Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 671397, 13 pages
doi:10.1155/2012/671397

Research Article

Contextual Hierarchical Part-Driven Conditional Random Field Model for Object Category Detection

Lizhen Wu, Yifeng Niu, and Lincheng Shen

College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, China

Correspondence should be addressed to Lizhen Wu, lzwu@nudt.edu.cn

Received 25 October 2012; Accepted 11 November 2012

Academic Editor: Sheng-yong Chen

Even though several promising approaches have been proposed in the literature, generic category-level object detection is still challenging due to high intraclass variability and ambiguity in the appearance among different object instances. From the view of constructing object models, the balance between flexibility and discrimination must be taken into consideration. Motivated by these demands, we propose a novel contextual hierarchical part-driven conditional random field (CRF) model, which is based on not only individual object part appearance but also model contextual interactions of the parts simultaneously. By using a latent two-layer hierarchical formulation of labels and a weighted neighborhood structure, the model can effectively encode the dependencies among object parts. Meanwhile, beta-stable local features are introduced as observed data to ensure the discriminative and robustness of part description. The object category detection problem can be solved in a probabilistic framework using a supervised learning method based on maximum a posteriori (MAP) estimation. The benefits of the proposed model are demonstrated on the standard dataset and satellite images.

1. Introduction

Object category detection is one of the most important problems in computer vision and is still full of challenges because of various factors such as object deformation, occlusion, and viewpoint change. To address these challenges, successful object detection methods need to strike the balance between being flexible enough to model intraclass variability and being discriminative enough to find objects with ambiguity appearance in complicate scenes [1–3].

Part-based object model, firstly proposed by Fischler and Elschlager [4] in 1973, has been proved as a powerful paradigm for object category detection and recognition in numerous researches [5–10], due to its advantages of intuitive interpretation and semantic expression. In such models, each part is generally represented by small templates or local image feature information, and the whole object is modeled as a collection of parts with or without geometric and cooccurrence constraints. The final discriminate of object is achieved by solving the probability density function or using a Hough vote mechanism. In the early researches on part-based approaches, parts are learned purely on the basis of their appearance by clustering visually similar image patches in the training images and do not exploit any spatial layout of the parts. Obviously, since the part appearance only reflects local image characteristics, these models cannot get enough spatial information support. The neglected contextual interactions that are used to capture geometric relationships between parts of an object should play a more crucial role in the part-based model to enhance the representational power of model.

On the other hand, most current part-based approaches can be roughly divided into two separate groups: generative and discriminative. Generative part-based models [5–9] have shown high flexibility because of their advantage of handling missing data (i.e., the correspondence between local features and parts) in a principled manner. So, each part can be interpreted in a semantically meaningful way. The most popular generative approach for part-based object detection was proposed by Fergus et al. [7] in 2003, in which objects are modeled as flexible constellations of parts and the appearance, spatial relations, and cooccurrence of local parts are learned in an unsupervised manner. Felzenszwalb and Huttenlocher [8] proposed a pictorial structure model, in which deformable configuration is represented by spring-like connections between pairs of parts. By integrating spatial relationships with “bag of features,” Sudderth et al. [9] developed a hierarchical probabilistic model to capture the complex structure in multiple object scenes. However, generative approaches often cannot compete with discriminative manner in the field of object category detection. The generative framework has natural drawback that it has to assume the independence of the observed data to make the model computationally. In contrast to the discriminative model, the generative model may be quite complex even though the class posterior is simple. Moreover, learning the class density models may become even harder when the training data is limited [10].

In this paper, we focus on the discriminative random field model, called conditional random field (CRF), which is originally proposed by Lafferty et al. [11] in 2001. Kumar and Herbert [12, 13] first introduced the extension of 1D CRFs to 2D graphs over image and applied it to object detection. By treating object detection problem as a labeling problem, CRF model cannot only flexibly utilize various heuristic image features, but also get the contextual interactions among image parts through its classic graphical structure. In order to deal with multiple labels for object parts, Kumar and Hebert presented a multiclass extension of CRF [14], and utilized fully labeled data where each object part is assigned a part label to train the model. By contrast, Quattoni et al. presented an expansion graph structure of CRF framework [15] that uses hidden variables, which are not observed during training, to represent the assignment of parts. Moreover, located CRF model [16], proposed by Kapoor and Winn, introduces global positions to the hidden variables and can model the long-range spatial configuration and local interactions simultaneously.

The goal of this paper is to introduce a novel contextual hierarchical part-driven CRF model for object category detection. The main novelty of our approaches lies in the use of a latent two-layer hierarchical formulation of labels and a weighted minimum spanning tree neighborhood structure. The model can effectively encode latent label-level context, as well as observation-level context. Meanwhile, beta-stable local features are also introduced as observed data to ensure the discriminative and robustness of part description. Such features provide a sparse and repeatable manner to express object parts and actually reduce the computation complexity of the model.

The remainder of this paper is organized as follows. Section 2 gives detailed introduction on the proposed contextual hierarchical part-driven CRF Model. The parameter learning and inference algorithms are introduced in Section 3. Experimental results are presented in Section 4. Finally, in Section 5 we draw the conclusions.

2. Contextual Hierarchical Part-Driven CRF Model

The conditional random field is simply a Markov random field (MRF) [17] globally conditioned on the observation. It is a discriminative model that relaxes conditional independence assumption by directly estimating the conditional probability of labels [18, 19].

In other words, let y be the observed data from an input image, where $y = \{y_i\},\ i \in S$, y_i is the data from the ith site, and S is the set of sites. The corresponding labels at image sites are given by $x = \{x_i\},\ i \in S$. For labeling problems, the general form of a CRF can be written as

$$\begin{aligned} P(x \mid y, \theta) &= \frac{1}{Z(\theta)} \exp\{\Phi(x, y, \theta)\} \\ &= \frac{1}{Z(\theta)} \prod_{i \in S} \varphi_i(x_i, y, \theta) \prod_{(i,j) \in E} \varphi_{ij}(x_i, x_j, y, \theta), \end{aligned} \tag{2.1}$$

where partition function $Z(\theta)$ is a constant normalization with respect to all possible values of x with parameters $\theta,\ E$ denotes the set of edges, and φ_i and φ_{ij} are the unary and pairwise potentials, respectively. Here, φ_i encodes compatibility of the label x_i with the observed image y and φ_{ij} encodes the pairwise label compatibility for all $(i, j) \in E$ that $j \in N_i$ conditioned on y.

2.1. Problem Formulation

For our object category detection problem, assume that we are given a training set of N images $Y = (y^1, \ldots, y^N)$, which contains objects from a particular class and background images. The corresponding labels can be denoted as $X = (x^1, \ldots, x^N)$, each x^n is a member of a set of possible image labels. Since in object detection we only focus on presence or absence of objects, the possible labels should be limited to binary data, that is, $x^n \in \{0, 1\}$ or $x^n \in \{\text{background, object}\}$. Now, our task is to learn a mapping from images Y to labels X. For simplicity of notation, we drop the superscript n indicating training instance.

According to the theory of part-based model, assume that each image y can be seen as a collection of parts $y = (y_1, \ldots, y_m)$, each part y_i corresponds to a local observation or local feature. In order to describe the relationship between these parts, similar to hidden random field approach [15], we introduce latent labels $h = (h_1, \ldots, h_m),\ h_i \in H$, where h_i corresponds to the "part-label" of part y_i, and H corresponds to the actually object parts, for example, $H = \{\text{nose, tail}, \ldots, \text{wing}\}$ for airplane objects.

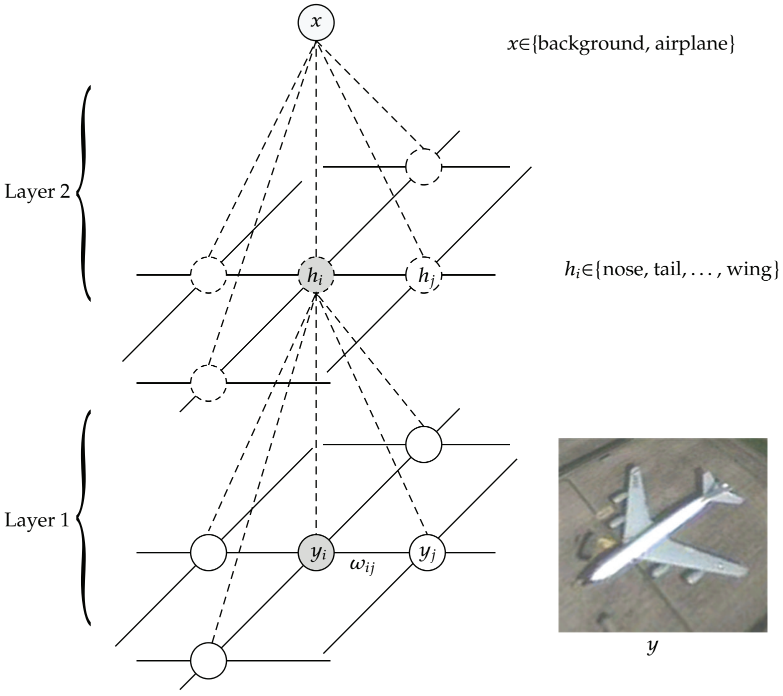

x
$x \in$ {background, airplane}
Layer 2
h_i
h_j
$h_i \in$ {nose, tail, . . . , wing}
Layer 1
y_i
ω_{ij}
y_j
y

Figure 1: The hierarchical graphical structure of proposed contextual hierarchical part-driven CRF model.

Now, we can model the posterior directly by marginalizing out the latent labels h, and the model can be defined as

$$P(x \mid y, \theta) = \sum_h P(x, h \mid y, \theta) = \sum_h \underbrace{P(x \mid h, \kappa)}_{\text{Layer 2}} \underbrace{P(h \mid y, \lambda)}_{\text{Layer 1}}, \tag{2.2}$$

where $\theta = \{\kappa, \lambda\}$ is the set of parameters.

Here, we assume that $P(x \mid h, \theta)$ is conditional independence of y given h. This means that the final object label only relies on the latent middle-level labels rather than on the original observations. The hypothesis makes sense because it is theoretically possible that we estimate the object or background occurrence by the spatial distribution of meaningful object parts in real world. Also, by doing this, we can build distinct two-layer structure of our proposed model. The hierarchical graphical structure of our contextual hierarchical part-driven CRF model is shown as Figure 1.

Obviously, both $P(x \mid h, \kappa)$ and $P(h \mid y, \lambda)$ can be modeled as CRFs. Thus, the whole model can be seen as the combination of two single-layer CRFs. By modeling the contextual interactions of these two layers, respectively, our model may have a high level of ability to describe different levels of context. Note that the latent label h cannot be observed during training (i.e., unlabelled), so we must learn the models in a unified framework to avoid the direct use of h. Detailed modeling approach and potential definitions for the two layers will be described below.

2.2. Model of Layer 1

Without considering the object label x, the distribution over the latent part labels h given the observations y may be modeled as a multiclass CRF. In such a model, observations are linked to local features located at certain spatial positions of the image. Therefore, the distribution of y may be arbitrary and disorganized due to the uncertainty of local feature extraction. Meanwhile, different observations may be associated with the same part label, which corresponded to the meaningful object component. Due to the fact that adjacent or relevant observations are more likely to have the same label, we should consider label-level context in our model in addition to observation-level context.

Furthermore, although we cannot use the part labels explicitly, we can theoretically use them to define the posterior to capture the context structure of layer 1. Considering only unary and pairwise potentials, the posterior distribution $P(h \mid y, \lambda)$ can be modeled as

$$P(h \mid y, \lambda) = \frac{1}{Z(\lambda)} \prod_{i \in S} \varphi_i^{(1)}(h_i, y, \mu) \prod_{(i,j) \in E} \varphi_{ij}^{(1)}(h_i, h_j, y, \nu), \tag{2.3}$$

where the set of parameters is given by $\lambda = \{\mu, \nu\}$, as shown in Figures 2(a) and 2(b), $\varphi_i^{(1)}(h_i, y, \mu)$ denote the unary potentials and are responsible for modeling part occurrences based on a single image feature, $\varphi_{ij}^{(1)}(h_i, h_j, y, \nu)$ denote the pairwise potentials and are responsible for modeling the cooccurrences of h_i and h_j based on the corresponding pairwise image feature. The connectivity of nodes (i, j), that is, neighborhood structure of the observations, is defined in Section 2.4.

Note that, different from multiclass CRF [14], the definitions of potentials here must consider the missing label data h. By using the parameter vector, the potentials can be denoted as

$$\varphi_i^{(1)}(h_i, y, \mu) = \exp\left[\mu(h_i)^T f_i(y)\right], \tag{2.4}$$

$$\varphi_{ij}^{(1)}(h_i, h_j, y, \nu) = \exp\left[\nu(h_i, h_j)^T g_{ij}(y)\right], \tag{2.5}$$

where $f_i(y)$ and $g_{ij}(y)$ refer to the unary feature vector and the pairwise feature vector, respectively. Parameter vectors of $\mu(h_i) \in \Re^d$ and $\nu(h_i, h_j) \in \Re^B$ have the same dimensions with the corresponding feature vectors.

In comparison with hidden CRF [15], our approach introduces the pairwise potentials $\varphi_{ij}^{(1)}$ to the model and can effectively capture the part label-level context by measuring the compatibility of different part labels.

2.3. Model of Layer 2

According to the conditional independence assumption mentioned in Section 2.1, the final image label x only depends on part labels h. In other words, the occurrence of an object can be estimated by the spatial distribution of object parts.

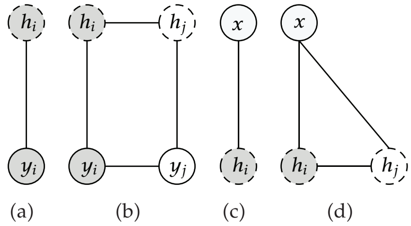


Figure 2: (a) Part evidence from single observation, (b) cooccurrence of connected parts, (c) compatibility between image label and single part label, and (d) compatibility between image label and connected part labels.

Particularly, part labels h should be regarded as observations in this layer, and the posterior distribution $P(x \mid h, \kappa)$ can be easily defined as

$$P(x \mid h, \kappa) = \frac{1}{Z(\kappa)} \prod_{i \in S} \varphi_i^{(2)}(h_i, x, \alpha) \prod_{(i,j) \in E} \varphi_{ij}^{(2)}(h_i, h_j, x, \gamma), \tag{2.6}$$

where $\kappa = \{\alpha, \gamma\}$ is the set of parameters, unary potentials $\varphi_i^{(2)}(h_i, x, \alpha)$ describe the compatibility between image label x and part label h_i, pairwise potentials $\varphi_{ij}^{(2)}(h_i, h_j, x, \gamma)$ describe the compatibility between image label x, part label h_i, and part label h_j, as in Figures 2(c) and 2(d).

Note that there is only one image label for an instance, so we do not need to model label-level context like in layer 1, and the potentials can be defined as

$$\varphi_i^{(2)}(h_i, x, \alpha) = \exp[\alpha(h_i, x)], \tag{2.7}$$

$$\varphi_{ij}^{(2)}(h_i, h_j, x, \gamma) = \exp[\gamma(h_i, h_j, x)], \tag{2.8}$$

where parameter vectors $\alpha(h_i, x) \in \Re$ and $\gamma(h_i, h_j, x) \in \Re$.

Now, we can give the complete expression of our part-driven CRF model with the specific potentials, which can be denoted as

$$\begin{aligned} P(x \mid y, \theta) &= \sum_h P(x \mid h, \kappa) P(h \mid y, \lambda) \\ &= \sum_h \left\{ \frac{1}{Z(\theta)} \prod_{i \in S} \varphi_i^{(1)}(h_i, y, \mu) \varphi_i^{(2)}(h_i, x, \alpha) \prod_{(i,j) \in E} \varphi_{ij}^{(1)}(h_i, h_j, y, \nu) \varphi_{ij}^{(2)}(h_i, h_j, x, \gamma) \right\}, \end{aligned} \tag{2.9}$$

where the set of parameters is given by $\theta = \{\mu, \alpha, \nu, \gamma\}$.

2.4. Neighborhood Structure

For probabilistic graphical models, the neighborhood structure is an important factor affecting the model capability. Moreover, as mentioned above, the observations y in our model are distributed over the image plane in an arbitrary layout. So, how to define the neighborhood structure becomes a question we have to consider during the model design phase.

In [15], Quattoni et al. evaluated a range of different neighborhood structure and come to the conclusion that the minimum spanning tree (MST) shows better performances than many other complex connected graph structures. Following this, we adopt MST as the basic structure and extend it to a novel weighted neighborhood structure (WNS). The basic idea is to exploit the edge cost, which is discarded in previous work, as heuristic information to reflect the degree of correlation between two nodes.

In other words, different edges should have different weights during calculating pairwise potentials. We denote the edge cost between node i and node j as ω_{ij} and modify (2.5) to

$$\varphi_{ij}^{(1)}(h_i, h_j, y, \nu) = \exp\left[\omega_{ij} \cdot \nu(h_i, h_j)^T g_{ij}(y)\right]. \tag{2.10}$$

Similarly, (2.8) is changed by

$$\varphi_{ij}^{(2)}(h_i, h_j, x, \gamma) = \exp\left[\omega_{ij} \cdot \gamma(h_i, h_j, x)\right]. \tag{2.11}$$

By doing this, the weighted neighborhood structure can not only describe the connectivity of nodes, but also encode the assumption that parts that are spatially close are more likely to be dependent.

3. Parameter Learning and Inference

Given N-labeled training images, the parameters $\theta = \{\mu, \alpha, \nu, \gamma\}$ can be learnt by using maximum A posteriori (MAP) estimation. Gaussian prior $p(\theta) \sim \exp(\|\theta\|^2/2\sigma^2)$ is introduced to prevent overfitting. So, parameter learning can be achieved by maximizing the following objective function:

$$L(\theta) = \sum_{n=1}^{N} \log P\left(x^{n'} \mid y^n, \theta\right) - \frac{1}{2\sigma^2}\|\theta\|_2. \tag{3.1}$$

We use gradient ascent to search for the optimal parameter values $\theta^* = \arg\max_\theta L(\theta)$. In our model, the derivatives of the log-likelihood $L(\theta)$ with respect to the model parameters

$\theta = \{\mu, \alpha, \nu, \gamma\}$ can be written in terms of local feature vectors, marginal distributions over individual part label h_i, and marginal distributions over pairwise labels h_i and h_j:

$$\begin{gathered}
\frac{\delta L(\theta)}{\delta \mu(h')} = \sum_{i \in s} f_i(y) \cdot \left[p\left(h_i = h' \mid x, y, \theta\right) - p\left(h_i = h' \mid y, \theta\right)\right], \\
\frac{\delta L(\theta)}{\delta \alpha(h', x')} = \sum_{i \in s} p\left(h_i = h' \mid x, y, \theta\right) - p\left(h_i = h', x' \mid y, \theta\right), \\
\frac{\delta L(\theta)}{\delta \nu(h', h'')} = \sum_{(i,j) \in E} \omega_{ij} \cdot g_{ij}(y) \cdot \ \left[p\left(h_i = h', h_j = h'' \mid x, y, \theta\right) - p\left(h_i = h', h_j = h'' \mid y, \theta\right)\right], \\
\frac{\delta L(\theta)}{\delta \gamma(h', h'', x')} = \sum_{(i,j) \in E} \omega_{ij} \cdot \left[p\left(h_i = h', h_j = h'' \mid x, y, \theta\right) - p\left(h_i = h', h_j = h'', x' \mid y, \theta\right)\right].
\end{gathered} \tag{3.2}$$

Note that, all the terms in the derivatives can be calculated using Belief Propagation (BP) algorithm [17], provided the graphical structure does not contain cycles. Otherwise, approximate methods, such as loopy BP could be considered. Here, BP is suitable for our case due to the tree-like neighborhood structure.

For the final class inference, we need to find the image label $\hat{x}$ that maximizes the conditional distribution $(x \mid y, \theta)$, given parameters θ^*. For this work, we can also use the max-product version of BP to find the MAP estimate $\hat{x} = \arg\max_x P(x \mid y, \theta^*)$.

4. Experiments

In this section, we demonstrate the capability of the proposed model on two different datasets: Caltech-4 standard dataset and airplane images collected from Google Earth. The aim of these experiments is to illustrate the performance of this detection framework using contextual hierarchical part-driven CRF model and compare with state-of-the-art models.

4.1. Image Features

For the object category detection task, the robustness and distinctiveness are basic requirements for local features in order to provide powerful expression ability for objects. On the other hand, the quantity of features corresponds to the quantity of observations and has an enormous influence on computational complexity. Taking these aspects into consideration, we should try to use the local features which have the characteristics of sparse, robust, and discriminative simultaneously.

In our experiments, we use the beta-stable feature extracting method [20] to locate local features and SIFT descriptor [21] to construct feature vectors. Rather than selecting features that persist over a wide interval of scales, beta-stable features are chosen at a scale so that the number of convex and concave regions of the image brightness function remains constant within a scale interval of length beta. As a result, the beta-stable features have stronger robustness than SIFT-like features and are better anchored to visually significant parts. The comparative feature-points detecting results are shown in Figures 3(a) and 3(b).

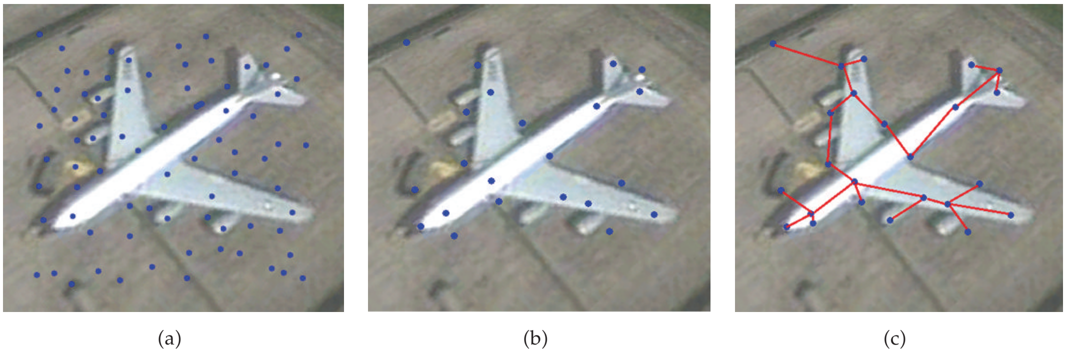

(a) (b) (c)

Figure 3: (a) SIFT feature detecting, (b) beta-stable feature detecting, (c) beta-stable features connected by MST.

The unary feature vector, $f_i(y)$, used in this work is represented by the combination of SIFT descriptor and relative location features. The pairwise feature vector, $g_{ij}(y)$, is just the joint of unary feature vector $f_i(y)$ and $f_j(y)$.

Given the locations of local observations, we construct graphical models using MST, as shown in Figure 3(c). The edge cost used in MST construction between two observations was computed by

$$\cos t_{ij} = \varepsilon_1 \times \text{2D distance } (i,j) + \varepsilon_2 \times \text{Distance of color histograms } (i,j), \tag{4.1}$$

where ε_1 and ε_2 are balance factors depending on the actual object, and $\varepsilon_1+\varepsilon_2 = 1$. If the object has richer shape information than appearance information, we think that the 2D distance might be more useful for discrimination, so we will take a bigger ε_1 than ε_2.

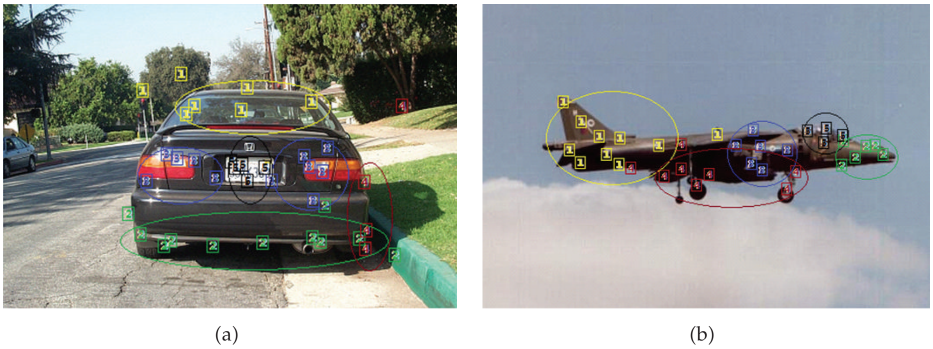
4.2. *Object Detection on Standard Database*

The first dataset that we used to test our model is a subset of the Caltech-4 standard dataset, which contains images for two object categories, car (rear view) and airplane (side view), and one background category. Each image contains at most a single instance of the objects in diverse natural background and, therefore, is suitable for our 2-class detection task. We randomly split the images into two equal separate subsets for training and testing.

Figure 4 shows the examples of the assignment of parts to local features for two object categories. It is apparent that the proposed model can effectively associate the mass of scattered and unordered observations with their corresponding object parts. Note that multiple observations may be assigned to the same part label with the premise that they physically belong to the same part. The number of parts can be empirically set according to the complexity of objects.

4.3. *Airplane Detection in Satellite Images*

In this section, we verify our model by 170 airplane (top view) satellite images taken from Google Earth. To gather a sufficiently large learning dataset, we acquire images from different heights and different directions. Furthermore, a few synthetic images with simulation



(a)

(b)

Figure 4: Examples of the assignment of parts to local features for car object category (a) and airplane object category (b), which are labeled by different number and colors. The number of parts is set to 5.

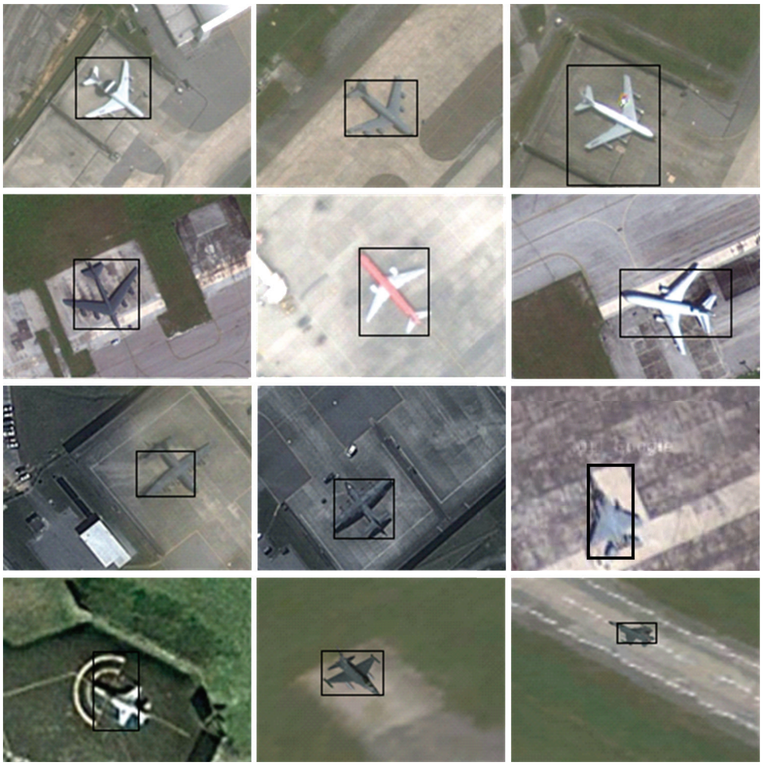


Figure 5: Examples of successful detections on satellite images and synthetic images. Note that simulation airplane models in the last two synthetic images are also correctly detected.

airplane models are also used for testing. All images are resized to $150 * 100$ pixels. The balance factor ε_1 used in the weighted neighborhood structure is set to 0.7 to encourage the use of shape information.

Due to space constraints, we provide a few examples of the detection results (as shown in Figure 5). We use a simple bounding box located at the center of the efficient observations to roughly label the detected objects in the test images.

Table 1: Comparisons of detection performance (EER).

Models	Car (rear)	Airplane (side)	Airplane (top)
Hidden CRF	91.0%	94.1%	93.4%
Located HRF	92.1%	95.6%	**97.0**%
Multiclass CRF	90.6%	93.8%	92.1%
Our part-driven CRF	**94.2**%	**97.3**%	**96.5**%
Part-driven CRF without WNS	93.4%	94.9%	93.7%

4.4. Performance Comparison

We compare the detection performance of our model with those of three existing models: hidden CRF [15] model, located HRF [16] model, and multiclass CRF [14] model. For fairness of comparison, the local features in these three models are also computed by SIFT descriptors. In order to measure the influence of neighborhood structures, we also investigate the performance of an equivalent model without weighted neighborhood structure. The object categories are car (rear), airplane (side), and airplane (top), which have been mentioned in previous sections. The equal error rate (EER) defined in [7] is adopted as evaluation criterion, in which higher EER values means better classification performance. The comparative results are summarized in Table 1.

As can be seen, our model consistently gives the best results for these three object categories for the car rear dataset and airplane side dataset. Note that the airplanes (side) are easier to be discriminated than cars due to their distinct shape structure. On the airplane top dataset, our model is exceeded slightly in accuracy only by the located HRF model. This may be caused by overfitting since our model has to use more parameters to encode more contextual dependencies.

From the results in the last row of Table 1, we can see that incorporating the weights of neighborhood structures is important since the performance of such a model dropped. Rather than hypothesizing that all the edges in MST are equally important, the weighted neighborhood structure uses weights to measure the degree of correlation between connected nodes and inherently have higher representational power.

Note that, since the local features are extracted automatically during training and testing, the quantity and structure (constructed by MST) of observations should be unpredictable. As a result, the computing time of our model is influenced by the object complexity and image quality. In this experiment, we use about 3 hours for training, and 1.2 second per image on the average for testing on a 2.8 GHz computer.

5. Conclusion

In this paper we presented a contextual hierarchical part-driven CRF model for object category detection. By incorporating two single-level models, the proposed model can effectively represent latent label-level context and observation-level context simultaneously. A weighted neighborhood structure is also introduced to capture the degree of correlation between connected nodes. Experimental results on challenging datasets with high intraclass variability have demonstrated that the proposed model can effectively represent multiple context information and give competitive detection performance. Our future researches will focus on the following directions: introducing more sparse and robust local features to reduce the computational complexity and utilizing high-order clique potentials to investigate more contextual dependencies in images.

Acknowledgments

This work is supported by a special financial grant from the China Postdoctoral Science Foundation (Grant no. 20100481512, 201104765). The authors would like to thank Kevin Murphy for publishing CRF2D toolbox, which is very helpful for this work. They also thank X. J. Xiang for data collection.

References

[1] P. Schnitzspan, S. Roth, and B. Schiele, "Automatic discovery of meaningful object parts with latent CRFs," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 121–128, San Diego, Calif, USA, June 2010.

[2] J. H. Zhang, J. W. Zhang, S. Y. Chen et al., "Constructing dynamic category hierarchies for novel visual category discovery," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '12)*, Vilamoura, Portugal, October 2012.

[3] S. Y. Chen, J. H. Zhang, Y. F. Li, and J. W. Zhang, "A hierarchical model incorporating segmented regions and pixel descriptors for video background subtraction," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 118–127, 2012.

[4] M. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 22, no. 1, pp. 67–92, 1973.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 832–838, Beijing, China, October 2005.

[6] D. Crandall, P. Felzenszwalb, and D. Huttenlocher, "Spatial priors for part-based recognition using statistical models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 10–17, San Diego, Calif, USA, June 2005.

[7] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, pp. II/264–II/271, San Diego, Calif, USA, June 2003.

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[9] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scenes using transformed objects and parts," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 291–330, 2008.

[10] S. Kumar, *Models for learning spatial interactions in natural images for context-based classification [Ph.D. thesis]*, The Robotics Institute, Carnegie Mellon University, 2005.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, pp. 282–289, 2001.

[12] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," *Advances in Neural Information Processing Systems*, pp. 1351–1358, 2004.

[13] S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification," in *Proceedings of the International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 1150–1157, Nice, France, October 2003.

[14] S. Kumar and M. Hebert, "Multiclass discriminative fields for parts-based object detection," in *Snowbird Learning Workshop*, March 2004.

[15] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1853, 2007.

[16] A. Kapoor and J. Winn, "Located hidden random fields: Learning discriminative parts for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV '06)*, pp. 302–315, Vienna, Austria, May 2006.

[17] S. Y. Chen, H. Tong, and C. Cattani, “Markov models for image labeling,” *Mathematical Problems in Engineering*, vol. 2012, Article ID 814356, 18 pages, 2012.

[18] C. Cattani, S. Y. Chen, and G. Aldashev, “Information and modeling in complexity,” *Mathematical Problems in Engineering*, vol. 2012, Article ID 868413, 3 pages, 2012.

[19] S. Y. Chen, H. Tong, Z. Wang, S. Liu, M. Li, and B. Zhang, “Improved generalized belief propagation for vision processing,” *Mathematical Problems in Engineering*, vol. 2011, Article ID 416963, 12 pages, 2011.

[20] S. Gu, Y. Zheng, and C. Tomasi, “Critical nets and beta-stable features for image matching,” in *Proceedings of the 11th European Conference on Computer Vision (ECCV ’10)*, vol. 6313 of *Lecture Notes in Computer Science*, pp. 663–676, 2010.

[21] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV’99)*, pp. 1150–1157, Kerkyra, Greece, September 1999.