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]

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