Detection of Artificial Structures in Natural-Scene Images Using Dynamic Trees

Sinisa Todorovic and Michael C. Nechyba
Department of Electrical and Computer Engineering,
University of Florida, Gainesville, FL 32611,
{sinisa, nechyba}@mil.ufl.edu

Abstract

We seek a framework that addresses localization, detection and recognition of man-made objects in natural-scene images in a unified manner. We propose to model artificial structures by dynamic tree-structured belief networks (DTSBNs). DTSBNs provide for a distribution over tree structures that we learn using our Structured Approximation (SVA) inference algorithm. Furthermore, we propose multiscale linear-discriminant analysis (MLDA) as a feature extraction method, which appears well suited for our goals, as we assume that man-made objects are characterized primarily by geometric regularities and by patches of uniform color. MLDA extracts edges over a finite range of locations, orientations and scales, decomposing an image into dyadic squares. Both the color of dyadic squares and the geometric properties of extracted edges represent observable input to our DTSBNs. Experimental results demonstrate that DTSBNs, trained on MLDA features, offer a viable solution for detection of artificial structures in natural-scene images.

1. Introduction

Generally speaking, recognition of man-made objects in natural-scene images entails three related components: (1) localization, (2) detection and, finally, (3) recognition of objects. A number of factors contribute to the difficulty of this problem including variations in camera quality and position, wide-ranging illumination conditions, and extreme scene diversity. We seek a framework that is sufficiently expressive to cope with this uncertainty, and jointly addresses the three sub-problems in a unified manner. Thus, we resort to a probabilistic framework, which offers a principled solution to the outlined challenges.

For the purposes of this paper, we assume that man-made objects are characterized primarily by geometric regularities, and that artificial structures are rigid and composed of smaller, uniformly colored sub-parts. In the literature, several techniques for extraction and subsequent statistical modeling of geometric regularities in images exist. For example, in [1] the authors examine the problem of grouping line segments, extracted from images of natural scenes, into geometrically significant components useful for image interpretation. Their algorithm groups extracted edges into larger geometric structures using geometric relations of collinearity, parallelness, relative angle and spatial proximity. Most importantly, they represent extracted lines as nodes of a graph, where the geometric relations between the lines are links in this graph. However, their graph structure accounts only for nearest neighbor relations, failing to capture more complex artificial structures. Recently, in [2] a multiscale graphical model – namely, the tree-structured belief network (TSBN) – has been used for detecting man-made structures in natural-scene images. Reported work on TSBNs demonstrates the powerful expressiveness of TSBNs, as they represent pixel neighborhoods of varying sizes, and the efficiency of their linear-time inference algorithms [3, 4]. Despite these successes, the fixed structure of nodes in TSBNs gives rise to "blocky" segmentations. Building off of the prior work, in this paper, we propose to model man-made objects by dynamic tree-structured belief networks (DTSBNs) [5, 6]. By providing a distribution over tree structures, as illustrated in Fig. 1, DTSBNs alleviate the shortcomings of TSBNs.

In [5], we showed that it was possible to assign physical meaning to DTSBN structures, such that root nodes model whole objects, while parent-child connections en-

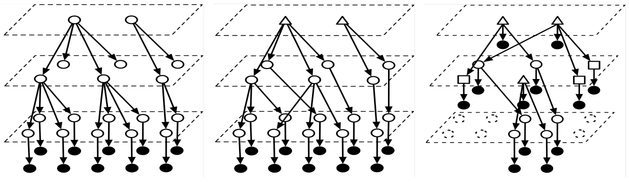

Figure 1. (a) TSBN; (b) DTSBN as in [5]; (c) our generalized-structure DTSBN.

code component-subcomponent relationships. Therefore, within the DTSBN framework, the treatment and recognition of object parts requires no additional training, but merely a particular interpretation of the tree/subtree structure. This property is very important for achieving our goal, as natural-scene images, for the most part, contain clutter and partially occluded object appearances. Thus, in the case of occlusion the recognition of an object part can lead to ultimate recognition of that object as a whole.

In this paper, we introduce even greater flexibility in the structure of DTSBNs, as compared to previously discussed models in [5, 6]. To accommodate for the specifics of our feature extraction, we allow for both leaf nodes and roots to occur at various layers, as illustrated in Fig. 1c. Moreover, we propagate observable information to higher levels of DTSBNs (black nodes in Fig. 1c).

Besides our model choice, the selection of a sufficiently discriminative set of image features is also critical for successful artificial-object detection. With respect to the aforementioned assumptions on man-made object appearances, we seek an efficient edge extraction method. In the literature, there are numerous heuristic solutions based on using local operators such as wavelets or Gabor filters. However, recent findings on human vision and natural image statistics report that cortical cells are not only highly sensitive to the location and scale, but also to the orientation and elongation of stimuli [7]. Moreover, the basis elements which best "sparsify" natural scenes are highly direction-specific, unlike wavelets. Finally, it is well known that wavelets do not economically represent even straight edges, let alone more complicated geometrical structures in images. Therefore, we suggest that there is a need for new image analysis methods that should exhibit, aside from the multiscale and localization properties of wavelets, also, characteristics that account for concepts beyond the wavelet framework. Herein, we contemplate that both geometric and color cues should be taken into account for optimal discrimination of man-made objects from other image classes. Thus, we propose multiscale linear-discriminant analysis (MLDA), which encodes both color and texture through a dynamic representation of image details [8]. MLDA extracts edges over a finite range of locations, orientations and scales, decomposing an image into dyadic squares of uniform color. Thus, MLDA appears perfectly suited for our goals, and in agreement with our assumptions that man-made objects exhibit geometric regularities and contain patches of uniform color. MLDA features represent the observable inputs to our DTSBNs.

Comparative studies with "standard" modeling paradigms (e.g., as in [2, 4, 6]) show that allowing for the more flexible structure of DTSBNs yields improved performance. While we defer discussion of ultimate object recognition for a forthcoming paper, for space reasons, herein, we address localization and detection of man-made objects; that is, we present DTSBN-based image segmentation in unsupervised settings. Our experimental results demonstrate that DTSBNs, modeling MLDA features, capture important structural information about man-made objects, and, therefore, offer a viable solution for artificial-object detection in natural-scene images.

2. Dynamic Trees

We define a DTSBN as a directed graphical model with V nodes, as depicted in Fig. 1c, representing both hidden (white) and observable (black) random variables. Below, we first introduce the set of hidden random variables.

The network connectivity is represented by a matrix Z, where entry $z_{ij}{=}1$ if there is a connection between nodes i and j. We define the distribution over tree connectivity as

$$P(Z) = \prod_{i,j\in V} [\gamma_{ij}]^{z_{ij}} \ , \tag{1}$$

where γ_{ij} is the probability of i being the child of j.

Each node i in the graph is characterized by the position r_i of the object part it represents relative to the position of its parent r_j, thereby explicitly expressing geometric component-subcomponent relationships. The joint probability of $R{=}\{r_i\}_{i\in V}$ is given by

$$P(R|Z){=}\prod_{i,j\in V}\left[\frac{\exp\left(-\frac{1}{2}(r_i{-}r_j)^T\Sigma_{ij}^{-1}(r_i{-}r_j)\right)}{2\pi|\Sigma_{ij}|^{\frac{1}{2}}}\right]^{z_{ij}} \tag{2}$$

where Σ_{ij} denotes the covariance matrix representing the size of object parts at various scales.

Further, each node i is characterized by an image class, represented by an indicator random variable, x_i^k, such that $x_i^k{=}1$ if node i is labeled as image class k. For the purposes of artificial/natural object detection, we define only two possible classes forming a finite class set M. The label k of node i is conditioned on the image class l of its parent j and is given by conditional probability tables P_{ij}^{kl}. The joint probability of $X{=}\{x_i^k\}_{i\in V}$ can be expressed as

$$P(X|Z) = \prod_{i,j\in V}\prod_{k,l\in M}\left[P_{ij}^{kl}\right]^{x_i^k x_j^l z_{ij}} \ . \tag{3}$$

The generative property of a DTSBN stems from treating observable random variables y_i conditionally independent given the hidden variables – in particular, image-class indicators x_i^k (see Fig. 1). Thus, we assume that the image class of node i determines (generates) the likelihood of observable y_i, such that the joint distribution of $Y{=}\{y_i\}_{i\in V}$

is given by,

$$P(Y|X) = \prod_{i\in V}\prod_{k\in M}\left[p(y_i|x_i^k)\right]^{x_i^k} \ , \tag{4}$$

where $p(y_i|x_i^k{=}1)$ is modeled as a G-component mixture of Gaussians.

Finally, our DTSBN is fully specified by the joint distribution $P(Z,X,R,Y){=}P(Z)P(X|Z)P(R|Z)P(Y|X)$.

3. Probabilistic Inference and Learning

Due to the complexity of DTSBNs, the exact computation of $P(X|Y)$, required, for example, for Bayesian pixel labeling of artificial and natural image classes, is intractable. Therefore, to compute $P(X|Y)$, we employ our Structured Variational Approximation (SVA), thoroughly discussed in [5]. SVA relaxes the poorly justified independence assumptions in prior work [6], such that SVA takes into account the statistical dependence between node positions and the model's structure and, thus, achieves faster convergence by an order of magnitude over currently available algorithms.

Variational-approximation inference methods can be viewed as minimizing a convex cost function known as *free energy*, which measures the accuracy of an approximate probability distribution [9]. Essentially, the idea is to approximate the true intractable posterior distribution, in our case $P(Z,X,R|Y)$, by a simpler distribution $Q(Z,X,R)$ closest to $P(Z,X,R|Y)$, by minimizing the free energy $J(Q,P)$:

$$J(Q,P) = \log P(Y) - KL(Q\|P) \ , \tag{5}$$

where Q denotes a variational distribution approximating $P(Z,X,R|Y)$ and $KL(Q\|P)$ denotes their Kullback-Leibler divergence.

We constrain the solution of the variational distribution to the form $Q{=}Q(Z)Q(X|Z)Q(R|Z)$, which enforces the aforementioned assumptions that both state-indicator variables X and position variables R should be statistically dependent on the tree connectivity Z. The forms of the approximating distributions are defined as follows:

$$Q(Z){=}\prod_{i,j\in V}\left[\delta_{ij}\right]^{z_{ij}} \ , \tag{6}$$

$$Q(X|Z){=}\prod_{i,j\in V}\prod_{k,l\in M}\left[Q_{ij}^{kl}\right]^{x_i^k x_j^l z_{ij}} \ , \tag{7}$$

$$Q(R|Z){=}\prod_{i,j\in V}\left[\frac{\exp\left(-\frac{1}{2}(r_i{-}\mu_j)^T\Omega_{ij}^{-1}(r_i{-}\mu_j)\right)}{2\pi|\Omega_{ij}|^{\frac{1}{2}}}\right]^{z_{ij}} \tag{8}$$

where δ_{ij} corresponds to γ_{ij}, Q_{ij}^{kl} is analogous to P_{ij}^{kl}, and μ_j and Ω_{ij} are the mean and the covariance of the parent j position, respectively.

Minimizing $J(Q,P)$ with respect to the model parameters, we derive the update equations for iterative computation of the variational distribution Q. The full derivation of the SVA update equations is given in [5]. Herein, for space reasons, we omit the details and continue with a brief summary of learning.

SVA presumes that the parameters that characterize $P(Z,X,R,Y)$ are available. In order to learn these parameters, initially, we build a balanced TSBN. Then, using Pearl's message passing scheme [3], we learn P_{ij}^{kl}. For learning the parameters of a mixture of Gaussians, we employ the EM algorithm [5]. Further, we initialize γ_{ij} to be uniform across all possible parents of i. We equate Σ_{ij} to the area of the corresponding dyadic square. Finally, we set all the variational parameters to be equal to the corresponding parameters of $P(Z,X,R,Y)$. After initialization, we optimize the parameters of Q according to the update equations, as reported in [5]. Once optimized, $\delta's$ specify the most likely tree connectivity. Unlike in [6], for each node i, we find the maximum probability δ_{ij}, $\forall j{\in}V$, and establish only that connection, deleting other candidate connections with lower probability. In this manner, we build a forest of new DTSBNs that are not balanced, yet preserve their tree structure. Finally, we close the learning loop, again performing Pearl's message passing scheme for each dynamic subtree.

4. Feature Extraction

Assuming that edges, belonging to artificial objects, exhibit a high degree of parallelism/collinearity or combine to yield perpendicular junctions, it is reasonable to investigate spatial interrelationships of extracted edges in the image as cues for man-made object detection. For this purpose, we employ multiscale linear-discriminant analysis (MLDA), thoroughly discussed in [8]. For completeness, herein, we briefly describe the main concepts behind MLDA.

The MLDA atom w is a piecewise constant function on either side of a linear discriminant d, which divides a square into two regions, as illustrated in Fig. 2a. A discriminant d is characterized by the maximum Mahalanobis distance $\max_d\{(\mu_0{-}\mu_1)^T(\Sigma_0{+}\Sigma_1)^{-1}(\mu_0{-}\mu_1)\}$, where μ and Σ denote the RGB mean and covariance of the two regions. Decreasing the size of squares, we achieve better piece-wise linear approximation of curves in the image, as illustrated in Fig. 2b. Thus, the image can be decomposed into dyadic squares, forming a MLDA tree $\mathcal{T}$, as depicted in Fig. 2c. The expansion of $\mathcal{T}$ is controlled by two competing criteria: *accuracy* and *parsimony*. The tree optimization procedure yields an *incomplete* tree structure, since atom generation stops at different scales for different locations in the image [8].

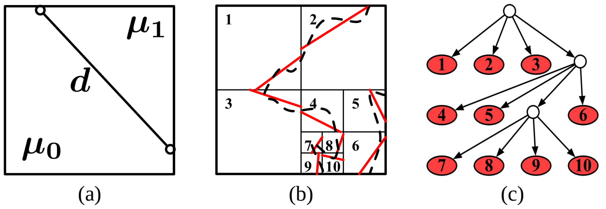

Figure 2. (a) MLDA atom; (b) dyadic decomposition; (c) corresponding MLDA tree.

To examine geometric properties of the extracted discriminants d, we first compute histograms over angles $\angle d$, measured from the x axis, for overlapping windows W_i centered at MLDA atoms w_i. The magnitude of the histogram, E_δ, for the δ-th bin, $\delta{\in}[1,\Delta]$, is smoothed using a Gaussian kernel function to alleviate the problem of hard binning of data. If W_i contains a structured image region then a few bins will have significant peaks in the histogram in comparison to the other bins. To measure the "spikiness" of the histogram, as an indicator of artificial structures in W_i, we compute the heaved central moment of the n-th order, S^n, as

$$S^n = \frac{\sum_{\delta=1}^{\Delta}(E_\delta-\bar{E})^n H(E_\delta-\bar{E})}{\sum_{\delta=1}^{\Delta}(E_\delta-\bar{E}) H(E_\delta-\bar{E})}\ ,\ n \geq 2\ , \tag{9}$$

where $\bar{E}{=}\frac{1}{\Delta}\sum_{\delta=1}^{\Delta} E_\delta$ is the mean magnitude of each histogram and $H(x)$ is a unit step function. Each bin value above the mean is linearly weighted by its distance from the mean so that the peaks far away from the mean contribute more to the proposed measure of "structuredness."

Then, to capture interscale and intrascale geometric properties, we introduce another two parameters. For each MLDA atom w_i with discriminant d, we compute $g_\delta{=}|\cos 2(\angle d{-}\delta)|$, where δ denotes a bin in the histogram of window W_i. Similarly, we compute $G_\delta{=}|\cos 2(\angle d{-}\delta)|$, where, now, δ denotes a bin in the histogram of window W_j, centered at the MLDA atom w_j, the parent of w_i. While both sets of parameters, g_δ and G_δ, point out either parallel/collinear structures or near right-angle junctions, g_δ accounts for relationships among discriminants belonging to one scale, whereas G_δ informs on continuity of geometric properties through scales.

5. Experiments

Herein, we present artificial-object detection using TSBNs and DTSBNs. The test data set consists of 100 256×256 natural-scene images with both natural and man-made objects, captured at medium to long distances from a ground-level camera, as illustrated in Fig. 3. Having computed the MLDA image representation, the ground truth was generated by hand-labeling each terminal MLDA atom as the *natural* or *artificial* image class. Despite quantization noise in classification, this coarse labeling was sufficient for our purposes, as we are interested in locating structured image regions without explicitly detecting object boundaries. Training of TSBNs and DTSBNs for the natural and artificial image classes was conducted on a distinct set of 120 256×256 images for each class. We employed Pearl's message passing scheme and our SVA as inference algorithms for learning TSBN and DTSBN parameters, respectively. The number of Gaussians (i.e. four) in the mixture model of likelihoods $p(y_i|x_i^k)$ was optimized using cross-validation. The cost of computing interscale and intrascale geometric features, g_δ and G_δ, was reduced by accounting only for the bin δ with the largest magnitude E_δ. Also, we used only the second order central moment S^2, which proved sufficient. The number of MLDA terminal nodes was set to 1000 for each image, which provided for sufficiently precise feature extraction at reasonable computational cost.

We experimented with two strategies regarding the observable variables of TSBNs and DTSBNs. In the first approach, to each node i of a model, we assign an observable vector $y_i{=}\{\bar{E}, S^2, g_\delta, G_\delta, \mu_0, \mu_1\}$ (see Fig. 1c). In the second approach, we do not propagate observable information to higher levels of the generative model. Rather, we form a long vector of the parameters of all MLDA atoms up the MLDA tree, following parent-child paths, and assign that vector to y_i of the leaf nodes, only (see Fig. 1a and 1b). Note that for TSBNs and "standard" DTSBNs, the MLDA tree is constructed without pruning. To emphasize the difference in the treatment of observable variables, we denote the models with observables propagated to higher levels as TSBN↑ and DTSBN↑, and the models with only terminal-node observables without arrows.

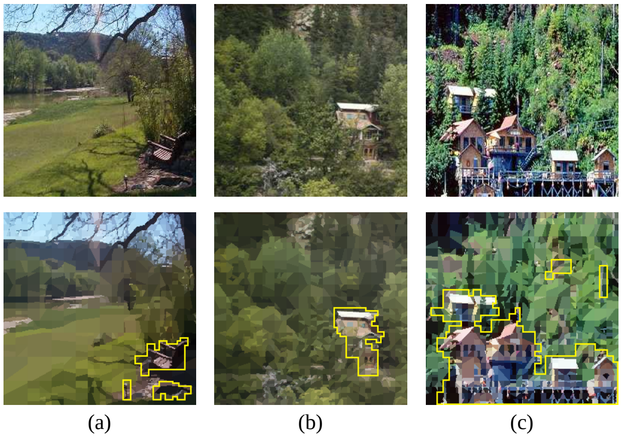

Figure 3. Artificial structure detection: (a) original images; (b) MLDA representation; and (c) MLDA atoms classified as artificial.

Several examples of natural/artificial ML classification results are demonstrated in Fig. 3. The marked MLDA atoms represent image regions classified as artificial. In Fig. 4, we report the confusion matrices of ML classification results for 100000 terminal MLDA atoms in 100 test images using DTSBN↑, DTSBN, TSBN↑, and TSBN models. The columns contain the ground truth, while the rows contain the detection results. The number of MLDA atoms that belong to natural and artificial image classes is 72345 and 27655, respectively. For a more complete comparison of detection performance, in Fig. 5, we show ROC curves for the four models, obtained for various decision boundaries between the likelihoods of natural and artificial image classes. Note that DTSBN↑ models outperform the other three models.

6. Conclusion

In this paper, we have successfully applied the DTSBN↑ model, trained on MLDA features, for man-made structure detection in natural-scene images. Through a set of experiments we have demonstrated that: (1) DTSBNs outperforms TSBNs; (2) the propagation of observable information to higher levels of generative models improves their ability to capture complex object appearances; and (3) the proposed generalized-structure DTSBN↑ models are superior to "standard" DTSBNs with respect to modeling spatial dependencies among image features.

The results presented in this paper raise a number of interesting points for further research. The capability of DTSBNs to perform unsupervised image segmentation into artificial and natural image regions can be used for localization and detection of man-made objects. Furthermore, DTSBNs, being parameterized graphical models, can be trained on those segmented regions using the SVA inference algorithm [5]. After a sufficiently large number of training samples, we would arrive at accurate statistical models of object appearances. Thus learned DTSBNs could be, then, used for Bayesian image classification, whereby we could perform man-made object recognition. Note that in the outlined procedure there is neither need for preparation of training samples nor for specification of the objects of interest. Therefore, DTSBNs could provide a unified framework for unsupervised unknown object registration – a principal topic of our future research.

References

[1] G. Reynolds and J. R. Beveridge, "Searching for geometric structure in images of natural scenes," in *Image Understanding Workshop Proceedings*, vol. 1, 1987, pp. 257–271.

[2] S. Kumar and M. Hebert, "Man-made structure detection in natural images using a causal multiscale random field," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, vol. 1, 2003, pp. I–119–26.

[3] J. Pearl, *Probabilistic reasoning in intelligent systems : networks of plausible inference*. San Mateo: Morgan Kaufamnn, 1988, ch. 4, pp. 143–236.

[4] X. Feng, C. K. I. Williams, and S. N. Felderhof, "Combining belief networks and neural networks for scene segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 4, pp. 467–483, 2002.

[5] S. Todorovic and M. C. Nechyba, "Interpretation of complex scenes using generative dynamic-structured models," in *(to appear) Proc. IEEE CVPR 2004, Workshop on Generative-Model Based Vision (GMBV)*, 2004.

[6] A. J. Storkey and C. K. I. Williams, "Image modeling with position-encoding dynamic trees," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 7, pp. 859–871, 2003.

[7] D. L. Donoho, "Wedgelets: Nearly-minimax estimation of edges," *Annals of Statistics*, vol. 27, no. 3, pp. 859–897, 1999.

[8] S. Todorovic and M. C. Nechyba, "Multiresolution linear discriminant analysis: efficient extraction of geometrical structures in images," in *Proc. IEEE Int'l Conf. Image Processing*, vol. 1, Barcelona, Spain, 2003, pp. 1029–1032.

[9] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

	A	N
A	26411	5787
N	1244	66558

DTSBN

	A	N
A	26327	6511
N	1328	65834

TSBN

	A	N
A	26548	4340
N	1107	68005

DTSBN↑

	A	N
A	26272	6149
N	1383	66196

TSBN↑

Figure 4. Confusion matrices for ML classification of artificial (A) and natural (N) image classes; columns contain the ground truth.

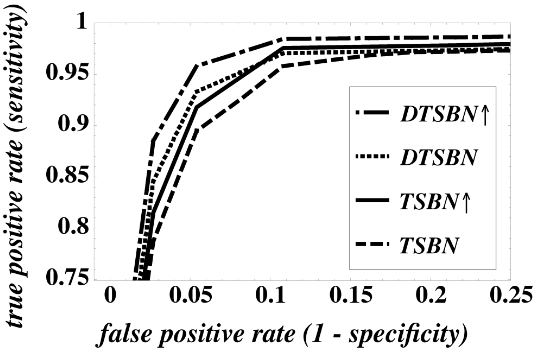

Figure 5. ROC curves for DTSBN↑, DTSBN, TSBN↑ and TSBN.