DATA-DRIVEN TREE-STRUCTURED BAYESIAN NETWORK FOR IMAGE SEGMENTATION

Kittipat (Bot) Kampa[†]*, Jose C. Principe*[†]*, Duangmanee (Pew) Putthividhya*[⋆]*, Anand Rangarajan*[†]

[†] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, 32611.
[⋆]eBay Inc., 2065 Hamilton Ave. San Jose, CA, 95125.

ABSTRACT

This paper presents Data-Driven Tree-structured Bayesian network (DDT), a novel probabilistic graphical model for hierarchical unsupervised image segmentation. Like[1, 2], DDT captures long and short-ranged correlations between neighboring regions in each image using a tree-structured prior. Unlike other previous work, DDT first segments an input image into superpixels and learn a tree-structured prior based on the topology of superpixels in different scales. Such a tree structure is refered to as data-driven tree structure. Each superpixel is represented by a variable node taking a discrete value of class/label of the segmentation. The probabilistic relationships among the nodes are represented by edges in the network. The unsupervised image segmentation, hence, can be viewed as an inference problem of the nodes in the tree structure of DDT, which can be carried out efficiently. We evaluate quantitatively our results with respect to the ground-truth segmentation, demonstrating that our proposed framework performs competitively with the state of the art in unsupervised image segmentation and contour detection.

***Index Terms*—** Unsupervised image segmentation, tree structure, Bayesian networks, graphical models, superpixels.

1. INTRODUCTION

Unsupervised image segmentation has long been an important subject of research in computer vision and image understanding. In the Bayesian formulation, a smoothing prior is employed to capture spatial correlations that exist between neighboring regions. Early work on Bayesian image segmentation focused on undirected models with lattice structure, e.g. Markov random fields (MRFs) [3] , whereby hidden units corresponding to true image labels are assumed connected using undirected lattice graph. Such a structure, however, has proved to be computationally expensive and inference on such a graph is generally NP-hard [1].

An alternative formulation using directed graphical models, i.e. Bayesian networks, was subsequently proposed. Feng *et al.* [1] used a tree-structured belief network (TSBN)—a quadtree Bayesian network [2] with 4 neighboring pixels sharing the same parent node. The hierarchical tree structure allows TSBN to capture multi-scale correlations (both short and long-ranged) similar to MRF while maintaining efficiency in inference and parameter estimation. One main drawback of TSBN, however, is that the fixed tree structure used to enforce the local homogeneity between neighboring pixels disregards the natural object boundaries, which often results in "blocky" segmentation output. Several approaches have been proposed to address this problem by introducing complex, cross-linked model [2], which in turn requires more complex inference algorithms such as junction tree algorithm or loopy-belief propagation .

Recently, the use of graphical models with adaptable structures has been proposed in [4] . By treating the model structure as a random variable and adapting the structure to fit each input image, the proposed models significantly mitigate the disagreement between the segmentation boundary and the natural object boundary. The need to re-estimate the model structure in every iteration indeed incurs significant computation cost—a major drawback for this line of approach.

In this work, we propose a novel probabilistic graphical model called Data-Driven Tree-Structured Bayesian Networks (DDTs) where all good merits of tree and hierarchical multi-scale structure are preserved. The tree structure of DDT is built according to the similarity of image regions in an input image, thus can describe approximately the orientation of objects in the image. As a result, the short- and long-ranged correlations encoded by DDT can be more precise than that of TSBN [1]. Unlike the flexible structure graphical models proposed in [4] , DDT does not need to adapt its structure in every iteration which drastically reduces the computation of the algorithm. In addition, instead of using pixel as the finest information scale, we use superpixel [5], a group of locally smooth labeling region, which contains more descriptive contextual information of the corresponding image region than at the pixel level. Experimental results demonstrate that our method performs competitively to the state of the art.

2. PROBABILISTIC MODEL OF DDT

DDT is a directed acyclic graph (DAG) with 2 disjoint sets of random variables; hidden and observed, graphically represented by round-shaped and rectangular-shaped nodes respectively, as depicted in Fig. 1. Hidden and observed nodes are associated with image sites which in general can be any

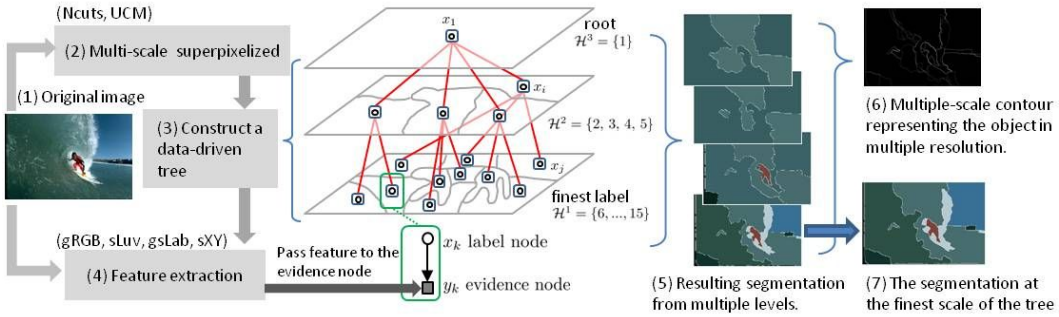

Fig. 1. The overview of Data-Driven Tree-Structured Bayesian Network (DDT) framework. The original image (1) is over-segmented in multi-scale hierarchical manner in process (2). (3) The corresponding DDT is built according to the superpixels in each level. (4) The features are extracted from the original image corresponding to each superpixel. (5)-(7) After learning and inference algorithm, the resulting hierarchical segmentation can be interpreted from level $\mathcal{H}^1$.

arbitrary image region, e.g. pixels, block of pixels, or superpixels. Throughout the paper, we associate image sites with superpixels as mentioned earlier in the text. All nodes in the structure are connected by directed edges indicating conditional independence assumption between them.

The hierarchical tree structure of DDT can be described as follows. There are L levels in the hierarchy, where $\mathcal{H}^l$ denotes the set of indices in the level $l \in \{1, ..., L\}$, and, when it is appropriate, the level $\mathcal{H}^l$ is used to denote the level l of the DDT. The root level $\mathcal{H}^L$ contains only a single superpixel, which is equivalent to the entire input image (the corresponding index set $\mathcal{H}^L = \{1\}$). As the level index l decreases along the depth of the tree, the number of the nodes containing in each level increases. The level $\mathcal{H}^1$ thus corresponds to the finest scale of the resulting segmentation of the DDT. Laying beneath the level $\mathcal{H}^l$ is the evidence level $\mathcal{E}^l$, whose nodes are the observed image features, which are connected to their corresponding parent nodes in the level $\mathcal{H}^l$ in a one-to-one manner. The indices of nodes in the tree are used in ascending order from the root to the leaf. The set of total indices in the structure is denoted by $\mathcal{H} \cup \mathcal{E}$, with the hidden and observed nodes denoted as $\mathcal{H} = \{\mathcal{H}^l\}_{l=1}^L$ and $\mathcal{E} = \{\mathcal{E}^l\}_{l=1}^L$ respectively. The number of nodes in the hidden and evidence level l is $|\mathcal{H}^l|$ and $|\mathcal{E}^l|$ respectively, and we denote the total number of nodes in the structure as $N = |\mathcal{H}| + |\mathcal{E}|$. Generally, we have $|\mathcal{H}^1| > |\mathcal{H}^2| > \cdots > |\mathcal{H}^L|$, and the same applies for $\mathcal{E}$, which constitutes a pyramidal tree.

The structure connectivity is characterized by an $N \times N$ adjacency matrix Z, where z_{ji} takes a value 1 if node $j \in \mathcal{H}^l$ and $i \in \{0, \mathcal{H}^{l+1}\}$ are connected. Connections are established under a constraint that a node in level l can only connect to a parent node in the adjacent upper level $l+1$ except the root node which connects to null. Therefore, a realization of *structure* matrix Z can have at most one entry equal to 1 in each row. Unlike [4], the structure Z in our framework is not a random variable as it is learned from each input image from a separate algorithm and remain fixed throughout the process of segmenting the image.

Each round-shaped node in Fig. 1 represents discrete random variable x_j, of an image site j in the level l, which takes a value from a label set $\mathcal{C}^l = \{1, ..., C^l\}$ such that $x_{jc} = 1$ and $x_{j\acute{c};\acute{c}\neq c} = 0$ if the image site j has class label $c \in \mathcal{C}^l$. This notation facilitates the use of the following expression for the multinomial conditional probability of hidden node x_j given its parent x_i as $p(x_j|x_i, \theta_{ji}; l) = \prod_{v=1}^{C^l} \prod_{u=1}^{C^l} \phi_{jivu}^{x_{jv}x_{iu}}$, where ϕ_{jivu} denotes the class transition probability $p(x_{jv} = 1|x_{iu} = 1)$ which is conventionally referred to as *conditional probability table* (CPT). For the sake of computational stability, let us assume that the CPT is shared among all the nodes in the same level, i.e. $\phi_{jivu} = \phi_{\acute{j}\acute{i}vu} = \phi_{lvu}$ for $l \in \{1, ..., L-1\}$. Consequently, $\phi = \{\phi_{lvu}\}$ collectively denotes the CPT of the model which is given by $p(x_j|x_i, \phi; l) = \prod_{v=1}^{C^l} \prod_{u=1}^{C^l} \phi_{lvu}^{x_{jv}x_{iu}}$.

Note that this framework is unsupervised, hence the probabilistic model of a class c is not provided a priori, and the number of labels allowed for each input image must be defined before executing the algorithm. Estimating the appropriate number of classes for each image (the model selection problem) is explained in more detail in Section 3.

We introduce observed variables represented by shaded square-shaped nodes in the structure as illustrated in Fig. 1. Each observed random variable $y_e \in \mathcal{R}^d$ of an image site $e \in \mathcal{E}$ represents the relevant image features such as color or texture which take on continuous values. Extensive details on our choice of features will be discussed in the experimental results section. We model the feature vector y_e using a multivariate Gaussian distribution given as: $p(y_e|x_i; l) = \prod_{c=1}^{C^l} \mathcal{N}(y_e|\mu_c, \Lambda_c^{-1}; l)^{x_{ic}}$, where $\mathcal{N}(y_e|\mu_c, \Lambda_c^{-1}; l)$ is the Gaussian distribution with μ_c and Λ_c are $D^l \times 1$ mean parameter and $D^l \times D^l$ precision matrix for class c in the level l respectively. Generally, the cardinality of the label set $\mathcal{C}^l$ decreases as the level l decreases.

Using the notation described above, we can now write the hidden labels collectively as $X = \{x_j\}_{j \in \mathcal{H}}$ and the observed

image features as $Y = \{y_e\}_{e\in\mathcal{E}}$. The log-likelihood of the complete data can be expressed as:

$$\begin{aligned}&\log p(X,Y|Z,\theta)\\&=\sum_{l=1}^{L}\left(\sum_{e\in\mathcal{E}^l}\sum_{i\in\mathcal{H}^l} z_{ei}\sum_{c\in\mathcal{C}^l} x_{ic}\log\mathcal{N}(y_e|\mu_c,\Lambda_c^{-1};l)\right)\\&+\sum_{l=1}^{L}\left(\sum_{j\in\mathcal{H}^l}\sum_{i\in\{0,\mathcal{H}^{l+1}\}} z_{ji}\sum_{v\in\mathcal{C}^l}\sum_{u\in\mathcal{C}^{l+1}} x_{jv}x_{iu}\log\phi_{lvu}\right).\end{aligned}$$

All the parameters of the joint can be grouped in the set $\theta = \{\phi_{lvu}, \mu_{c\in\mathcal{C}^l}, \Lambda_{c\in\mathcal{C}^l}\}$, $\forall l \in \{1,...,L-1\}$, $\forall c,v,u \in \{1,...,C^l\}$. Note that the connectivity structure matrix Z is assumed known from each input image, hence is excluded from the parameter set.

We present a maximum likelihood estimation algorithm for DDT model parameter θ using Expectation-Maximization (EM) algorithm. In M-step, by maximizing the expectation of the log-likelihood w.r.t. θ, we derive closed-form update equations for μ_c, Λ_c and ϕ_{lvu} as follows:

$$\mu_{c\in\mathcal{C}^l} = \frac{\sum_{e\in\mathcal{E}^l}\sum_{i\in\mathcal{H}^l} z_{ei}\left\langle x_{ic}\right\rangle_{p(X|Y,Z,\theta^{t-1})} y_e}{\sum_{\acute{e}\in\mathcal{E}^l}\sum_{\acute{i}\in\mathcal{H}^l} z_{e\acute{i}}\left\langle x_{\acute{i}c}\right\rangle_{p(X|Y,Z,\theta^{t-1})}}$$

$$\Lambda^{-1}_{c\in\mathcal{C}^l} = \frac{\sum_{e\in\mathcal{E}^l}\sum_{i\in\mathcal{H}^l} z_{ei}\left\langle x_{ic}\right\rangle_{p(X|Y,Z,\theta^{t-1})}(y_e-\mu_c)(y_e-\mu_c)^\top}{\sum_{\acute{e}\in\mathcal{E}^l}\sum_{\acute{i}\in\mathcal{H}^l} z_{e\acute{i}}\left\langle x_{\acute{i}c}\right\rangle_{p(X|Y,Z,\theta^{t-1})}}$$

$$\phi_{lvu} = \frac{\hat{\phi}_{lvu}}{\sum_{\acute{v}}\hat{\phi}_{l\acute{v}u}},$$

where $\hat{\phi}_{lvu} = \sum_{j\in\mathcal{H}^l}\sum_{i\in\{0,\mathcal{H}^{l+1}\}} z_{ji}\left\langle x_{jv}x_{iu}\right\rangle_{p(X|Y,Z,\theta^{t-1})}$ denotes the unnormalized class-transition CPT and $\left\langle f(x)\right\rangle_{q(x)}$ denotes the expectation of function $f(x)$ respect to the distribution $q(x)$. The update equations above require that we compute the following expectation terms: $\left\langle x_{ic}\right\rangle_{p(X|Y,Z,\theta^{t-1})}$ and $\left\langle x_{jv}x_{iu}\right\rangle_{p(X|Y,Z,\theta^{t-1})}$, which can be done efficiently using a sum-product algorithm.

We can infer the distribution at each node of the the hidden level ($\mathcal{H}^l$) using *maximum posterior marginal* (MPM) which is optimal for sitewise $0-1$ loss function . That is, we label image site j with label $c^* = \arg\max_{c\in\mathcal{C}^l} p(x_j = c|Y,Z,\theta^*;l)$, which can be computed using the sum-product algorithm as $p(x_j = c|Y,Z,\theta^*) = \left\langle x_{jc}\right\rangle_{p(X|Y,Z,\theta^*)}$.

3. EXPERIMENT AND RESULTS

We report our results from experiment on segmentation and matching of boundary images from Berkeley Segmentation Data Set and Benchmarks 300 (BSDS300), reported in [6].

Although the main focus of this framework is for unsupervised image region segmentation, we also report our results as a contour detection problem. The quantitative comparison of the region segmentation is based on two performance measures: 1) Probabilistic Rand Index (PRI) measures the consistency of segmentation between the calculated and the groundtruth segmentation and 2) Variation of Information (VoI) metric measures the distance between two segmentations.

The performance of contour detection is based on the precision-recall framework for image segmentation developed by [7]. The measure evaluates the contour detection performance in terms of precision (P) and recall (R), whose harmonic mean, namely F-measure, at the optimal detector threshold, summarizes the detection quality. All the experiments are run in MATLAB r2010a on an Intel(R) Core(TM)2 Duo CPU E8400 @ 3.00GHz machine with Ubuntu operating system.

As opposed to our previous work [8] on DDT, our proposed multiscale evidence DDT (meDDT) has evidence in all the levels of the tree structure built from the superpixels generated by ultrametric contour map (UCM) [6]. The number of tree levelis fixed at 5 throughout the experiment, and the number of superpixels in each level is 1, 10, 20, 150, and 300 respectively. The data-driven structure is built from majority-overlapping criteria as proposed in [8]. The number of labels in each input image is estimated by GMM segmentation with Bayesian information criteria (BIC). More specifically, we first randomly picking 10% of the points in each superpixel of the input image, then applying GMM segmentation with BIC on the feature space. The number of classes is each level is determined separately using the same method. In this experiment, we use only CIELuv and the location features.

The meDDT is compared against 4 contour detection algorithms made available publicly: MeanShift [9], NCuts[10], Felzenszwalb and Huttenlocker (FH) [11] and complete ultrametric contour mapping (comp_UCM) [6] which is the state of the art. The results are listed in Table 1 in the descending order with respect to PRI, top is the best. In terms of contour detection, meDDT outperforms all other candidate algorithms due to its bigger F-measure, except for comp_UCM. When compared using region segmentation, meDDT also outperforms MeanShift and NCuts, but not comp_UCM and FH, because of its larger PRI and smaller VoI.

The comp_UCM is obtained by accumulating several hundreds of hierarchical contours of an input image together, hence the resulting contour is concentrated on the objects appearing in several scales. However, meDDT only uses 5 levels of UCM, as opposed to several hundreds of them, to form multiscale contour. We shall evaluate the selected 5-level contours and name it as init_UCM as we use it to initialize the meDDT. The contour detection performance of init_UCM is slightly degraded from the comp_UCM, but still competitively outperforms meDDT in contour detection, but not in region segmentation.

We also compare meDDT against Gaussian Mixture models for INDependent superpixel (GMiND) which is meDDT with all the tree-structured prior removed. Visually, GMiND segmentation results are noisier than that of meDDT as shown

	Contour			Region		
algorithms	P	R	F	PRI	VoI	run-time
meDDT	0.68	**0.69**	0.68	**0.77**	**2.16**	1 min
init_UCM	**0.70**	0.69	0.69	0.76	3.07	-
GMiND	0.70	0.68	0.69	0.76	2.19	30 sec
reDDT	0.67	0.73	**0.70**	-	-	-
comp_UCM	-	-	0.70	0.81	1.65	-
Meanshift	-	-	0.63	0.76	2.48	-
NCuts	-	-	0.62	0.72	2.93	-
FH	-	-	0.58	0.78	2.66	-

Table 1. Contour detection and region segmentation on BSDS300. The algorithm is in descending order with respect to PRI, top is the best. The optimal single scale for region segmentation is $l = 2$.

in Fig. 2, however, that can occasionally be an advantage in some images containing a lot of details. Quantitatively, meDDT is superior to GMiND in region segmentation evaluation, but is inferior to GMiND in contour detection because BSDS300's human-made ground-truth seem to prefer multiscale contours with significant details.

Additionally, we observe that the segmentations from meDDT seem to distinguish the object contour well, but does not seem to preserve the details. On the other hand, the init_UCM seems to be noisy and preseve well the details of the contour. Therefore, we reinforce init_UCM by combining with the multiscale contours result from meDDT, resulting in reinforced DDT (reDDT) which outperforms all of its descendant contours.

4. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a probabilistic graphical model framework for unsupervised image segmentation called data-driven tree-structured Bayesian network (DDT). The proposed framework improves the segmentation result obtained from fixed quad-tree structure method with the less computational cost than the dynamic structure methods, hence combines good merits of both extremes. We have introduced a robust and inexpensive way to build such a tree structure which is applicable to any superpixel algorithms in general. The meDDT using only color and location features outperforms existing algorithms and can perform competitively with the state of the art. This motivates us to extend our experiment to richer set of features.

In this paper, we have focused our study on unsupervised image segmentation, however, the proposed framework can also be extended to supervised and semisupervised framework as well. Furthermore, the spirit of this framework is probabilistic model containing nodes and edges, so this framework is applicable to not only images, but also to any structured data.

Fig. 2. Segmentation results. The original images (top) and corresponding ground-truth multiscale contours (2^{nd} row). The multiscale contours produced by our meDDT (3^{rd} row) and GMiND (4^{th} row). The optimal scale segmentation of meDDT (5^{th} row) and GMiND (6^{th} row).

5. ACKNOWLEDGEMENTS

This research is partially supported by the Office of Naval Research (ONR), Code 321OE. The authors would like to thank Bilal Fadlallah for his help on evaluation code on Ubuntu.

6. REFERENCES

[1] Xiaojuan Feng, C.K.I. Williams, and S.N. Felderhof, "Combining belief networks and neural networks for scene segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 4, pp. 467 –483, Apr. 2002.

[2] C.A. Bouman and M. Shapiro, "A multiscale random field model for bayesian image segmentation," *Image Processing, IEEE Transactions on*, vol. 3, no. 2, pp. 162 –177, Mar. 1994.

[3] S.Z. Li, *Markov random field modeling in computer vision*, Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1995.

[4] S. Todorovic and M.C. Nechyba, "Dynamic trees for unsupervised segmentation and matching of image regions," *IEEE transactions on pattern analysis and machine intelligence*, pp. 1762–1777, 2005.

[5] Xiaofeng Ren and Jitendra Malik, "Learning a classification model for segmentation," *Computer Vision, IEEE International Conference on*, vol. 1, pp. 10, 2003.

[6] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 898–916, 2011.

[7] D.R. Martin, C.C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 5, pp. 530 –549, may 2004.

[8] Kittipat Kampa, Duangmanee Putthividhya, and Jose Principe, "Irregular tree-structured bayesian network for image segmentation," in *Proceedings of the 2011 International Workshop on Machine Learning for Signal Processing (MLSP 2011)*, 2011, vol. x, pp. xxx–xxx.

[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 603–619, 2002.

[10] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888 –905, Aug. 2000.

[11] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.