaries in the image and use this to successfully guide segmentation. The work in [15] exploits the fact that when a 2D image is a projection of a 3D scene, perspective effects result in an uneven distribution of the sizes of object classes, and therefore an uneven distribution in the density of object boundaries across the scene. This observation is illustrated in Figure 1. This suggests learning a non-stationary model for boundary priors. To achieve this we draw on two methods that have proved effective for image labeling problems [9, 10] and learn a mixture of multiscale conditional random fields.

The layout of the paper is as follows: In Section 2 we outline the boundary distribution model of [15], which we take as our model of boundary patches. In Section 3 we extend this to a hierarchical model, where we learn clusters on the distribution of boundaries across an image. In Section 4.1 we benchmark a variety of boundary detection methods, along with our new prior, on a fully labeled image database with high scene complexity.

2 A Generative Model for Boundary Patches

We begin by describing a generative model for boundary patches. This is a Clustered Latent Trait model (CLT), as has been used previously in [15] to model boundary distributions at the image level. We represent the n'th boundary patch, which includes P pixels, by a vector of discrete variables $\mathbf{x}_n = [x_{n1} \ldots x_{nP}]^T$, where x_{np} takes the value 1 where a boundary is present, and 0 otherwise. $\mathbf{x}_{np}$ is taken to be produced by a generative process as depicted in Figure 2a. Each patch is assigned first to a cluster c_n, which may take 1 of K_c values, and is drawn from

[1] Dataset[#images]: CMU[30] motion boundary dataset [23], MSRC[519] object dataset [21], PASCAL Challenge 07[422] segmentation competition [4] and LabelMe database (Boston houses 2005)[218] [20].

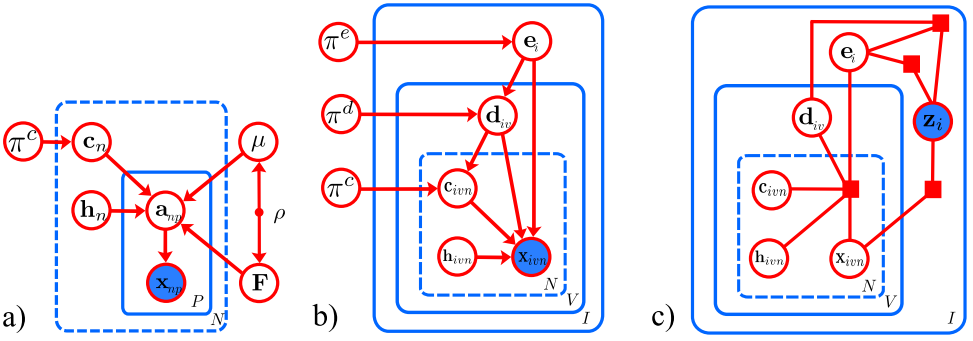

Figure 2: Graphical Models. a) A Clustered Latent Trait (CLT) model for patches. The plates denote a set of N discrete image patches each with P pixels. b) Hierarchical CLT model. The plates denote a set of I images each with a set of V vistas (a grid of non-overlapping large image regions), each with a set of N discrete image patches. The CLT model for image patches can be seen marked with a dashed plate in each model. c) Embedding the same variables in a CRF conditioned on the observations $\mathbf{z}$ (see Equation 6).

a multinomial distribution with parameters π^c. Each patch is also assigned a position $\mathbf{h}_n$ in a continuous subspace of dimension J, where $\mathbf{h}_n$ is drawn from a zero mean unit covariance Gaussian distribution, and may be taken to be a parameterizations of the kinds of variation that can be applied to the patch. Having chosen $c_n = k$ and $\mathbf{h}_n$, a continuous activation $\mathbf{a}_n$ is produced via a linear combination of the mean μ_k for cluster k, and factors $\mathbf{f}_{1k} \ldots \mathbf{f}_{Jk}$, for the same cluster (which form the rows of $\mathbf{F}_c$), where the latter are weighted by the variable $\mathbf{h}_n$: $\mathbf{a} = \mu_c + \mathbf{F}_c\mathbf{h}$. These activations are then converted to probabilities by passing them through the logistic sigmoid function, and $\mathbf{x}_n$ is generated by taking independent samples at each pixel. We can thus write the entire generative process as:

$$Pr(c_n = k) = \pi_{ck} \quad (1)$$

$$Pr(\mathbf{h_n}) = \mathscr{G}_{\mathbf{h}}[\mathbf{0}, \mathbf{I}] \quad (2)$$

$$Pr(\mathbf{a_n}|\mathbf{h_n}, c_n = k) = \delta_{\mathbf{a_n}}(\mu_k + \mathbf{F}_k\mathbf{h_n}) \quad (3)$$

$$Pr(\mathbf{x_n}|\mathbf{a_n}) = \prod_{p=1}^{P} \mathrm{Bin}_{x_{np}}[\sigma(a_{np})] \quad (4)$$

where $\mathscr{G}_\alpha[\beta, \Gamma]$ represents a Gaussian in variable α with mean β and covariance Γ. The function $\delta_\alpha(\beta)$ denotes a probability distribution over α where all of the mass is at β and hence describes a deterministic relationship. The function $\mathrm{Bin}_\alpha[\beta]$ denotes the binomial likelihood of observing value α given binomial parameter β, and σ is the logistic sigmoid function, $\sigma(a) = 1/(1+\exp(-a))$.

The CLT model can be learnt from a training set of example binary patches by using an E-M algorithm. In the E-step, the patches are assigned to the MAP cluster c_n and position in the latent space $\mathbf{h}_n$ given the current estimates of the parameters $\theta = \{\pi_{1\ldots k}, \mu_{1\ldots k}, \mathbf{F}_{1\ldots k}\}$. In the M-step, these parameters are then updated by using a quasi-Newton method to estimate the MAP of the data likelihood and a hyperprior, ρ, set to influence expected smoothness.

3 A Hierarchical Multiscale CRF Model

The preceding section described a model for boundary patches. In this section, we show how this model can be embedded in a hierarchical multiscale CRF to provide a model for entire scene boundary images. We describe the model in two stages, first introducing the hierarchical aspects in section 3.1, and then the multiscale ones in 3.2. Sections 3.3 and 3.4 then describe how we perform learning and inference. An overview of the different levels of the model is provided in Figure 3.

3.1 Hierarchy

We begin by enlarging our generative model as shown in Figure 2b. Here, we imagine that we have I images, each containing V 'vistas' (defined as a 4×4 grid of non-overlapping image regions),[2] each containing N independent patches. We introduce further latent variables d and e into the generative process, where e_i represents which 1 of K^e scene-clusters the image as a whole belongs to, and d_{iv} represents which 1 of K^d vista-clusters vista v in image i belongs to. These are drawn from multinomial distributions parameterized by π^e and π^d respectively, where π^d includes a separate set of multinomial parameters for each value of e^i. As before, each patch also has its own latent variable taking 1 of K^c values and position in a continuous latent space, which are now written c_{ivn} and $\mathbf{h}_{ivn}$ to indicate that they refer to the n'th patch of the v'th vista in image i. We now assume that we have learned separate μ and $\mathbf{F}$ parameters for every combination of values $\{e_i, d_{iv}, c_{ivn}\}$, and generate $\mathbf{x}_{ivn}$ by first forming the activation $\mathbf{a}_{ivn} = \mu_{c_{ivn},d_{iv},e_i} + \mathbf{F}_{c_{ivn},d_{iv},e_i}\mathbf{h}_{ivn}$, before passing it through the logistic sigmoid function and sampling. The entire generative model can be written as follows, where Pr_{pr} indicates that we have here a model for the prior (as opposed to Pr_{un}, which will be used later to indicate a unary term) and $\text{Mult}_\alpha[\beta]$ is the multinomial likelihood of observing α given parameters β:

$$\begin{aligned} Pr_{pr}(e_i, \mathbf{d}_i, \mathbf{c}_i, \mathbf{h}_i, \mathbf{x}_i) &= Pr_{pr}(e_i) Pr_{pr}(\mathbf{d}_i|e_i) Pr_{pr}(\mathbf{c}_i|e_i, \mathbf{d}_i) Pr_{pr}(\mathbf{h}_i) Pr_{pr}(\mathbf{x}_i|e_i, \mathbf{d}_i, \mathbf{c}_i, \mathbf{h}_i) \\ &= \text{Mult}_{e_i}[\pi^e] \cdot \prod_v \text{Mult}_{d_{iv}}[\pi^d_{e_i}] \cdot \prod_v \prod_n (\text{Mult}_{c_{ivn}}[\pi^c_{e_i,d_v}] \cdot \mathscr{G}_{\mathbf{h}_{ivn}}[\mathbf{0}, \mathbf{I}]) \cdot \\ &\quad \prod_v \prod_n \prod_p \text{Bin}_{x_{ivnp}}[\sigma(a_{ivnp})] \end{aligned} \tag{5}$$

The model so far developed is a generative prior. However, we are interested in the conditional probability of a boundary map given an observed image, which we shall call $\mathbf{z}_i$. One possibility would be to extend our generative model of Figure 2b to include $Pr(\mathbf{z}_i|\mathbf{x}_i)$, and then perform inference via Bayes theorem. This is inefficient though, since we do not require a fully generative model of images from boundary maps for our purposes. Instead, we embed the generative model outlined in a conditional random field model (CRF), where the conditional probabilities learned become potential terms in the CRF, and a set of unary terms is placed between $\mathbf{z}_i$ and each x_{ivnp} derived from estimates by a discriminative classifier of $Pr(x_{ivnp}|\mathbf{z}_i)$. The chosen form of the CRF is illustrated in Figure 2c in plate notation. The model can be written:

$$Pr(e_i, \mathbf{d}_i, \mathbf{c}_i, \mathbf{h}_i, \mathbf{x}_i|\mathbf{z}_i) \propto \underbrace{Pr_{un}(\mathbf{x}_i|\mathbf{z}_i)}_{\text{unary term}} \cdot \underbrace{Pr_{pr}(e_i, \mathbf{d}_i, \mathbf{c}_i, \mathbf{h}_i, \mathbf{x}_i)^\lambda}_{\text{prior term}} \cdot \underbrace{\phi_1(e_i|\mathbf{z}_i) \cdot \phi_2(\mathbf{d}_i|e_i, \mathbf{z}_i)}_{\text{scene and vista constraints}} \tag{6}$$

[2] The idea being that a vista is a region "seen through a long, narrow avenue or passage" and it therefore represents the boundary distribution over a sub-region of the full image. This may be assumed to account for the specific distribution of objects that may vary within one particular scene.

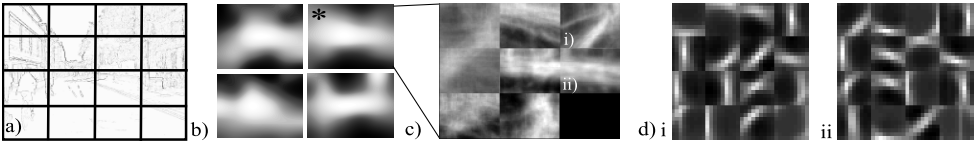

Figure 3: Clusters learnt at scene, vista and patch levels with the hierarchical model (boundary probability maps for the means of each cluster, $\sigma(\mu)$, shown). a) Example boundary map divided into vistas. b) Possible scene clusters ($\sigma(\mu^e)$). The cluster * is the one to which the image belongs (see Equation 7). c) Possible vista clusters ($\sigma(\mu^d)$) corresponding to scene cluster * (see Equation 8). d) Possible patch clusters ($\sigma(\mu^c)$) corresponding to vistas i and ii (see Equations 4 and 5). Note the varying distribution of orientations of the patch clusters.

Here, $Pr_{un}(\mathbf{x}_i|\mathbf{z}_i)$ denotes the conditional probability derived from a unary classifier. $Pr_{pr}(e_i,\mathbf{d}_i,\mathbf{c}_i,\mathbf{h}_i,\mathbf{x}_i)$ is the prior probability, as in Equation 5. We note that this has an extra weighting term, λ, to control its influence. In addition, we add two extra 'constraining' terms ϕ_1 and ϕ_2, which directly link the scene clusters and vista clusters with the observed image. We model these terms again using a CLT model, but now learnt over subsampled boundary maps of entire images or vistas. The value of these terms can be found in general by setting an arbitrary threshold $\tau = 0.5$ to the unary response map, and then finding the posterior value for the particular scene and vista clusters e_i and $\mathbf{d}_i$ given that label map. In practice, we adopt deterministic forms for these potentials, setting them to 1 for the maximum likelihood scene and vista cluster values, and 0 for other values:

$$\phi_1(e_i|\mathbf{z}_i) = \delta_{e_i}(\text{argmax}_k \max_h Pr(\psi_\tau(\mathbf{z}_i)|k,h,\{\mu^e,\mathbf{F}^e\})) \tag{7}$$

$$\phi_2(\mathbf{d}_i|e_i,\mathbf{z}_i) = \prod_v \delta_{d_{iv}}(\text{argmax}_k \max_h Pr(\psi_\tau(\mathbf{z}_i)|k,h,\{\mu^d_{e_i},\mathbf{F}^d_{e_i}\})) \tag{8}$$

where $\psi_\tau(\mathbf{z})$ calculates the unary classifier response thresholded at τ, and $Pr(\psi_\tau(\mathbf{z}_i)|k,h,\{\mu^e,\mathbf{F}^e\})$ and $Pr(\psi_\tau(\mathbf{z}_i)|k,h,\{\mu^d_{e_i},\mathbf{F}^d_{e_i}\})$ are further CLT distributions learnt at the image and vista levels (and evaluated using Equations 1 to 4). Figure 3 illustrates the way the CLTs at these and the patch level capture boundary information at different resolutions.

Because we are no longer treating the model as fully generative, there is now no reason for the boundary patches $\mathbf{x}_{ivn}$ to be fully independent, and in general we can consider overlapping these across the image. Figure 4a shows a possible 1-d expansion of the graph in Figure 2c, where we have five overlapping patches (c_{111} - c_{122}) each containing 2 pixels. The graph, contains 2 miniature vistas of three pixels each, controlled by d_{11} and d_{12}.

3.2 Multiscale

The model in Figure 2c can be considered as a hierarchical mixture of CRFs (see [10]), since the settings of the latent variables $\mathbf{d}$ and $\mathbf{e}$ control the forms of the patch potentials over the $\mathbf{x}$'s. This allows us to tailor our expectations concerning local boundary shapes depending on the overall appearance of the image, or subregion (vista). In addition though, we would like to embody expectations of boundary shape at multiple scales within a vista.[3] For this

[3] Particularly, we found when experimenting with synthetic data that one scale was often an insufficient cue for boundary completion, but with two or more combined, the larger scales could help the smaller ones.

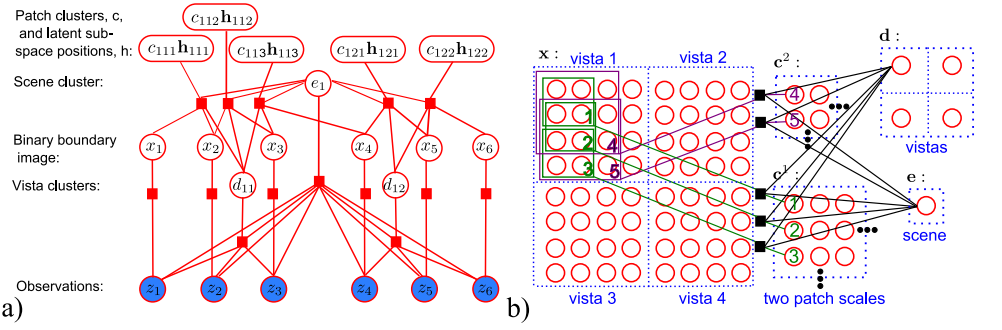

Figure 4: Full CRF model with a) overlapping patches and b) multiple scales. a) 1-d model of a 6 pixel image. The one scene, e_1, includes two non-overlapping vistas, d_{11}, d_{12}, of three pixels each. Overlapping patches of size two pixels are associated with variables $\mathbf{c}$ and $\mathbf{h}$. The various cliques in Equation 6 are shown by the interconnections between these variables. b) 2-d multiscale model. The 4 vistas are associated with latent variables $\mathbf{d}$, and the scene as a whole with $\mathbf{e}$. The model includes overlapping patches at two scales of 2×2 and 3×3 pixels, where examples are marked respectively as $\{1,2,3\}$ and $\{4,5\}$ in their bottom right corners in $\mathbf{x}$. These patches are associated with latent variables in $\mathbf{c}^1$ and $\mathbf{c}^2$, and all belong to vista 1. Note that the $\mathbf{h}$'s are not included and no cliques involving $\mathbf{z}$ are shown.

purpose, we let our patches within a vista $\mathbf{c}_{iv}$ range over levels $j = 1...L$, encompassing varying numbers of pixels. We thus write c^j_{ivn} to denote the latent variable associated with the n'th patch at level j. In addition, we assume that the parameters $\mathbf{F}$ and μ are tied across each level, but not between levels, so we must now consider that we have not only separate versions of these parameters for every scene and vista cluster, but also every level within the vista (i.e. $K^c \times K^d \times K^e \times L$ parameter sets). Further, we can now split the prior term in Equation 6 into separate terms for each level j, and introduce a separate weighting λ_j for each. We thus change the prior term in equation 6 to:

$$\underbrace{\prod_j Pr_{pr}(e_i, \mathbf{d}_i, \mathbf{c}^j_i, \mathbf{h}_i, \mathbf{x}_i)^{\lambda_j}}_{\text{prior term}} \tag{9}$$

The overlapping cliques of several sizes over the $\mathbf{x}$'s produces a multiscale CRF model as in [9]. As a whole then, it may be described as a hierarchical mixture of multiscale CRFs. Figure 4b provides an illustrative example of the CRF formed where we have 4 vistas per image, and 2 levels of overlapping patches within each vista (of sizes 2×2 and 3×3 pixels respectively). For simplicity, only cliques not involving the observations $\mathbf{z}$ are shown.

3.3 Learning

We adopt a piecewise approach to training, as in [21]. We thus assume we have a discriminative classifier which can be trained independently to provide the unary potentials. The CLT models for the constraint potentials, ϕ_1 and ϕ_2, are simply trained on sub-sampled images and vistas from the training set ground truth boundary labelings. This involves first training the image-level CLT, assigning image clusters to all training images, and then sampling vistas from each group to train the vista-level CLT for that cluster. Image and vista level clusters can then be assigned to each training example, and patches at sizes $1...L$ drawn from

the ground truth to train the CLT patch models for every combination. An assumption of independent patches is thus made during training, so that the model is as in Section 2.

3.4 Inference

Our goal during inference on image i is either to draw a sample from the posterior $Pr(\mathbf{x}_i|\mathbf{z}_i)$, or more generally to estimate this posterior (or its marginals at each pixel), providing a discrete or a continuous boundary value at each pixel respectively. For this task, we use a version of block Gibbs sampling which is made possible by the restricted Boltzmann machine form of the model [9], where the latent variables are independent of each other given the labels, and vice-versa. The sampling process can be initialized by thresholding the results of the unary classifier to set $\mathbf{x}_i$. Since we adopt a deterministic form of the constraint potentials ϕ_1 and ϕ_2, we can then directly set the image cluster e_i, and the vista clusters $\mathbf{d}_{i,1\ldots V}$ by choosing the MAP values from their respective CLT models. This implicitly chooses for us the CLT models that will be used for each patch within the image, and we can then sample from the joint distribution by alternately picking c and $\mathbf{h}$ values for each patch (across all levels j), and then re-sampling the $\mathbf{x}$'s. We choose to take the MAP values for the c's and $\mathbf{h}$'s, although we note that we could probabilistically sample from these as well. Further, if we chose to use soft constraining potentials on e and $\mathbf{d}$, these could also be re-sampled each iteration, making it possible to rectify an initial cluster misassignment. Finally, a discrete sample from the posterior is generated by outputting $\mathbf{x}_i$ after a fixed number of iterations, and a continuous boundary map can be generated by finding the marginal for $\mathbf{x}$ at each pixel given the final setting of the latent variables via Equation 6 (alternatively we could record the $\mathbf{x}_i$ samples drawn across a large number of iterations, and then find the frequencies of boundary positives at each location). The process is summarized in Algorithm 1.

Algorithm 1 Hierarchical CLT Inference Algorithm

Input : Observed image features, $\mathbf{z}_i$
Initialize : $\mathbf{x}_i$ unary classifier with threshold
Set e_i to the MAP image cluster given $\mathbf{x}_i$
Set $\mathbf{d}_{i,1\ldots V}$ to the MAP vista clusters given $\mathbf{x}_{iv}$ and e_i
for all iterations **do**
  for all levels j **do**
    for all patches $\mathbf{x}_{ivn}$ at level j **do**
      Find the likelihood of patch $\mathbf{x}_{ivn}$ for each setting of c^j_{ivn} using the MAP value for $\mathbf{h}^j_{ivn}$, and the CLT model given the values of e_i and d_{iv}.
      Set c^j_{ivn} to its MAP value, along with its accompanying $\mathbf{h}^j_{ivn}$
    end for
  end for
  Re-sample the $\mathbf{x}_i$'s given the current latent variables and unary terms
end for
Output : The final sample $\mathbf{x}_i$ and estimate of the marginal image $Pr(\mathbf{x}_i)$

4 Evaluation

Our evaluation is based on video sequence stills and human-labeled ground truth from the CamVid database [3]. This consists of road scene sequences taken from the passenger seat

of a moving car. This is a challenging dataset that includes 32 classes and ego-motion. We follow [4] and use sequences 06R0 and 16E5 for training and 05VD for testing.

To learn the model presented in the previous sections we construct training data from the set of 406 binary ground truth images. For learning scene and vista clusters we down-sample the original image data from 720×960 and 180×240 to 36×48 respectively using an **OR** operation. As these clusters only learn coarse distributions of boundaries it is possible to learn them at the reduced scale. We learn the following CLT models: Images: 3 factors, 4 clusters, ρ=100; Vistas: 3 factors, 8 clusters, ρ=10; Patches at three scales 9×9,19×19 and 39×39 with 6 factors and 16 clusters, ρ=5. In a similar manner to the images, patch scales 19 and 39 are scaled to 10×10 for training and testing. These parameter settings were set by hand and involve no validation set. During training we separate the boundaries for each object class into separate training patches. This allows us to mitigate double boundaries created by the presence of a void label at the expense of modeling boundary junctions and close parallel edges between different classes. However the proportion of boundary junctions in the training set is small [13] and a proper treatment of junctions is left for future work. We also learn the weighting for the three scales setting $\lambda = \{0.035, 0.007, 0.007\}$ (small, medium, large) using a subset of of 62 images of the training data for validation.

Performance on the dataset is evaluated using the precision-recall framework of [14]. The objects in the CamVid dataset are of very uneven sizes and the objects at a distance can be separated by only a few pixels, moreover unlike the datasets in a recent comparative study [18] the ground truth is fully labeled. We therefore show P/R curves using a distance tolerance for boundary matching of 0.4% of the image diagonal, slightly stricter than that of 0.6% used in [18], resulting in a tolerance of 5 pixels. Again to mitigate the effect of the void class, where a boundary can effectively be labeled twice from either side of the object, we separate the boundaries in masks for each class and use the multi-subject framework [14] to match between boundaries.

Results showing the improvement gained by combining our boundary distribution prior with the BEL classifier [6] as the unary term can be seen in Figure 5f. We can see a modest improvement along the PR curve, which is most noticeable in the low-recall/high-precision range. There is also an improvement in the maximum f-measure score (from 0.46 to 0.47). Figure 5 gives a qualitative comparison of unary and full CRF results for selected images.

4.1 Comparison to other algorithms

To assess the the effect of alternative techniques on this new dataset we benchmark six other boundary detection algorithms. These include simple edge detection algorithms like MATLAB's implementation of the Canny detector and another well know scale space method [12][4]. In contrast to other datasets there is a noticeable difference in the performance of competing algorithms. For instance, the current best performance on the Berkeley Segmentation Database is the gPb algorithm [13]. However, it performs poorly here suggesting that there is less useful information in these street scenes to be found in the spectral components of the image, except in the low recall range where it outperforms other variants of the Pb algorithm. While the maximum f-measure is achieved by mPb [13] its performance in the mid recall range is low. Conversely, while the f-measure score is low the Canny detector provides very reasonable performance in certain ranges. Overall the performance of all competing methods is low which suggests there is room for significant improvement when considering specific complex scenes.

[4]Code provided by Mark Dow at http://lcni.uoregon.edu/~mark/SS_Edges/SS_Edges.html

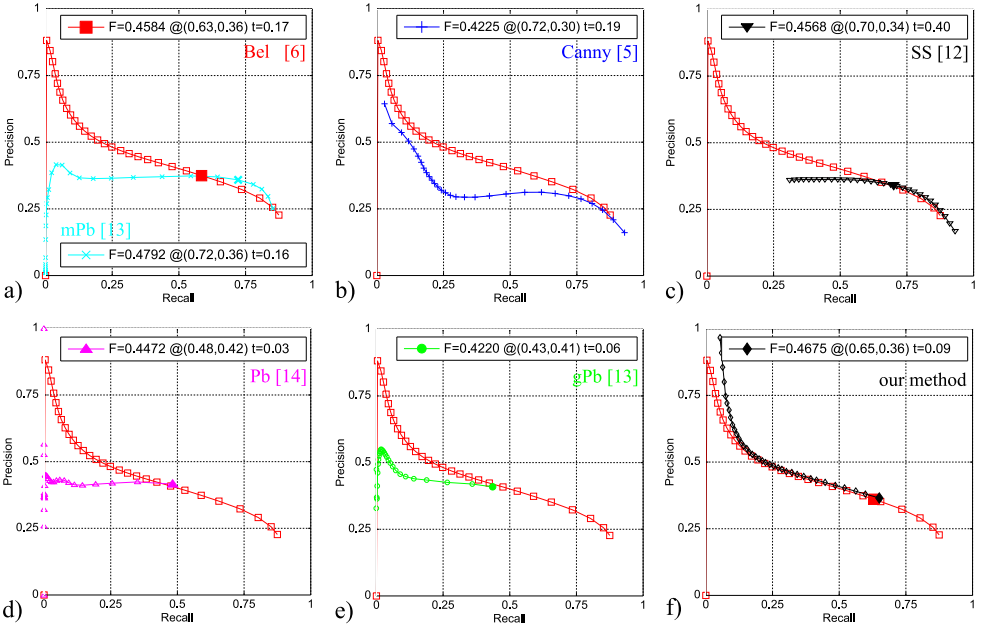

Figure 5: Precision Recall curves. a)-e) Precision/Recall curves for 6 competing boundary detection methods with BEL [6] for reference. f) Precision recall curve for our method using the BEL classifier as the unary term. Note a modest improvement along the length of the PR curve with our CLT prior having greater effect in the regions of lower recall.

5 Summary

In this paper we have introduced a novel prior for learning boundary distributions. Our model exploits these distributions at different scales in the image to learn local models for boundaries. We have performed boundary analysis on a new dataset and shown that state-of-the-art algorithms perform poorly compared to other more frequently used databases. Furthermore, we have shown for this particular database that our novel prior provides a useful source of extra information to existing state-of-the-art unary classifiers.

However, the increase in performance is only modest and this suggests two notable directions of enquiry: Firstly, improving the learning in our model. We note that the distribution of data to clusters is uneven with 75% of scenes in the test set being drawn from one image cluster. The scenes in the test data are less varied than the training data and the benefit of clustering in the model may be underestimated using this partition of the data. It is also possible to learn the full model rather than pursue piece-wise learning as we have done here, and extend our inference algorithm to sample probabilistically from all of the latent variables. It would also be interesting to continue our analysis for other unary classifiers and investigate if similar improvements in performance could be achieved. Secondly, it is likely that a full solution to the problem of boundary detection in complex scenes will use object class information [24]. This may include both class specific unary detectors as well as novel priors based on the estimation of objects in the scene. Lastly, in future work we would like to see if our model can be profitably extended to datasets without consistent scenes.

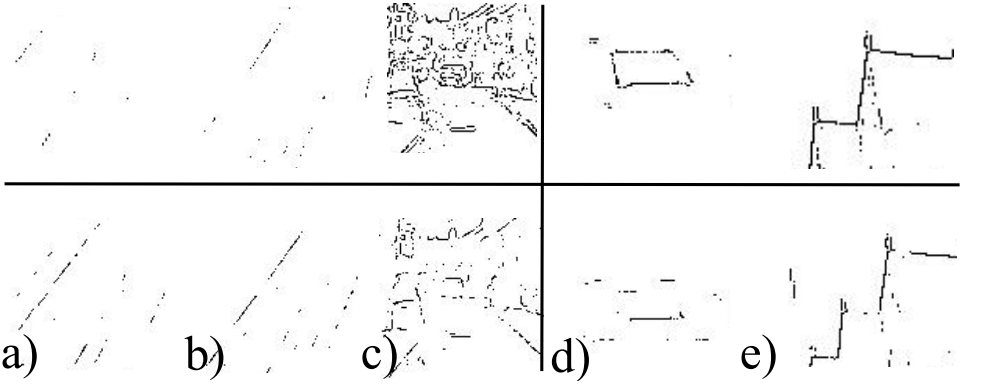

Figure 6: Good and Bad examples. Top row thresholded unary term. Bottom row incorporating prior. a)-c) Example patches where the prior has either completed boundaries or removed clutter from the patch window. a)-b) show examples of strong boundary completion that result in better performance in the low recall range. d)-e) Examples where the prior has removed useful boundary information.

Acknowledgements The first two authors contributed equally to the work in this paper. We acknowledge the support of the EPSRC Grant No. EP/E013309/1. We thank Gabriel Brostow for help with the CamVid dataset.

References

[1] Pablo Arbelaez. Boundary extraction in natural images using ultrametric contour maps. *Workshop on Perceptual Organization in Computer Vision*, 2006.

[2] Serge Belongie, Jitendra Malik, and J. Pusicha. Matching shapes. *International Conference of Computer Vision*, 1:454–461, 2001.

[3] Gabriel Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth databse. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[4] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. *European Conference on Computer Vision*, 1:44–57, 2008.

[5] John Canny. A computational approach to edge detection. *PAMI*, 8(6):679–698, 1986.

[6] Piotr Dollár, Zhuowen Tu, and Serge Belongie. Supervised learning of edges and object boundaries. *Computer Vision and Pattern Recognition*, 2:1964–1971, 2006.

[7] Mark Everingham, Luc Van Gool, Chris Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge 2007 (voc2007). URL `http://www.pascal-network.org/challenges/VOC/voc2007/workshop`.

[8] Pedro Felzenszwalb and David McAllester. A min-cover approach for finding salient curves. *Workshop on Perceptual Organization in Computer Vision*, 2006.

[9] Xuming He, Richard D. Zemel, and Miguel A. Carreira-Perpinan. Multiscale Conditional Random Fields for Image Labeling. *CVPR*, 2:695–702, 2004.

[10] Xuming He, Richard S. Zemel, and Debajyoti Ray. Learning and incorporating top-down cues in image segmentation. *ECCV*, 1:338–351, 2006.

[11] Derek Hoiem, Andrew N. Stein, Alex A. Efros, and Martial Hebert. Recovering occlusion boundaries from a single image. *ICCV*, 1:1–8, 2007.

[12] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30:77–158, 1998.

[13] Michael Maire, Pablo Arbelaez, Charless Fowlkes, and Jitendra Malik. Using contours to detect and localize junctions in natural images. *CVPR*, 2008.

[14] David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.

[15] Alastair P. Moore, Simon J. D. Prince, Jonathan Warrell, Umar Mohammed, and Graham Jones. Scene shape priors for superpixel segmentation. *ICCV*, 2009.

[16] X. Ren and J. Malik. A probabilistic multi-csale model for contour completion based on image statistics. *European Conference on Computer Vision*, 1:312–327, 2002.

[17] X. Ren and J. Malik. Learning a classification model for segmentation. *International Conference on Computer Vision*, 1:10–17, 2003.

[18] Xiaofeng Ren. Multi-scale improves boundary detection in natural images. *European Conference on Computer Vision*, 2008.

[19] Xiaofeng Ren, Charles C. Fowlkes, and Jitendra Malik. Scale-invariant contour completion using conditional random fields. *ICCV*, 2:1214–1221, 2005.

[20] B. C. Russell, Antonio Torralba, Kevin Murphy, and W. Freeman. Label me: a database and web-based tool for image annotation. Technical report, MIT, 2005.

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *European Conference on Computer Vision*, 1:1–15, 2006. doi: 10.1007/11744023_1.

[22] Jamie Shotton, Andrew Blake, and Roberto Cipolla. Contour-based learning for object detection. *International Conference of Computer Vision*, 1, 2005.

[23] Andrew Stein, Derek Hoiem, and Martial Hebert. Learning to find object boundaries using motion cues. *International Conference of Computer Vision*, 2007.

[24] Paul Sturgess, Karteek Alahari, Lubor Ladicky, and Philip H. S. Torr. Combining appearance and structure from motion feature for road scene understanding. *British Machine Vision Conference*, 2009.

[25] Stella Yu. Segementation induced by scale invariance. *CVPR*, 1:444–451, 2005.