Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 3846125, 15 pages
http://dx.doi.org/10.1155/2016/3846125

Review Article

Conditional Random Fields for Image Labeling

Tong Liu, Xiutian Huang, and Jianshe Ma

Division of Advanced Manufacturing, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

Correspondence should be addressed to Xiutian Huang; 709794457@qq.com

Received 25 November 2015; Accepted 28 March 2016

Academic Editor: Alessandro Gasparetto

With the rapid development and application of CRFs (Conditional Random Fields) in computer vision, many researchers have made some outstanding progress in this domain because CRFs solve the classical version of the label bias problem with respect to MEMMs (maximum entropy Markov models) and HMMs (hidden Markov models). This paper reviews the research development and status of object recognition with CRFs and especially introduces two main discrete optimization methods for image labeling with CRFs: graph cut and mean field approximation. This paper describes graph cut briefly while it introduces mean field approximation more detailedly which has a substantial speed of inference and is researched popularly in recent years.

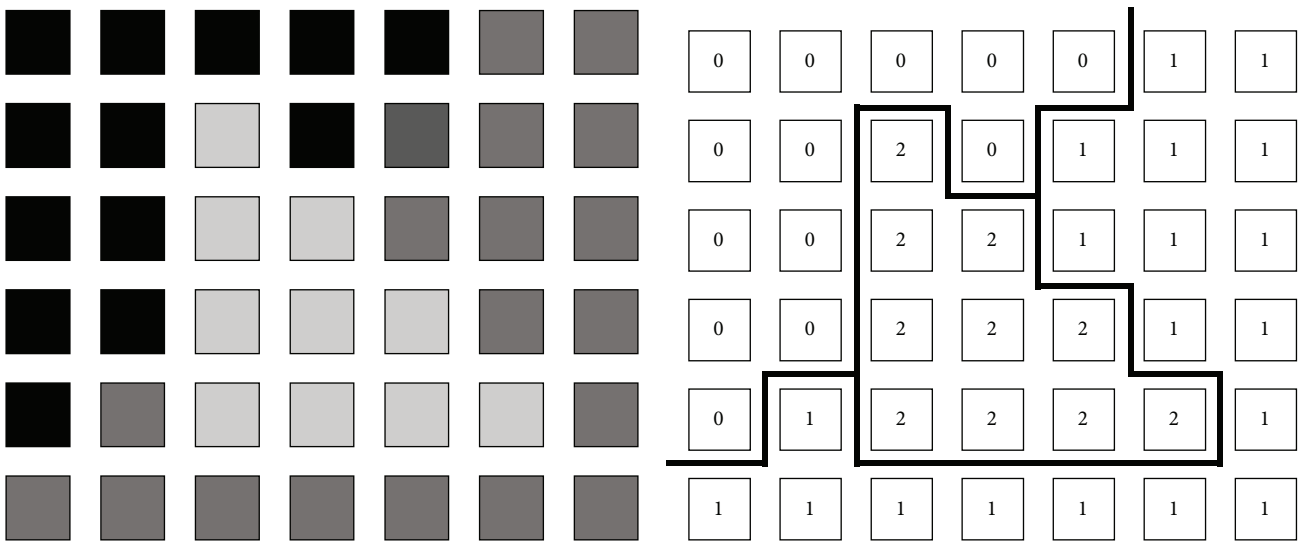
1. Introduction

Recognizing and labeling objects and properties in a given image is an important task in computer vision. The goal of image labeling is to label every pixel or groups of pixels in the image with one of several predetermined semantic object or property categories, for example, "dog," "building," and "car." It is a natural ability for human beings to perform object recognition effortlessly, but it is not straightforward for a computer to do so. Researchers [1–4] are still trying to improve the image labeling technique to reach a better result in terms of speed and accuracy. Figure 1 is an example of label image labeling.

Image labeling usually includes several issues: first we should set up a model and train it; then we will make inference of labeling for a new image. The state-of-the-art of algorithmic solution to image labeling is yet to reach a satisfactory state, especially for the process of inference. Graph cut method [5–8]was popular previously. But the speed of graph cut method is very slow, especially when there are many labels. In [1], Vineet et al. are able to achieve remarkable speed-ups and improvements in accuracy with graph cut base inference techniques comparing with the baseline method in both joint stereo-object labeling and object class segmentation. However, their method [9] has two limitations: the first is the fact that mean field approximation assumes complete factorization over the individual variables; the second limitation relates to the form of the pairwise weights in the formula which are a linear combination of Gaussian kernels. See Section 3.2 for more details of these two limitations.

Naturally, human beings understand a scene mainly by using the spatial and visual information assimilated through their eyes. Inversely, given an image or several images, this information, such as boundary or object, is extremely necessary for scene interpretation. What we hope is to capture the full interaction between pixels. Due to the sensor noise and complexity of the real world, researchers realize that the solution of vision problems can be transformed to some equivalent optimization process as exact interpretation is unapproachable for computers.

In the early history of computer vision, Markov random field (MRF) was popularly used in both low-level and high-level vision perception after it was first introduced into vision by S. Geman and D. Geman in 1984 [10]. The MRF provides a mathematical framework to find optimal solutions by using the contextual visual information in the images. Recently, the MRF model regained attention in the field of computer vision thanks to the progress in powerful energy minimization algorithms [3] such as graph cut [6], belief propagation [11], dual decomposition [12], fusion move [13], and iterated conditional modes. The MRF has been applied to image problems such as restoration, matting [14], segmentation, optical flow, object classification [15, 16], face recognition [17], and text recognition [18].



(a) An image (b) A labeling

Figure 1: An example of image labeling. An image in (a) is a set of pixels P with observed intensities I_p for each $p \in P$. A labeling L shown in (b) assigns some label $L_p \in \{0, 1, 2\}$ to each pixel $p \in P$. Such labels can represent depth (in stereo), object index (in segmentation), original intensity (in image restoration), or other pixel properties. Thick lines in (b) show labeling discontinuities between neighboring pixels [5].

(a) Stereo matching (b) Image segmentation (c) Image restoration

Figure 2: Some examples of labeling problems in computer vision. For stereo matching, the goal is to find the corresponding pixel in one image given a pixel in another image. Its label set is the differences (disparities) between corresponding pixels. For image segmentation, its goal is to partition an image into multiple disjoint regions with region IDs as its label set. For image restoration, it tries to "compensate for" or "undo" defects which degrade an image, and its label set is restored intensities or color.

Object classification can be formulated as a pixel labeling problem; that is, the correct label is to be assigned to each pixel or clique where the label of a pixel represents some property in the real scene, such as the same object or disparity. In [3], Chen et al. introduced the background, basic concepts, and fundamental formulation of image labeling with MRF. They discussed two distinct types of discrete optimization method, that is, belief propagation and graph cut. And they further applied them to the solutions of two classical vison problems: stereo and binary image segmentation using MRF model. Figure 2 shows some examples of labeling problems in computer vision.

It was later recognized that the image labeling problem can be naturally described with a Conditional Random Fields (CRFs) model [1]. The CRF model was first proposed by John Lafferty et al. [19] in 2001. In their work they present iterative parameter estimation algorithms for Conditional Random Fields and compare the performance of the resulting models to HMMs and MEMMs on synthetic and natural-language data. The CRF model is brought to image labeling by Shotton et al., Peng and McCallum, and Kristjansson et al. [20–22].

The use of CRFs was originally restricted in the area of Information Extraction [22–25], in which, given a dataset, the problem is to extract relevant information that belongs to some predefined types. Since the datasets are mostly linguistic, imposing a chain structure on the texts is both effective in capturing temporal relations and efficient in inference and learning for texts is inherently sequential. Therefore, CRFs have been quickly adopted in a wide range of text processing applications, such as part-of-speech tagging (POS), chunking [26, 27], and semantic role labeling [28]. Later on, the application of CRFs has been expanded to word alignment [29], question answering [30], and document summarization [31].

Recently, the research of the CRF model in computer vision has been very popular, as it can be solved by efficient energy minimization algorithms. The efficiency of inference is a critical issue for CRFs in training and predicting the labels on new inputs. After training a CRF model, the marginal

distribution over subsets of labels is computed so as to estimate the parameters of the model. As a result, it can be used to predict the labels of a new input such as a new image using the most likely labels. A lot of inference algorithms have been deployed to solve the CRF optimization problems, such as iterated conditional modes [32], Monte Carlo methods [33], graph cut methods [5–8], and message passing methods, in which mean field inference [1, 34] and belief propagation [35] are the two most popular ways, and people also developed many extensions around the methods.

Local information is well captured by the standard form of a CRF [6, 36]. Since it is not effective for modeling global information as it often fails to capture global consistency in image recognition, researches on how to capture global information of images in CRF with different forms [5–7, 37] become a hot area. To capture both local and global information of images makes the learning and inference very tough; we should not only focus on the accuracy of the method, but also consider the efficiency which turns out to be very poor with the increasing number of the input, such as the dimensions of the feature captured, or the number of input images. Therefore, many methods [38–41] have been proposed to solve such a problem. Recently, a number of cross bilateral Gaussian filter-based methods have been proposed for problems such as object class segmentation [34], denoising [42], and stereo and optical flow [2]; all of these permit substantially faster inference, which maintains or improves accuracy as well. On the basis of [6], Vineet et al. [1] show how higher-order terms can be formulated such that filter-based inference remains possible and demonstrate their techniques on joint stereo and object labeling problems, as well as object class segmentation. In fact, they show that they are able to speed up inference in these model around 10–30 times with respect to competing graph cut methods.

In this paper, we review the progress in the inference of image labeling with CRF models. As mentioned above, a good inference method algorithm is critical in both predicting a new label with a new input and learning the parameters of the model to satisfy the goals of accuracy and efficiency which are two main aspects that we pursue.

Section 2 gives the model of CRFs and their extensions. In Section 3, we mainly introduce two inference methods: graph cut and mean field approximation which are widely used in recent years. And we conclude this paper in Section 4.

2. The Model of CRFs

A CRF is a discriminative undirected probabilistic graphical model that can represent relationships between different variables [20, 43]. The structure of a CRF model helps to estimate the unobserved ones given the observed ones. The classical CRF model is described as follows [34].

Denote by X the input variable and $Y = (y_1, y_2, \ldots, y_N)$ the joint output variable. The input variable X represents our knowledge about the domain such as color and texture. The output Y can be continuous or discrete, but, in most cases, all the labels we set are discrete.

We would like to model the mapping from X to Y via the conditional distribution $P(Y \mid X)$. As a result, we are only interested in the output structure conditioned on the input. CRFs approach the modeling of $P(Y \mid X)$ by representing Y as a Markov random field. More precisely, let $G = (V, E)$ be an undirected graph, where V is the set of nodes in the graph and each node corresponds to a variable y_i, and E is the set of edges. Let $n = |V|$ denote the number of nodes in the graph. Define X as the set of input random variables and $Y = \{y_v\}_{v \in V}$ as the set of output random variables where $V = X \cup Y$ and each y_v ($v \in V$) takes a value from a range of possible discrete labels. In a conditional random field, we assume that each random variable y_v obeys the *Markov property* when conditioned on X, such that the conditional probability distribution of y_v given its adjacent nodes is independent of the rest of the nodes in the graph. That is, if G is such a graphical model that

$$P\left(y_v \mid X, y_w,\ w \neq v\right) = p\left(y_v \mid X, y_w,\ w \in N\left(v\right)\right), \quad (1)$$

where $N(v)$ is the set of adjacent nodes of v, the (Y, X) is conditional random field (CRF). Let $N = \{N_v \mid \forall v \in V\}$ represent the neighbor system to indicate the interrelationship between nodes or the order of CRF. The edges are added between one node P_v and its neighbors N_v. Usually, the neighbor system should satisfy the following:

(1) A site does not neighbor with itself: $i \notin N_i$.

(2) The neighboring relationship is mutual: $i \in N_j \Leftrightarrow j \in N_i$.

The definition of the neighbor system is important because it reflects how far the contextual constraint is. For regular data, as in Figure 3, the neighbors of i are defined as the set of sites within a radius of sqrt(r) from i where r is the order of the neighbor system. One has $N_i = \{i, j \in V \mid [\mathrm{dis}(\mathrm{pixel}_i, \mathrm{pixel}_j)]^2 \leq r,\ i \neq j\}$, where dis($i$, j) measures the Euclidean distance between a and b.

5	4	3	4	5
4	3	2	3	4
3	2	1	2	3
4	3	2	3	4
5	4	3	4	5

Figure 3: An example of 5th-order neighbor system.

In object recognition problems, the observations X are often the image data themselves, or extracted visual features, and Y correspond to the outputs of vision system, for example, possible class labels of the image to be classified, which is shown in Figure 4.

To make the concept clear, we only consider the case when each variable in V takes a value from a range of possible

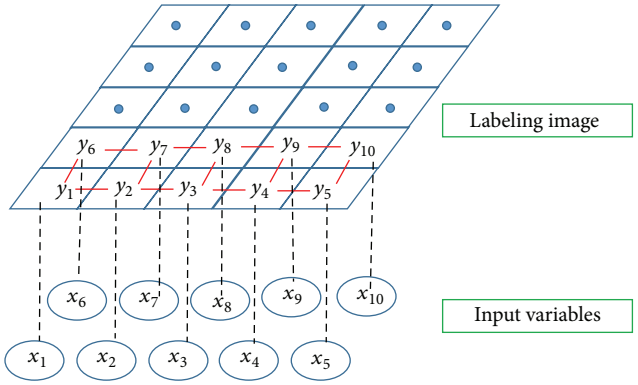


Figure 4: The model of CRF in image labeling. y_i represents the label in ith pixel, and x_i is the features of the corresponding pixel such as color and texture. The red lines in this figure only connect neighboring pixels which means each random variable y_i obeys the *Markov property*.

discrete labels, although they can be either continuous or discrete in a more general case. The paper will describe it in two aspects: probabilistic and energy function.

Under probabilistic understanding, it gets the set of all maximal cliques Λ of G, by using x and y to denote the values assigned to variables X and Y, respectively. The conditional probability distribution of a CRF can be written as

$$P(y \mid x) = \frac{1}{Z(x)} \prod_{A \in \Lambda} \Psi_A (x_A, y_A), \tag{2}$$

where the so-called *potential function* or *compatibility function* Ψ_A is a nonnegative potential function defined over A which is a maximal clique in G. Z is a normalization factor which is also called *partition function* depending on the observed values of input variable x and is defined as

$$Z(x) = \sum_{y} \prod_{A \in \Lambda} \Psi_A (x_A, y_A). \tag{3}$$

We also assume that the conditional distribution over graph G is an *exponential family* [44]; thus we require each potential function Ψ_A to have the form

$$\Psi_A = \exp \left\{ \sum_{k} w_{Ak} f_{Ak} (x_A, y_A) \right\}, \tag{4}$$

where w_A is a real-valued parameter vector and $\{f_{Ak}\}$ is a set of feature functions defined on the potential Ψ_A.

To simplify the solution to the energy function (see (2)), one can take the negative logarithm of the left hand side and right side of (2), and the problem of maximizing the conditional probability becomes an energy minimization problem. In practice, we usually model structures using pairwise constraints, since inference is easier in this case and the model parameters are easy to learn. For example, in computer vision problems, we often see CRFs with maximal cliques of size 2. In this case we can write down the energy as

$$E(y \mid x) = \sum_{i} D(y_i \mid x) + \sum_{i,j} V(y_i, y_j \mid x), \tag{5}$$

where we call D the unary potential and V the pairwise potential. Occasionally we also use high-order cliques and there are special types of high-order clique potentials that are useful in a few applications.

Probabilistic models need to be normalized properly and in many cases require evaluating intractable integrals over the space of all possible variable configurations. While energy functions have no such normalization requirement, thus they provide more flexibility in designing the architecture of the underlying graphical model.

The standard form [1, 25] of a CRF is good for modeling local information. We can write down the form of the standard CRF as follows:

$$P_{\text{std}}(L \mid X) = \frac{1}{Z_{\text{std}}} \exp \left\{ \sum_{i \in S} f_i (l_i \mid X) + \alpha \sum_{i \in S} \sum_{j \in N_i} f_{ij} (l_i, l_j \mid X) \right\}, \tag{6}$$

where X is an input image, $L = \{l_i\}_{i \in S}$ represents labeling, and l_i is a category label at size i. S is a set of sites in the image, N_i is a set of neighbors of i, and α is a coefficient that modulates the effects of the potentials.

In fact, the unary potential f_i represents relations between labels and local image features. It predicts label l_i based on the local features at site i. And the pairwise potential f_{ij} represents relationships between labels of neighboring sites. It means if neighboring sites have similar image features, f_{ij} favors the same category label for them; if not, they might be assigned different category labels. So the pairwise potential f_{ij} works for data-dependent smoothing. What is important is that both potentials represent only local information, as a result, the global information was lost, and some intuitive mistakes can happen; for example, a "dog" might appear in the water [43]. Using the global information, some classification mistakes in image labeling will be avoided which is shown in Figure 5.

Later on, the multiscale CRF [43] (mCRF) was invented to use regional and global label features that encode particular label patterns at local and global scales. The form of mCRF can be presented below by multiplicatively combining component conditional distributions that capture statistical structure at different spatial scale s:

$$P(L \mid X) = \frac{1}{Z} \prod_{s} P_s (L \mid X). \tag{7}$$

Although the mCRF uses regional and global label features, it has massive variables and parameters to be estimated. And it also involves inefficient stochastic sampling for learning and label inference. So the overwhelmingly large dataset size and number of classes are its limitations in practical application.

The boosted random fields [37] model long-range interactions learned by using a boosting algorithm [45]. The hierarchical CRF [23] (hCRF) uses a hierarchical structure of CRFs to model long-range interaction (e.g., relative configurations of objects or regions) and short-range interactions (e.g., pixel-wise label smoothing) in a tractable manner. Its

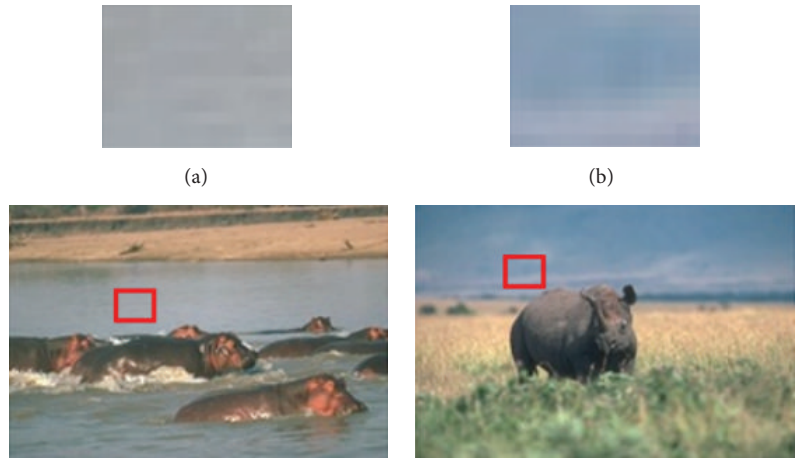


Figure 5: (a, b) Two small image patches that are difficult to label based on local information. (c, d) Images containing the patches. We will usually make mistakes in such classification problems if we use only the local information because the color and texture features in these two patches are too similar. However, as described in (c, d), the global context makes it clear what the patches are ((a, c) water; (b, d) sky) [43].

two-layer formulation to exploit different levels of contextual information in images for robust classification is general enough to be applied to different domains ranging from pixel-wise image labeling to contextual object detection. Both of these two methods do not incorporate global information of the image and thus make the labeling highly dependent on local information.

The random field model proposed by Toyoda and Hasegawa [46] explicitly models local information and global information in conditional random field. The method extracts global image features as well as local ones and uses them to predict the scene of the input image. The form is

$$P(L \mid X) = \frac{1}{Z} \exp \left\{ \sum_{i \in S} f_i (l_i \mid X) + \alpha \sum_{i \in S} \sum_{j \in N_i} f_{ij} (l_i, l_j \mid X) + \beta \sum_{i \in S} g_i (l_i \mid X) + \lambda \sum_{i \in S} \sum_{j \in N_i} g_{ij} (l_i, l_j \mid X) \right\}, \tag{8}$$

where g_i and g_{ij} are global unary potential and global pairwise potential, respectively, α, β, and λ are coefficients that modulate the effects of the potentials, and Z is the partition function for normalization. The global unary potential g_i represents relationships between labels and global image features. It predicts the spatial configuration of labels according to the scene of the input image. The global pairwise potential g_{ij} represents the compatibility of all pairs of labels. This method not only incorporates the local information and global information, but also enables rapid processing by using the global image features. However, it will not do the classification well if there are too many classes (there are only 7 classes in their experiments) because the relationship between classes becomes substantially complex.

Some researchers [47–49] move their research point to the higher-order cliques. In fact, most energy minimization methods for solving computer vision problems assume that the energy can be represented in terms of unary pair-wise clique potentials. As a result, this assumption severely restricts the representational power of these models making them unable to capture the rich statistics of natural scenes [50], while higher-order clique potentials have the capability to model complex interactions of random variables and thus could overcome this problem. The initial work with high-order potentials [36, 50–52] has been quite promising but their use has been limited due to the unavailability of efficient algorithms for minimizing the resulting energy functions. Kohli et al. [49] extend the class of energy functions for which the optimal α-expansion and $\alpha\beta$-swap moves can be computed in polynomial time. In the paper, they propose the P^n Potts model for which the optimal move can be found by solving a st-mincut problem. They define the P^n Potts model potential for cliques of size n as

$$\psi_c (x_c) = \begin{cases} \gamma_k, & \text{if } x_i = l_k, \ \forall i \in c, \\ \gamma_{\max} & \text{otherwise,} \end{cases} \tag{9}$$

where $\gamma_{\max} > \gamma_k, \forall l_k \in L$. For a pairwise clique this reduces to the P^2 Potts model potential defined as $\psi_{ij}(a, b) = \gamma_k$ if $a = b = l_k$ and $\gamma_{\max}$ otherwise. The Gibbs energy of the CRF with high-order cliques is as follows in this paper:

$$E(x) = \sum_i \psi_i (x_i) + \sum_i \sum_{j \in N_i} \psi_{ij} (x_i, x_j) + \sum_{c \in C} \psi_c (x_c), \tag{10}$$

where c is a clique which represents the path $D_c = \{D_i, i \in c\}$ of the frame D and C is the set of all cliques. The example in the paper demonstrates the importance of enforcing label consistency over homogeneous regions for object class segmentation. However, the inference speed is inefficient comparing to mean field inference method.

The P^n Potts model potential is a particular case of the pattern-based potentials [48] which is defined as

$$\psi_c^{\text{pat}}(x_c) = \begin{cases} \gamma_{x_c}, & \text{if } x_c \in P_c, \\ \gamma_{\max} & \text{otherwise,} \end{cases} \tag{11}$$

where $P_c \subset L^{|c|}$ is a set of recognized patterns (i.e., label configurations for clique) each associated with an individual cost γ_{x_c}, while a common cost $\gamma_{\max}$ is applied to all other patterns. If we set P_c to be the L configurations with constant labels, then we will get the P^n Potts model as described.

Cooccurrence relations capture global information about which classes tend to appear together in an image and which do not. And to model object class cooccurrence statistics a new term $K(x)$ is added to the energy:

$$E(x) = \sum \psi_c(x_c) + K(x). \tag{12}$$

Torralba et al. [53] proposed the use of additional unary potentials to capture scene based occurrence priors. Their costs took the form:

$$K(x) = \sum_{i \in V} \phi(x_i). \tag{13}$$

However, the complexity of inference over such potentials scales linearly with the size of the graph; they are prone to overcounting costs and it also requires an initial hard decision of scene type before inference.

Rabinovich et al. [54, 55] proposed cooccurrence as a soft constraint that took the form:

$$K(x) = \sum_{i,j \in V} \phi(x_i, x_j), \tag{14}$$

where ϕ is some potential which penalizes labels that should not occur together in an image. It can capture the global information, however, because it is on the basis of a fully connected graph; the memory requirements of inference scale badly with the size of a fully connected graph. It grows with complexity $O(|V|^2)$ rather than $O(|V|)$ with the size of the graph.

To improve these methods, Ladicky et al. [40] proposed a new form of $K(x)$:

$$K(x) = C(L(x)), \tag{15}$$

where $L(x) = \{l \in L : \exists x_i = l\}$ which guarantees invariance to the size of an object and $C(L(x))$ can be seen as a particular higher-order potential defined over a clique which includes the whole of V, that is, $\psi_V(x)$. And the restriction is placed on $C(L(x))$ that it should be nondecreasing with respect to the inclusion relation; that is, $L_1, L_2 \in L$, and $L_1 \in L_2$ imply that $C(L_1) \leq C(L_2)$. By incorporating these potentials, they got a quantitatively better and visually more coherent labelings. But it carries a comparable higher computer cost comparing to mean field inference.

Similar to Ladicky et al.'s form of $K(x)$, Vineet et al. [47] proposed the form of $C(\Lambda(x))$:

$$C(\Lambda) = \sum_{l \in L} C_l \cdot \Lambda^l + \sum_{l_1, l_2 \in L} C_{l_1, l_2} \cdot \Lambda^{l_1} \cdot \Lambda^{l_2}, \tag{16}$$

where $\Lambda^l = [l \in \Lambda]$, where [] is 1 for a true condition and 0 otherwise. They used filter-based mean field inference to solve the energy with higher-order terms and showed that they are able to spend up inference in relative models about 10–30 times with respect to competing graph cut methods [43].

Joint optimization for object class segmentation is another important area of research in image labeling, such as combining objects and attributes for image segmentation [56], or joint optimization for object class segmentation and dense stereo reconstruction [4]. In [57], Farhadi et al. proposed a method to shift the goal of recognition from naming to description; for example, we not only recognize a basketball as a basketball, but also describe its attributes such as round. Therefore, the method allows them not only to name a familiar object, but also to report unusual aspects of a familiar object and to learn how to recognize new objects with few or no visual examples. The attributes in the paper consist of two aspects: semantic and discriminative. Since the concepts of objects and attributes are both important for describing images precisely, in [57], they formulated the problem of joint visual attribute and object class image segmentation as a dense multilabeling problem, where each pixel in an image should be associated with both an object class and a set of visual attributes labels. In the paper, they proposed a factorial multilabel CRF model which combines the multiclass CRF model and the multilabel model.

The multiclass CRF for objects can be defined in terms of an energy function:

$$E^O(x) = \sum_{i \in V} \psi_i^O(x_i) + \sum_{\{i,j\} \in E} \psi_{ij}^O(x_i, x_j), \tag{17}$$

where ψ_i^O and ψ_{ij}^O are unary potential and pairwise potential functions, respectively, and $E = \{(i, j) \mid i, j \in V,\ i \neq j\}$.

The multilabel CRF for attributes is defined as

$$E^A(y) = \sum_{i \in V} \psi_i^A(y_i) + \sum_{\{i,j\} \in E} \psi_{ij}^A(y_i, y_j), \tag{18}$$

where $y = \{Y_1, Y_2, \ldots, Y_n\}$ are a set of random variables and $A = \{a_1, a_2, \ldots, a_m\}$ are a set of random attribute labels. Rather than taking values directly in A though, the Y_i's take values in the power-set operator.

They also defined a joint CRF in terms of a pairwise energy over the Z_i ($Z_i = (X_i, Y_i)$):

$$E^J(z) = \sum_{i \in V} \psi_i^J(z_i) + \sum_{\{i,j\} \in E} \psi_{ij}^J(z_i, z_j), \tag{19}$$

where

$$\begin{aligned} \psi_i^J(z_i) &= \psi_i^O(x_i) + \psi_i^A(y_i) + \sum_{l,a} \psi_{i,l,a}^{OA}(x_i, y_{i,a}), \\ \psi_{ij}^J(z_i, z_j) &= \psi_{ij}^O(x_i, x_i) + \psi_{ij}^A(y_i, y_i). \end{aligned} \tag{20}$$

Using a two-level hierarchical model, where labeling object classes and attributes is done not only at the pixel level but also at a regional level, they gave the following energy:

$$E^H(z) = \sum_{i \in V_{\text{pix}}} \psi_i^J(z_i) + \sum_{\{i,j\} \in E} \psi_{ij}^J(z_i, z_j) + \sum_{i \in V_{\text{reg}}} \psi_i^{J'}(z_i) + \sum_{\substack{\{i,j\} \in E \\ i \in V_{\text{pix}}, j \in V_{\text{reg}}}} \psi_{ij}^{J'}(z_i, z_j). \quad (21)$$

It was recognized that the problems of dense stereo reconstruction and object class segmentation can both be transformed as one CRF model based labeling problem, in which every pixel in the image is assigned a label corresponding to either its disparity, or an object class. This inspires [4, 46] to provide an energy minimization framework that unifies the two problems. In their paper, the energy function of object class segmentation using a CRF took the form

$$E^O(x) = \sum_{i \in V} \psi_i^O(x_i) + \sum_{i \in V, j \in N_i} \psi_{ij}^O(x_i, x_j) + \sum_{c \in C} \psi_c^O(x_c). \quad (22)$$

And the problem of dense stereo reconstruction using a CRF can be written as

$$E^D(x) = \sum_{i \in V} \psi_i^D(y_i) + \sum_{i \in V, j \in N_i} \psi_{ij}^D(y_i, y_j). \quad (23)$$

Thus the energy of the CRF for joint estimation can be written as

$$E(x) = \sum_{i \in V} \psi_i^J(z_i) + \sum_{i \in V, j \in N_i} \psi_{ij}^J(z_i, z_j) + \sum_{c \in C} \psi_c^O(x_c). \quad (24)$$

Using the fact that certain objects occupy a certain range of real world heights, they jointed unary potentials successfully by

$$\psi_i^C([x_i, y_i]) = -\log(H(h(y_i, i) \mid x_i)), \quad (25)$$

where $h(y_i, i)$ is the corresponding height above the ground plane and $H(h \mid l)$ is a histogram based measure of the naïve probability that a pixel taking label l has height h in the training set. So the combined unary potential can be written as

$$\psi_i^J([x_i, y_i]) = w_O^w \psi_i^O(x_i) + w_D^w \psi_i^D(y_i) + w_C^u \psi_i^C(x_i, y_i), \quad (26)$$

where w_O^u, w_D^u, and w_C^u are the corresponding weights.

For pairwise interactions, we know that an object classes boundary is more likely to occur if the disparity of two neighboring pixels differs significantly. Taking it into account, they chose tractable pairwise potentials of the form

$$\psi_{ij}^J([x_i, y_i], [x_j, y_j]) = w_O^p \psi_{ij}^O(x_i, x_i) + \psi_D^p \psi_{ij}^D(y_i, y_i) + \psi_C^p \psi_{ij}^O(x_i, x_i) \psi_{ij}^D(y_i, y_i), \quad (27)$$

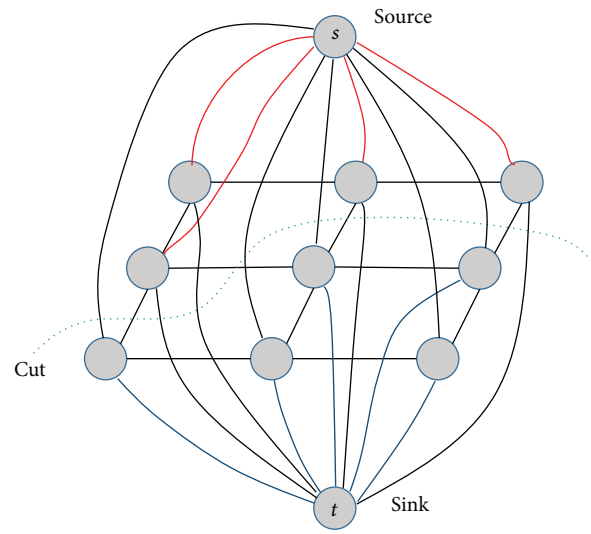


Figure 6: An example of min-cut graph cut. The circles represent the pixels, and the lines including curves represent the edges between nodes including t-links and n-links. The dotted line indicates a cut of graph partition [3].

where w_O^p, ψ_D^p, and ψ_C^p are the weights of the pairwise potential.

Although the two models described as above need more parameters to learn which makes the processes of learning and inference more complicated, they achieved a better scene understanding comparing to other models before.

3. Inference Methods

Over the years, a large number of inference algorithms have been developed; although exact inference in such CRFs is intractable, much attention has been paid to developing fast approximation algorithms, including graph cut approaches [6], variants of belief propagation [11, 35, 50], and a number of Gaussian filter-based methods [1, 39]. In this section, we briefly introduce two inference methods for approximating energy minimums; one is the classical method, graph cut, and the other is mean field approximation which has been popular in recent years.

3.1. Graph Cut. Greig and Porteous [59] first applied the graph cut in computer vision which describes a large family of MRF inference algorithms based on solving min-cui/max-flow problem. If a type of computer vision problems can be formulated in terms of an energy function, then we can use graph cut to get the minimum energy configuration that corresponds to the MAP theory. Figure 6 is an example of min-cut graph cut.

In this method, we set a directed weighted graph $G = (V, E)$ which consists of a set of nodes V and a set of directed edges E and the edge weight is nonnegative. The nodes correspond to pixels in image labeling problem. There are two additional nodes which are called terminals, that is, the source s and the sink t. In computer vision, terminals correspond to the set of labels that can be assigned to pixels. All

(a) Observed image (b) Restored image at convergence

Figure 7: Restoration using the α-expansion algorithm [6].

edges in the graph are assigned some weight or cost. In fact, it is very important to assign edge weights for many graph-based applications in vision. And there are two types of edges in the graph: n-links and t-links. The former connect pairs of neighboring pixels so they can represent a neighborhood system in an image. The latter connect pixels with terminals; thus a t-link connecting a pixel and a terminal corresponds to a penalty for assigning the corresponding label to the pixel. A cut $C \subset E$ is a set of edges such that the terminals are separated in the induced graph $G(C) = (V, E-C)$. In addition, no other proper subset of C separates the terminals in $G(C)$. And the weight of a cut C is the sum of its edge weights, for example, $|C|$. The minimum cut problem is to find the cut with the smallest cost.

Boykov et al. [6] proposed the two most graph cut algorithms: α-expansion and $\alpha\beta$-swap. $\alpha\beta$-swap is described as follows: for a pair of labels α, β, it exchanges the labels between an arbitrary set of pixels labeled α and another arbitrary set labeled β. The algorithm generates a labeling such that there is no swap move that decreases the energy. As for α-expansion: for a label α, this move assigns an arbitrary set of pixels to the label α. This algorithm is ended when there is no expansion move that decreases the energy. In their paper they define two concepts: semimetric and metric. Suppose V is the interaction potentials of the energy, for example, $V_{\{p,q\}}(f_p, f_q)$ with features $f_{(\cdot)}$. V is called a semimetric on the space of labels L if, for any pair of labels $\alpha, \beta \in L$, it satisfies two properties: $V(\alpha, \beta) = V(\beta, \alpha) \geq 0$ and $V(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$. If V also satisfies the triangle inequality $V(\alpha, \beta) \leq V(\alpha, \gamma) + V(\gamma, \beta)$ in L, then V is called a metric. Although α-expansion is more accurate and efficient and can produce a result with lower energy, the interaction potential must be a metric when using α-expansion, while for $\alpha\beta$-swap, it must be semimetric.

The main idea of the α-expansion algorithm is to successively segment all α and non-α pixels with graph cuts and the algorithm will change the value of α at each iteration. The algorithm will iterate through each possible label for α until it converges. At each iteration, the α region P_α can only expand. This changes somehow the way to set the graph weights. Also when two neighboring nodes do not currently have the same label, an intermediate node is inserted and links are weighted so they are relative to the distance of the α label.

The main idea of the $\alpha\beta$-swap algorithm is to successively segment all α pixels from β pixels with graph cuts and the algorithm will change the α-β combination at each iteration. The algorithm will iterate through each possible combination until it converges. Within an iteration the graph is constructed in a normal way so it can segment efficiently between the α region and the β region. Special care must be taken with nodes that are neither in the α nor in the β region. That means, for a pixel, the terminal link weight is the data term plus the sum of all links to neighbors which are neither in the α region nor in the β region.

In [6], the energy formula was described as $E(y) = \sum_{i \in V} D_i(d_i, y_i) + \sum_{i,j \in N} V_{i,j}(y_i, y_j)$. The first term is known as the data term. It ensures that the current labeling y is coherent with the observed data d_i. It penalizes a label y_i to pixel i if it is too different from the observed data d_i. The second term is the smooth term. To make it clear for algorithms used in [6], a quick implementation of the α-expansion algorithm for image restoration is shown in Figure 7. Here an image with embedded squares is of intensity values 255, 191, 128, and 64. Noise was added to the original image so intensity is $i' = i \pm 10$. Possible labels are all integers between 0 and 255. The algorithm will perform α-$\overline{\alpha}$ segmentations until it converges. Note that $\overline{\alpha}$ means non-α labels. The data term used here is a simple squared difference $D(d_i, y_i) = (d_i - y_i)^2$. The smoothing term used here is Potts model $V(y_i, y_j) = \lambda T(y_i \neq y_j)$, where $T(x) = 1$ if x is true, or zero otherwise.

For more details about $\alpha\beta$-swap and α-expansion, one can go to [6]. In addition, Kolmogorov and Rother [60] wrote a survey about graph cut and pointed out that graph cut can be applied to both submodular and nonsubmodular functions. Other more recent developments in graph cut include order-preserving graph cut [61] and combination graph cut [3, 62].

3.2. The Mean Field Approximation. Recently, a number of mean field approximations in computer vision have been proposed, such as object class segmentation [8, 9, 11, 34]. The mean field algorithm finds the distribution Q, which is closest to P which is the exact distribution by minimizing the KL-divergence $D(Q \mid P)$ within the class of distributions representable as a product of independent marginal, $Q(X) = \prod_i Q_i(X_i)$ [63]. Although the approximation of P as a fully factored distribution is likely to lose a lot of information in the distribution, this approximation is computationally attractive. The mean field approximation can be formulated as follows:

$$\begin{aligned} \text{Find} \quad & \{Q_i(X_i)\} \\ \text{maximize} \quad & F \\ \text{subject to} \quad & Q(X) = \prod_i Q_i(X_i) \\ & \sum_{x_i} Q_i(x_i) = 1, \ \forall i, \end{aligned} \tag{28}$$

where F is the energy functional. See [63] for more details.

The approach of [34] of provides a filter-based method for performing fast approximate maximum posterior marginal (MPM) inference; for example, the solution satisfies $x_i^{\text{MPM}} \in \arg\max_i \sum_{\{x \mid x_i = l\}} P(x \mid I)$, in multilabel CRF models with fully connected pairwise terms, where the pairwise terms have the form of a weighted mixture of Gaussian kernels. We can express the fully connected pairwise CRF as

$$\begin{aligned} P(X \mid I) &= \frac{1}{Z(I)} \exp(-E(X \mid I)) \\ E(X \mid I) &= \sum_{i \in N} \psi_u(x_i) + \sum_{i<j \in N} \psi_p(x_i, x_j), \end{aligned} \tag{29}$$

where $E(X \mid I)$ is the energy associated with a configuration X conditioned on I and ψ_u and ψ_p are unary and pairwise potential functions, respectively. And, in [34], the pairwise potentials take the form of a weighted mixture of Gaussian kernels:

$$\psi_p(x_i, x_j) = u(x_i, x_j) \sum_{m=1}^{M} w^{(m)} k^{(m)}\left(\vec{f}_i, \vec{f}_j\right), \tag{30}$$

where u is a label compatibility function, $k^{(m)}(\cdot,\cdot)$, $m = 1 \cdots M$ are Gaussian kernels, and $w^{(m)}(\cdot,\cdot)$, $m = 1 \cdots M$ are the corresponding weight of the kernels. We briefly deduce the whole process of the iterative update equation:

$$\begin{aligned} Q_i(x = l) = \frac{1}{Z} \exp\Bigg\{ & -\psi_u(x_i) \\ & - \sum_{l' \in L} u(l, l') \sum_{m=1}^{M} w^{(m)} \sum_{j \neq i} k^{(m)}\left(\vec{f}_i, \vec{f}_j\right) Q_j(l') \Bigg\}. \end{aligned} \tag{31}$$

First, we can write the KL-divergence $D(Q \parallel P)$:

$$\begin{aligned} D(Q \parallel P) &= \sum_x Q(x) \log\left(\frac{Q(x)}{P(x)}\right) \\ &= -\sum_x Q(x) \log P(x) + \sum_x Q(x) \log Q(x) \\ &= -E_{U \sim Q}[\log P(U)] + E_{U \sim Q}[\log Q(U)] \\ &= -E_{U \sim Q}\left[\log \widetilde{P}(U)\right] + E_{U \sim Q}[\log Z] \\ &\quad + \sum_i E_{U_i \sim Q}[\log Q(U_i)] \\ &= E_{U \sim Q}[E(U)] + \sum_i E_{U_i \sim Q_i}[\log Q_i(U_i)] \\ &\quad + \log Z, \end{aligned} \tag{32}$$

where $E_{U \sim Q}$ refers to the expected value under the distribution Q. Since $Q(X) = \prod_i Q_i(X_i)$ and linearity of expectation $E_{U \sim Q}[\log Q(U)] = \sum_i E_{U_i \sim Q_i}[\log Q_i(U_i)]$, one has

$$\begin{aligned} E_{U \sim Q}[E(U)] &= E_{U \sim Q}\left[\sum_i \psi_u(U_i) + \sum_{i<j} \psi_p(U_i, U_j)\right] \\ &= \sum_i E_{U_i \sim Q_i}[\psi_u(U_i)] \\ &\quad + \sum_{i<j} E_{U_i \sim Q_i, U_j \sim Q_j}\left[\psi_p(U_i, U_j)\right], \end{aligned} \tag{33}$$

where

$$\begin{aligned} \sum_{i<j} E_{U_i \sim Q_i, U_j \sim Q_j}\left[\psi_p(U_i, U_j)\right] &= \frac{1}{2} \\ \cdot \sum_i E_{U_i \sim Q_i}\left[\sum_{j \neq i} E_{U_j \sim Q_j}\left[\psi_p(U_i, U_j)\right]\right] &= \frac{1}{2} \sum_{m=1}^{M} w^{(m)} \\ \cdot \sum_i E_{U_i \sim Q_i}\left[\sum_{j \neq i} k^{(m)}\left(\vec{f}_i, \vec{f}_j\right) E_{U_j \sim Q_j}\left[u(U_i, U_j)\right]\right]. \end{aligned} \tag{34}$$

The marginal $Q_i(x_i)$ which we need is found by minimizing a Lagrangian that consists of all terms in $D(Q \mid P)$ plus Lagrange multipliers assuring that the marginal $Q_i(X_i)$ are probability distributions. The detailed derivations will be presented below:

$$\begin{aligned} L_i(Q) &= E_{U \sim Q}[E(U)] + \sum_i E_{U_i \sim Q_i}[\log Q_i(U_i)] \\ &\quad + \log Z + \lambda\left(\sum_{x_i} Q_i(x_i) - 1\right). \end{aligned} \tag{35}$$

So we can get

$$\begin{aligned}\frac{\partial L_i(Q)}{Q_i(x_i)} &= \frac{\partial}{Q_i(x_i)}\left\{E_{U\sim Q}[E(U)]\right.\\ &+\sum_i E_{U_i\sim Q_i}[\log Q_i(U_i)]+\log Z\\ &\left.+\lambda\left(\sum_{x_i}Q_i(x_i)-1\right)\right\}=\frac{\partial}{Q_i(x_i)}\left\{\sum_i E_{U_i\sim Q_i}\right.\\ &\cdot[\psi_u(U_i)]+\sum_{i<j}E_{U_i\sim Q_i,U_j\sim Q_j}\left[\psi_p\left(U_i,U_j\right)\right]\\ &+\sum_i E_{U_i\sim Q_i}[\log Q_i(U_i)]+\log Z\\ &\left.+\lambda\left(\sum_{x_i}x_i-1\right)\right\}=\psi_u(x_i)+\frac{\partial}{Q_i(x_i)}\left\{\frac{1}{2}\right.\\ &\cdot\sum_i E_{U_i\sim Q_i}\left[\sum_{j\neq i}E_{U_j\sim Q_j}\left[\psi_p\left(U_i,U_j\right)\right]\right]\Bigg\}\\ &+\log Q_i(x_i)+1+\lambda=\psi_u(x_i)+\frac{1}{2}\\ &\cdot\sum_{j\neq i}E_{U_j\sim Q_j}\left[\psi_p\left(x_i,U_j\right)\right]+\log Q_i(x_i)+1+\lambda.\end{aligned} \tag{36}$$

Setting the derivative to 0,

$$\begin{aligned}&\psi_u(x_i)+\frac{1}{2}\sum_{j\neq i}E_{U_j\sim Q_j}\left[\psi_p\left(x_i,U_j\right)\right]+\log Q_i(x_i)+1\\ &\quad+\lambda=0\end{aligned} \tag{37}$$

and rearranging terms, we get that

$$\begin{aligned}&\log Q_i(x_i)=-\psi_u(x_i)-\frac{1}{2}\sum_{j\neq i}E_{U_j\sim Q_j}\left[\psi_p\left(x_i,U_j\right)\right]-1\\ &\quad-\lambda\Longrightarrow\\ &Q_i(x_i)=\exp\left\{-\psi_u(x_i)\right.\\ &\left.-\frac{1}{2}\sum_{j\neq i}E_{U_j\sim Q_j}\left[\psi_p\left(x_i,U_j\right)\right]-1-\lambda\right\}\Longrightarrow\\ &Q_i(x_i)=\frac{1}{Z_i}\exp\left\{-\psi_u(x_i)\right.\\ &\left.-\sum_{j\neq i}E_{U_j\sim Q_j}\left[\psi_p\left(x_i,U_j\right)\right]\right\},\end{aligned} \tag{38}$$

where Z_i is the corresponding partition function.

Substituting the definition of the pairwise potential above into the mean field update in (38) yields the following formulation of the update equation:

$$\begin{aligned}Q_i(x_i)&=\frac{1}{Z_i}\exp\left\{-\psi_u(x_i)\right.\\ &\left.-\sum_{j\neq i}E_{U_j\sim Q_j}\left[\psi_p\left(x_i,U_j\right)\right]\right\}=\frac{1}{Z_i}\exp\left\{-\psi_u(x_i)\right.\\ &\left.-\sum_{m=1}^{M}w^{(m)}\sum_{j\neq i}E_{U_j\sim Q_j}\left[u\left(l,U_j\right)k^{(m)}\left(\overrightarrow{f_i},\overrightarrow{f_j}\right)\right]\right\}\\ &=\frac{1}{Z_i}\exp\left\{-\psi_u(x_i)\right.\\ &\left.-\sum_{m=1}^{M}w^{(m)}\sum_{j\neq i}\sum_{l'\in L}Q_j\left(l'\right)\left[u\left(l,l'\right)k^{(m)}\left(\overrightarrow{f_i},\overrightarrow{f_j}\right)\right]\right\}\\ &=\frac{1}{Z_i}\exp\left\{-\psi_u(x_i)\right.\\ &\left.-\sum_{l'\in L}u\left(l,l'\right)\sum_{m=1}^{M}w^{(m)}\sum_{j\neq i}\left[k^{(m)}\left(\overrightarrow{f_i},\overrightarrow{f_j}\right)\right]Q_j\left(l'\right)\right\}.\end{aligned} \tag{39}$$

In fact, the general form of the mean field update equations (see [52]) is

$$\begin{aligned}&Q_i(v_i=v)\\ &\quad=\frac{1}{Z_i}\exp\left\{-\sum_{c\in C}\sum_{\{v_c|v_i=v\}}Q_{c-i}(v_{c-i})\cdot\psi_c(v_c)\right\},\end{aligned} \tag{40}$$

where v is a value in the domain of random variable v_i, v_c denote an assignment of all variables in clique c, and v_{c-i} is an assignment of all variables in c apart from V_i, and Q_{c-i} denotes the marginal distribution of all variables in c apart from V_i derived from the joint distribution Q. Thus $\sum_{\{v_c|v_i=v\}}Q_{c-i}(v_{c-i})\cdot\psi_c(v_c)$ evaluates the expected value of ψ_c over Q given the condition that V_i takes the value v. When we set $v_i=x_{1\cdots N}$ and $v=1\cdots L$ by evaluating (40) across the unary and pairwise potentials defined in [34], we will directly get (39).

In [34], it is shown that parallel updates for (39) can be evaluated by convolution with a high dimensional Gaussian kernel using any efficient bilateral filter, for example, the permutohedral lattice method of [39]. It is achieved by the following transformation:

$$\begin{aligned}\widetilde{Q}_i^{(m)}&=\sum_{j\neq i}k^{(m)}\left(\overrightarrow{f_i},\overrightarrow{f_j}\right)Q_j(l)\\ &=\left[G_m\otimes Q(l)\right]\left(\overrightarrow{f_i}\right)-Q_i(l),\end{aligned} \tag{41}$$

(a) (b) (c) (d) (e)

FIGURE 8: Results of [1] on Leuven dataset. From (a–e): input image, ground truth, object labeling for [4] (using graph cut + range-moves for inference), object labeling, and stereo outputs from dense CRF with higher-order terms and extended cost-volume filtering [1].

where G_m is a Gaussian kernel corresponding to the mth component of (30), and $\otimes$ is the convolution operator. The following algorithms are the algorithms used in [34].

Algorithm 1 (mean field in fully connected CRFs).

while not converged do

$$\begin{aligned}
\widetilde{Q}_i^{(m)}(l) &\longleftarrow \sum_{j\neq i} k^{(m)}\left(\vec{f}_i, \vec{f}_j\right) Q_j(l) \quad \forall m \\
\widehat{Q}_i(x_i) &\longleftarrow \sum_{l\in L} u^{(m)}(x_i, l) \sum_m w^{(m)} \widetilde{Q}_i^{(m)}(l) \\
Q_i(x_i) &\longleftarrow \exp\left\{-\psi_u(x_i) - \widehat{Q}_i(x_i)\right\}
\end{aligned} \tag{42}$$

end while

In [34], the permutohedral lattice [39] was used for the filter-based inference; the recently proposed domain transform filtering approach [58] has certain advantages over the permutohedral lattice. Since domain transform filtering approach does not subsample the original signal, its complexity is independent of the filter size, while the complexity and filter size are inversely related using the permutohedral lattice. In [47], it was demonstrated that the domain transform approach achieves even faster inference times than using the permutohedral lattice for accurate object/stereo labeling. On the basis of [34, 47] the mean field approximation to the inference of models with higher-order terms was further applied.

In [47] the pattern-based potentials $\psi_c^{\text{pat}}(x_c)$ were added, which is described in Section 2, to the energy function; the required expectation for the mean field updates (39) can be calculated:

$$\begin{aligned}
&\sum_{\{x_c | x_i = l\}} Q_{c-i}(x_{c-i}) \cdot \psi_c^{\text{pat}}(x_c) \\
&= \sum_{p\in P_{c|i=l}} \left(\prod_{j\in c, j\neq i} Q_j\left(x_j = p_j\right)\right) \gamma_p \\
&\quad + \left(1 - \left(\sum_{p\in P_{c|i=l}} \left(\prod_{j\in c, j\neq i} Q_j\left(x_j = p_j\right)\right)\right)\right) \gamma_{\max},
\end{aligned} \tag{43}$$

where $P_{c|i=l}$ is the subset of patterns in P_c for which $x_i = l$.

A particular case of the pattern-based potential is the P^n-Potts model,

$$\psi_c^{\text{potts}}(x_c) = \begin{cases} r_l & \text{if } \forall i \in c,\ x_i = l \\ r_{\max} & \text{otherwise,} \end{cases} \tag{44}$$

and the required expectations can be expressed as

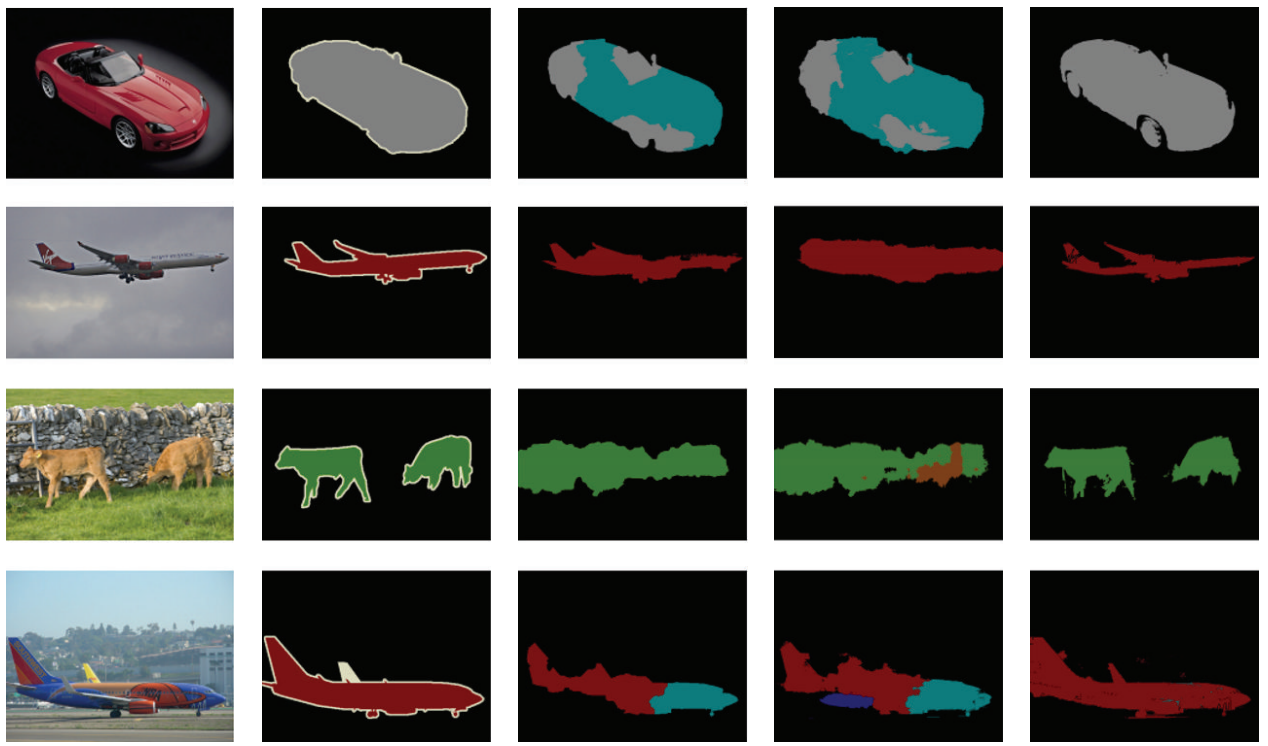
$$\begin{aligned}
&\sum_{\{x_c | x_i = l\}} Q_{c-i}(x_{c-i}) \cdot \psi_c^{\text{potts}}(x_c) \\
&= \left(\prod_{j\in c, j\neq i} Q_j\left(x_j = l\right)\right) r_l \\
&\quad + \left(1 - \left(\prod_{j\in c, j\neq i} Q_j\left(x_j = l\right)\right)\right) \gamma_{\max}.
\end{aligned} \tag{45}$$

The paper [1] also added coconcurrence potentials (see [47] for more details) which is over the entire image clique with a defined form and tested their approach on object class segmentation. As a result, they showed substantial improvements in inference speed with respect to graph cut based methods, particularly by using recent domain transform filtering techniques, while also observing similar or better accuracies. Figures 8 and 9 are the results of [1] in both stereo and image labeling. All the experiments in [1] are based on an Inter® Xeon® 3.33 GHz processor, and they fixed the number of full mean field update iterations to 5 for all models.

In Figure 8, [1] applied their model to the Leuven dataset, consisting of stereo images of street scenes, with ground truth labeling for 7 object classes, and manually annotated ground truth stereo labeling quantized into 100 disparity labels. In their model they used JointBoost classifier responses to form the object unary potentials. A truncated l_2-norm of the intensity differences is used to form the disparity potentials. For the densely connected pairwise terms, identical kernels and weightings and Ising model for the label compatibility function were used. For the P^n-Potts potentials, $\gamma_l = 0$ for all $l = 1, \ldots, L$ was set and $\gamma_{\max}$ was set by cross-validation. Figure 9 is the results of [1] on PascalVOC-10 dataset.

Table 1: Quantitative comparison on Leuven dataset of [1]. The table compares the average time per image and performance (object and stereo labeling accuracy) of joint object and stereo algorithms, using graph cut + range-move (GC + Range (x)), an extension of cost-volume filtering, and [1]'s dense CRF with higher-order terms and filter-based inference (with and without cost-volume filtered unary, and using different approaches). HO means higher-order terms of [1] in the table.

Algorithm	Time (s)	Object (% correct)	Stereo (% correct)
GC + Range (1) [4]	24.6	95.94	76.97
GC + Range (2) [4]	49.9	95.94	77.31
GC + Range (3) [4]	74.4	95.94	77.46
Extended CostVol ([39] filter)	4.2	95.20	77.18
Dense + HO ([39] filter)	3.1	95.24	78.89
Dense_HO ([58] filter)	2.1	95.06	78.21
Dense + HO + CostVol ([58] filter)	6.3	94.98	79.00



(a) (b) (c) (d) (e)

Figure 9: Results of [1] on PascalVOC-10 dataset. From (a–e): input image, ground truth, output from [40] (AHCRF + cooccurrence), output from [34] (dense CRF), and output from dense CRF with Potts and cooccurrence terms [1].

From Figure 8 and Table 1, we note that the densely connected CRF with higher-order terms (Dense + HO) achieves comparable accuracies to [4], and that the use of domain transform filtering methods [58] permits an extra speed-up, with inference being almost 12 times faster than the least accurate setting of [4] and over 35 times faster than the most accurate. The Dense + HO + CostVol approach achieves the best overall stereo accuracies. Although the improved stereo performance appears to generate a small decrease in the object labeling accuracy in [1]'s full model, the former remains at an almost saturated level.

Figure 9 and Table 2 compare timing and performance of [1]'s approach (final 2 lines) against two baseline. The importance of higher-order information is confirmed by the better performance of all algorithms compared to the basic dense CRF of [34]. Further, the filter-based inference is able to improve substantially on the inference time and class-average performance of the AHCRF [40], with P^n-Potts and cooccurrence potentials each giving notable gains.

Table 2: Quantitative results of [1] on PascalVOC-10 dataset.

Algorithm	Time (s)	Overall (% correct)
AHCRF + Cooc [40]	36	81.43
Dense CRF [34]	0.67	71.63
Dense + Potts	4.35	79.87
Dense + Potts + Cooc	4.4	80.44

Although the mean field algorithm is an easy approximation method, it still has several limitations. As mentioned in [9], the first limitation is related to the fact that the mean field approximation assumes complete factorization over the individual variable. As a result, the mean field inference methods are usually sensitive to initialization although the simplified model leads to efficient and tractable models for learning and inference. Another limitation relates to the form of the pairwise weights in (30) which are a linear combination of Gaussian kernels. In fact, they allow each Gaussian component to take only zero mean and use the same combination of Gaussian kernels for each label pair. Although these are improved in [9], they are still lead to unsatisfactory results. Therefore, in the future, we hope to find some other methods which have not only substantial speed of inference but also considerable accuracies.

4. Conclusion

Recently, CRF is accepted as one of the popular approaches for solving the image labeling problem in computer vision and image analysis. An important issue in CRF models is to develop an efficient inference algorithm to find the most appropriate labels especially when considering the global information of an image.

In this paper we review the research development and status of object recognition with CRFs, especially the two main discrete optimization methods for image labeling with CRFs: graph cut and mean field approximation. We describe graph cut briefly while we introduce mean field approximation more detailedly which has a substantial speed of inference and is popular in recent years. Compared to the graph cut method, the mean field inference improves speed substantially for its simplified model.

In the application of image labeling problem in computer vision, one typical problem is that there are too many nodes. For example, for an image with the size of $i \times j$, supposing each node takes L possible labels, the computation space is $L^{i \times j}$. Thus the computation space expands exponentially with the growth of image's size. It is very clear that the inference algorithm plays a very important role in these problems. Another key issue is to construct reasonable CRF models as Section 2 introduces. Learning the parameters of a CRF model efficiently from images instead of being manually or empirically chosen is also an important issue, though it is not the focus of this paper.

Nowadays, many tasks in computer vision and image analysis can be formulated as a labeling problem where the correct label has to be assigned to each pixel or clique. However, computational expense of training is still a computational burden for the need to perform inference repeatedly during training process. In the future, we hope to improve the accuracy of mean field inference for image labeling while maintaining its efficiency. Solving these problems will greatly influence some technology such as driverless car. On the other hand, with the development of the skills for capturing image depth information such as Kinect, depth information of an image is easily obtained like color features. So it is considerable to combine these properties with CRF models and efficient inference approaches for image labeling and stereo reconstruction in 3-dimensional space. Moreover, using these theories for facial action labeling research may be another strategy.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the Special Fund Project of "Industry-Education-Academy" Cooperation in Guangdong Province in 2013 (2013A090100002), the National High Technology Research and Development Program ("863" Program) of China (2015AA043302), and the Key Scientific Projects of Guangzhou Huadu in 2014 (HD14ZD004).

References

[1] V. Vineet, J. Warrell, and P. H. S. Torr, "Filter-based mean-field inference for random fields with higher-order terms and product label-spaces," *International Journal of Computer Vision*, vol. 110, no. 3, pp. 290–307, 2014.

[2] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 504–511, 2013.

[3] 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.

[4] L. Ladický, P. Sturgess, C. Russell et al., "Joint optimization for object class segmentation and dense stereo reconstruction," *International Journal of Computer Vision*, vol. 100, no. 2, pp. 122–133, 2012.

[5] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, vol. 2134 of *Lecture Notes in Computer Science*, pp. 359–374, Springer, Berlin, Germany, 2001.

[6] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[7] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[8] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, 2004.

[9] V. Vineet, J. Warrell, P. Sturgess, and P. H. S. Torr, "Improved initialisation and Gaussian mixture pairwise terms for dense random fields with mean-field inference," in *Proceedings of the 23rd British Machine Vision Conference (BMVC '12)*, Surrey, UK, September 2012.

[10] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[11] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, 2006.

[12] N. Komodakis, N. Paragios, and G. Tziritas, "MRF energy minimization and beyond via dual decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 531–552, 2011.

[13] V. Lempitsky, C. Rother, S. Roth, and A. Blake, "Fusion moves for markov random field optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1392–1405, 2010.

[14] S.-Y. Lin and J.-Y. Shi, "A Markov random field model-based approach to natural image matting," *Journal of Computer Science and Technology*, vol. 22, no. 1, pp. 161–167, 2007.

[15] J. Verbeek and B. Triggs, "Region classification with Markov field aspect models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[16] D. Larlus and F. Jurie, "Combining appearance models and Markov random fields for category level object segmentation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–7, Anchorage, Alaska, USA, June 2008.

[17] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[18] S. España-Boquera, M. J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martinez, "Improving offline handwritten text recognition with hybrid HMM/ANN models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 767–779, 2011.

[19] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random field: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289, 2001.

[20] J. Shotton, J. Winn, and C. Rother, "Joint appearance, shape and context modeling for multi-class object recognition and segmentation[M]//Leonardis," *Lecture Notes in Computer Science*, p. 15, 2006.

[21] F. C. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields," in *HLT-NAACL 2004: Main Proceedings*, pp. 329–336, Association for Computational Linguistics, Boston, Mass, USA, 2004.

[22] T. Kristjansson, A. Culotta, P. Viola, and A. McCallum, "Interactive information extraction with constrained conditional random fields," in *Proceedings of the 19th National Conference on Artifical Intelligence (AAAI '04)*, pp. 412–418, San Jose, Calif, USA, July 2004.

[23] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2003.

[24] F. C. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields," *Information Processing & Management*, vol. 24, no. 4, pp. 963–979, 2006.

[25] F. Peng, F. Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," in *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, pp. 562–568, Geneva, Switzerland, 2004.

[26] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 1, pp. 134–141, Edmonton, Canada, May 2003.

[27] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data," *Journal of Machine Learning Research*, vol. 8, pp. 693–723, 2007.

[28] T. Cohn and P. Blunsom, "Semantic role labeling with tree conditional random fields," in *Proceeding of the 9th Conference on Natural Language Learning (CoNLL '09)*, pp. 169–172, 2005.

[29] P. Blunsom and T. Cohn, "Discriminative word alignment with conditional random fields," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 65–72, Sydney, Australia, July 2006.

[30] A. Hickl and S. Harabagiu, "Enhanced interactive question-answering with conditional random fields," in *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL (IQA '06)*, pp. 25–32, Association for Computational Linguistics, 2006.

[31] D. Shen, J. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07)*, pp. 2868–2873, Hyderabad, India, January 2007.

[32] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society—Series B: Methodological*, vol. 48, no. 3, pp. 259–302, 1986.

[33] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA, 2006.

[34] P. Krahenhuhl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," *Advances in Neural Information Processing Systems*, vol. 24, pp. 109–117, 2011.

[35] N. Piatkowski and K. Morik, "Parallel loopy belief propagation in conditional random fields," in *Proceedings of the KDML Workshop of the LWA*, Magdeburg, Germany, 2011.

[36] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, no. 2, pp. 179–201, 2006.

[37] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Advances in Neural Information Processing Systems*, vol. 17, pp. 1401–1408, 2005.

[38] A. Adams, N. Gelfand, J. Dolson, and M. Levoy, "Gaussian KD-trees for fast high-dimensional filtering," *ACM Transactions on Graphics*, vol. 28, no. 3, article 21, 2009.

[39] A. Adams, J. Baek, and M. A. Davis, "Fast high-dimensional filtering using the permutohedral lattice," *Computer Graphics Forum*, vol. 29, no. 2, pp. 753–762, 2010.

[40] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Graph cut based inference with co-occurrence statistics," in *Computer Vision—ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., Lecture Notes in Computer Science, pp. 239–253, Springer, Berlin, Germany, 2010.

[41] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854–869, 2007.

[42] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: theory and applications," *Foundations and Trends in Computer Graphics and Vision*, vol. 4, no. 1, pp. 1–73, 2009.

[43] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, pp. II695–II702, Washington, Wash, USA, July 2004.

[44] E. B. Andersen, "Sufficiency and exponential families for discrete sample spaces," *Journal of the American Statistical Association*, vol. 65, no. 331, pp. 1248–1255, 1970.

[45] S. Abney, R. E. Schapire, and Y. Singer, "Boosting applied to tagging and PP attachment," in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 38–45, College Park, Md, USA, 1999.

[46] T. Toyoda and O. Hasegawa, "Random field model for integration of local information and global information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1483–1489, 2008.

[47] V. Vineet, J. Warrell, and P. H. Torr, "Filter-based mean-field inference for random fields with higher-order terms and product label-spaces," *International Journal of Computer Vision*, vol. 110, no. 3, pp. 290–307, 2014.

[48] N. Komodakis and N. Paragios, "Beyond pairwise energies: efficient optimization for higher-order mrfs," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 2985–2992, Miami, Fla, USA, June 2009.

[49] P. Kohli, M. P. Kumar, and P. H. S. Torr, "P^3 & beyond: solving energies with higher order cliques," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[50] X. Y. Lan, S. Roth, D. Huttenlocher, and M. J. Black, "Efficient belief propagation with learned higher-order Markov random fields," in *Computer Vision—ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., Lecture Notes in Computer Science, pp. 269–282, Springer, Berlin, Germany, 2006.

[51] S. Roth and M. J. Black, "Fields of experts: a framework for learning image priors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, C. Schmid, S. Soatto, and C. Tomasi, Eds., vol. 2, pp. 860–867, San Diego, Calif, USA, June 2005.

[52] S. Roth and M. J. Black, "Fields of experts: a framework for learning image priors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 860–867, San Diego, Calif, USA, June 2005.

[53] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, vol. 1, pp. 273–280, IEEE, Nice, France, 2003.

[54] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, Rio de Janeiro, Brazil, October 2007.

[55] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[56] S. Zheng, M.-M. Cheng, J. Warrell et al., "Dense semantic image segmentation with objects and attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 3214–3221, IEEE, Columbus, Ohio, USA, June 2014.

[57] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1778–1785, Miami, Fla, USA, June 2009.

[58] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM Transactions on Graphics*, vol. 30, no. 4, article 69, 2011.

[59] M. D. Greig and T. B. Porteous, "Exact maximum a-posteriori estimation for binary images," *Journal of the Royal Statistical Society Series B-Methodological*, vol. 51, no. 2, pp. 271–279, 1989.

[60] V. Kolmogorov and C. Rother, "Minimizing nonsubmodular functions with graph cuts-a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1274–1279, 2007.

[61] X. Liu, O. Veksler, and J. Samarabandu, "Order-preserving moves for graph-cut-based optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1182–1196, 2010.

[62] N. Komodakis and G. Tziritas, "Approximate labeling via graph cuts based on linear programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1436–1453, 2007.

[63] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.