J. Vis. Commun. Image R. 28 (2015) 83–96

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Hybrid graphical model for semantic image segmentation ☆

Li-Li Wang *, Nelson H.C. Yung

Laboratory for Intelligent Transportation Systems Research, Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 1 April 2014
Accepted 21 January 2015
Available online 30 January 2015

Keywords:
Semantic segmentation
Conditional Random Field
Bayesian Network
Graphical model
Spatial relationship
Hybrid model
Sub-scene
Contextual interaction

ABSTRACT

To make full use of both non-causal and causal cues in natural images, we propose a hybrid hierarchical Conditional Random Field (HCRF) and Bayesian Network (BN) model for semantic image segmentation in this paper. The HCRF is used to capture non-causal relationship, such as appearance features and inter-class co-occurrence statistics, to produce initial semantic sub-scene predictions. Whereas, the BN is used to model contextual interactions for each semantic sub-scene in the form of class statistics from its neighboring regions, of which its conditional probabilities are learned automatically from training data. The learned BN structure is then used to encode the structure of contextual dependencies for sub-scenes in the initial predictions to generate final refined predictions. Experiments on the Stanford 8-class dataset and the LHI 15-class dataset show that the hybrid model outperforms pure CRF models by 2–4% in average classification accuracy.

1. Introduction

Semantic image segmentation aims at labeling pixels in an image according to predefined classes, such as building, grass, cow or face. It offers rich knowledge to wide range of computer vision tasks, such as content based image retrieval [1], foreground/background extraction [2], face recognition [3], human pose estimation [4], and scene categorization [5].

Many approaches have been proposed for solving the semantic image segmentation problem. The most popular methods at present include a list of probabilistic graphical model (PGM) based methods [6–16] and deep learning methods [17–20]. In the PGM methods, both discriminative and generative models have been investigated. Markov Random Field (MRF) [13–16] is a typical representative of the generative model. It models the joint probability of an image and corresponding labels. To estimate the parameters of the MRF model, a large number of labeled images are required. As a result, it makes parameter estimation and inference fairly complexity. Different from MRF, Conditional Random Field (CRF) [6–12], as a discriminative model, estimates the posterior probability over labels. Compared with MRF, CRF model learns more effectively. Constructing CRF models for segmentation problems is a hot research topic at present. Among the CRF models, the Associative Hierarchical Random Fields (AHRF) model presented in [7] achieves excellent segmentation results. In the AHRF model, a variety of cues in the form of hierarchies are taken into account. It first extracts appearance information, such as color, texture, to decide on the labeling of pixels in the lowest level of semantic understanding. In the middle level, region continuity is considered for labeling regions. In the high level, co-occurrence statistics of inter-classes is encoded to suppress impossible combinations of objects in the same scene. Semantic image segmentation is thus achieved through optimizing an energy function that is defined by the AHRF model. However, both MRF and CRF models are unidirectional graphs and only non-causal relationships are captured by them. As for causal relationships, such as spatial relationship, they are also important for enhancing semantic image segmentation performance [21]. In order to capture such relationships, directional graphs, such as Bayesian Networks, are preferred. Recently, deep learning [17–20] has attracted much attention in the field of machine learning. It represents a set of algorithms that attempt to learn high-level abstract representation in data through multiple layers of perceptions. In computer vision, Convolutional Neural Network (CNN) [22,23] as a deep learning algorithm has been developed and applied for scene labeling successfully. A typical architecture of CNN [22] usually composes of an input layer, several convolutional layers, several pooling layers and a fully connected layer. Weights connecting two adjacent layers are learned based on back propagation [24]. Consequently, more discriminative features can be directly extracted from pixels for recognition tasks. Note that CNN cannot capture high-level semantic cues, such

☆ This paper has been recommended for acceptance by M.T. Sun.
* Corresponding author.

http://dx.doi.org/10.1016/j.jvcir.2015.01.014
1047-3203/

as co-occurrence, long range pixel interaction and contextual information among labels at present. Although pleasing semantic image segmentation results can be obtained based on CNN, object boundaries in segmented space are typically noisy. To compensate for this shortcoming, one possible approach is to incorporate a CRF model into CNN. However, when noise or clutters appear in scenes, global contextual information [8,21,25,26] is vital in assisting to disambiguate object identities. For example, global co-occurrence statistics are incorporated in [8] to encode the frequency of two objects appearing in an image together. By introducing co-occurrence potential term in the energy function, certain impossible combinations of object classes can be prohibited. On the other hand, when visual occlusions occur between objects, local context [27–32] is more robust to identify objects. In [27], relative location prior was considered for capturing spatial relationships between two classes to improve semantic segmentation accuracy. However, we note that the relative location prior is learned based on over-segmentation. In another word, an image is required to be first partitioned into many segments without supervision for statistics. Such statistics are not unique due to uncertain number of segments in an image and different unsupervised segmentation algorithms. It becomes difficult to generate accurate relative location maps to guide scene labeling. Another problem is that complexity is high due to over segments required for both training and testing. Note that contextual information aforementioned is represented as interaction of only two object classes.

To address the above problems, we propose to extract contextual information for each sub-scene from its spatial layout in an image. This representation can provide more specific information about the configuration of objects in natural scenes. Furthermore, to take advantage of both appearance features and contextual information in natural images, we develop a hybrid model that combines both hierarchical CRFs (HCRFs) and Bayesian Networks (BNs) for semantic image segmentation in this paper. The hybrid model incorporates different types of cues in a hierarchical manner. More specifically, the HCRFs consist of three layers to capture non causal relationships among the random variables, such as pixels, pair of pixels, segments and classes by taking into account appearance features and global context cues. By optimizing the HCRFs part, initial predictions are produced. Subsequently, contextual constraint on spatial layout of each sub-scene is incorporated through a naïve BN (NBN) model with conditional probabilities to disambiguate the initial prediction results. The main contributions of this paper are summarized as follows.

- The new hybrid model incorporates both non causal (knowledge from low level to high level information) and causal relationships (spatial layout of objects) for semantic image segmentation. It produces globally consistent labeling results and rational spatial layouts of scenes.
- As K-Means [33] focuses on extracting foreground details, and Meanshift [34] segments background well, the proposed hybrid model achieves finer and balanced segmentation for both background and foreground. As a result, accurate labeling of both background and foreground is achieved under the guidance of the finer unsupervised segments.
- Different from published methods [27,29], this paper shows an alternate way of capturing contextual relationships for each semantic sub-scene, i.e. spatial layout of an object is described in the form of class statistics from its neighboring nine regions. The first advantage is that the collection of local contextual information in the unit of sub-scene avoids hard segmentation. Secondly, no side effect is introduced due to inaccurate segmentations. Thirdly, the local contextual information is scale invariant due to relative statistics.
- Embedded in the hybrid model, a two-stage inference is also proposed. Through imposing the causal relationship constraint as the second stage inference, the computational complexity is reduced significantly while semantic image segmentation with reasonable spatial layouts is achieved.

Experimental results show that the proposed hybrid model can provide more logical labeling results and substantially improve classification accuracy when compared with pure CRF models. For the Stanford dataset [35], the proposed method achieves a global accuracy of 81.0% and average accuracy of 71.4%. For the LHI dataset [36], the global and average accuracy are 82.2% and 62.1% by using the proposed method, respectively.

The rest of the paper is organized as follows. In Section 2, we review the AHRF model and their shortcoming for semantic image segmentation. In Section 3, we describe the details of the proposed method. Experimental results are given in Section 4, and the paper is concluded in Section 5.

2. Related work

In recent years, semantic image segmentation has made significant advances. Generally speaking, a good approach usually consists of three steps. The first step is to select versatile cues from low level to high-level, such as appearance cues, region continuity cues, co-occurrence cues, to describe image contents. The second step is to build a model to incorporate the selected cues. The third step is to determine the optimal segmentation through minimizing an energy function defined by the model.

Extracting discriminative features is fundamental but crucial for identifying different classes. Up to now, methods for extracting features can be grouped into two groups based on either engineered features or trained features. Engineered features mean that features are extracted by using fixed descriptors, such as texton [10,37], Scale-Invariant Feature Transform (SIFT) [38], local binary patterns (LBP) [39]. Whereas, trained features are learned directly from raw pixels based on methods such as CNN [23], from which a powerful representation of input data is generated. Based on the extracted features, unary prediction in one-layer CRF model can be obtained. However, as the one-layer CRF model does not produce good enough contours of objects [40] and the prediction is typically noisy, higher-order potentials such as co-occurrence statistics [8] need to be taken into account. A PGM is then constructed to capture interactions between random variables. In the PGM, an energy function is defined on a discrete random field X. Each random variable $X_i \in X$ corresponds to a node in the graphical model. The indexes of all basic nodes consist of a set of $V = \{1, 2, \ldots N_b\}$, where N_b denotes the number of basic nodes. The value x_i of each random variable X_i represents the class label which takes a value from the label set $L = \{l_1, l_2, \ldots\}$. Thus the semantic image segmentation problem is to find a label for each node in the PGM from the label set. Typically, an energy function defined by the AHRF model [7] is expressed as a sum of unary, pairwise and higher-order potentials as follows

$$E(x) = \sum_{i \in v} \phi_i(x_i) + \sum_{i \in v, j \in N_i} \psi_{ij}(x_i, x_j) + \sum_{c \in C} \psi_c^h(x_c^h) + \kappa(L), \quad (1)$$

where V denotes the set of pixels in an image, N_i denotes the set of neighboring pixels of pixel i in a two-dimensional image space, and C denotes a set of cliques (super pixels or segments). In Eq. (1), the first three terms are typically evaluated as follows [7]

$$\phi_i(x_i \in L) = -\alpha \log(p(x_i)), \quad (2)$$

$$\psi_{ij}(x_i, x_j) = \beta_0 + \beta_1 \exp\left(-\frac{W||I_i - I_j||^2}{\beta_2}\right), \quad j \in N_i, \tag{3}$$

$$\psi_c^h(x_c^h) = \min_{l \in L}\left(\gamma_c^{\max}, \gamma_c^l + \sum_{i \in c} w_i k_c^l \Delta(x_i \neq l)\right), \tag{4}$$

where $\phi_i(x_i)$ denotes the unary potential, which is defined on a pixel i. It can be evaluated by the probability $p(x_i)$ that pixel i is assigned to label x_i as shown in Eq. (2), and α is a model parameter. This probability can be estimated from the output of a joint boosting classifier [7,41] based on low level appearance features of each pixel in an image. In addition, softmax regression model [42] can also be used to estimate $p(x_i)$ as follows

$$p(x_i) = \begin{bmatrix} p(x_i = 1|f_i, \theta) \\ p(x_i = 2|f_i, \theta) \\ \vdots \\ p(x_i = L|f_i, \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{L} e^{\theta_j^T f_i}} \begin{bmatrix} e^{\theta_1^T f_i} \\ e^{\theta_2^T f_i} \\ \vdots \\ e^{\theta_L^T f_i} \end{bmatrix} \tag{5}$$

where $\theta_1, \theta_2, \ldots, \theta_L \in \mathbf{R}^{n+1}$ are the parameters of softmax regression model, and $\boldsymbol{f}_i$ denotes feature vector of pixel i.

The pairwise potential $\psi_{ij}(x_i, x_j)$ defined in Eq. (3) encodes a smoothness prior between the neighboring random variables X_i and X_j, and it is used to penalize neighboring pixels that have been assigned different classes. β_0, β_1 and β_2 are model parameters whose values are learned from the training set. I_i and I_j are the color vectors of pixel i and j respectively, and W is the weight vector corresponding to the three color components. The higher-order potential term $\psi_c^h(x_c^h)$ is defined over a segment c that takes the form of Eq. (4), where $\gamma_c^{\max}$ denotes the maximum penalty for segment c when it is labeled as one class with the smallest probability, γ_c^l represents the potential cost if segment c takes a dominant label $l \in L$, and $w_i(k_c^h)$ is used for calculating an additional penalty on each pixel in segment c without taking the label l. The weight w_i is usually set to 1. Segment c can be generated from any unsupervised image segmentation methods. For each segment, one dominant label l is assigned based on the output of a joint boosting classifier [41] by using only appearance information. This potential term is employed to capture long range pixel interactions to interpret structural dependencies between pixels in segments. As a result, fine segmented contours of objects are obtained by incorporating higher order potentials in the AHRF model.

To further improve segmentation accuracy, contextual information is introduced into the AHRF model by taking the form of co-occurrence potential $\kappa(L)$ as shown in Eq. (1). Co-occurrence statistics are used to encode the frequencies of the pair of object classes appearing in an image together. For uncommon combinations of object classes from high-level statistics, a larger penalty is imposed. As a result, it has the potential to prohibit certain impossible combinations in natural images. By minimizing the energy function in Eq. (1), consistent semantic image segmentation results can be achieved.

From the above analysis, we can see that the AHRF model incorporates both appearance features and object co-occurrence statistics for semantic image segmentation. It should be noted that object co-occurrence statistics in the AHRF model only captures the frequency that two object classes appear in the same image, while the spatial layout of object classes is ignored. To further improve the classification performance, we integrate spatial layout of objects with a hierarchical CRF to form a hybrid model. Note that the CRF model is an undirected graphical model which is used to model non-causal relationships among random variables, such as pixels and segments. However, spatial layout of object classes is generally collected based on causal interactions between objects. To express the causal relationships, a directed graphical model is preferred. In the following section, we will illustrate how to build the structure of a directed graphical model, learn parameters of the model, and how to fuse the learned model with CRF to jointly perform inference.

3. Proposed hybrid graphical model

3.1. HCRF to capture non causal relationship

Similar to the method presented in [7], the energy function used in an HCRF model takes the form of Eq. (1). The unary and pairwise potential terms are defined by Eqs. (2) and (3), respectively. In order to capture the fine contours of objects, higher-order potentials defined over a set of segments are incorporated into the energy function of the HCRF model. To generate segments in the middle-level of the HCRF model, unsupervised segmentation methods, such as K-Means [33] and Meanshift [34], are commonly used. In general, fine segmentation can extract more accurate object boundaries. Higher-order potentials, also called segment consistency potentials, have the capability to smooth regions belonging to the same object in the labeling space under the guidance of the fine segmentation. From the inference point of view, it helps the labeling process recover from false unary predictions. However, if unsupervised segments are too coarse, the inferred objects become incomplete under the guidance of inaccurate segments, which obviously results in an incorrect labeling. In order to avoid this problem, we adopt a new criterion to produce fine image segments in the middle layer of the HCRF model.

Given an image I, it is over segmented to produce multiple segments $I = \{C_1, C_2, \ldots, C_M\}$, where M denotes the number of segments in an image. In this paper, we adopt the morphological method of watersheds [43,44] to partition an image into primitive regions. The initial regions are then hierarchically (bottom-up) merged to achieve the predefined number (M) of segments. Suppose the number of initial segments is N. In order to produce M ($M < N$) segments, the most similar pair of adjacent regions is merged at each step. Similar to the methods in [45,46], the most similar pair of regions is the one that minimizes the following dissimilarity function.

$$\delta(R_i, R_j) = \frac{||R_i|| \cdot ||R_j||}{||R_i|| + ||R_j||}\left[\mu(R_i) - \mu(R_j)\right]^2, \tag{6}$$

where $||R_i||$ represents the cardinality of region i, and $\mu(R_i)$ denotes mean intensity value of pixels in region i. R_i and R_j are adjacent regions.

To obtain more consistent semantic image labeling with segments, three-level spatial pyramid segments are generated by merging the initial fine regions into increasingly coarser regions. The number of segments in the three levels is set to M, $M/2$ and $M/4$, respectively. Here, M denotes the number of segments used in the first layer. The higher-order potentials are then defined over the three levels of segments as in Eq. (4) by taking the form of a robust P^n Potts model [6,7].

3.2. Bayesian Networks to model local contextual relationships of sub-scenes

In order to incorporate spatial contextual information, we use a Bayesian Networks to model the conditional distribution over the class labeling given a sub-scene. In our paper, sub-scene means a collection of pixels assigned the same label. We model contextual information based on sub-scenes instead of unsupervised segments, and an image in the labeling space is represented by an ensemble of its sub-scenes. This idea has a number of effects (i)

no engineered segmentation is required, and the labeling of segments is thus saved; (ii) the statistical information of spatial layout is easily collected due to the fact that the number of sub-scenes is usually less than that of segments in an image used for relative location statistics, (iii) the inference thereby becomes simpler. The reason is that the number of sub-scenes in natural images is usually stable and relatively smaller.

3.2.1. Contextual information of sub-scenes

Given a training set $T = \{I_1, I_2, \ldots\}$ of images with ground truth, an image I in the training set is automatically decomposed into a set of sub-scenes $I = \{S_1, S_2, \ldots\}$ where each sub-scene S_i is generated by a collection of pixels assigned with the same label based on the ground truth in T. The contextual information for each sub-scene of interest (subSOI) is depicted by regions of eight directions and one central region as shown in Fig. 1.

To interpret the contextual structure in Fig. 1, a box is used to bound each subSOI in the two-dimensional labeling space. Nine regions from R_0 to R_8 as shown in Fig. 2 are thereby formed as the contextual model of each subSOI in an image, of which the model is similar to the one used in [47]. Also note that in Fig. 2, subSOI denotes the area marked as *foreground*, and the bounding area minus subSOI is denoted by R_0, hence nine regions as a result. The statistical information in these nine regions is used as the contextual descriptor of the subSOI. In order to identify these nine regions for each subSOI in an image, the coordinates of the four vertices (from P_0 to P_3) of the bounding box are first determined. As shown in Fig. 2, $P_0 = (h_{\min}, v_{\min})$, $P_1 = (h_{\max}, v_{\min})$, $P_2 = (h_{\min}, v_{\max})$ and $P_3 = (h_{\max}, v_{\max})$, where $h_{\min}$ and $h_{\max}$ denote the minimal and maximal horizontal coordinate values of *foreground*, respectively. Similarly, $v_{\min}$ and $v_{\max}$ denote the minimal and maximal vertical coordinate values of *foreground*, respectively. These four values are easily calculated by searching all points in *foreground*. Once the four values are determined, nine regions can be constituted

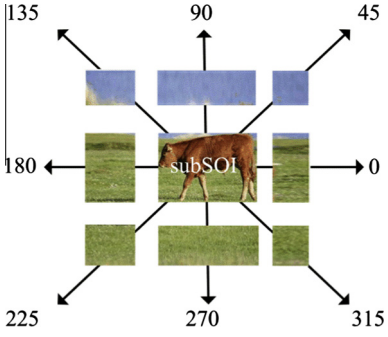

Fig. 1. Contextual structure of subsoil.

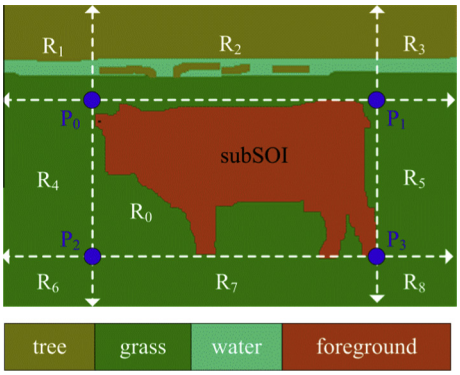

Fig. 2. Spatial layout of contextual information for subSOI (scale invariant).

Table 1
Nine contextual regions for one subSOI.

$R_0 = \{(h, v) | h_{\min} \leqslant h \leqslant h_{\max}, v_{\min} \leqslant v \leqslant v_{\max}, (h, v) \notin \text{subSOI}\}$
$R_1 = \{(h, v) | h < h_{\min}, v < v_{\min}\}$
$R_2 = \{(h, v) | h \geqslant h_{\min}, v < v_{\min}\}$
$R_3 = \{(h, v) | h > h_{\max}, v < v_{\min}\}$
$R_4 = \{(h, v) | h < h_{\min}, v_{\min} \leqslant v \leqslant v_{\max}\}$
$R_5 = \{(h, v) | h > h_{\max}, v_{\min} \leqslant v \leqslant v_{\max}\}$
$R_6 = \{(h, v) | h < h_{\min}, v > v_{\max}\}$
$R_7 = \{(h, v) | h \geqslant h_{\min}, v > v_{\max}\}$
$R_8 = \{(h, v) | h > h_{\min}, v > v_{\max}\}$

as described in Table 1 for the *foreground*. The spatial contextual relationships for *foreground* are thus described by the conditional probability distributions $\{p_j(l|foreground)|\ j = 0, 1, \ldots 8$ and $l = 1, \ldots, |L|-1\}$, where j denotes region index and l denotes object class. These distributions are encoded as the spatial contextual descriptor of *foreground*. As such, the local contextual descriptors for other subSOIs can be learned by accumulating statistical information in their nine corresponding regions from images in a training set. This contextual descriptor is scale invariant since the statistical information is collected from nine relative regions of each subSOI. Fig. 3 lists contextual information of 8 sub-scenes on the Stanford dataset in terms of conditional distributions of object classes. From Fig. 3, we can see that each sub-scene has its unique contextual information. Take subSOI *sky* as an example, the contextual information $\{p_j(l|sky)|\ j = 0, 1, \ldots 8$ and $l = 1, \ldots, |L|-1\}$ is shown in the first column of Fig. 3. From the seven figures, it can be observed that all classes including *tree*, *road*, *grass*, *water*, *building*, *mountain* and *foreground* have dominant probabilities in R_7, namely, these classes are usually located below *sky*. In general, it is impossible for class *tree* to appear in R_1–R_3 when subSOI is *sky*. As a result, $p_j(tree|sky)$ equals to 0, where $j = 1, 2, 3$ as shown in the first column of Fig. 3.

3.2.2. Bayesian Networks to model the contextual information of subSOI

To incorporate spatial relationship into semantic image segmentation, we construct a Naïve Bayesian Network (NBN) model as depicted in Fig. 4 to capture contextual information of each sub-scene from nine spatial regions and encode it as a local feature. In the NBN model, the father node G_i denotes the subSOI. The contextual information from nine spatial regions including eight directional regions and one central area are considered as nine child nodes. Each child node Z_{ij} in the NBN model is associated with one region R_j in Fig. 2. They have the same father node, and are independent. The value z_{ij} of random variable Z_{ij} denotes the label assigned to R_j. The Bayesian parameters, namely the conditional probability distributions $\{p(Z_{ij}|G_i)|\ j = 0, 1, \ldots, 8,$ and $i = 1, 2, \ldots, |L|\}$, are learned given image sub-scenes, where $|L|$ denotes the total number of classes in a dataset. It is a simple training problem since the Bayesian Network is defined over a relatively small number of sub-scenes rather than pixels or segments. The learned model is then used in the test images.

By using the NBN, the posterior probability can be calculated by Eq. (7).

$$p(G_i|Z_{i0}Z_{i1}Z_{i2}\ldots Z_{i8}) = \frac{p(Z_{i0}Z_{i1}Z_{i2}\ldots Z_{i8}|G_i)p(G_i)}{p(Z_{i0}Z_{i1}Z_{i2}\ldots Z_{i8})} = \frac{p(Z_{i0}Z_{i1}Z_{i2}\ldots Z_{i8}|G_i)p(G_i)}{\sum_{j=1}^{C} p(Z_{i0}Z_{i1}Z_{i2}\ldots Z_{i8}|G_j)p(G_j)} \tag{7}$$

Since variables $Z_{i0}, Z_{i1}, \ldots, Z_{i8}$ are independent, $p(Z_{i0}Z_{i1}\ldots Z_{i8}|G_i)$ is given by Eq. (8) according to Chain Rule,

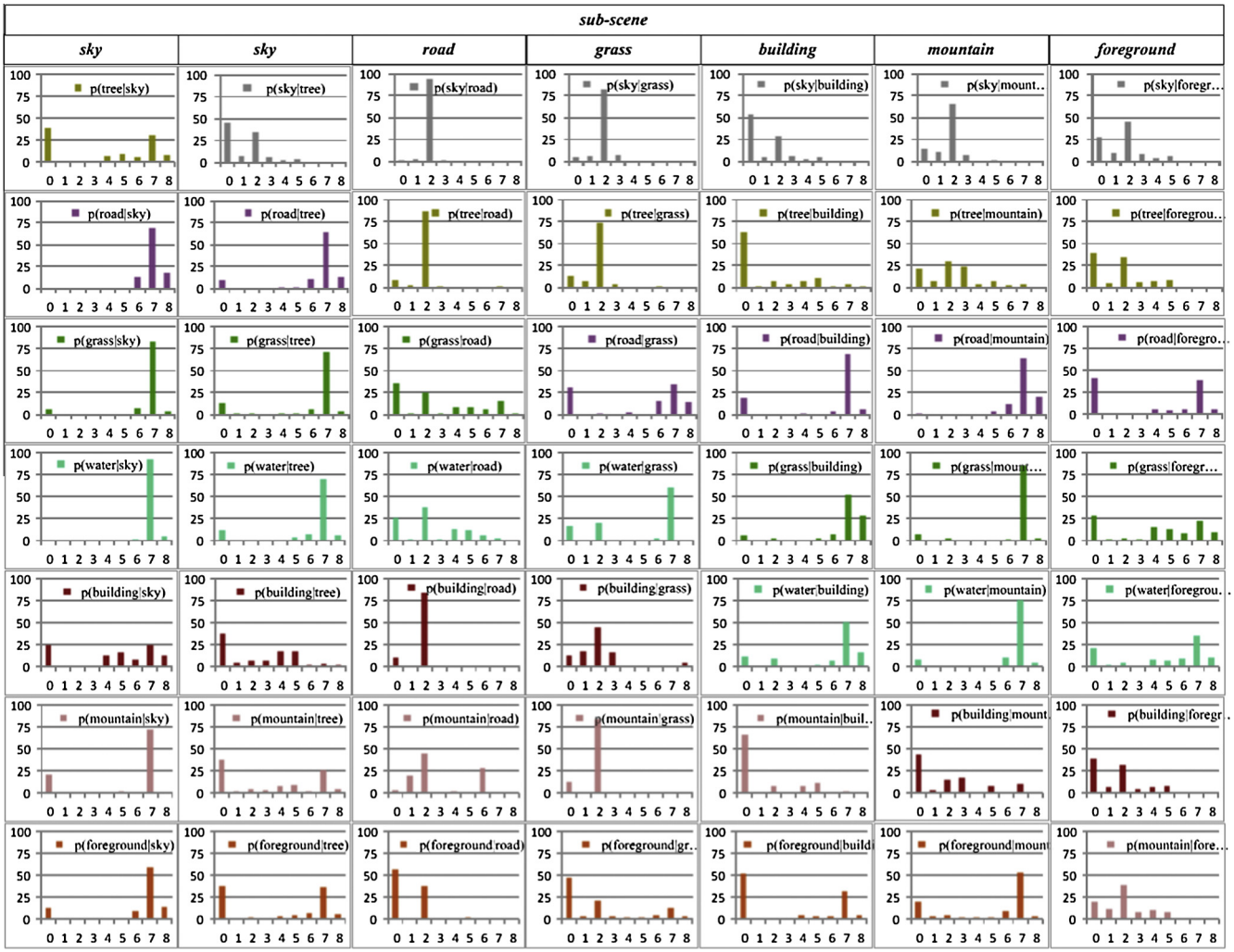

Fig. 3. Contextual information of 8 sub-scenes in the Stanford dataset in terms of conditional distributions (vertical axis: probability; horizontal axis: region index as shown in Fig. 2).

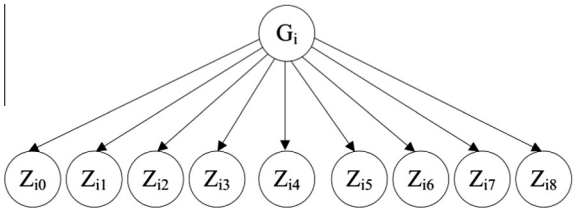

Fig. 4. NBN to model contextual information of the subSOI.

$$p(Z_{i0}Z_{i1}Z_{i2}\ldots Z_{i8}|G_i)=\prod_{j=0}^{8}p(Z_{ij}|G_i), \tag{8}$$

where the class conditional probability $p(Z_{ij}|G_i)$ denotes Bayesian parameter. Given one subSOI as shown in Fig. 2, the class conditional probabilities $\{p(Z_{ij}|G_i)|\ j=0,\ 1,\ \ldots,\ 8\}$ that denote the frequency of object classes appearing in regions $\{R_j|\ j=0,\ 1,\ \ldots,\ 8\}$ as given in Table 1 are usually different, which can be learned from the ground truth of the training set T. It describes what we expect to observe in a certain direction or center area of the subSOI. Note that we use the regional statistics to represent contextual information in a certain direction, $p(Z_{ij}|G_i)$ is thereby approximated as

$$p(Z_{ij}|G_i)=\prod_{k=1}^{N_{ij}}p_j(l_k|G_i), \tag{9}$$

where N_{ij} denotes the number of classes available in R_j when subSOI is G_i, and $p_j(l_k|G_i)$ can be learned from the training set T as follows:

$$p_j(l_k|G_i)=\frac{n_j(l_k|G_i)}{\sum_{m=1}^{|L|}n_j(m|G_i)},\ \text{for } j=0,1,\ldots,8 \text{ and } k=1,2,\ldots|L| \tag{10}$$

$$n_j(l_k|G_i)=\sum_{G_i\subset T}\sum_{p\in R_j}\delta(X_p=l_k|G_i). \tag{11}$$

In Eq. (7), $p(G_i)$ denotes the prior probability of sub-scene S_i, which can be approximated as follows:

$$p(G_i=l)=\frac{\sum_{j\in S_i}p(x_j=l)}{|S_i|},\ \text{for } l\in[1,2,\ldots,L], \tag{12}$$

where $|S_i|$ represents the number of pixels in sub-scene S_i. $p(x_j=l)$ denotes the probability that pixel j in sub-scene S_i is labeled as l, and it can be estimated from the output of an joint boosting unary classifier. By substituting Eqs. (8) and (12) into Eq. (7), the posterior probability $p(G_i|Z_{i0}Z_{i1}\ldots Z_{i8})$ is determined. This means that the probability of each subSOI in an image can be estimated based on the contextual information from eight directional regions and one central area. In the following section, we will describe how to obtain

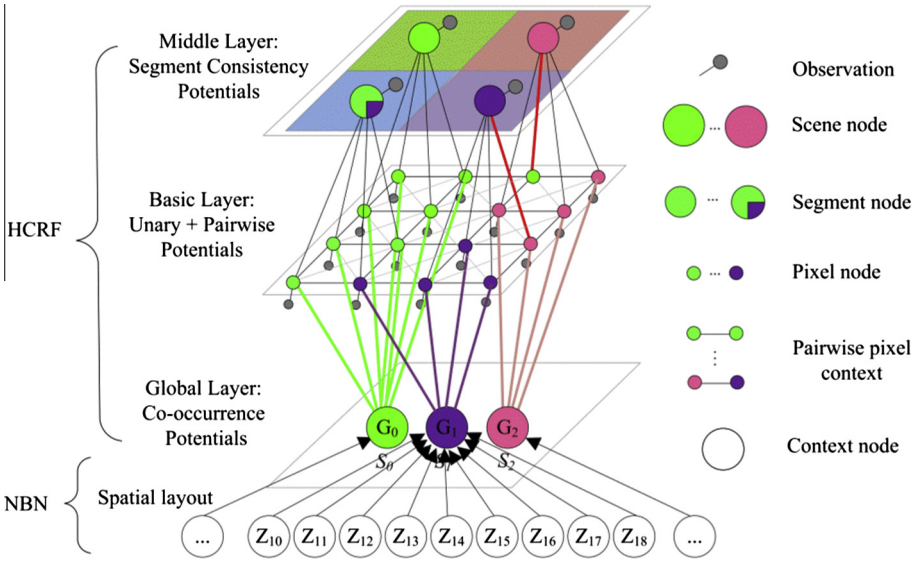

Fig. 5. Proposed hybrid graphical model.

the final segmentation by fusing the HCRF model and the NBN model.

3.3. New hybrid graphical model

Fig. 5 depicts the structure of the proposed hybrid graphical model. It incorporates the HCRF model with the NBN model to express non-causal and causal relationships existing in natural images, respectively. As shown in Fig. 5, the HCRF model is composed of three layers which are used to model local interactions between pixels, super-pixels or inter-objects, respectively. At the basic layer of the HCRF model, each node represents one pixel. Different types of features, such as texton, color, color SIFT, LBP, are extracted as the observations of each pixel appearance. Based on the observations, each pixel is assigned a list of labels according to confidence. Unary potentials are then defined based on the observations and adaptive boosting classifiers to produce baseline predictions, and pairwise potential defined between two neighboring pixels is adopted to smooth the baseline predictions. The middle-layer is defined on a large collection of three-level spatial pyramid segments resulted from the morphological method of watersheds. Each node in the middle-layer of the HCRF model is associated with one segment. The value of random variable denotes the label assigned to the segment. For each segment, a list of candidate labels from L according to confidence is assigned based on its local appearance. Segment consistency potentials in the middle-layer are thus used to refine the contours of the labeled objects obtained from the baseline predictions at the bottom-layer. Finally, the global layer is defined on a labeling space. Each node G_i in the global layer of the HCRF model corresponds to one sub-scene appearing in the bottom layer. Global co-occurrence potentials are used to measure the cost that two sub-scenes occur in the same image together. Based on the HCRF model, fine intermediate predictions are obtained, and sub-scenes are thus formed by collecting pixels that are assigned the same label. The spatial contextual relationships of each sub-scene are then modeled by an NBN as shown in Fig. 5. In the NBN model, G_i denotes sub-scene node, and $Z_i = \langle Z_{i0}, Z_{i1}, \ldots, Z_{i8}\rangle$ denotes nine contextual feature nodes of sub-scene G_i. The spatial contextual cues have the potential to disambiguate the identities of objects. The two models are unified through sub-scene nodes to form a hybrid graphical model. Semantic image segmentation is thus solved by considering both non-causal and causal cues in natural images. Consequently, globally consistent labeling results and rational spatial layouts of scenes can be achieved through optimization of the proposed hybrid model.

3.4. Inference for the hybrid model

Note that the pixel, pairwise and segment potentials can be evaluated by lower-level features. By contrast, local contextual interactions of each sub-scene require a high-level description, namely label distribution over an image. However, this information is not available until an initial prediction of an image is achieved. Therefore, the inference process of the hybrid model is divided into two stages, as depicted in Fig. 6. In the first stage, appearance features of a test image I, including texton, multi-scale dense SIFT, color SIFT and LBP, are extracted. The trained joint boosting unary classifier is used to calculate the probability of each label given the local features of each pixel. Meanwhile, image I is over-segmented by Eq. (6). The trained joint boosting segment classifier is used to specify the segment consistency potentials in the HCRF. The energy function in Eq. (1) defined by the HCRF is first minimized by using α-*expansion* move making algorithm based on graph cut [6,7,40] to make an initial prediction $\mathbf{x} = \{x_i\}_{i=1}^{|I|}$ for image I. From the initial solution, sub-scenes $\mathbf{S} = \{S_i\}_{i=1}^{N}$ in an image are then formed by collecting pixels assigned the same label together at the second stage, where N denotes the total number of sub-scenes appearing in the current labeling space. The NBN is then used to model the spatial contextual interactions for each sub-scene S_i. Through imposing spatial contextual constraints on initial prediction, global consistent labeling is finally achieved.

From the local contextual point of view, we are able to interpret each sub-scene S_i as a series of object classes with the costs defined as follows

$$J(S_i = l_k) = \varphi_{S_i} + \varphi_{S_i -> \{S_j | j \neq i\}} \text{ for } k = 1, \ldots, |L|. \tag{13}$$

In Eq. (13), the first term φ_{S_i} denotes the cost that sub-scene S_i is assigned label l_k based on the spatial contextual information, and it can be calculated by Eq. (14).

$$\begin{aligned} \varphi_{S_i} &= -\log p(S_i = l_k | Z_{i0} Z_{i1} Z_{i2} \ldots Z_{i8}) \\ &= -\sum_{m=0}^{8} \log p(S_i = l_k | Z_{im}) \end{aligned}, \tag{14}$$

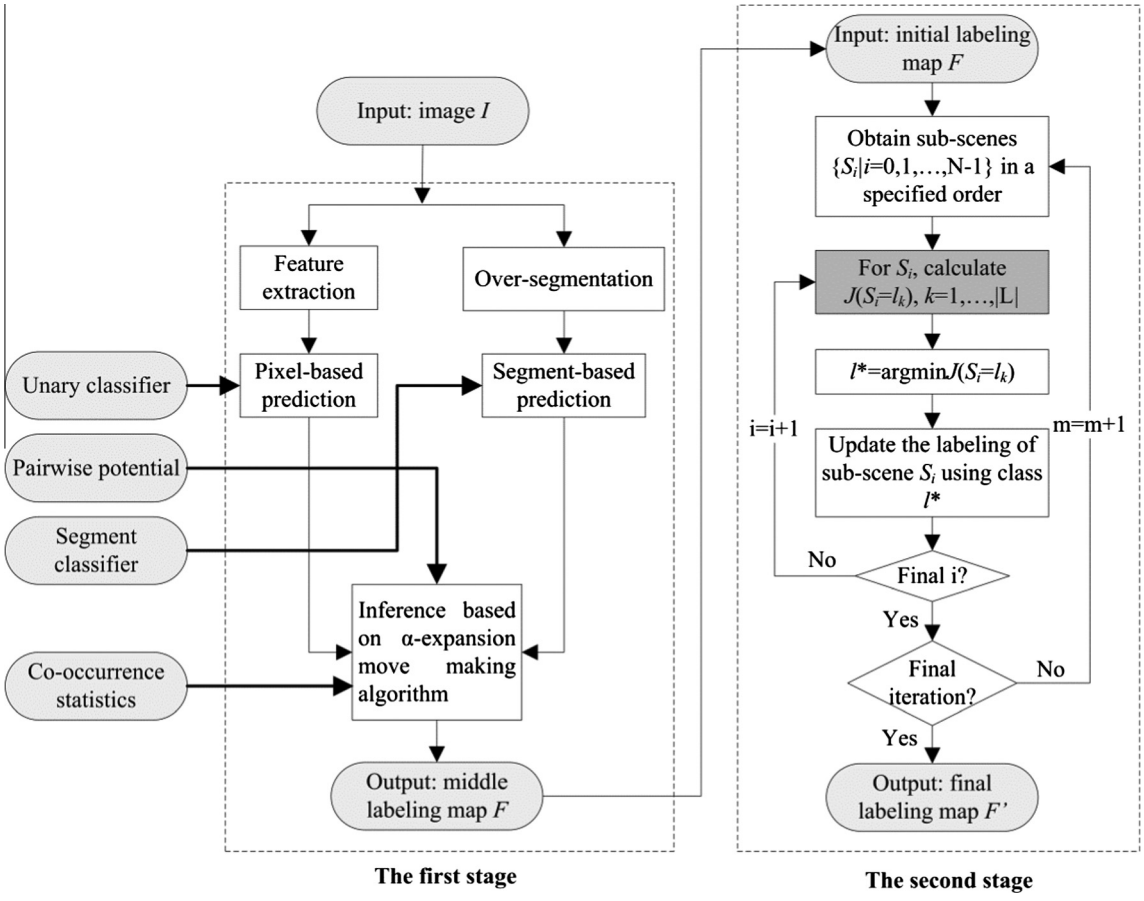

Fig. 6. Flowchart of inference process for the hybrid model.

where $p(S_i|Z_{i0}Z_{i1}Z_{i2}\ldots Z_{i8})$ denotes the posterior probability of sub-scene S_i conditioned on the contextual information from neighboring eight directions and one central area, and it can be calculated by Eq. (7).

The second term $\varphi_{S_i->\{S_j|j\neq i\}}$ in Eq. (13) represents the cost of other sub-scenes $\{S_j|S_j \subset \mathbf{S},\ j\neq i\}$ appearing in the current locations when region belonging to S_i has label l_k, and it takes the form of

$$\varphi_{S_i->\{S_j|j\neq i\}} = -\sum_{S_j\neq S_i, S_j\subset \mathbf{S},} \sum_{r} \log p(S_j = l_0^j | Z_{jr} = l_k), \tag{15}$$

where $\boldsymbol{S}$ denotes the set of sub-scenes in the mth iteration labeling space of image I, l_0^j represents the initial class label of sub-scene S_j in the mth iteration, and r denotes the index as shown in Fig. 2 of neighboring region where label l_k is available when considering sub-scene S_j.

For each sub-scene, $|L|$ costs should be calculated. The problem of inferring the most probable label l^* for sub-scene S_i is equivalent to minimizing the cost function in Eq. (13) as

$$l^* = \arg\min_{l_k\in L} J(S_i = l_k) \tag{16}$$

If l^* does not equal to original label l_0^j for sub-scene S_i, the labeling state of S_i is updated by using l^*. Similar to the inference process of S_i, the other sub-scene can be updated sequentially until all sub-scenes are checked. The inference process is terminated if there is no change to be found or the iteration reaches the maximal value.

During the inference process at the second stage, one point we need to stress is that the sub-scenes in an image should be checked one after another. Since sub-scenes interact with each other from the contextual point of view, the change of one sub-scene influences the contextual descriptor of another sub-scene in the image. In the proposed method, we check the sub-scenes according to the following order.

$$p_{S_1}(l_0^1) < p_{S_2}(l_0^2) < \ldots < p_{S_N}(l_0^N), \tag{17}$$

where $p_{S_i}(l_0^i)$ denotes the confidence that sub-scene S_i is labeled as l_0^i. This means that we first check the sub-scene with the minimal confidence. This confidence is estimated by Eq. (12) as $p_{S_i}(l_0^i) = p(G_i = l_0^i)$.

3.5. Computational complexity analysis

Similar to most supervised learning algorithms in computer vision, the computational bottleneck of the hybrid model is due to training, which includes feature extraction, feature clustering, unsupervised image segmentation, the training of adaptive boosted classifiers for both pixels and segments, and compiling co-occurrence statistics. In this case, feature extraction incurs a computational complexity of O($sptdk$) where s denotes the size of the patch, p denotes the number of training patches in an image, t denotes the number of feature types used to represent one patch, d denotes the dimension of feature vector, and k denotes the number of training images. After features are extracted, clustering is used to simplify feature representation. The computational complexity of feature clustering is O(tds_1c_1) where s_1 denotes the number of feature vectors for clustering, c_1 denotes the number of clusters. To incorporate middle-level semantic information, images are segmented firstly by using unsupervised methods. The

Table 2
Computational complexity of the hybrid model.

Process	Procedure	Complexity
Training	Feature extraction	$O(sptdk)$
	Feature clustering	$O(tds_1c_1)$
	Unsupervised segmentation	$O(7.25nk + kr + k(r - M/4))$
	Adaptive boosted classifier for pixels or segments	O(wfTC)
	Co-occurrence statistics	$O(kC^2)$
Testing	Feature representation and unary potential	$O(k(tdc_1 + f + w_1))$
	Pairwise potential	$O(rp)$
	Unsupervised segmentation and higher-order potential	$O(7.25n + r + r - M/4 + w_2)$
	Co-occurrence costs	$O(C^2)$
	Graph cut	$O(iC(k + p_o + 1.75M + C^2))$
Post-processing	NBN	$O(N)$

complexity of the unsupervised segmentation [48] is at most $O(7.25nk + kr + k(r - M/4))$ when the number (r) of over-segments is larger than M, where n denotes the number of pixels in an image, and M denotes the number of unsupervised segments used in the first segment layer. To train the adaptive boosted classifier for pixels or segments, the computational complexity is $O(wfTC)$ where w denotes the number of weak classifiers, f denotes the dimension of feature vector to represent a training sample, T denotes the number of training samples, and C denotes the number of classes in the dataset. The co-occurrence statistics has the computational complexity of $O(kC^2)$. The computational complexity for training is summarized in Table 2.

After training, the learned model is then used to interpret an image. The computational complexity for testing is also listed in Table 2. To sum up, the computational complexity of an image interpretation is $\mathbf{O}(\sum_{\mathbf{j}=1}^{5}\mathbf{g_j})$, where $g_1 = k(tdc_1 + f + w_1)$, $g_2 = rp$, $g_3 = 7.25n + r + r - M/4 + w_2$, $g_4 = C^2$, $g_5 = iC(k + p_o + 1.75M + C^2)$. g_1 denotes the loops in feature representation and unary potential calculation of an image, g_2 denotes the loops in pairwise potentials calculation, g_3 denotes the loops in unsupervised segmentation and higher-order potentials calculation, g_4 denotes the loops in co-occurrence potentials calculation, and g_5 denotes the loops in energy minimization based on graph cut. Parameter i denotes the number of α-expansion iterations, and p_o denotes the number of pair pixels in 8-connected Pott model.

It should be noted that the NBN model is proposed as a post-processing strategy to disambiguate objects in an image. It performs on sub-scenes $\mathbf{S} = \{S_i\}_{i=1}^{N}$, which is collected from the initial prediction based on the HCRF stage. Its computational complexity is $O(N)$ as shown in Table 2, where N is the number of sub-scenes in a labeled image obtained from the HCRF stage. As depicted in Fig. 7, it can be seen that the number of sub-scenes available in an image is much smaller than the resolution of an image. As a result, it has little overall impact on the complexity of the hybrid model.

4. Experimental results

We have evaluated the performance of the proposed hybrid model on two fully labeled datasets for semantic image segmentation: the Stanford background dataset [35] and the LHI dataset [15,49] in terms of the global and average-per-class recall criteria as defined in [10]. The Stanford background dataset includes 715 outdoor images with the resolution of 320 × 240 pixels. The whole dataset is fully labeled with 8 classes including 7 background classes and one foreground class by hand. Another dataset is provided by Lotus Hill Institute (LHI). For the LHI dataset, 370 of 375 images are public [36] with fully labeled ground truth of 15 classes. In our experiments, three-layer unsupervised segments are generated for each image and used for semantic image segmentation.

4.1. Results for the Stanford 8-class dataset

The 8-labeled classes in the Stanford dataset are *sky*, *tree*, *road*, *grass*, *water*, *building*, *mountain* and *foreground*. In our experiments, the Stanford dataset is randomly split into the training and testing sets with equal size. The experiments were performed on a platform using Intel(R) Core(TM)2 Quad CUP Q9400 @2.66 GHz and RAM of 6.0 GB. For the Stanford background dataset without any code optimization, around 2125 ms is required to label an image by using the hybrid model.

Table 3 lists the quantitative results of the proposed method and other published methods on the same dataset. Note that in the proposed method, we utilized two algorithms, a deep learning (DL) based algorithm and a joint boosting classifier based algorithm as described in Section 2, to produce baseline predictions in the hybrid model. For DL based algorithm, we use the same architecture of CNN as in [23] but with less number of training patches (280,000 patches with size of 46 × 46 pixels) for an immediate observation. Of course, "Deep learning has the property that if you feed it more data, it gets better and better," as noted by Ng [50]. To compensate for the shortcoming of the DL method discussed in Section 1, CRF with smoothness potential, Co-occurrence statistics and BN are incorporated to an enhanced DL method. From the experimental results in Table 3, we can see that the fusion of CRF and BN models can improve the prediction accuracy obtained by using deep learning model only. When compared with other algorithms listed in the table, we can see that the proposed method achieves the best performance of 81.0% and 71.4% for the global and average classification accuracy, respectively. In terms of individual classes, the proposed method performs better in 6 classes (except *sky* and *mountain*) when compared with all other methods. When compared with the pixel-based CRF (PCRF) and HCRF methods [51], the proposed method is obviously superior.

Some of the successful classification results are depicted in Fig. 7 for qualitative evaluation. From columns (c) and (e) in Fig. 7, we can see that the classification results based only on PCRF or DL are noisy and ambiguous. By considering high order pixel interactions, smoothed labeling results are obtained as shown in Fig. 7(d), (f) and (g). By comparing results from the proposed method with the PCRF results, it can be observed that some uncommon spatial layout can be suppressed effectively. For example, in the first row, *building* which unexpectedly appears in the scene has been successfully suppressed by the proposed method. Instead, AHRF (column (f)) wrongly labels a large area of *building* next to the foreground. In the second row of Fig. 7, *road* and *water* in the middle of *foreground* are also removed by the proposed method, while AHRF eliminates the *road* unnecessarily. Furthermore, the benefit of the proposed method is again manifested in the other images in Fig. 7, where *building*, *foreground*, *mountain*, *water* can be clearly identified by the proposed method while AHRF tends to over-simplify the scene and PCRF tends to complicate the scene. Note that *foreground* in the Stanford dataset is a special class which includes more than one object, such as car, aeroplane, boat with or without person, bicycle or motor with person. Therefore, it has no uniform appearances. For this reason, it is often confused with other classes. By using AHRF, for example, *foreground* and *building* are often mistakenly identified as each other, and due to similar inter-class appearances, *foreground* is misclassified as *sky*, *road* and *water*, respectively. On the contrary, the proposed method produces more appropriate results for *foreground*.

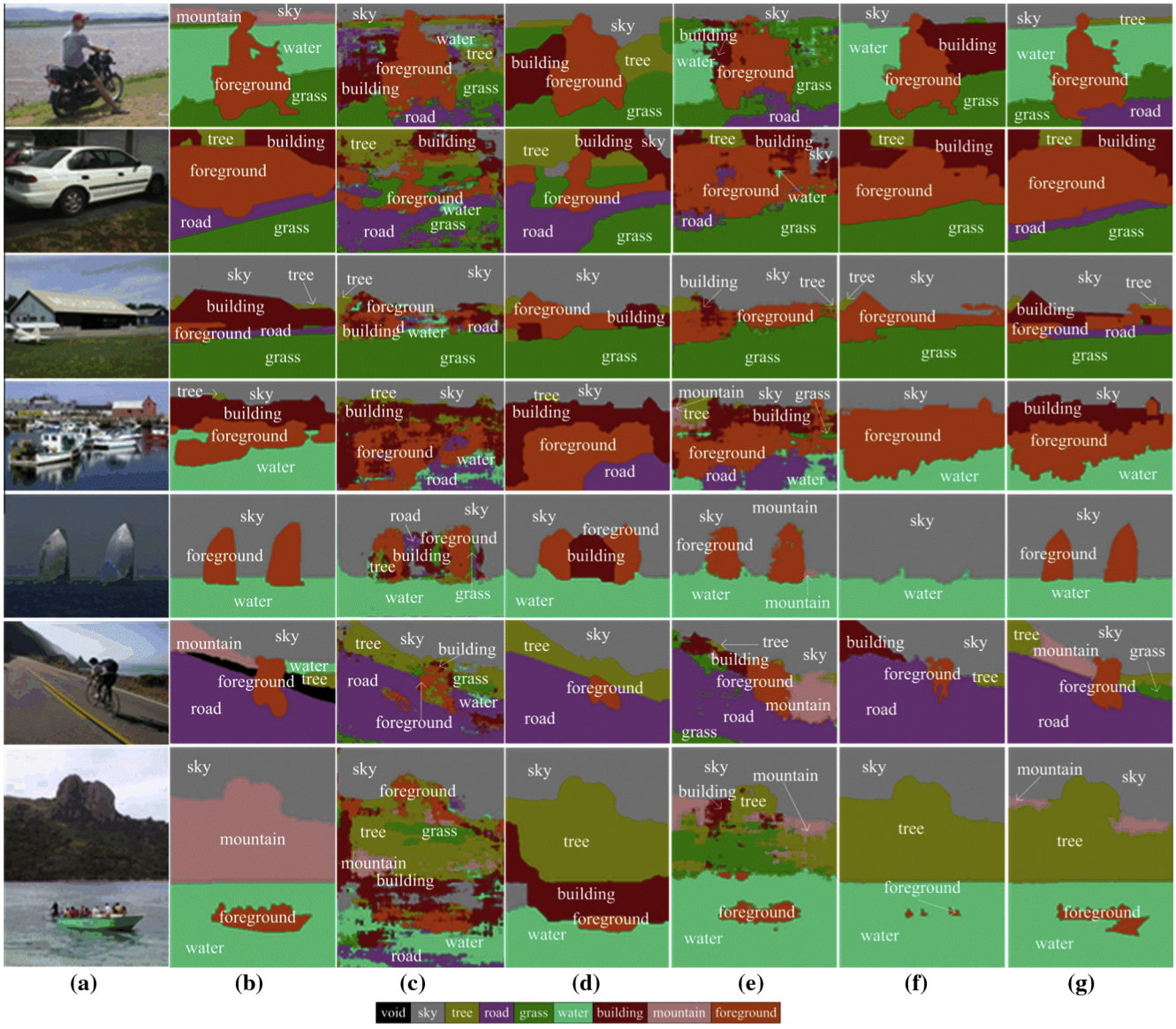

Fig. 7. Some successful cases on the Stanford dataset: (a) original image, (b) ground truth, (c) DL method, (d) enhanced DL method, (e) pixel-based RF, (f) AHRF, and (g) proposed method.

Table 3
Classification accuracy on the Stanford dataset in terms of percentage.

		Global	Average	Sky	Tree	Road	Grass	Water	Building	Mountain	Foreground
DL		73.8	61.5	84.2	75.5	84.1	67.6	45.7	78.0	2.2	54.7
Enhanced DL		76.4	63.1	85.7	77.5	88.1	70.7	45.2	82.6	0.0	54.7
Gould et al. [35]		76.4	65.5	92.6	61.4	89.6	82.4	47.9	82.4	13.8	53.7
Munoz et al. [51]		76.9	66.2	91.6	66.3	86.7	83.0	59.8	78.4	5.0	63.5
Ladicky et al. [7]	PCRF	77.8	68.2	90.3	73.2	87.5	79.5	65.2	76.0	8.4	65.4
	AHRF	79.0	68.2	94.8	72.9	90.6	84.5	64.8	80.3	2.4	55.2
Proposed		81.0	71.4	94.0	74.8	90.7	84.5	67.0	80.9	13.6	66.1

Table 4 lists the confusion matrix of the proposed method on the Stanford dataset. The values in each row represent the percentages of 8 classes which the object in the first column is assigned to. First, *mountain* is investigated due to the lowest classification accuracy (13.6%) as shown in Table 4. From Table 4, we can see *mountain* is frequently misclassified as *tree* with a percentage of 39.8%. The primary reason is that some mountains are often covered with trees as shown in the first row of Fig. 8. In such a case, it is easy to classify *mountain* as *tree*. Besides, mountains in a distance are usually ambiguous as shown in the second row of Fig. 8. As a result, 10.6% of mountains are misclassified as *sky*. Moreover, it is often wrongly identified as *foreground* (12.2%), *building* (9.4%) or *grass* (6.5%) with a relatively high probability due to similar inter-class appearance. In general, the proposed method shows a slightly better recognition rate for this class as shown in the last two rows of Fig. 7 than HCRF. In contrast, *sky* has a high classification accuracy of 94.0%. However, it can be misclassified as neighboring object classes or classes with similar appearance. For example, *sky* may

Table 4
Confusion matrix of 8 classes on the Stanford dataset in terms of percentage.

	Inferred class							
True class	Sky	Tree	Road	Grass	Water	Building	Mountain	Foreground
Sky	94.0	3.1			0.2	2.2	0.1	0.5
Tree	4.5	74.8	0.9	1.8		12.2	1.8	4.1
Road		0.7	90.7	1.0	1.0	1.5	0.1	5.1
Grass	0.4	5.7	5.1	84.5	0.7	1.0	0.6	2.0
Water	6.8	0.3	12.7	5.2	67.0	1.6	0.6	5.8
Building	2.0	6.2	2.6	0.8	0.1	80.9	0.1	7.3
Mountain	10.6	39.8	5.3	6.5	2.7	9.4	13.6	12.2
Foreground	2.1	5.8	8.1	2.5	0.9	14.4	0.1	66.1

be identified as *tree*, *building*, *foreground*, *water* or *mountain*, although the percentage is low. Similar observations can also be made for other classes from Table 2 and Fig. 8. Generally speaking, major misclassifications are ascribed to two aspects. One is being misclassified as one of the adjacent object classes, and the other is mistakenly classified as the class with similar appearance. Both problems are inherent in the HCRF method and the proposed method, because they are both derivatives of the CRF concept. If these problems can be resolved, higher classification accuracy is expected for both methods.

4.2. Results for the LHI 15-class dataset

The LHI dataset used in our experiments was download from [36], which contains 370 images. The dataset is fully labeled with

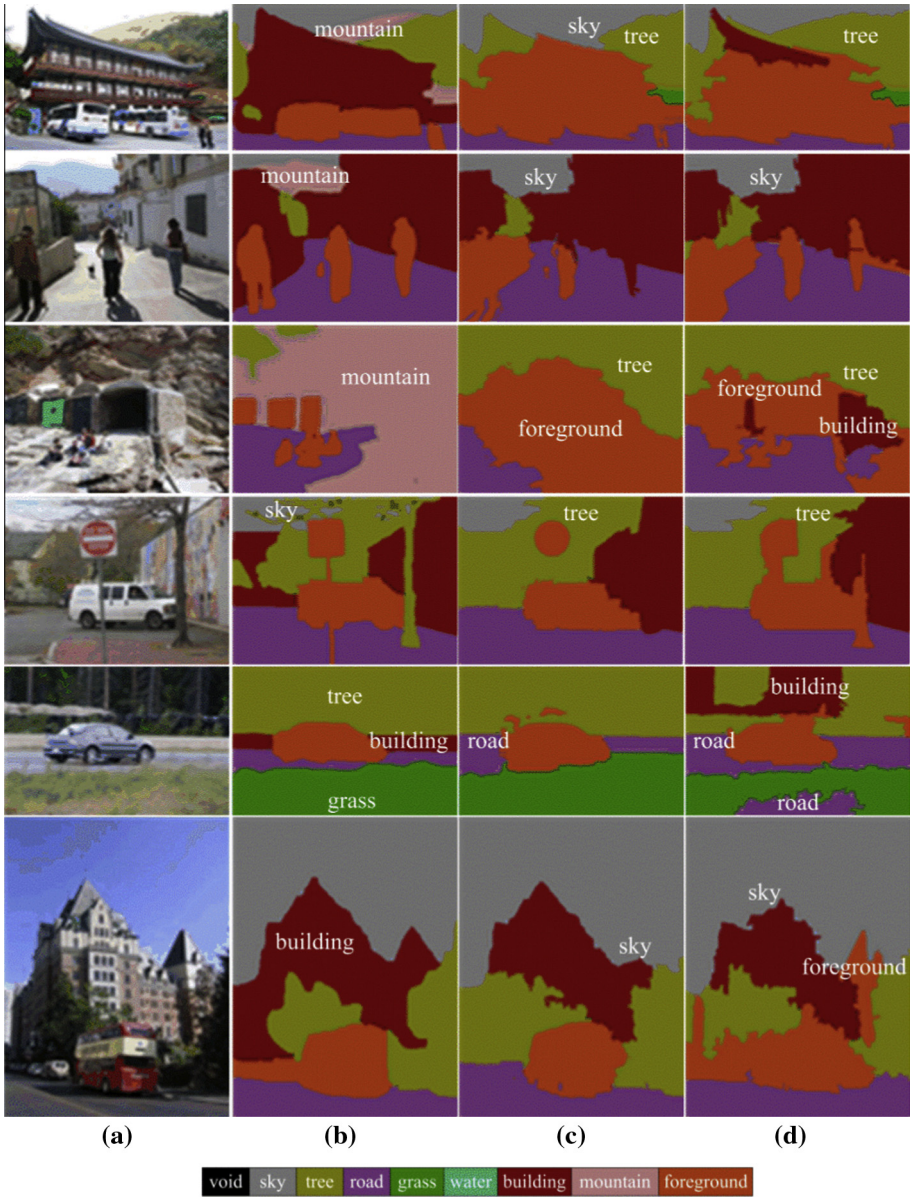

Fig. 8. Some failure cases on the Stanford dataset (a) original image, (b) ground truth, (c) AHRF, and (d) proposed method.

Table 5
Classification accuracy on the LHI dataset in terms of percentage.

		Global	Average	Building	Grass	Tree	Sky	Mountain	Water	Car	Road	Cow	Sheep	Horse	Rhinoceros	Plane	Motorbike	Elephant
DL		75.5	53.3	79.0	87.0	66.4	92.6	27.9	9.0	63.7	78.7	48.7	58.8	37.1	8.0	55.8	64.3	22.5
Enhanced DL		78.0	55.5	81.7	89.4	69.0	95.6	28.4	1.0	65.7	80.5	58.5	68.1	39.4	9.2	62.8	70.7	12.0
Ladicky et al. [7]	PCRF	78.7	59.7	80.0	84.5	74.1	94.2	34.3	11.2	72.9	85.0	56.4	67.5	60.6	13.8	60.0	74.9	25.6
	AHRF	81.6	57.8	82.3	91.0	76.5	98.0	23.1	15.1	58.9	88.6	62.1	70.2	56.7	00.0	44.5	79.0	21.3
Hybrid model		82.2	62.1	84.3	90.3	75.6	98.4	36.0	0.00	75.9	86.9	64.6	72.9	58.7	14.8	76.7	74.3	21.7
Occurrence in training set (total = 185)				84	120	116	99	19	13	62	100	21	19	29	7	18	17	3

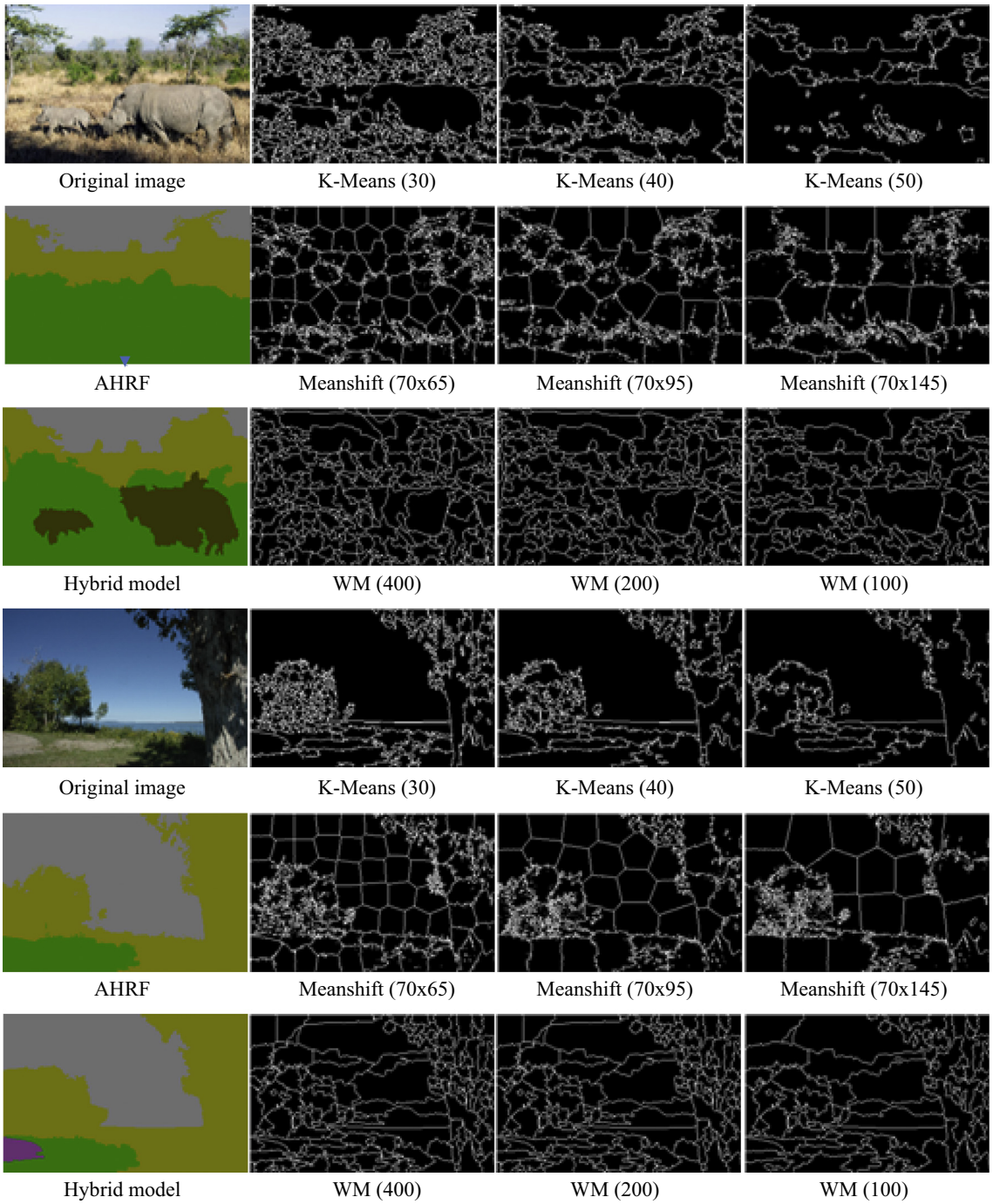

Fig. 9. Impact of unsupervised segmentation: K-Means (a) where a denotes the size of initial cluster; Meanshift ($b \times c$) where b and c denote kernel bandwidths in spatial and range domains, respectively; and WM (d) where d denotes the number of segments required.

15 classes: *building*, *grass*, *tree*, *sky*, *mountain*, *water*, *car*, *road*, *cow*, *sheep*, *horse*, *rhinoceros*, *plane*, *motorbike* and *elephant*. In our experiments, the LHI dataset was randomly split into training and testing sets of equal size.

In Table 5, the classification accuracy of the LHI dataset from DL, enhanced DL, PCRF, AHRF and the proposed method are presented. The last row in the table shows the frequency of each class appearing in the training set. Once again, the experimental results show that enhanced DL performs better than DL in both global and average measurements. The proposed method achieves a global accuracy of 82.2% (1% better than AHRF) and an average accuracy of 62.1% (4% better than AHRF). In terms of individual classes, the proposed method performs better in 8 classes when compared with the two CRF methods. Note that *water* and *rhinoceros* have rather poor classification accuracy (i.e. 11.2% and 13.8%) in the PCRF method. When segment-consistency potential is included, they are not even recognized, i.e. *water* 0% by the proposed algorithm and *rhinoceros* is also 0% by the AHRF. In addition to similar appearances between classes, another reason is that the learning of segment-based classifiers is not sufficient for the two classes due to small number of training examples available as shown in Table 5. Furthermore, the method to produce unsupervised segments also influences the classification results when higher-order potentials are considered. In Fig. 9, unsupervised segmentation results from K-Means, Meanshift and Watershed under different segmentation parameters are shown. From the segmentation results, it can be seen that the K-Means focuses on extracting foreground details. As size of the initial cluster increases, segments from K-Means become coarser. At value of 50, the contours of rhinoceros are not discernible. By contract, Meanshift segments background well but not the foreground objects. As discussed in Section 3.1, semantic segments are strongly influenced by quality of unsupervised segments. If the unsupervised segments are too coarse, the inferred objects are incomplete or incorrect. Obviously, this results in incorrect labeling using six-layer segments in the AHRF method, as depicted in Fig. 9. Instead of just using K-Means and Meanshift, we also used watershed-merging (WM) as described in Section 3.1. WM is able to provide finer segmentation for both background and foreground, and the boundaries of objects are also more accurately segmented. As such, more accurate image labeling can be obtained. Note that some unsupervised segments produced by K-Means and Meanshift include both *sky* and *water* in the second scene of Fig. 9. As a result, *water* is misclassified as *sky* by AHRF. The proposed hybrid model is not able to correctly classify the two sub-scenes too. From the experiments, it is found that it is highly likely to misclassify *water* as *sky* (37.3%), *grass* (29.2%) or *road* (18%). Therefore, it would be necessary to select one or more better performed unsupervised segmentation methods in the high-order potentials. Another approach is to select reliable unsupervised segments instead of using all segments involved in the hybrid model.

More classification results for qualitative evaluation are depicted in Fig. 10. The labeling result based on DL or PCRF is quite rough and noisy as shown in Fig. 10 (c) and (e), since ambiguity exists in an individual pixel's preference for each label based on unary potential. On the other hand, enhanced DL, AHRF and the proposed method resolve the ambiguity quite effectively, as depicted in Fig. 10(d), (f) and (g) respectively. In general, the proposed method achieves more accurate and logical results in spatial layouts of scenes when compared with those achieved by AHRF. For example, the small elephant in between the two larger ones is missed by AHRF while the proposed method segmented it well. For the objects with similar appearance features, such as the

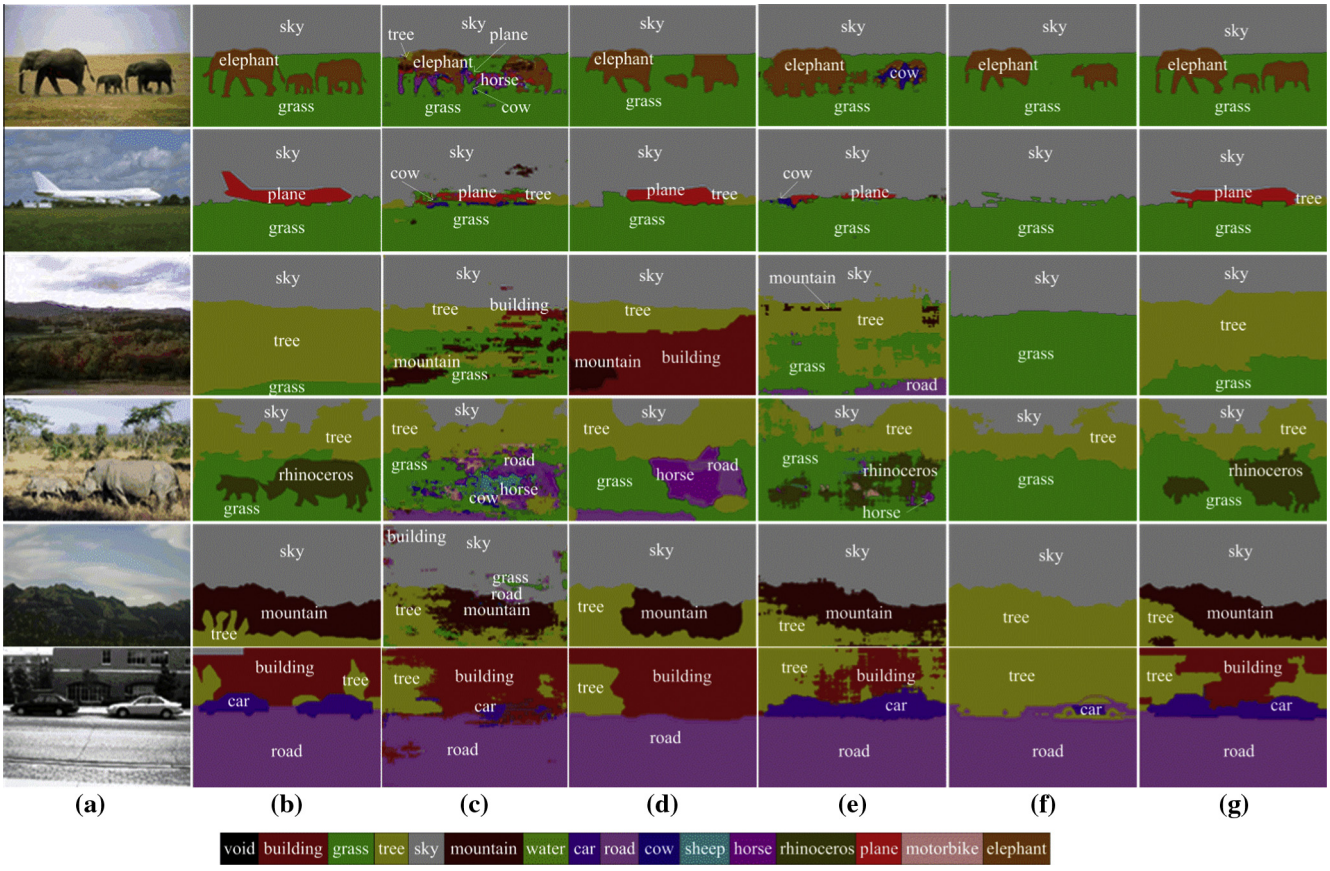

Fig. 10. Some classification cases on LHI dataset: (a) original image, (b) ground truth, (c) DL method, (d) enhanced DL method, (e) pixel-based RF, (f) AHRF, and (g) proposed method.

middle four examples, the proposed method can successfully segment them when compared with PCRF and AHRF. For gray images as shown in Fig. 10, the proposed algorithm also achieves better labeling results.

5. Conclusion

In this paper we have presented a hybrid model combining both HCRF and BN for semantic image segmentation. HCRF captures non-causal relationships among random variables for image labeling. In contrast, local contextual information is incorporated through the BN to exploit causal relationship within natural images. This effectively helps disambiguate the identity of object classes, and provide more accurate and reliable labeling results. The merit of the proposed hybrid model is that it allows us to better fuse a variety of cues based on appearance features and contextual information to jointly perform semantic image segmentation. We have evaluated the proposed model on the Stanford dataset and the LHI dataset, and experimental results show that it improves recognition of objects in scenes. When compared with other published models, the proposed model has also achieved better global and average accuracies. Future research will focus on improving the basic predictions by considering more discriminative features for object classes, and we will also focus on integrating more causal relationships to improve the labeling results produced by CRF models.

Acknowledgments

This research was supported by a Grant from the Research Grant Council of the Hong Kong Special Administrative Region, China, under Project HKU718912E.

References

[1] Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recogn. 40 (2007) 262–282.
[2] C. Rother, V. Kolmogorov, A. Blake, Grabcut: interactive foreground extraction using iterated graph cuts, in: ACM Transactions on Graphics (TOG), 2004, pp. 309–314.
[3] G.B. Huang, M. Narayana, E. Learned-Miller, Towards unconstrained face recognition, in: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, 2008, pp. 1–8.
[4] L. Ladicky, P.H. Torr, A. Zisserman, Human pose estimation using a joint pixel-wise and part-wise formulation, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 3578–3585.
[5] S.-s. Zhu, N.H. Yung, Improve scene categorization via sub-scene recognition, Mach. Vis. Appl. (2014) 1–12.
[6] P. Kohli, P.H. Torr, Robust higher order potentials for enforcing label consistency, Int. J. Comput. Vision 82 (2009) 302–324.
[7] L. Ladicky, C. Russell, P. Kohli, P.H. Torr, Associative hierarchical crfs for object class image segmentation, in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 739–746.
[8] Ľ. Ladický, C. Russell, P. Kohli, P.H. Torr, Inference methods for CRFs with co-occurrence statistics, Int. J. Comput. Vision (2012) 1–13.
[9] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of Machine Learning, 2001, pp. 282–289.
[10] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: Computer Vision–ECCV 2006, 2006, pp. 1–15.
[11] S. Kumar, M. Hebert, A hierarchical field framework for unified context-based classification, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 2005, pp. 1284–1291.
[12] X. He, R.S. Zemel, M.A. Carreira-Perpinán, Multiscale conditional random fields for image labeling, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2, 2004, pp. II-695–II-702.
[13] W. Pieczynski, A.-N. Tebbache, Pairwise Markov random fields and segmentation of textured images, Machine Graphics Vision 9 (2000) 705–718.
[14] S.Z. Li, S. Singh, Markov Random Field Modeling in Image Analysis, vol. 26, Springer, 2009.
[15] Q. Zhou, J. Zhu, W. Liu, Learning Dynamic Hybrid Markov Random Field for Image Labeling, 2013.
[16] D. Larlus, F. Jurie, Combining appearance models and markov random fields for category level object segmentation, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1–7.
[17] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, 2013.
[18] D. Grangier, L. Bottou, R. Collobert, Deep convolutional networks for scene parsing, in: ICML 2009 Deep Learning Workshop, 2009.
[19] H. Schulz, S. Behnke, Learning object-class segmentation with convolutional neural networks, in: 11th European Symposium on Artificial Neural Networks (ESANN), 2012.
[20] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, Y.L. Cun, Learning convolutional feature hierarchies for visual recognition, in: Advances in Neural Information Processing Systems, 2010, pp. 1090–1098.
[21] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1–8.
[22] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[23] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, Pattern Anal. Machine Intell., IEEE Trans. 35 (2013) 1915–1929.
[24] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Cognitive Model. (1988).
[25] C. Galleguillos, S. Belongie, Context based object categorization: a critical survey, Comput. Vis. Image Underst. 114 (2010) 712–722.
[26] S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, M. Hebert, An empirical study of context in object detection, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1271–1278.
[27] S. Gould, J. Rodgers, D. Cohen, G. Elidan, D. Koller, Multi-class segmentation with relative location prior, Int. J. Comput. Vision 80 (2008) 300–316.
[28] B. McFee, C. Galleguillos, G. Lanckriet, Contextual object localization with multiple kernel nearest neighbor, Image Process., IEEE Trans. 20 (2011) 570–585.
[29] C. Galleguillos, A. Rabinovich, S. Belongie, Object categorization using co-occurrence, location and appearance, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1–8.
[30] M. Blaschko, C. Lampert, Object localization with global and local context kernels, 2009.
[31] L. Wolf, S. Bileschi, A critical view of context, Int. J. Comput. Vision 69 (2006) 251–261.
[32] A. Torralba, K.P. Murphy, W.T. Freeman, Contextual models for object detection using boosted random fields, in: Advances in Neural Information Processing Systems, 2004, pp. 1401–1408.
[33] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, p. 14.
[34] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, Pattern Anal. Machine Intell., IEEE Trans. 24 (2002) 603–619.
[35] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 1–8.
[36] http://www.imageparsing.com/LHI_SceneParsing15Classes/index.html.
[37] J. Malik, S. Belongie, T. Leung, J. Shi, Contour and texture analysis for image segmentation, Int. J. Comput. Vision 43 (2001) 7–27.
[38] D.G. Lowe, Object recognition from local scale-invariant features, in: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, 1999, pp. 1150–1157.
[39] T. Ojala, M. Pietikainen, D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, in: Pattern Recognition, 1994. Vol. 1 – Conference A: Computer Vision & Image Processing, Proceedings of the 12th IAPR International Conference on, 1994, pp. 582–585.
[40] Y.Y. Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in ND images, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, 2001, pp. 105–112.
[41] A. Torralba, K.P. Murphy, W.T. Freeman, Sharing features: efficient boosting procedures for multiclass object detection, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2, 2004, pp. II-762–II-769.
[42] B.B. Le Cun, J. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Handwritten digit recognition with a back-propagation network, in: Advances in Neural Information Processing Systems, 1990.
[43] F. Meyer, S. Beucher, Morphological segmentation, J. Vis. Commun. Image Represent. 1 (1990) 21–46.
[44] L. Vincent, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 583–598.
[45] Z. Tan, N.H. Yung, Image segmentation towards natural clusters, in: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008, pp. 1–4.
[46] K. Haris, S.N. Efstratiadis, N. Maglaveras, A.K. Katsaggelos, Hybrid image segmentation using watersheds and fast region merging, Image Process., IEEE Trans. 7 (1998) 1684–1699.

[47] S. Skiadopoulos, M. Koubarakis, Composing cardinal direction relations, Artif. Intell. 152 (2004) 143–171.
[48] A.S.a.P. Anandhakumar, An Improved Fast Watershed Algorithm Based on Finding the Shortest Paths with Breadth First Search, 2012.
[49] B. Yao, X. Yang, S.-C. Zhu, Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks, in Energy Minimization Methods in Computer Vision and Pattern Recognition, 2007, pp. 169–183.
[50] N. Jones, The learning machines, in: Nature Publishing Group Macmillan Building, 4 Crinan St, London N1 9XW, England, 2014.
[51] D. Munoz, J.A. Bagnell, M. Hebert, Stacked hierarchical labeling, in: Computer Vision–ECCV 2010, 2010, pp. 57–70.