DETECTING MULTIPLE OBJECTS USING BOOSTING AND CONDITIONAL RANDOM FIELD

P. Dhivya[1], D. Chitra[2]

ABSTRACT

The contextual information is exploited to detect and localize multiple object categories in an image. The context model incorporates global image features, dependencies among object categories and output of local detectors into one probabilistic framework. The performance benefit of context models has been limited because Markov Random Field technique was tested on data sets with only a few object categories. The Sun 09 dataset is used with images that contain many instances of different object categories. The coherent structure among object categories models the object co-occurrences and spatial relationships using tree structured graphical model. The context model and spatial relationship improves object recognition performance and provides coherent interpretation of scene, enables reliable image querying system by multiple object categories. Boosted Random Field (BRF) technique is introduced to combine both Boosting and Conditional Random Field for improving the accuracy and speed. BRF provides better performance and requires fewer computations. BRF searches objects in an image and detects stuff things in an office.

[1]Assistant Professor, Department of Computer Science and Engineering, PA College of Engineering and Technology, Pollachi, Coimbatore, Tamil Nadu, India - 642002
E-mail : dhivyapd1989@gmail.com

[2] Professor and Head, Department of Computer Science & Engg., PA College of Engg and Technology, Pollachi, Coimbatore, Tamil Nadu, India - 642002
E-mail : chitrapacet@gmail.com

Key words - Boosted Random field, Contextual information, Markov Random Field, Spatial relationship.

I. INTRODUCTION

In this paper, the tree structured model is used to capture contextual information of a scene and apply it to object recognition and scene understanding problems. The previous Markov Random Field technique consists of AdaBoost, Texon Boost and Bag of Texons classifier are the detectors focus on locally identifying a particular object category. In order to detect multiple object categories in an image, a separate detector is running for each object category at every spatial location and scale. In order to improve the accuracy and performance of object recognition, the contextual information and spatial relationship is exploited to detect global features of an image such as street scene and dependencies among object categories such as road and cars co-occur often are captured in addition to local features. Many false alarms appear on the image provides an incoherent scene interpretation.

Even if perfect local detectors correctly identify all object instances in an image, some tasks in scene understanding requires an explicit context model and cannot be solved with local detectors alone. One example is finding unexpected objects that are out of their normal context require modeling expected scene configurations. These scenes attract a human's attention and it didn't occur often in daily settings.

Object dependencies in a typical scene can be represented in a hierarchy [16, 23]. For example, it is important to model the outdoor objects like sky and mountain and indoor objects like desk and bed do not co-occur in a scene. Instead of encoding this negative relationship for all possible pairs of outdoor and indoor objects.

The tree model is more efficient for implementing all outdoor objects are in one subtree, all indoor objects are in another subtree and the two trees are connected by an edge with a strong negative weight. Similarly, in order to capture the contextual information that kitchen-related objects such as a sink, a refrigerator, and a microwave co-occur often, all kitchen-related objects can be placed in one subtree with strong positive edge weights. Motivated by such inherent structure among object categories, the object co-occurrences and spatial relationships are modeled using a tree-structured graphical model.

The tree structured dependencies among objects allows learning the model for more than a hundred object categories and applying it to images efficiently. The prior model of object relationships are combined with local detector outputs and global image features to detect and localize all instances of multiple object categories in an image.

Feature extraction is a dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features. The input image for feature extraction and classification is shown in the Fig. 1

In machine learning, pattern recognition is the assignment of a label to a given input value. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of *classes* (for example, determine whether a given email is "spam" or "non-spam"). However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values (for example, part of speech tagging, which assigns a part of speech to each word in an input sentence); and parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence.

(a) Input image

(b) Gray scale image

Figure 1 : Feature extraction (a) & (b)

Feature classification is used to classify the entities into different categories are shown in the Fig. 2.

Figure 2 : Feature Classification

The previous PASCAL 07 data sets [7] were originally designed to evaluate single-object detectors, and most of the images have no co-occurring instances. The new SUN 09 dataset [30] is used with more than 200 object categories in a wide range of scene categories. Each image contains instances of multiple object categories with a wide range of difficulties due to variations in shape, sizes and frequencies. The SUN 09 dataset contains richer contextual information and is more suitable to train and evaluate context model than PASCAL 07.

In this paper, the contextual correlations are exploited between the object classes by introducing Boosted Random Fields (BRFs). BRF build on both boosting [6, 17] and Conditional Random Fields (CRFs) [22]. Boosting is a simple way of sequentially constructing strong classifiers from weak components and has been used for single class object detection with great success [11]. CRFs are a natural way to model the correlation between labels. The main problem with CRFs is learning the correlation structure of the model. A4-nearest neighbor grid structure is successful in low-level vision but fails in capturing important long distance dependencies between whole regions and across classes.

In BRFs, the graph structure is learned by using boosting to select from a dictionary of connectivity templates. The boosting is also used to learn the local evidence potentials of the model. The traditional sliding window approach to object detection does not work well for detecting stuff things. Instead, object detection and image segmentation is combined for improving the performance.

II. Related Works

A simple form of contextual information is a co-occurrence frequency of pair of objects. Rabinovich et al. [24] use local detectors to first assign an object label to each image segment and adjusts these labels using CRF. This approach is extended in [10] and [11] to encode spatial relationships between a pair of objects. In [10], spatial relationships are quantized to four prototypical relationships like above, below, inside, and around, whereas in [11], a nonparametric map of spatial priors is learned for each pair of objects. Torralba et al. [27] combine boosting and CRFs to detect easy objects like monitor and pass the contextual information to detect other difficult objects like keyboard, mouse. Tu [29] uses both image patches and their probability maps estimated from classifiers to learn a contextual model and iteratively refines the classification results by propagating the contextual information.

Desai et al. [5] combine individual classifiers by using spatial interactions between object detections in a discriminative manner. Contextual information may be obtained from coarser and the global features. Torralba [28] demonstrates global image feature called "gist" can predict the presence or absence of objects and their

locations without running an object detector. This is extended in [21] to combine patch-based local features and the gist feature. Heitz and Koller [13] combine a sliding window method and unsupervised image region clustering to leverage "stuff" such as the sea, the sky or a road to improve object detection.

A cascaded classification model [14] links object detection, multiclass image segmentation, scene categorization and 3D reconstruction. Hierarchical models can be used to incorporate both local and global image features. In [15] Image understanding requires not only estimating elements of the visual world but also capturing the interplay among them. Multiscale conditional random field [12] combines local classifiers with regional and global features. [19] uses lexical semantic networks to extend the state-of-the-art object recognition techniques and the semantics of image labels to integrate prior knowledge about inter-class relationships into the visual appearance learning. In [20] graphical model relating features, objects and scenes. detects objects by scene classification and exploits contextual information, patch based local features and gist features.

Sudderth et al. [26] model the hierarchy of scenes, objects and parts using hierarchical Dirichlet processes encourage scenes to share objects, objects to share parts and parts to share features. Parikh and Chen [22] learn a hierarchy of objects in an unsupervised manner, under the assumption that each object appears exactly once in all images. Hierarchical models are common within grammar models for scenes [16], [23] and it is flexible to represent complex relationships. Bayesian hierarchical models [18] also provide a powerful mechanism to build generative scene models. In [1] the pattern recognition recognize the different object categories in an image. Dalal and Triggs [3] models the histograms of oriented gradients for human detection.

In [4] Deng. J et al. proposed large-scale hierarchical image database. The explosion of image data on the internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. In [8] Felzenszwalb. P et al. proposed Object Detection with discriminatively trained part based models. The problem of detecting and localizing generic objects from categories like people or cars in static images are considered. This is difficult problem because objects in such categories can vary greatly in appearance.

In [9] the authors present a method to learn, detect and recognize object class models from unlabeled and unsegmented cluttered scenes in a scale invariant manner. The main issues are representation, detection and learning have to be tackled in designing a visual system for recognizing object categories.

In the previous work, a fully connected model [24] is computationally expensive for modeling relationships among many object categories and may overfit with limited number of samples. In the scene-object model [21], objects are assumed to be independent conditioned on the scene type does not capture direct dependencies among objects.

In this work, a tree-structured model provides a richer representation of object dependencies while maintaining a number of connections. In addition, it allows efficient integration of different sources of contextual information results in improved object detection performances [6].

III. Tree Based Context Model

A tree-structured graphical model [21] is used to capture dependencies among object categories. In the scene-object model [21], objects are assumed to be independent conditioned on the scene type and it does not capture direct dependencies among objects. Ada Boost, Texon Boost, Conditional boosting, Edge detection, Bag of Texons classifiers are the local detectors detect multiple object categories but it does not detect scene understanding tasks.

The model allows efficient integration of different sources of contextual information which results in improved object detection performances. The tree-structure is learned from the co-occurrence statistics of object categories and it is different from the semantic hierarchies used.

The probabilistic framework admits efficient learning and inference algorithms. The new SUN 09 dataset describes context model that incorporates global image features is shown in the Fig 3 and 4.

Figure 3 : Input image

Figure 4 : Context model

Tree structure dependencies are shown in the Fig 5. The model evaluates object recognition and scene understanding performances of the context model.

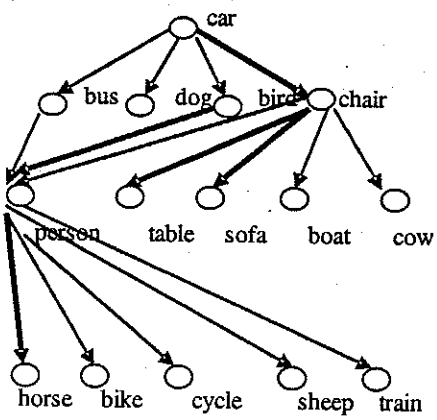


Figure 5 : Object dependency structure. Red edges correspond to negative correlations between categories. The thickness of edge represents the strength of the link.

A. Contextual information

The context of an image encapsulates rich information about natural scenes and objects are related to each other. Detecting out-of-context images is different from detecting changes in surveillance applications because the goal in surveillance is to identify the presence or absence of certain objects in a known scene. Object co-occurrence statistics provide strong contextual information. Each object category is associated with a binary variable representing the object is present or not in the image and a Gaussian variable representing its location. Each node b_i in a tree represents the corresponding object i is present or not in an image. The joint probabilities of all binary variables are factored according to the Equation (1).

$$p(b)=p(b_{root}) \, \eth_i \, p(b_i \mid b_{pa(i)}) \qquad (1)$$

Where pa(i) is the parent of node i. A subscript i is used to denote a variable corresponding to object i and a symbol without a subscript denotes a collection of all corresponding variables. A parent-child pair may have either a positive relationship like floor, wall co-occur often and negative relationship like floor seldom appears with the sky.

B. Spatial location representation

Objects often appear at specific relative positions to one another. A computer screen typically appears above a keyboard and a mouse. The location variables are added to the tree model by using Spatial relationships [11]. Instead of using the segmentation of an object, a bounding box is used in the minimum enclosing box for all the points in the segmentation to represent the location of an object instance.

Let Lx, Ly be the horizontal and vertical coordinates of the centre of the bounding box and Lw, Lh be the width and height of the box. The model assumes that the image height is normalized to one and that Lx=0, Ly=0 is the centre of the image.

The expected distance between centers of objects depends on the size of the objects. If a keyboard and a mouse are small, the distance between the centers should be small as well. The constellation model achieves scale invariance by transforming the position information to a scale invariant space. The coordinate transformations is applied to represent object locations in the 3D-world coordinates are:

$$Lx=H.lx/lh,\ Ly=H.ly/lh,\ Lz=H.lz/lh \quad (2)$$

Where Lz is the distance between the observer and the object and H_i is the physical height of an object i. The heights H_i of each object category could be inferred from the annotated data using the Equation (2).

The model assumes the variable b, the dependency structure of the L_{is} has the same tree structure as the binary tree is illustrated in the Equation (3).

$$P(L|b)=p(L_{root}|b_{root})\,\eth_i\,p(L_i|L_{pa(i)},b_i,b_{pa(i)}) \quad (3)$$

Where each edge potential $p\ (Li \mid L_{pa(i),\,bi},\ b_{pa(i)})$ encodes the distribution of a child location conditioned on its parent location and the presence or absence of both child and parent objects.

C. Measurement model

i) Incorporating Global Image Features

The gist descriptor is low-dimensional representation of an image, capturing coarse texture and spatial layout of a scene. The gist as a measurement for each presence variable b_i to incorporate with global image features into the model. This allows the context model to infer a scene category implicitly and is particularly helpful in predicting the indoor objects or outdoor objects. The gist is used with local detectors to enhance the context inferring power of the co-occurrence tree model. A gist descriptor is effective in classifying scenes and metaobjects [18].

ii) Integrating Local Detector Outputs

An off-the-shelf single-object detector [13] is applied to detect and localize object instances in an image and a set of candidate windows for each object category is obtained. Let i denote an object category and k index candidate windows generated by baseline detectors. Each detector output provides a score S_{ik} and a bounding box.

The coordinate transformation is applied to get the location variable using the Equation (4).

$$W_{ik} = (Ly, \log Lz) \quad (4)$$

Binary variable C_{ik} is assigned to each window to represent a correct detection (C_{ik} ¼ 1) or a false positive (C_{ik} ¼ 0). It shows the measurement model for object i to integrate gist and baseline detector outputs into the prior model used to plate notations to represent K_i different candidate windows. A candidate window is correct detection of object i (C_{ik} ¼ 1), its location W_{ik} is a Gaussian vector with mean L_i, the location of object i and if the window is a false positive (C_{ik} ¼ 0), W_{ik} is independent of L_i and has a uniform distribution.

The Chow-Liu [2] algorithm is simply selecting strong pairwise dependencies by minimizing the Bayes error rate using the Equation (5).

$$\min H(w|X) = \max \sum_{\infty} \sum_{i=1 \text{ to } n} I_w(X_i, X_{j(i)}) \quad (5)$$

Markov Random Field [22] is applied to detect object categories by using the Equation (6).

$$P(x, y) \alpha \Pi_i \Phi_i(x_i) \Pi_{j \in E} \phi_{ij}(x_i, x_j) \quad (6)$$

iii) ***Learning object dependency structure***

The model learns the dependency structure among objects from a set of fully labeled images. The Chow-Liu algorithm is simple and efficient way to learn a tree model. The algorithm computes empirical mutual information of all pairs of variables using their sample values and it finds the maximum weight spanning tree with edge weights equal to the mutual information between the variables connected by the edge. It learns the tree structure using the samples of b_{is} in a set of labeled images.

More than hundred objects and thousands of training images using a tree model can be learned in a few seconds in Matlab.

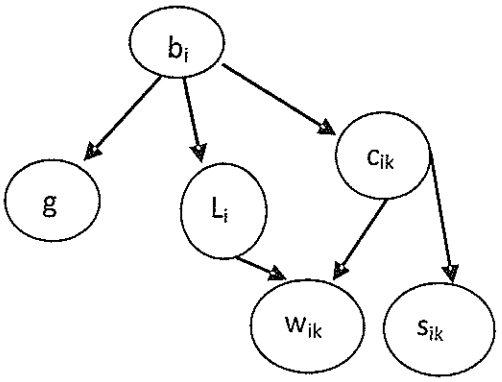


Figure 6 : Graphical model representations for parts of the context model.

All nodes are observed during training and the shaded nodes are observed during testing. Prior model relating object presence variables b_{is} and location variables L_{is}. The gist descriptor g represents global features and local detector provides candidate window locations W_{ik} and scores S_{ik}.The binary variable C_{ik} checks the window is a correct detection or not is shown in the Fig 6.

This algorithm is an undirected tree; the model has selected sky to be the root of the tree to obtain a directed tree structure. The algorithm does not use any information regarding the inherent hierarchical structure among object categories is simply selecting strong pairwise dependencies.

The learned tree structure organizes objects in a natural hierarchy. A subtree rooted at building has many objects appear in street scenes and the subtree rooted at sink contains objects that commonly appear in a kitchen. The

learned tree structure captures the inherent hierarchy among objects and scenes resulting in better object recognition and scene understanding performances.

IV. DETECTING MULTIPLE OBJECTS

The proposed algorithm consists of Boosted Random Field [6, 17]. It is used to overcome the disadvantages of Markov Random Field. The key idea is to detect the multiple objects like stuff things [14] in an image and captures the dependencies among object categories.

BRF uses Boosting to learn the graph structure and local evidence of a CRF. **CRF** models the correlation between the labels, learn the graph structure and classify the object categories by using the Equation (7).

$$P(S|x) = \frac{1}{Z} \Pi_i \Phi_i(S_i) \Pi_{j \in N_i} \phi_{ij}(S_i, S_j) \quad (7)$$

Where x is the input and N_i are the neighbours of node.

The graph structure is learned by assembling graph fragments in an additive model. The connections between individual pixels are not very informative but by using dense graphs, the information can be pooled from large regions of the image and dense models also support efficient inference. The contextual information from other objects also improves detection performance, both in terms of accuracy and speed.

The BRF technique is applied to detect stuff things in office and street scenes are shown in the Fig 7.

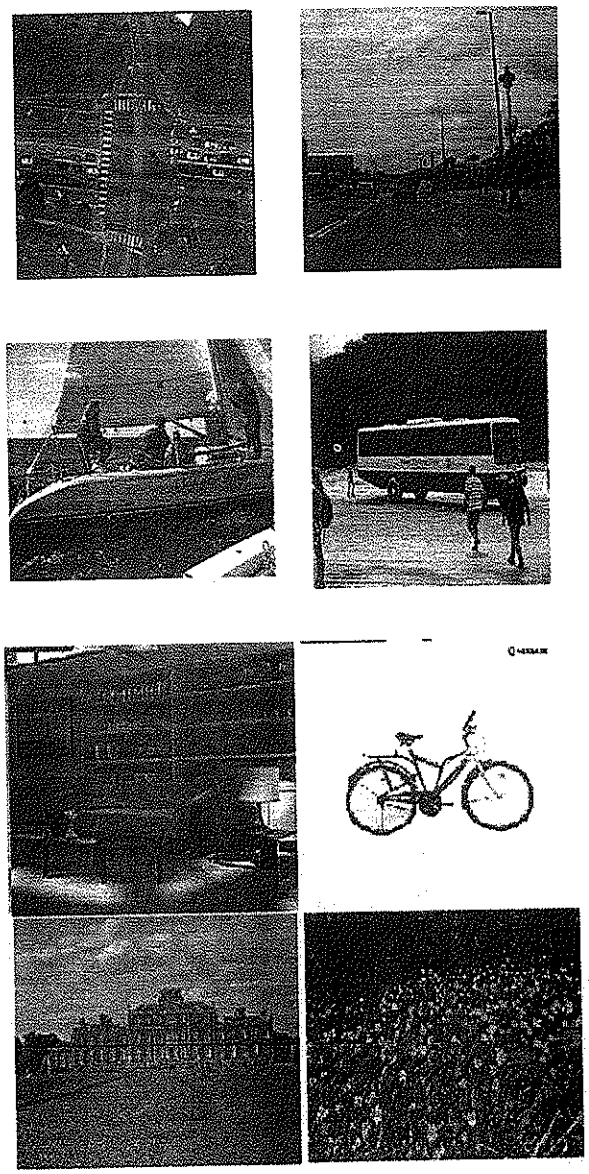


Figure 7 : Sample images from SUN09 dataset

The SUN 09 data set is suitable for leveraging contextual information. The data set contains 12,000 annotated images covering a large number of indoor and outdoor scenes with more than 200 object categories and 152,000 annotated object instances.

The resulting annotations have a higher quality than that by LabelMe or Amazon Mechanical Turk. This SUN09

data set can be used both for training and performance evaluation.

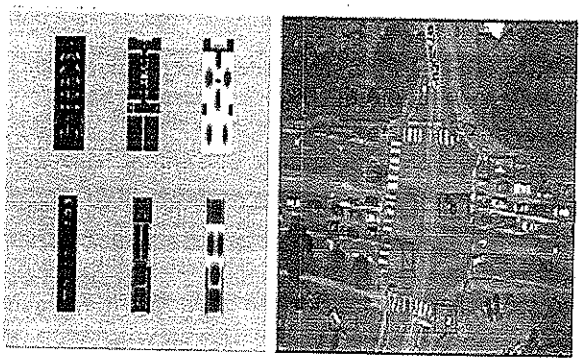
The images were collected from multiple sources such as Google, Flickr, Altavista, LabelMe and any closeup of an object or images with white backgrounds were removed to keep only images corresponding to scenes in the collection. The annotation procedure was carried out by a single annotator over one year using LabelMe [25]. The labeled images were carefully verified for consistency and synonymous labels were consolidated.

V. Results

A. Object recognition

The SUN 09 data set presented in this paper has richer contextual information than PASCAL 07 and is more suitable for training and evaluating context models. The context model is learned from SUN 09 significantly improves the accuracy of object recognition and image query results. It can be applied to find objects out of context.

The effective contextual information is the co occurrence frequency of pair of object pairs. The co-occurrence statistics using a binary tree model is encoded. The edge detection of object and the tree structure are shown in Fig. 8 (a) & (b).



(a) (b)

Figure 8 : Object recognition (a) and (b)

B. Experimental Results

To evaluate categorization accuracy of the proposed model and the relative importance of spatial context in this task, we consider MSRC and PASCAL 2007 and SUN09 datasets. Table 1 summarizes the performance of average categorization per category.

Table 1 : Comparison of recognition accuracy between the models for MSRC , PASCAL and SUN09 categories. Results in bold indicate an increase in performance by this model.

Categories	MSRC	PASCAL 07	SUN 09
Building	0.85	**0.91**	**0.92**
Grass	0.94	**0.95**	**0.95**
Sky	0.89	**0.97**	**0.971**
Airplane	0.73	0.73	0.73
Car	0.95	0.95	0.95
Flower	0.65	0.65	**0.66**
Road	0.94	**0.95**	**0.96**
Person	0.43	0.43	**0.44**
Bicycle	0.22	0.22	**0.23**

The Boosting and Conditional Random Field maximize the object label agreement in the scene according to spatial and coocurrence constraints. The spatial information that captures the relative object location in an image.

However, unlike simple co-occurrence relationships, which can be learned from auxiliary sources such as Google Sets, spatial context must be learned directly from the training data. As our experiments have shown, spatial context learned from both MSRC, PASCAL and SUN datasets is highly nonuniform.

VI. Conclusion

The Boosted Random Field algorithm combines boosting and CRF is easy for both training and inference. The model has demonstrated object detection in cluttered scenes by exploiting contextual relationships between

Boosted Random Fields," Advances in Neural Information Processing Systems, MIT Press, 2005.

[27] A. Torralba, *"Contextual Priming for Object Detection,"* Int'l J. Computer Vision, vol. 53, pp. 169-191, 2003.

[28] Z. Tu, *"Auto-Context and Its Application to High-Level Vision Tasks,"* Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

[29] J. Winn, A. Criminisi, and T. Minka, *"Object Categorization by Learned Universal Visual Dictionary,"* Proc. IEEE Int'l Conf. Computer Vision, 2005.

[30] Myung Jin Choi, Antonio Torralba and Alan S. Willsky, (2012), A Tree-Based Context Model for Object Recognition IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 34, NO. X, XXXXXXX 2012.

AUTHOR'S BIOGRAPHY

P. Dhivya was born on September 22, 1989. She received her BTech,degree from Sri Ramakrishna institute of Technology, Coimbatore (Anna University) in the year 2011. She completed her ME, degree from PA College of Engineering and Technology, Pollachi (Anna University, Chennai) in the year 2013. Currently she is working as a Assistant Professor in P.A College of Engineering and Technology, Pollachi. Her area of interest is Image Processing.

Dr. D. Chitra is a Professor in the Department of Computer Science and Engineering at P. A. College of Engineering and Technology, Pollachi, Coimbatore. She received her M.E. Degree in CSE from Anna University, Chennai and PhD degree in CSE from Anna University of Technology, Coimbatore. Her resource interests include image analysis, Pattern recognition and Computer Vision. She is a life member of ISTE.