Randomly-connected Non-Local Conditional Random Fields

by

Mohammad Javad Shafiee

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Prof. Greg Mori School of Computing Science, Simon Fraser University
Supervisor	Prof. Alexander Wong Systems Design Engineering, University of Waterloo
Supervisor	Prof. Paul Fieguth Systems Design Engineering, University of Waterloo
Internal Member	Prof. Stacy Scott Systems Design Engineering, University of Waterloo
Internal-external Member	Prof. Dana Kulic Electrical & Computer Engineering, University of Waterloo
Internal-external Member	Prof. Zhou Wang Electrical & Computer Engineering, University of Waterloo

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Content from 9 papers are used in this thesis. I was the co-author with major contributions on designing the methods, implementation and writing the papers:

M. J. Shafiee, A. Wong, P. Siva, and P. Fieguth, "Efficient Bayesian Inference Using Fully Connected Conditional Random Fields with Stochastic Cliques", IEEE International Conference on Image Processing (ICIP), 2014.
This paper is incorporated in Chapter 3 of this thesis.

M. J. Shafiee, A. G. Chung, A. Wong, and P. Fieguth, "Improved fine structure modeling via guided stochastic clique formation in fully connected conditional random fields", IEEE International Conference on Image Processing (ICIP), 2015.
This paper is incorporated in Chapter 4 of this thesis.

M. J. Shafiee, A. Wong, and P. Fieguth, "Deep Randomly-connected Conditional Random Fields For Image Segmentation", IEEE Access Journal, 2016.
This paper is incorporated in Chapter 4 and 5 of this thesis.

M. J. Shafiee, A. Wong, and P. Fieguth, "Forming A Random Field via Stochastic Cliques: From Random Graphs to Fully Connected Random Fields", arXiv:1506.09110, 2015.
This paper is incorporated in Chapter 6.

M. J. Shafiee, P. Siva, and A. Wong, "Stochasticnet: Forming deep neural networks via stochastic connectivity", IEEE Access Journal, 2016.
This paper is incorporated in Chapter 6.

M. J. Shafiee, P. Siva, P. Fieguth, and A. Wong, "Efficient Deep Feature Learning and Extraction via StochasticNets", IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), 2016.
This paper is incorporated in Chapter 6.

M. J. Shafiee, P. Siva, C. Scharfenberger, P. Fieguth, and A. Wong, "NeRD: A Neural Response Divergence Approach to Visual Saliency Detection", IEEE Signal Processing Letters (SPL), 2016.
This paper is incorporated in Chapter 6.

M. J. Shafiee, P. Siva, P. Fieguth, and A. Wong, "Embedded Motion Detection via Neural Response Mixture Background Modeling", IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), 2016.
This paper is incorporated in Chapter 6.

M. J. Shafiee, and A. Wong, "Evolutionary Synthesis of Deep Neural Networks via Synaptic Cluster-driven Genetic Encoding", Neural Information Processing Systems Workshops (NIPS), 2016. **Best paper award.**
This paper is incorporated in Chapter 6.

Abstract

Structural data modeling is an important field of research. Structural data are the combination of latent variables being related to each other. The incorporation of these relations in modeling and taking advantage of those to have a robust estimation is an open field of research. There are several approaches that involve these relations such as Markov chain models or random field frameworks. Random fields specify the relations among random variables in the context of probability distributions. Markov random fields are generative models used to represent the prior distribution among random variables. On the other hand, conditional random fields (CRFs) are known as discriminative models computing the posterior probability of random variables given observations directly.

CRFs are one of the most powerful frameworks in image modeling. However practical CRFs typically have edges only between nearby nodes. Utilizing more interactions and expressive relations among nodes make these methods impractical for large-scale applications, due to the high computational complexity. Nevertheless, studies have demonstrated that obtaining long-range interactions in the modeling improves the modeling accuracy and addresses the short-boundary bias problem to some extent. Recent work has shown that fully connected CRFs can be tractable by defining specific potential functions. Although the proposed frameworks present algorithms to efficiently manage the fully connected interactions/relatively dense random fields, there exists the unanswered question that fully connected interactions are usually useful in modeling. To the best of our knowledge, no research has been conducted to answer this question and the focus of research was to introduce a tractable approach to utilize all connectivity interactions.

This research aims to analyze this question and attempts to provide an answer. It demonstrates that how long-range of connections might be useful. Motivated by the answer of this question, a novel framework to tackle the computational complexity of a fully connected random fields without requiring specific potential functions is proposed. Inspired by random graph theory and sampling methods, this thesis introduces a new clique structure called stochastic cliques. The stochastic cliques specify the range of effective connections dynamically which converts a conditional random field (CRF) to a randomly-

connected CRF. The randomly-connected CRF (RCRF) is a marriage between random graphs and random fields, benefiting from the advantages of fully connected graphs while maintaining computational tractability. To address the limitations of RCRF, the proposed stochastic clique structure is utilized in a deep structural approach (deep structure randomly-connected conditional random field (DRCRF)) where various range of connectivities are obtained in a hierarchical framework to maintain the computational complexity while utilizing long-range interactions.

In this thesis the concept of randomly-connected non-local conditional random fields is explored to address the smoothness issues of local random fields. To demonstrate the effectiveness of the proposed approaches, they are compared with state-of-the-art methods on interactive image segmentation problem. A comprehensive analysis is done via different datasets with noiseless and noisy situations. The results shows that the proposed method can compete with state-of-the-art algorithms on the interactive image segmentation problem.

Acknowledgments

I would like to express my gratitude to my two co-supervisors, professors Alexander Wong and Paul Fieguth for their tremendous supports and guidance on my development as a scientist and a researcher. Prof. Wong thank you for involving me in different research and industrial projects while guiding me in the right direction, encouraging me to explore different ideas and helping me shape them properly; most important of all, thanks for giving me the insight that research is fun. Prof. Fieguth thank you for your meticulous comments on how to develop the ideas and the elucidation of several different topics. I learnt how to formulate my problems mathematically which helped me provide strong theory to back my experiments. Thank you both, you taught me there is no fear in research, just need to have faith, dive in and be persistent.

I would sincerely like to thank my Ph.D. committee members Prof. Stacy Scott from Systems Design Engineering department, Prof. Dana Kulic and Prof. Zhou Wang from Electrical and Computer Engineering department for their time and commitment. I also would like to thank Prof. Greg Mori from School of Computing Science, Simon Fraser University for his time to accept reviewing the thesis.

Many thanks to my colleagues in Vision and image Processing Research Group at University of Waterloo, for always being helpful and bring the productivity atmosphere into the lab.

Special thanks to my friends Hamed (Dr. Shahsavan), Mohammad (Dr. Mohammadi) and Amir-Hossien (to be Dr. Karimi in future years) for cheering me up during the lows and highs, no matter I was down in the dumps or jumped for a joy (the Ph.D. life situations)!

Finally but the most importantly, I would like to thank my family for their huge supports and encouragement. To my beloved parents for always encouraging me to pursue the academic pathway and their non-stop supports, to my sister and brothers for their enthusiasm in my success and their encouragement in defeat.

Dedication

To my beloved parents,
My dear sister and brothers,
Those who their love and support never get the stochastic manner in my life!

Table of Contents

List of Tables

List of Figures

Nomenclature

Randomly-connected random Field

γ	Sparsity factor
σ_p	Controling paramter
σ_q	Controling paramter
L	Number of layers
$P^s_{i,j}$	Spatial probability
$Q^d_{i,j}$	Color similarity probability

Random Field

$\hat{Y}^t$	Sub-optimal result at layer t
$\mathbb{E}(\cdot)$	Expectation
$\mathcal{C}$	Clique structures
$\mathcal{S}^t$	Graph cut at layer t
ω	Weigth of feature fucntion
$\psi(\cdot)$	Potential function

$\psi_p(\cdot)$ Pairwise potentials

$\psi_u(\cdot)$ Unary potentials

θ The set of weights

φ A clique

$C_p(i)$ The set of pairwise cliques corresponding to node i

$D(\cdot)$ KL-divergence

$E(\cdot)$ Energy function

x_i Measurement of node i

Y Probabilistic model, Random field

$Y^\star$ The optimal result

y_i Associated random variable to node i

Z Normalization constant

Graph

$\mathcal{E}$ The set of edges in graph

$\mathcal{V}$ The set of nodes in graph

e_{ij} Edge between nodes i and j

G Undelying graph

n Number of nodes

u_i The coordinates of node i

v_i Node i in graph

Chapter 1

Introduction

Graphical models [84, 170] are one of the most important field of statistical machine learning which try to encode the statistical computational models dealing with structural data via a graph representation. The random variables are represented by nodes and the interactions among them are visualized via weighted edges in the graph. Nowadays, several fields of machine learning such as computer vision [13, 89], speech recognition [53, 129] and natural language processing [28] utilize graphical models to formulate and solve problems. Computer vision due to dealing with images is the field with most usage of graphical models specially random fields [13, 45, 96].

Computer vision applications have been usually proposed to provide a solution to natural images such as human body images [4, 107] (MRI, OCT) or natural scenes [94, 150] and man-made objects. The most important intrinsic property of those images is the relations among pixels in the image, specially when the pixels are close to each other. This property implies that neighbor pixels should have same color intensity. Due this fact, probabilistic graphical models have been trying to take advantage of this property and to address various type of problems [65, 128]. Incorporating different size of interactions is the focus in this challenge [89] and researchers have been trying to provide more interactions in their models by introducing tractable approaches. This research aim at stepping toward analyzing those approaches and answers the question that there is any way to determine how to configure

the underlying graph representation of a graphical model to produce an optimal model. Here we restrict our focus to conditional random fields (CRFs) [80, 98] and we study the effect of long-range non-local connectivity in modeling accuracy. Motivated by the answer of the question, we propose an efficient framework to address the computational complexity in the inference step of long-range non-local random field models. In this dissertation, the idea of randomly-connected non-local conditional random fields is explored to address the aforementioned issues of local random fields.

1.1 Problem Definition

Probabilistic graphical models are the way to visualize probabilistic models which ease the understanding of these models, unify and generalize them in the context of graph theory. Random fields are a type of graphical models utilized to model 2-D problems such as computer vision applications [89, 96, 185]. Random fields model the problem by taking advantage of the relations among different pixels in the image. Each pixel or super-pixel is represented by a random variable in the random field and it models the image by computing the relations among the random variables. Those relations are expressed as energy over the random field. Based on the Gibbs theory [45] the energy propagates in the random field to reach an equilibrium state. The relations among random variables are simulated by the Gibbs energy and the estimation is obtained by finding the equilibrium over the random field.

Conventionally, local interactions are applied to define the energy function. This is because of computational complexity of inference step. The computational complexity is increased by adding more interactions in the model and it has quadratic relation with the number of interactions. The ability of the local CRFs[1] is limited to model long-range connections and generally leads to excessive smoothing of object boundaries.

[1]We use the terms of "random field" and "conditional random field" sometimes in the text interchangeably. Although the main focus of this thesis is on CRF models, it is possible to extend all algorithms to other types of random fields as well.

In order to improve modeling accuracy and providing the longer-range of connections, researchers have expanded the local framework to incorporate hierarchical connectivity and higher-order potentials [43, 65, 94] defined on image regions. Unsupervised image segmentation usually provides the higher order connectivity structure in the majority of these frameworks. Although these approaches incorporated higher order connectivity in modeling, their accuracy is necessarily restricted by the effectiveness of unsupervised image segmentation.

Global connectivities can prevent smoothness in inference and labeling step. In addition, fully connected random fields can encode both color contrasts and spatial arrangements of different nodes by edge potentials instead of grid random field (locally random fields) where the edges only serve the contrast sensitive smoothness. Therefore, new researches have been conducted to address this problem recently [89, 185].

Although the proposed frameworks [89, 185] presented algorithms to manage the fully connected interactions, there exists a question that fully connected interactions are always useful in modeling. To the best of our knowledge, no research has been conducted to answer this question and the focus of researches was to introduce a tractable approach to utilize all connectivity interactions. Here we aim to take advantage of random graph theory and model the dense and non-local random fields via a stochastic process. By use of this approach, the underlying dense graph of random field is modeled by a sparse graph which addresses the computational complexity of the inference to some extent.

1.2 Challenges and Objective

Despite the fact that several methods and algorithms have been proposed to tackle different aspects of long-range non-local random fields, it is still considered as an open-field of research and there are many unanswered questions in this area yet. Although it has been shown that the long-range non-local random fields can improve the modeling accuracy compared to local random fields, generating the long-range interactions in the underlying graph and also managing the computational complexity of the new model are still the big

challenges of these approaches:

- Although it has been shown that fully-connected random fields improve modeling accuracy in several problems, there is no comprehensive answer to the question that how many connections are needed to have an accurate model.
- Increasing the number of interactions in the random field modeling has a significant effect on computational complexity, therefore proposing a strategy to make a trade-off between modeling accuracy and computational complexity is another challenge in this field.

Several algorithms have been conducted to tackle the aforementioned challenges either addressing only one aspect or proposing a method to solve them simultaneously. These methods can be divided in different categories depends on their focus in formulating the problem which mainly are proposing new penalty functions, reformulating the potential functions to reduce the computational complexity and applying higher order cliques or longer-range of connectivities via a less computational complexity burden framework. The first and second categories usually are associated with some limitations. Here in this dissertation, our main objective is to tackle the aforementioned challenges via the third category and we follow these objectives through this thesis:

- As the first objective, we will study whether utilizing larger number of long-range connectivities always helps to improve the modeling accuracy or not. We will demonstrate the advantages and disadvantages of longer-range of connectives via different examples.
- Motivated by those examples, the main objective will be a new representation for the underlying graph of random fields such that while it utilizes the long-range connectivities it maintains the computational complexity as well.
- It has been shown that hierarchical models have the ability to capture global and local information in modeling random fields in several problems. As the last objective,

we aim to take advantage of hierarchical models to obtain more useful information combined with the achievement of the second objective to model the underlying graph of random fields more efficiently and accurately.

The proposed methods based on the above objectives will be examined within the context of interactive object segmentation problem and they will be analyzed in terms of performance and behavioral in random fields modeling.

1.3 Contribution

The described objectives in this research lead to the following contributions:

- Proposing a new clique structure, stochastic clique, [146] which is the marriage of random graph theory and random field theory. The stochastic clique structure takes advantage of long-range connectivities via random graph theory within a stochastic process. The stochastic process determines which set of cliques are more informative in the inference step.

- Randomly-connected conditional random field (RCRF) [139, 146]; applying the stochastic clique structure into the random field modeling leads to a new representation for the long-range non-local random fields which the interactions among nodes are randomly determined and as the result, the structure of underlying graph is random. The proposed RCRF model creates a random field where only the most useful long-range and non-local interactions are incorporated in the modeling. The computational complexity of modeling procedure is maintained while long-range connections are involved in the modeling using this approach.

- Deep structure randomly-connected conditional random field (DRCRF) [145]; incorporating stochastic cliques makes the use of non-local and long-range connectivities applicable, however there is a limitation to the number of interactions which is mostly related to the inference algorithm. Although utilizing stochastic cliques decreases the

number of active cliques in the inference step, inference methods are not capable of maintains huge number of connectivities in the random field modeling yet. To address this problem, we aim to apply a deep structure model which utilizes a subset of connectivities in each level of hierarchy and it helps the inference method to be able to handle the inference process easily.

All the described contributions are combined together at the end to work as a unique framework to model the problem via the context of non-local and long-range random field.

1.4 Thesis Structure

This thesis is organized in six chapters. Chapter 2 introduces graphical models and various type of random fields. This chapter provides a mathematical definition for random fields specially discriminative model and conditional random fields. Then, an overview of fully connected random fields and proposed tractable approaches are presented. In Chapter 3 we analyze the effect of long-range interactions in random field modeling for different set of simulated examples. Motivated by those examples, we introduce the concept of stochastic cliques and we propose randomly-connected conditional random field model. Our proposed deep structure randomly-connected conditional random field is explained in Chapter 4 and we analyze its behavioral in this chapter. We examine the proposed methods along with several state-of-the art approaches in Chapter 5 and we evaluate them in different situations. Finally, thesis is concluded in chapter 6 and future directions for this dissertation are provided at the end.

Chapter 2

Background & Related Work

"Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering – ***uncertainty and complexity*** *– and in particular they are playing an increasingly important role in the design and analysis of machine learning algorithms. Fundamental to the idea of a graphical model is the notion of modularity – a complex system is built by combining simpler parts. Probability theory provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data. The* ***graph theoretic*** *side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms."*

Michael I. Jordan [74]

2.1 Graphical Models

Probabilistic graphical models [74, 84, 170] are powerful tools to combine graph theory and probability theory in a unified formalism for statistical multivariate problems. Probabilistic graphical models (usually called graphical models) are utilized to represent probabilistic models in a visual way to make them easier to interpret. There are several probabilistic models such as hidden Markov models (HMMs) [128], Markov random fields [27, 40, 104], Kalman filters [40, 175] and Bayesian networks [10, 161] which have been utilized in different applications. The role of a graphical model is to ease the understanding of these models, to unify them and to generalize them in the context of graph theory. For instance, HMMs and Kalman filters can be described with a common graphical model (Figure 2.1). A graphical model is a natural tool to formulate the variations of these classical models, especially when a large numbers of interacting variables are being studied together.

Graphical models are the combination of probabilistic models and graph theory. Graph theory plays an important role as a language to formulate probabilistic models. Furthermore, it is a useful tool to express computational complexity and feasibility when designing a model. In particular, the running time of an algorithm or the magnitude of an error bound can often be characterized in terms of structural properties of a graph.

A graph $G = (\mathcal{V}, \mathcal{E})$ is formed as a collection of vertices $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$ and the

Figure 2.1: The graphical model realization of a HMM and a Kalman filter. The difference between (a) the HMM and (b) the Kalman filter is that the number of states y_i is indefinite in the Kalman filter while it is finite in HMM. x_i represents the measurement at time i where y_i is the associated random variable of time i.

set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ where each edge is encoded via a pair of vertices $\{v_i, v_j\} \in \mathcal{V}$ (its two end nodes). The edge e_{ij} may be undirected or directed, such that edge $(v_i \rightarrow v_j)$ is different from $(v_j \rightarrow v_i)$ in directed graphs.

Each random variable y_i of a probabilistic model $Y = \{y_1, y_2, \cdots, y_n\}$ is represented as a node $v_i \in \mathcal{V}$. The dependency between two random variables y_i and y_j is expressed by the edge e_{ij} or e_{ji} when it is modeled by a directed graph. The first subscript represents the starting point (parent node) while the second one shows the ending point (child node). The edges e_{ij} and e_{ji} are identical since there is no direction in an undirected graph.

There are causal relations among variables in applications such as temporal problems. For example when you want to predict the weather condition based on some previous days, you have a set of random variables y_t that are dependent on each other and the condition of a new day y_{t+1} is a consequence of previous days $y_{1:t}$. Therefore, in such applications the interactions among random variables must be characterized by direct relations, such that the state of the new day is a child of earlier ones. The directed graphical model is the target for these types of applications since there are causal interactions among the random variables. Hidden Markov models [128] and other types of Bayesian networks [161] are examples of directed graphical models.

Mathematically speaking, a graphical model is a collection of marginal probability distributions factorized according to the underlying structure of the graph. The key idea in the graphical model is the factorization that will be explained more in Sections 2.2, 2.3 and 2.4. For a directed graphical model, $P(Y)$ can be computed based on the chain rule where each variable is marginalized given its ancestors (parent) variables (nodes):

$$P(Y) = P(y_1, y_2, \ldots, y_n) = P(y_1|y_2, \ldots, y_n) \cdot P(y_2|y_3, \ldots, y_n) \cdot \ \ldots \ \cdot P(y_n) \tag{2.1}$$

the set of all random variables is shown by Y and $n = |Y|$ is the number of random variables. To reduce the complexity of (2.1), Markov assumption is usually applied. Theoretically speaking (i.e., statistical definition) Markov assumption refers to the property of conditional probability distribution of future states of the process depends only on the present state or the limited number of past states. Based on this definition each random

variable can be modeled given a specific amount of information from the past:

$$P(y_{t+1}|Y) = P(y_{t+1}|y_t) \tag{2.2}$$

here it is assumed the future state of y_{t+1} only depends on the preset. Incorporating the Markov assumption (Section 2.3.1), the marginal probability of each variable y_i given other variables can be formulated as

$$P(y_i|Y) = P(y_i|y_{\pi(i)}) \tag{2.3}$$

where random variable $y_i \in Y$ and $y_{\pi(i)} \subset Y$ is the set of parental nodes for random variable y_i. In other words, $y_{\pi(i)}$ determines the set of random variables that node i is dependent on. Substituting (2.3) into (2.1), the joint probability distribution of all random variables can be factorized in the following way

$$P(y_1, y_2, \ldots, y_n) = P(Y) = \prod_{i=1}^{n} P(y_i|y_{\pi(i)}). \tag{2.4}$$

There are several applications such as image modeling where no specific causality among the random variables can be determined. Each pixel or super-pixel (Section 2.4.1) is modeled by a random variable. Since there does not exist any causal relation among the pixels (nodes) in the image, it is not possible to specify any direction between two nodes (i.e., edge direction) or random variables in those applications. Therefore, undirected graphical models are the best choice of modeling in such applications. Based on the Markovianity assumption, the probability distribution is factorized via a set of functions defined via the parent nodes of each random variables and the conditional independence assumption that will be explained in Section 2.3.1.

The probability distribution of an undirected graphical model is factorized via a set of functions $\psi(y_c) : R^d \rightarrow R^+$

$$P(y_1, y_2, \ldots, y_m) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi(y_c) \tag{2.5}$$

where y_c is a sub-set of random variables which are connected based on the Markovianity property of the graph. Function $\psi(\cdot)$ must be non-negative and it is called potential function. Parameter Z is a normalization constant to enforce the output value as a probability (i.e., $P(y_1, y_2, \ldots, y_m) \in [0, 1]$), since $\psi(\cdot)$ can be any non-negative arbitrary function. The normalization constant Z is the summation over all configurations to represent the output value as a probability.

In summary, there are two most common types of graphical models: Bayesian networks (or belief network) [11, 161] and Markov networks (or Markov random fields) [14, 45, 172]. The former ones are represented by directed graphs while the latter ones are mostly visualized via undirected graphs.

Over all mentioned types of network structures, computational complexity is an important aspect in the graphical model approaches [85]. The goal of modeling usually is to find the joint distribution $P(\cdot)$ over a some set of random variables. Let us assume that the variables are binary-valued and there are n variables, therefore, a joint distribution requires the specification of 2^n numbers of different assignments of values. However, there are some relational structures among random variables and they can be illustrated in the factorization step (2.5) which reduces the required number of calculations. The conditional independence property can be utilized to represent such high-dimensional distributions much more compactly. The marginal probability of each random variable is only represented by conditioning on its neighbors, therefore, there is no need to compute all interactions.

2.2 Probabilistic Models

For many applications in machine learning the problem is to predict the value of a vector Y given the value of a vector X of input features [12]. Most machine learning applications are divided into classification and regression [11]. In classification applications, Y is single variable which is expressed by a discrete class label, whereas in a regression problem it corresponds to one or more continuous variables.

The goal of probabilistic models is to find $P(Y|X)$. There are two different approaches [11] to formulate $P(Y|X)$:

- Direct solution to this problem is to represent the conditional distribution using a parametric model. The model must be trained to find the required parameters based on training set consisting pairs of $\{X, Y\}$. The trained model can be used to predict Y for a new input vector X. These types of methods are called discriminative models since they discriminate directly between the different values of Y.
- Joint distribution of $P(Y, X)$ is counted as the second approach. The joint distribution is utilized to evaluate the conditional probability $P(Y|X)$ implicitly in order to make predictions of Y for new values of X. This is known as a generative approach since by sampling from the joint distribution it is possible to generate synthetic examples of the feature vector X.

The following sections explain these two frameworks with more details.

2.2.1 Generative Models

In the context of probability and statistics, generative models [74, 33] are a type of probabilistic model that can generate synthetic examples randomly given hidden parameters Y. As mentioned before, a generative model specifies the joint probability distribution of observation and hidden states (i.e., labels or class) while the conditional probability of states given observations is formed via Bayes' rule. Although generative models obtain the conditional probability of states given observations by use of the joint probability, an assumption needs to be made to be able to compute the conditional probability. Bayes' rule states that:

$$P(Y, X) = P(X|Y)P(Y) \tag{2.6}$$

where $P(X|Y)$ is the likelihood model and $P(Y)$ is the prior one. The exact computation of the likelihood models is usually intractable, therefore, a conditional independence assumption [80] is taken into account to make the likelihood model tractable. The conditional

independence assumption assumes that each observation is independent from other observations given its state. Mathematically speaking, the likelihood model is approximated as the product of independent probabilities given the conditional independence assumption,

$$P(X|Y) = \prod_{i=1}^{N} P(x_i|y_i), \qquad |X| = N. \tag{2.7}$$

However natural data are usually high dimensional and several dimensions are correlated [40], therefore, the conditional independence assumption on X is not a proper assumption to make. Thus the modeling accuracy of generative models created by this assumption is lower than discriminative model [12] since the discriminative model provides a direct conditional probability.

Although generative models have some drawbacks compared to discriminative approaches in modeling accuracy, they have some advantages as well [164]:

- The generative model can handle missing data since the data are modeled for each class separately. This model can augment small set of labeled data with a large set of unlabeled data when the labeled data are expensive.
- A new class of data easily can be incremented to the whole classification problem since the data are modeled class dependently.

Discriminative models are preferable versus generative ones when dealing with classification problems, according to the reason asserted by Vapnik [165], the problem should be solved directly rather than be solved by the general formulation and computing the likelihood model.

2.2.2 Discriminative Models

Discriminative approaches model the problem as a conditional probability of states Y or unseen variables given observations X directly. The common applications of machine learning are classifications or regressions, therefore, it is more proper to model the problem directly

($P(Y|X)$) rather than using joint probability distribution (i.e., $P(Y, X)$). As mentioned in the last section, it is assumed that observations, X, are conditionally independent given the states, Y, to compute the likelihood model in the generative model. Therefore, the generalization performance of discriminative models is more than generative ones due the differences between the model and the true distribution of the data [12], in practice.

Generative models were more common before the advent of the maximum entropy models (MEM) [80]. The main problem of discriminative models was a way to represent the parametric form of these models. MEM asserts that the only unbiased distribution given the incomplete available data is the one that maximizes the conditional entropy of states given observations. Based on MEM, the probabilistic discriminative models are usually represented as a factorization of exponential family functions[1] [80].

In other words, MEM is based on the Principle of Maximum Entropy [72] stating that the only unbiased assumption can be made to define a conditional distribution is that the distribution is as uniform as possible given the available information when information regarding the probability distribution is incomplete [80].

Two advantages of discriminative models are:

- Discriminative models are faster [84, 164] than generative ones since they predict the new data point directly instead of an iterative procedure to find probability of the data point given each class.
- It is expected that discriminative methods have better predictive performance than generative approaches [84, 164] because discriminative models are trained to predict the class label rather than the joint distribution of input vectors and targets.

Although our main focus here is the general representation of a graphical model, the experimental results have been done in the context of discriminative frameworks specially conditional random fields. Following sections present two well-known graphical models:

[1]The general formulation of MEM and conditional random fields (discriminative probabilistic models) is identical which will be explained in Section 2.4.

Markov random fields (MRFs) known as a generative model and conditional random fields (CRFs) which are one of the well-know discriminative models.

2.3 Markov Random Fields

Markov random fields (MRFs) or Markov networks [40, 45, 78] are a set of random variables having the Markov characteristic based on an undirected graph. The foundation of Markov random field theory came from statistical physics [78]. Each random variable is represented as a node in an undirected graph and dependencies between random variables are expressed by an undirected edge.

MRFs are generative models [172] since they are usually utilized to model the prior part of generative models. They are usually based on the notion of conditional independence and Gibbs theory; the probability distribution is formulated as a factorization of an energy function (2.5). The energy function is the combination of feature functions (2.9) specifying the relations among random variables.

Since these models are not associated with a topological ordering, it is not possible to apply chain rule to expand $P(Y)$. Instead a potential function or a factorization is utilized to formulate $P(Y)$ upon the maximal cliques. A potential function can be any non-negative function of its arguments. The joint distribution is then defined to be proportional to the product of clique potentials. Since exponential functions have non-negative manner, the potential functions in MRF are usually formulated in the context of exponential equation which the exponent of the formulation is called energy function. This type of expansion derives the MRF model to equate with Gibbs distribution [112]. Hammersley-Clifford [57] has been proved that the Gibbs distribution is equal to the MRF model.

In other words, the conditional independence assumption is characterized by the cliques on the represented graph. The probability distribution is formulated as a factorization of

energy function based on clique structures:

$$P(Y) = \frac{1}{Z} \exp(-E(Y, \theta)) \tag{2.8}$$

$$E(Y, \theta) = \sum_{c \in \mathcal{C}} \psi(y_c, \theta_c) \tag{2.9}$$

where c is a clique template in the set of clique structures $\mathcal{C}$, θ determines the weight of each potential function $\psi(\cdot)$ to construct energy function $E(\cdot)$ and Z is the normalization constant which is the summation over all configurations. Finding the most probable configuration based on the MRF, is to minimize the energy function $E(\cdot)$.

It is worth to mention that since the proposed framework has been applied within a conditional random field model, MRFs are not our main focused and we just explained it in a very short description here.

2.3.1 Clique Structures

Generally, a clique [110] is a subset of vertices $v_c \subset \mathcal{V}$ of graph $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex-set and $\mathcal{E}$ is the set of edges of the graph G. The sub-set v_c is a fully connected undirected graph:

$$v_c = \big\{ v_i | \forall (v_i, v_j) \in v_c, \quad v_i \quad \textit{neighbor} \quad v_j \big\}. \tag{2.10}$$

Although cliques have a long history in graph theory [50, 55], it is a well-known terminology to define a random field model as well. The connectivity and relation among nodes in a random field is defined according to the neighborhood size (i.e., Markov order) to have a tractable inference approach in an undirected graphical model. The number of cliques and the shape of them are determined by the neighborhood size. Cliques are specified by their sizes and orientations. The position of a clique in the random field is another property which characterizes it as well.

The first-order Markov assumption is a common neighborhood size. Each node in the random field is connected to the nearest four other nodes in the graph. As seen in

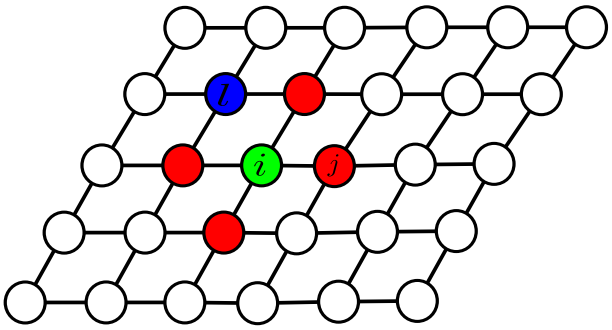


Figure 2.2: Each node v_j (red) is the neighbor of node v_i (green) via the first-order Markov assumption. Node v_l is not explicitly the neighbor of node v_i.

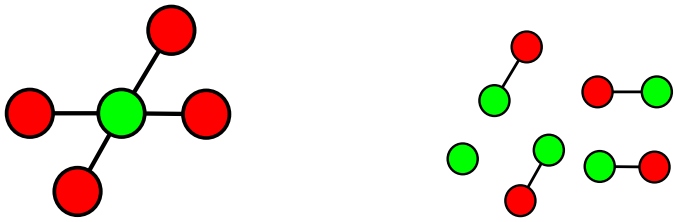
Figure 2.2, the green node v_i is in relation with four red nodes such as v_j. The relation between node v_i and v_l is not considered explicitly in the random field by the first-order Markov assumption.

Figure 2.3 demonstrates the clique structures for the first-order Markov connectivity. As seen in Figure 2.3 (b), the connectivity in each orientation is specified by a distinct clique. Each node neighbors with eight closest nodes based on the second-order Markov connectivity. Figure 2.4 shows cliques based on the second-order Markov connectivity.

There are two assumptions [32] to merge different cliques into a same category:

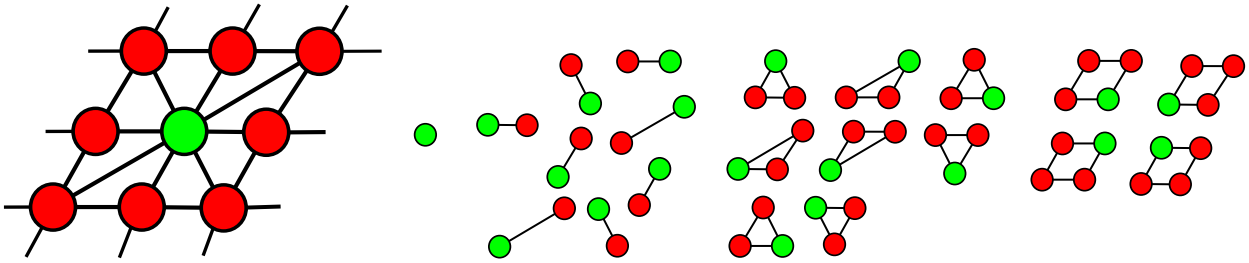
- Homogeneous assumption: if the identically oriented cliques are the same in the whole random field.
- Isotropic assumption: if all orientations of a specific size of clique (i.e., number of nodes in a clique) are identical.

Our first contribution is to present a new type of clique structure to tackle the computational complexity of fully connected networks. Inspired by random graph theory [16, 24], the proposed clique structure takes advantage of a randomness among variables' interactions to reduce the computational complexity of the inference step.



(a) (b)

Figure 2.3: Clique structures for the first-order Markov neighbors. (a) shows a four neighborhood connectivity and (b) demonstrates all clique structures related to the first-order Markov connectivity. There are five different clique structures based on the size and the orientations of cliques. A clique is a complete graph and based on the definition (3.4), all nodes in the cliques must be connected to each other. As a result, the maximum size of a clique is two when the random fields is defined based on four neighborhood size. It is obvious that if we assume we have maximal clique of greater than 2 (e.g. 3) there is at least a node which is not connected to the interested node and, therefore, the maximal clique is two.



(a) (b)

Figure 2.4: Clique structures corresponding to the second-order Markov neighbors. (a) shows the second-order Markov neighbor connectivity and (b) demonstrates all possible clique structures. As seen in (a), the maximal subgraph which all nodes are connected to the interested node is with four nodes, therefore, the maximal clique is four while the maximal clique in Figure 2.3 was two.

2.3.2 Random Graphs

Random graph theory [16, 34, 69, 183] defines graphs based on a probability distribution which a graph is generated via a probability distribution defined over the graph. This theory is a marriage between graph theory and probability theory. A random graph is obtained by starting with n isolated vertices that will be connected at random, iteratively.

There are several studies and approaches to generate a random graph. Gilbert [16] denoted the graph as $G(n, p)$ such that each edge connectivity is determined independently based on the probability p. Erdös–Rényi model [24] represents the graph as $G(n, m)$ where m determines the number of connected edges of the graph. The selection probability is calculated such that it provides the exact m edges for the graph. The Erdös–Rényi model is an effective model for extracting the essential behaviors of various graph properties which are explained in this section. We define a random graph as $G_{n,p}$ where n is the number of nodes and p represents the connection probability.

The generated random graph illustrates specific structure [24] based on the range of p[2]:

- $G_{n,p}$ is the disjoint union of trees if $p = o(\frac{1}{n})$ where[3]

$$f(x) = o\Big(g(x)\Big) \qquad \text{iff} \quad |f(x)| < \epsilon |g(x)| \quad \forall x \geq x_0 \quad , \ \epsilon \in R^+$$

- $G_{n,p}$ contains cycles with different sizes if $p \sim \frac{c}{n}$ for $0 < c < 1$. All connected components are either trees or unicyclic components and almost all nodes $(n - o(n))$ are in components that are trees.

- There is an amazing fact that $G_{n,p}$ is dramatically different when $p < \frac{1}{n}$ from when $p > \frac{1}{n}$. The largest component has size $O(\log n)$ for the former one, while most of the small components merge to a giant component with the size $O(n)$. The remaining components are of size $O(\log n)$. It is called double jump when $p \sim \frac{1}{n} + \frac{\mu}{n}$.

[2]We are only using these properties and the mathematical backgrounds of them are not the concern of this thesis, therefore, we accept them as facts.

[3]It is worth to note that ϵ is a very small real-value factor, therefore, this case is different from the second case.

- $G_{n,p}$ almost surely becomes connected if $p = c\frac{\log n}{n}$ with $c \geq 1$.
- $G_{n,p}$ is not only almost surely connected, but the degrees of almost all vertices are asymptotically equal when $p \sim \omega(n)\frac{\log n}{n}$ where $\omega(n) \to \infty$.

The new proposed clique structure is based on a stochastic approach which takes advantage of random graph theory. The effect of p on the behavior of the random graph structure is an interesting property that we want to observe in probabilistic random field models. The proposed framework is applied on conditional random fields in this thesis.

2.4 Conditional Random Fields

Conditional Random Fields [80, 84, 94, 98] are one of the most effective discriminative tools developed in the last decade. The idea of CRFs was first proposed by Laffety *et al.* [98]. CRFs directly model the conditional probability of labels given the measurements, without specifying any sort of underlying prior model, and they relax the conditional independence assumption of measurement given label commonly used by generative models.

Formally, let $G = (\mathcal{V}; \mathcal{E})$ be an undirected graph such that $y_i \in Y$ is indexed by the vertex $v_i \in \mathcal{V}$ of G. Then $(Y; X)$ is said to be a conditional random field if, when globally conditioned on X, each random variable y_i obeys Markov property with respect to the graph G. In other words, $P(y_i|X; Y/\{y_i\}) = P(y_i|X; y_{N_i})$ where $Y/\{y_i\}$ is the set of all nodes in G except node i, N_i is the set of neighbors of node i in G, X are input variables that are observed, and Y is the set of output variables that we aim to predict. According to the Hemmersley-Clifford theorem [57, 172] the interesting distribution is a Gibbs distribution which can be factorized into the following form based on G

$$P(Y|X) = \frac{1}{Z(X)} \prod_{c \in C} \Psi(Y_c, X_c) \tag{2.11}$$

where $Z(X)$ is the normalization constant and $\Psi(\cdot)$ denotes the potential function of over clique c. $\Psi(\cdot)$ is a non-negative real-valued function. Based on the principle of Maximum

Entropy [72, 80] (Section 2.2.2), the optimization of the Lagrange equation [80] constructed by that assumption leads to the same formulation as (2.11).

The objective function must be optimized to find the best distribution

$$P(y|X)^{\star} = \arg\max_{P(y|X)\in\bar{P}} H(y|X) \tag{2.12}$$

$$H(y|X) = -\sum_{y} P(y,X)\log P(y|X) \tag{2.13}$$

where $\bar{P}$ is the set of all possible distributions. The optimization problem is evaluated by assigning some constraints:

- The feature function, f_i, provides arbitrary relations among states and observations.
- $P(y|X) \geq 0$ for all y, X.
- $\sum_{y\in Y} P(y|X) = 1$.
- $\mathbb{E}(f_i) = \hat{\mathbb{E}}(f_i)$[4].

Finding $P(y|X)^{\star}$ (2.12) under these constraints can be formulated as a constrained optimization problem [80]

$$\Lambda(P(y|X), \vec{\lambda}) = H(y|X) + \sum_{i}^{m} \lambda_i(\mathbb{E}(f_i) - \hat{\mathbb{E}}(f_i)) + \lambda_{m+1}\sum_{y\in\mathcal{Y}} P(y|X) - 1 \tag{2.14}$$

resulting in

$$P(y|X)^{\star} = \frac{1}{Z(X)}\exp\Big(\sum_{i=1}^{m}\lambda_i f_i(y,X)\Big) \tag{2.15}$$

where λ determines the weight of each feature function in the distribution according to available training data. λ is calculated in the training step [80]. The derivative of log-likelihood of the energy function is computed regarding to each λ and the optimal weights are obtained by a gradient ascent procedure.

[4] $\hat{\mathbb{E}}(f_i)$ is the expected value of f_i based on the empirical distribution $\hat{P}(y,X)$. $\hat{P}(y,X)$ is obtained by simply counting how often the different values of the variables occur in the training data. $\mathbb{E}(f_i)$ is the expected value of feature f_i on the model distribution.

(2.15) is applicable to non-sequential data where y is only single output variables, and is the basic equation of Maximum Entropy classification [9]. The general form of a CRF to model the sequential random variables Y is

$$P(Y|X) = \frac{1}{Z(X)} \exp\Big(-\psi(Y|X)\Big) \tag{2.16}$$

$\psi(\cdot)$ can be divided into two distinct functions

$$\psi(Y|X) = \sum_{i=1}^{n} \psi_u(y_i, X) + \sum_{\varphi \in \mathcal{C}} \psi_p(y_\varphi, X) \tag{2.17}$$

here $y_i \in Y$ is a single state in the set $Y = \{y_i\}_{i=1}^n$, $y_\varphi \in Y$ is the subset of states (clique) from the set of clique template $\mathcal{C}$ (Section 2.3.1), and $X = \{x_j\}_{j=1}^n$ is the set of observations. The relations among nodes in the random fields are characterized using the concept of clique structure. Each subset of nodes constructing a clique, are involved in the non-unary feature functions to represent the relation among the random variables.

Basic CRFs utilize unary and pairwise potentials on local neighborhoods and it is usually first-order Markov (Fiqure 2.5, Section 2.3.1). Let node v_i is in the position of (k, l) of the graph and $|N(i)| = 4$ is the number of neighbors. Adjacency CRF can be formulated based on the clique structures that are single and pairwise cliques

$$C = \Big\{C_p(i)\Big\}_{i=1}^{n} \cup c_s(i) \tag{2.18}$$

$$C_p(i) = \Big\{(i_{k,l}, j_{k,l-1}), (i_{k,l}, j_{k,l+1}), (i_{k,l}, j_{k-1,l}), (i_{k,l}, j_{k+1,l})\Big\} \tag{2.19}$$

where $C_p(i)$ is the set of pairwise cliques corresponding to node v_i, and $c_s(i)$ expresses the single clique. The ability of the adjacency CRF is limited to model long-range inter-node connections and, therefore, generally leads to excessive smoothing of object boundaries. According to Gibbs distributions [45] energy is distributed in the random field until it reaches an equilibrium state. Based on the definition of energy in local random fields, the energy transformation would stop when close nodes have the same state value. In this situation there is no extra energy in one node that can be transformed to its neighbors. It means that close nodes have approximately the same label, which makes

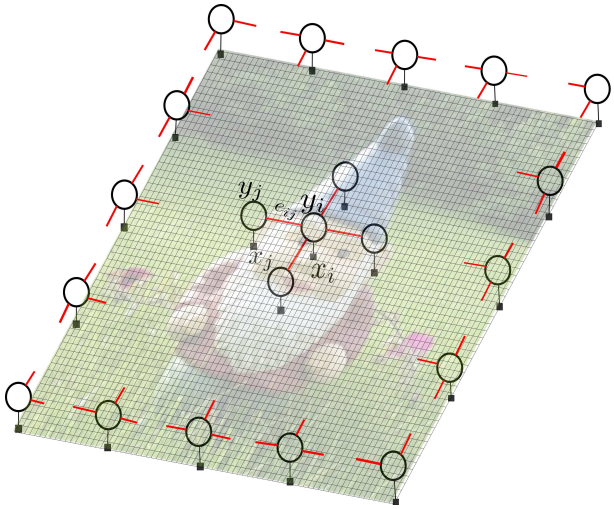


Figure 2.5: Adjacency CRF graphical model; each node i is connected to its local neighbors corresponding to the first-order Markov.

the boundaries smoother than desire. This is like transforming heat energy through a metal. Heat propagates through the metal until the whole surface of metal reach to a same temperature. But in the long-range connections the definition of energy can be changed such that, a same label is assigned to the random variables with the same characteristics. In addition to color similarity, the spatial information are incorporated in modeling. In these models the boundaries are preserved better in optimization step.

However fully connected random fields and global connectivity can model different relations among data, modeling of long-range or fully connected graph is not tractable. The computational complexity of long-range connections is the most challenging part of this problem.

In order to improve modeling accuracy and providing the longer connections range, researchers have expanded the basic CRF framework [65, 96] to incorporate hierarchical connectivity and higher-order potentials defined on image regions which will be discussed in section 2.4.2.

2.4.1 Local Random Fields

2D CRFs are useful tools in image processing applications [65, 96] due to the image structure. Pixels are related to their neighbors based on the color intensities and statistical features. The color intensity changes smoothly in natural images except at image boundaries; therefore Gibbs distributions and Markov random fields are appropriate approaches to model image processing and computer vision problems.

Conventional approaches [65, 174] utilized local neighborhood relations in statistical modeling such as CRFs, since the computational complexity increases by increasing the number of connections. In those studies each pixel is represented by a random variable and each random variable is related to others based on the first-order Markov.

The main role of CRFs in an image modeling problem is to incorporate contextual information and spatial relations among variables. Kumar and Hebert [94] proposed a discriminative framework based on the CRF to classify the man–made structure from natural scenes. A logistic classifier was applied as the associative potential while a simple Ising model was incorporated to extract interaction relations and to penalize every dissimilar pair of labels by a cost.

Shotton *et al.* [150] proposed the textonboost method to do object segmentation and object classification simultaneously. The CRF is utilized to capture the spatial interactions among neighbor pixels and also improves the segmentation of specific object instances. To overcome the ambiguities of local appearance of an image patch, they incorporated longer range information such as the spatial layout of an object and also contextual information from the surrounding image by 4-connected grid structure of the CRF. The use of CRF allows them to incorporate texture, layout, color, location, and edge features in a single unified model. Unary potential functions are provided by use of an adapted Joint Boost algorithm [163] which is a type of Adaboost classifier.

Modeling based on 4–connected CRFs imposes some issues in modeling accuracy since the spatial relations are defined at pixel levels and those relations are extracted only based on four neighboring pixels. On the other hand, incorporating larger connections has a

big impact on computational complexity [122]. Due to this fact, some studies have been proposed to incorporate super-pixel [43, 133] instead of pixels as random variables in the random field. In those approaches, a group of close pixels having the same color or statistics are obtained as one node in the random field instead of each pixel.

Fulkerson *et al.* [43] addressed the image class segmentation and localization by formulating the CRF on super-pixels. They obtained super-pixels from a conservative over-segmentation method. The local redundancies of data are captured by aggregating pixels into super-pixels and also it allows the model to measure feature statistics. A bag-of-features classifier was constructed by use of image descriptors such as SIFT. The classifier merges pixels to a region based on extracted features. Since the selected segmentation algorithm is an over-segmented type, boundaries are preserved in the output image. The CRF was utilized to reduce misclassifications that occur near objects' boundaries. The unary potentials are defined directly by the probability provided by the SVM classifier while, the interaction potentials are a combination of dissimilarity distance on color of super-pixels.

The modeling accuracy of the super-pixel methods is dependent on the accuracy of over-segmentation algorithm. However, there are some other studies which tried to capture both local and global relationships and address this issue [65, 96]. Hierarchical algorithm is a common approach involving global features as well as local features in modeling. The next section will analyze these methods in more details.

2.4.2 Hierarchical Random Fields

Common CRF approaches are based on the quantization of an image space-pixel [96]. The simplest ones utilize each pixel as a random variable in the random field while more complex methods use segments or a group of segments [130]. The goal of segment based representation of pixels as one random variable is to capture global information in order to improve the classification, localization or segmentation accuracy. However segment based (superpixel) approach relaxes the computational complexity, the final accuracy relies upon the initial quantization (superpixel or initial segmentation) over the image space.

Hierarchical approaches are a way that can compensate the effect of initial quantization [96]. Thanks to the structure of random field modeling, different relations can be incorporated in an unified model. In other words, the local and global interactions can be involved in modeling, simultaneously. Hierarchical models allow the integration of features computed at different levels of the quantization hierarchy.

Hierarchical or multi-scale approaches have been utilized for different applications. Fieguth [38] proposed a multi-scale framework to do the posterior sampling when there exist sparse measurements. The proposed model takes advantage of multi-scale properties to relax the computational complexity of computing the covariance matrix. He also used a hierarchical approach to model the characteristic of a porous media [39] locally and non-locally to sample a new image.

Besides the multi-scale methods proposed in the context of Markov random fields, hierarchical methods have been utilized within the context of conditional random field frameworks as well. The unary and pairwise potentials are computed in different layers. Each layer is constructed based on a quantization level on observations. The CRF plays the role of an unification procedure to combine all potential functions. The final step is the same as common CRFs which is to minimize the energy to find the best state configuration.

He *et al.* [65] utilized local and global features in a unified hierarchical CRF model to label different regions in an image. Local features are the result of statistical classifiers, such as a neural network. The image is divided into non-overlapped patches to extract the global features. All features are associated with a learned weight in the training step which determines the impact of that feature on the classification and the decision. Inference and finding the optimal label configuration has been done in the lower level (fine level) in order to minimize the proposed energy function.

Ladický *et al.* [96] applied a hierarchical CRF on object class image segmentation. Most of the hierarchical approaches follow the same procedure to define hierarchical levels. The main difference between [65] and [96] is that [96] incorporated additional term in the energy function that represents the label consistency between different layers. Also the unary features were provided with the same method as [150].

Due the weakness of local CRFs, studies have intended to incorporate more global interactions in modeling. It was mentioned that decision just by local features leads to mis-classification [43] in some problems. Classification of "sky" and "water" in an image is a good example for this situation. These two type of objects have the same characteristics locally while they can be distinct by use of global characteristics. Due to this fact, a wide variaty of research [89, 169, 185] have been conducted to utilize more interactions in modeling and the goal is to take the advantage of all interactions. The ultimate goal is to define a fully connected interactions to acquire an accurate model.

The main challenge of utilizing long-range interactions in random fields is the computational complexity since the model complexity increases exponentially by increasing the number of connections in the modeling procedure. To make more intuition about the computational complexity of those models and know how they work, the following section explains fully connected random fields generally and reviews some efficient proposed approaches to tackle this problem.

2.4.3 Fully Connected Random Fields

The general framework of CRFs is identical for different sizes of connectivities and usually the only difference from a local CRF is the design of interactions among random variables in the field, therefore the formulation of a fully connected CRF (2.20) is the same as adjacency CRFs (2.16)

$$P(Y|X) = \frac{1}{Z(X)} \exp\Big(-\psi(Y|X)\Big) \tag{2.20}$$

where $Z(\cdot)$ is the partition function and $\psi(\cdot)$ is the combination of unary and pairwise potential functions. The main differences of a fully connected CRF and an adjacency one as shown in Figure 2.6 are the neighborhood size, the number of cliques and their structures. Since the graph is fully connected, each node is in the neighborhood set of all other nodes in the graph,

$$N(i) = \Big\{ j | j = 1 : n, j \neq i \Big\} \tag{2.21}$$

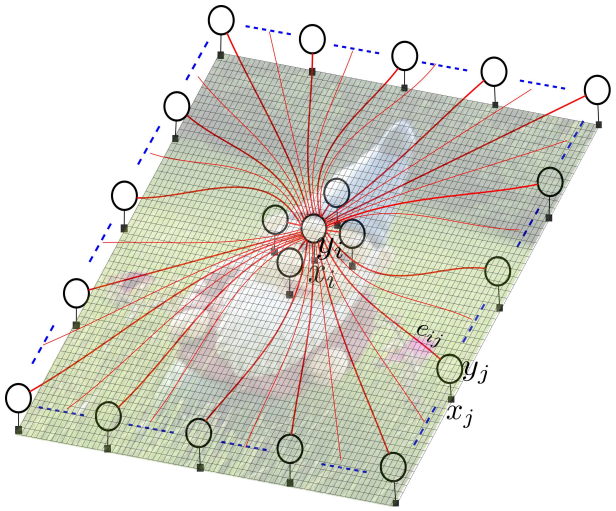


Figure 2.6: Fully connected conditional random fields; each node is connected to all other nodes in the random field. The connectivities of node i are only shown for the visualization purposes. Compared to Figure 2.5 the interested node can be connected to all other nodes in the random fields.

where $|N(i)| = n - 1$. Based on the neighborhood size and the clique structure [160], the number of possible cliques are varied. Different clique structures can be utilized in fully connected modeling but using only the pairwise clique structures is the most common approach [89, 185]. The main reason is that the number of cliques increases the computational complexity of the model in addition to the number of connectivities. Here, without loss of generality the specified clique structure C is assumed to be pairwise clique

$$C = \Big\{C_p(i)\Big\}_{i=1}^{n} \tag{2.22}$$

$$C_p(i) = \Big\{(i,j)|j \in N(i)\Big\} \tag{2.23}$$

The inverse of the covariance matrix among random variables (i.e., with computational complexity $O(N^3)$ [134]) must be evaluated to find the solution and, therefore, the computational complexity of exact inference on a fully connected graph is $O(N^3)$.

Due to the high computational complexity of the exact inference, approximation methods have been proposed to tackle this problem [84, 176]. Mean field inference [84] is one

of the tractable frameworks performing a message passing algorithm by approximating each marginal probability. The message passing procedure updates each approximated marginal distribution until convergence. The computational complexity of single iteration of updating the marginal distribution for one random variable is $O(N)$ and the complexity of updating all marginals is $O(N^2)$.

The recent method proposed by Krähenbühl and Koltun [89] reduces the computational complexity of the inference from quadratic to linear by extending the mean field approximation. Their inference is highly efficient since a linear combination of Gaussian kernels in an arbitrary feature space is defined as pairwise feature function.

2.5 Efficient Inference Approach

Incorporating long range connectivities is a challenging part of random field modeling. The main advantage of random fields in modeling problems such as image segmentation and image classification is that it facilitates the use of spatial information in the model. Although utilizing long range connectivities benefits the model, increasing the size of interactions and spatial relations in the model have a significant impact on computational complexity of the inference step. There are some approximation techniques [84] to relax the computational complexity but they are not helpful when the size of image or observation is very large. There are several methods [89, 186] reformulated the inference problem as filtering and solved the inference step by fast convolutions. Those algorithms are mainly divided into two folds based on the convolution implementation to compute the interactions. The next sections explain those methods in more detail.

2.5.1 Permutohedral Lattice Based Method

Krähenbühl and Koltun [89] proposed a tractable inference procedure by incorporating specific potential functions. They modeled the multi-class image segmentation by a fully connected CRF where the edge potentials were obtained using Gaussian kernels. By use

of these new feature functions, they reformulated the inference procedure as a filtering problem via a mean-field approximation.

The exact distribution of $P(Y|X)$ is approximated by $Q(Y|X)$. The mean field [84] approximation computes the distribution $Q(Y|X)$ among all distributions Q minimizing the KL-divergence of Q and P, $D(Q||P)$:

$$D(Q||P) = \sum_i Q_i \ln \left(\frac{Q_i}{P_i} \right). \tag{2.24}$$

In other words, the product of independent marginals $Q(y_i|X)$ over each variable is computed as an approximation of $P(Y|X)$ via KL-divergence. However mean field approximation relaxes the computational complexity of the inference to $O(N^2)$, the computation is not tractable yet when working with large data (e.g., high resolution images). To address the computational complexity of the fully connected random field, the proposed method [89] utilized a bilateral filtering and a Gaussian smoothness kernel as pairwise feature functions to be able to model the random field in a Permutohedral lattice which can compute the convolution in linear time complexity.

The utilized feature function $k(\cdot)$ is the combination of two contrast-sensitive kernel filters. The first kernel in (2.25) is a bilateral filter which is a nonlinear filter smoothing while preserves strong edges [121] and the second one is a smoothness kernel to remove small isolated regions:

$$k(i,j) = \omega^{(1)} \exp\Big(- \frac{\|u_i - u_j\|}{2\theta_\alpha^2} - \frac{\|x_i - x_j\|}{2\theta_\beta^2} \Big) + \omega^{(2)} \exp\Big(\frac{\|u_i - u_j\|}{2\theta_\gamma^2} \Big) \tag{2.25}$$

where u_i and u_j are the coordinates of node i and j in the image; x_i and x_j are corresponding features for each node according to the observation. $\omega^{(1)}$ and $\omega^{(2)}$ are the weights of each feature function determined by cross validation. Since they used Gaussian kernels as pairwise feature functions, the message passing procedure can be seen as a filtering problem. Thanks to a novel data structure (Permutohedral lattice) where reduces the computational complexity of convolution to linear [1] the convolution can be implemented efficiently.

The feature space is represented with simplices arranged along $d+1$ axes where d is the feature space dimension. The procedure is divided into four stages: generating position vectors, splatting, blurring, and slicing. The position of each sample must be generated and embedded in the high-dimensional space $d+1$. After that we must identify its enclosing simplex and compute barycentric weights so-called splatting. The third stage is to perform a regular Gaussian blur within that subspace. To do that, splatted data are convolved by the kernel [1 2 1] along each lattice direction. Slicing is identical to splatting, except that it uses the barycentric weights to gather from the lattice points instead of scattering to them.

The proposed method was examined by a multi-label image segmentation problem. The actual computational complexity of the proposed method is $O(dN)$ where d is the dimension of feature space and N is the number of nodes in the random field.

2.5.2 FFT Based Method

Zhang and Chen [185] proposed an alternative method to address the computational complexity of the fully connected CRF. Their approach is similar to [89] in the context of filtering. The proposed method modifies the inference of the fully connected CRF to a filtering problem and provides an efficient procedure by doing convolution.

The pairwise feature functions are divided into color contrast and spatial relation:

$$\psi_{ij}(y_i, y_j) = \phi_{i,j}(u_i, u_j)\varphi_{i,j}(y_i, y_j, x_i, x_j) \tag{2.26}$$

where $\phi_{i,j}(u_i, u_j)$ is the spatial relation of node i and j and $\varphi_{i,j}(y_i, y_j, x_i, x_j)$ is color contrast. The color contrast term encourages a same label when the colors of nodes i and j (x_i and x_j) are similar, and different labels otherwise. The spatial relation represents the log-likelihood of the spatial distribution of two categories (i.e., the probability that two

categories co-occur at positions u_i and u_j):

$$\phi_{i,j}(u_i, u_j) = \mathcal{G}(-\theta \|u_i - u_j\|) \tag{2.27}$$

$$\varphi_{i,j}(y_i, y_j, x_i, x_j) = \begin{cases} \mathcal{G}(x_i - x_j), & y_i = y_j \\ 1 - \mathcal{G}(x_i - x_j), & \text{otherwise.} \end{cases} \tag{2.28}$$

There is also a binary labeling function $\mu_i(y_i)$ indicating that the node i is labeled by y_i or not:

$$V_{ij}(y_i, y_j) = \phi_{i,j}(u_i, u_j)\varphi_{i,j}(y_i, y_j, x_i, x_j)\mu_j(y_j). \tag{2.29}$$

However implementing the feature function computation via convolution is not applicable because $\varphi_{i,j}(y_i, y_j, x_i, x_j)$ can take several values. By ignoring the color contrast part of feature function, the problem can be viewed as the filtering of an image with value of $\mu_i(y_i)$ for the pixel i where the filter is $\phi_{i,j}(u_i, u_j)$ depending on the relative position of u_i and u_j. Image filtering can be greatly accelerated with fast Fourier transform (FFT), which reduces the complexity to $O(NlogN)$. Although the convolution is not applicable by involving the color feature function since there are both spatial and color variations in the equation, the problem is addressed in [121]. Image filtering on the space and color dimensions simultaneously has been studied in the context of bilateral filtering. Based on the proposed idea of [121], if the color value x_i of node i is fixed to x_c, (2.29), it turns the original convolution to the convolution of $\varphi_{i,j}(y_i, y_j, x_c, x_j)\mu_j(y_j)$. Therefore, the color values are discretized into C clusters $\{X^c\}$ and filtering problem is computed for each x_c. The computational complexity of the proposed method is $O(CNlogN)$ as the convolution must be applied for every C clusters. However like the previous approach they take the advantage of down-sampling to reduce the computational complexity.

Both methods used the advantage of the stationary property in image modeling. They assumed that cliques are stationary and, therefore, potential functions are identical in the whole random field. Due to this fact, those solutions are not appropriate for non-stationary problem. The Gaussian feature functions are the second drawback of those methods. It is only possible to select the feature functions in the form of Gaussian functions to be able to utilize the mentioned approaches.

2.5.3 Related Methods

Following the work of Krähenbühl and Koltun [89], new studies have been done on different aspects of the proposed model in [89]. Vineet *et al.* [169] presented a new initialization for this approach. Since the mean field method is sensitive to initialization, they proposed a hierarchical mean-field approach where labeling from the coarser level is propagated to the finer level for better initialization. They used a two-level hierarchies which a variable of the coarser level is the parent of four variables in finer level. The result of inference in coarser level is an initialization for inference of the finer level.

The permutohedral lattice based dense CRF [89] is restricted to take the pairwise feature function from a weighted Gaussian kernel with zero mean rescaled with a single value. This issue was addressed in [169] by learning the mean and the covariance matrix of general Gaussian mixture model. They added some Gaussian mixture feature functions in addition to the zero-mean Gaussian one. However they asserted that the reported accuracy is 0.3% more than [89], their method is slower than permutohedral lattice based dense CRF since they utilized the Gaussian mixture which must evaluate the filtering step separately for each of the mixture components in the model.

The other drawback of [89, 169] is that they are restricted to some Gaussian parametric feature functions. Campbell *et al.* [21] generalized the pairwise potential to a conditional non-parametric model learnt from training data and can also be conditioned on the input data at test time. They applied this procedure in three steps; first, the conditional pairwise probability densities were learnt from training data conditioned on a test image as feature function. Secondly, the probability was expressed as a dissimilarity measure between nodes in the CRF. An efficient approximate embedding technique was applied to find a set of feature spaces that encode the dissimilarity measure as the Euclidean distance and thus the desired pairwise potential under a Gaussian kernel in this space for the last step. A similar approach to [89] was provided in the inference step since the pairwise feature functions were represented as Gaussian. The permutohedral lattice was provided to do the inference step with linear computation.

The conditional density probability of pairwise feature functions are estimated directly

from a set of training data $\mathcal{T}$. For each node i, first the local area (an image patch) around a particular node i in the test image X is considered and then similar patches in the training images are found. For each label l the conditional probability

$$P(y_j = l | y_i = l; X; \mathcal{T}) \tag{2.30}$$

for every node j around node i is estimated (i.e., the conditional local density distribution of the label l). In other words, the kernel density estimation is applied for each label with the mean of node j. Therefore, they utilized a prior σ_ω to indicate the range of useful information in the pairwise potential

$$g_{win}(i,j) = \exp\left(\frac{-\|u_i - u_j\|}{\sigma_\omega}\right) \tag{2.31}$$

where u_i and u_j are the pixel positions. To denote the potential function in the form of dissimilarity measure, the logarithmic value of the conditional density is computed by

$$d(i,j,I,\mathcal{T}) = -\log\Big(g_{win}(i,j) \cdot P(x_j = l | x_i = l; X; \mathcal{T})\Big). \tag{2.32}$$

Finding $P(x_j = l | x_i = l; X; \mathcal{T})$ has high computational complexity[5]. An uniform sampling procedure was applied and $\mathcal{C}$ sample points were picked and the dissimilarity was measured for $i \in \mathcal{C}$ and all j. The proposed method was applied to the in-painting of binary images, a collection of handwritten Chinese characters that are occluded by a centered rectangular region

The authors examined the model accuracy against different parameters. The experimental results demonstrated a noticeable decrease in performance for small windows. There was also a drop-off in performance with large window sizes. This is to be expected since the relations among nodes are not valid over very long-range connections. The reported results imply that adding long-range connections increases the accuracy of model instead of local

[5]The conditional probability table must be computed to have all conditional probabilities. The table determines the conditional probability of various configurations, therefore, it is needed to find a lot of conditional probabilities to fill out the table entities. The number of table entities depends on the number of labels.

interactions; however, several problems such as inpainting problem do not need the fully connected iterations for having the best model. The size of connections and interactions among nodes is an intrinsic of the problem that must be discovered.

Apart from classification applications of the dense CRF, Ristovski *et al.* [134] proposed a new method to handle continuous values and to be useful for regression problems. The unary potentials (2.33) were formulated as the quadratic relation of prediction of unstructured models based on observation and output

$$\psi_k(y_i, X) = -(y_i - R_k(X))^2 \qquad k = 1, \ldots, K \tag{2.33}$$

where the result of feature functions are large when predictions and outputs are similar. The interaction potential is derived as

$$\psi(y_i, y_j, X) = -\mathcal{K}_l(F_i, F_j)(y_i - y_j)^2 \qquad l = 1, \ldots, L \tag{2.34}$$

where F_i and F_j are feature vector corresponding to node i and j. $\mathcal{K}_l$ is some similarity measures between feature vectors. The probability is approximated under mean-field theory as same as [89]. The independent approximated marginals Q_i of the mean-field theory were expressed as a Gaussian distribution with closed-form mean and variance. Since the potential functions were represented as Gaussian shapes, fast high-dimensional Gaussian filtering can be applied by use of permutohedral lattice as well.

The dense CRF approach [89] was also utilized in semantic image segmentations with objects and attributes [186]. The task is to label each pixel with:

1. An object label such as car or bicycle
2. Visual attribute like wood or metal
3. Surface properties such as shiny or glossy.

The authors used a factorial CRF [77] framework to handle the segmentation and attributes selection simultaneously. They defined the CRF in an hierarchical framework in which both objects and attributes are labeled at two levels of pixels and regions.

The potential functions were formulated as a Gaussian framework to be able to use the fast convolution framework in the inference step. The multi-class segmentation result was obtained by the same framework as [89]. The multi-label attribute selection is similar to the segmentation part except that the random variables take the sub-set of labels instead of single labels. The attribute labels are assigned by an element of the power-set of the defined attributes set $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$.

2.6 Deep Conditional Random Fields

In Sections 2.4 and 2.5 we discussed about standard pairwise random field approaches and the smoothness problem (i.e., short-boundary bias) associated to local random fields which can be resolved by increasing the number of connectivities and taking advantage of long-range interactions in the random field modeling. The smoothness drawback of standard random field models results from the fact that they penalize the assignment of different labels to neighboring pixels [15, 18]. These types of potential functions penalize the long object boundaries equivalently which results to smoothen the boundaries.

Parallel to development of long-range connectivities and higher order cliques to address smoothness issue of standard pairwise random fields, there have been other frameworks which reformulate the pairwise potentials to resolve this drawback. Jegelka & Bilmes [73] proposed a cooperative graph cut (Coop-cut) model where instead of penalizing the number of label discontinuities, they penalized the number of types of label discontinuities. They observed that the graph cut energy function is improved if the usual cut cost (the sum of edge weights) is replaced and the cost of the edges is not only based on the sum the edge weights.

Motivated by the promising results of Coop-cut [73] introduced by Jegelka & Bilmes, Kohli *et al.* [82] proposed a deep structure framework to take advantage of the new potential function in a multi-layered random field to preserve boundaries while allowing for fast MAP inference. They reformulated the model of [73] in a context of a hierarchical structure via a transformation of higher order potentials. The proposed Coop-cut framework made

the MAP inference an NP-hard problem, and the approximation algorithms were utilized to optimize the model. By use of this new hierarchical framework, they were able to derive an exact yet practical algorithm for MAP inference. The new potential function (penalizing the diversity of cuts in the graph) is considered as a higher-order potential which makes the optimization an NP-hard problem. Kohli *et al.* transformed the higher-order potential function into that of a pairwise potential by introducing additional auxiliary random variables. However the number of auxiliary variables grows exponentially with the arity of the function and it makes the approach infeasible. Due to this fact, they took advantage of the inherent structure properties of the data and solved the problem within a multi-layered structure factor graph.

Deep-structured graphical models have been attracting researchers in the past which took a different approach to improving inference performance by introducing intermediate state layers, where there is a dependency of each higher layer on its previously layer, and inference is carried out in a layer-by-layer manner from bottom to top. Prabhavalkar and Fosler-Lussier [126] and Peng *et al.* [123] both introduced multi-layer conditional random field models where the local factors in linear-chain conditional random fields are replaced by multi-layer neural networks and trained via back-propagation. Ratajczak *et al.* [132] introduced a context-specific deep conditional random field model by replacing the local factors in linear-chain conditional random fields with sum-product networks. Yu *et al.* [180, 179] proposed a deep-structured conditional random field model which consists of multiple layers of simple CRFs where each layer's input consists of the previous layer's input and the resulting marginal probabilities. While such deep-structured graphical models are good at handling high observational uncertainties such as measurement noise and outliers by characterizing different information at the different layers, they only implicitly take advantage of long range relationships and are more limited in this aspect when compared to fully-connected graphical models.

2.7 Summary

The proposed dense CRFs formulated the inference of a fully connected CRF as a fast filtering approach. The message passing stage of the inference step is designed as the filtering of the whole image by each potential function. They utilized the fast convolutional procedure to address the computational complexity of the filtering problem which can be formulated as either I) fast FFT method or II) the permutohedral lattice.

However the main drawback of those approaches is that potential functions must be in the shape of a Gaussian function to be applicable in convolution procedure. Although some studies addressed the restriction of the Gaussian function to a non-parametric conditional probability [21], in the convolution step the potential functions must be reformulated as the Gaussian function again. This approach limits the well-known advantage of CRFs which is the ability to select arbitrary feature functions.

Furthermore, those approaches assumed that the problem is spatially stationary since they used convolution and the potential function is identical in different parts of the image. However there are several problems which are not homogenous and modeling by the stationary feature functions results to attenuate the accuracy.

Those proposed methods usually used truncated Gaussians and asserted that the accuracy decreases when the standard deviation of Gaussian potential functions are very large. It means that we do not need fully connected interactions in some problems and the efficient number of connections depends on the observation and it is problem related.

To address those drawbacks and have a general framework with reasonable computational complexity in the inference step, we present an efficient Bayesian inference method in next chapters. Inspired by random graph theory, we propose a new stochastic clique structure, which allows the computational complexity of the fully connected graph to be reduced without limiting the CRF with specific feature functions. Based on the provided flexibility, the new framework preserves the merits of the standard CRF, such that any arbitrary function can be selected as the potential function. Our proposed method is a mixture between random graph theory and random fields theory, obtaining the inference

based on a Bayesian approach which samples cliques of the fully connected random field while allowing for computational tractability.

Chapter 3

Randomly-Connected Random Fields

"God may not play dice with the universe, but something strange is going on with the prime numbers."

Paul Erdös

The theory of **random graphs** was founded by Erdös and Rényi in 1960. In mathematics, random graph is the general term referring to probability distributions over graphs. Random graphs may be described simply by a probability distribution, or by a random process which generates them. The theory of random graphs lies at the intersection between graph theory and probability theory.

3.1 Introduction

As we discussed in Chapter 2, random fields configured with local connectivities usually result with excessive smoothness on image boundaries in application such as image segmentation. The incorporation of long range connectivities and global features have a significant impact on the accuracy of modeling. However, the main issue with the use of long range interactions, particularly in the case of fully connected random fields, is the high computational complexity associated with the inference step of the random field, which was explained in Section 2.4.3. Thanks to the approximation inference algorithms [84, 113, 170], the computational complexity of inference is decreased to a quadratic, dependent on the number of connections and the number of nodes (random variable) in the underlying graph of a random field (i.e., $O(N^2)$) where N is the number of random variables in the random field. There are several state-of-the-art frameworks [89, 185] (analyzed in Section 2.5) which tackled this problem by making the stationary assumption among the random variables in the random field (i.e., by utilizing the Gaussian pairwise potential (2.25)) and obtaining inference by a filtering and convolutional procedures (Section 2.5.1).

Although the computational complexity of fully connected random fields became tractable [89, 185] by using Gaussian pairwise potentials as mentioned in Section 2.5, they are associated with some limitations (e.g., the potential functions must be formulated by Gaussian equations). Furthermore, considering the mentioned trend in sections 2.4.2 and 2.4.3 between the number of random variables' interactions and modeling accuracy the primarily focus on this field have been incorporating more interactions in modeling and the only considered challenge has been reducing the computational complexity of the inference approximation. To the best of our knowledge, the challenge of determining optimal interaction sizes and configurations has been largely left unexplored. For example, the authors of [21, 89] proposed how to address the computational complexity of a fully connected random field for image segmentation. However the experimental results of [21, 89] illustrated that the best interaction range is problem dependent.

3.2 Problem Definition

As we discussed in Section 2.5, a popular approach to image segmentation problems [150] is to model the framework using random fields and to formulate the problem as Maximum a Posteriori (MAP) [18, 159]. These methods incorporate prior information regarding interactions between neighboring pixels into the model.

The most common strategy is to train a unary potential (i.e., $\psi_u(\cdot)$ at (2.17)) based on information provided by the user as training data or marked area of the object and the background as interactive image segmentation as shown in Figure 3.1, and then to introduce pairwise potentials (i.e., $\psi_p(\cdot)$ at (2.17)), to refine the result of the unary segmentation which is shown in Figure 3.1. Although successful proposed methods [159, 20] using this framework reported efficient MAP inference using graph cut [18], they have been restricted in their ability to handle complex object structures [81] since they utilize local MRF (adjacency) models that incorporate pairwise potentials on neighboring pixels [44, 150]. These structures are limited in their ability to take advantage of long-range connections within the image and generally result in excessive smoothing of object boundaries [52, 168].

As seen in Figure 3.1, we are motivated to maximize the conditional probability of the segmentation result given a real image, $P(Y|X)$, as the observation with two objectives which must be satisfied: I) accurate region segmentation and II) object boundary preservation. The first objective asserts that the intersection of the segmented object and its ground truth must be maximized, while the second objective enforces the optimization to preserve the objects' boundaries as much as possible. The conventional local random field approaches [65, 174] usually considered only the first objective and maximized the conditional probability $P(Y|X)$. Here in this thesis we are aiming to find the best configuration of Y maximizing the conditional probability $P(Y|X)$ by considering both objectives at the same time. Although utilizing non-local and long-range connectivities is the simple but effective solution to this problem, increasing the number of interactions in the model increases the computational complexity. Therefore, we are introducing a novel algorithm to address this issue.

Figure 3.1 demonstrates the problem more intuitively. Here we want to formulate

the image segmentation problem within a random field framework and maximize it via a Maximum a Posteriori (MAP) method. We address the drawbacks of local random fields (i.e., short-boundary bias problem [82]) by utilizing non-local interactions in the random field. However to resolve the computational complexity of the produced non-local random field, we introduce a new clique structure which takes advantage of most useful long-range cliques instead of utilizing all possible clique structures in the random field model. In other words, the cliques for each node participating in the inference step are determined by use of a stochastic approach. To formulate the problem more accurately, the goal is to maximize the conditional probability of $P(Y|X)$ such that

$$\begin{aligned} P(Y|X) &= \frac{1}{Z(X)} \exp\Big(-\psi(Y|X)\Big) \\ \psi(Y|X) &= \sum_{i=1}^{n} \psi_u(y_i, X) + \sum_{\varphi \in \mathcal{C}} \psi_p(y_\varphi, X) \end{aligned} \tag{3.1}$$

where $\mathcal{C}$ is the set of cliques in the random fields selected based on a stochastic procedure and are involved in the inference step. Introducing the best set of clique structures $\mathcal{C}$ leads to find the sub-optimal solution easier. For more information regarding the notations readers are referred to Section 2.4 and (2.16) and (2.17). We are aiming to maximize the conditional probability of $P(Y|X)$ via a MAP approach (so-called inference and more details will be provided in Section 3.4) such that the relations among random variables are modeled by a random field structure that is non-local and the underlying graph structure is designed by a stochastic algorithm. Therefore, we want to find the best configuration of Y such that it maximizes the conditional probability of $P(Y|X)$:

$$Y^* = \arg\max_{Y'} P(Y|X) = \arg\min_{Y'} \Big(\sum_{i=1}^{n} \psi_u(y_i, X) + \sum_{\varphi \in \mathcal{C}} \psi_p(y_\varphi, X) \Big) \tag{3.2}$$

where y_i is a random variable in the set of Y and y_φ is a subset of random variables from Y specified by clique φ. Here the clique $\varphi \in \mathcal{C}$ is not a regular clique and it is determined by a stochastic process which specifies that which sets of two-node cliques can be involved in the inference process. By use of this approach, the random field can be considered as a fully connected random field since each possible clique in the random field has a chance

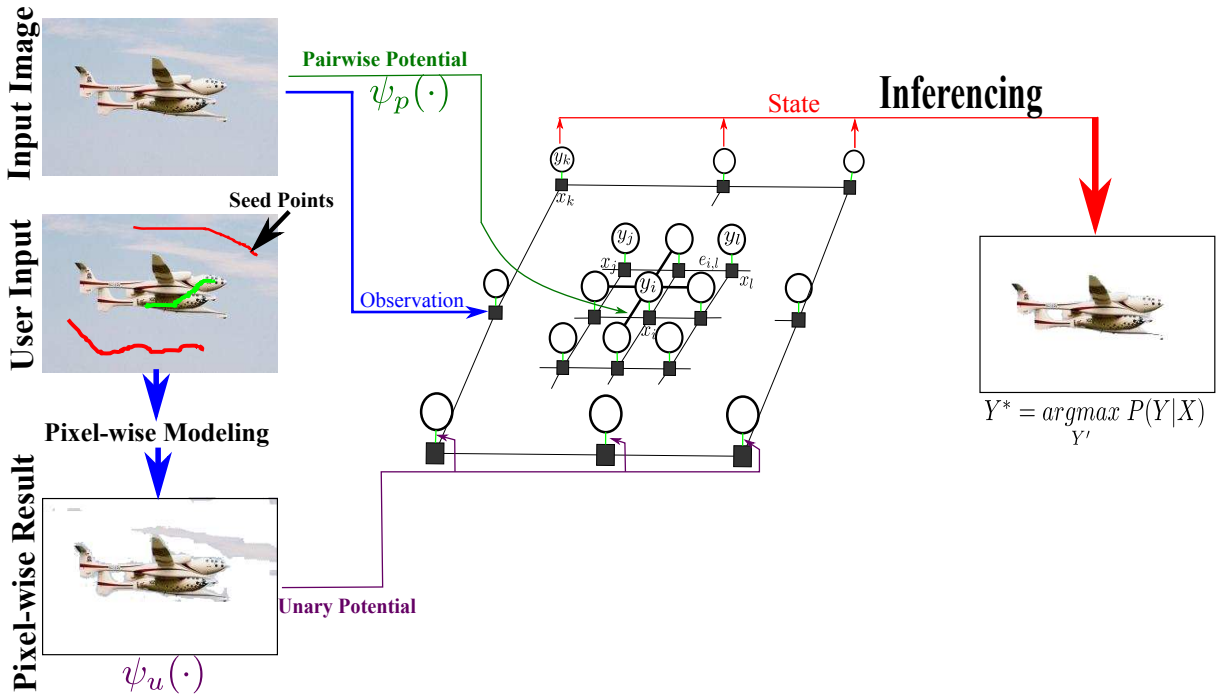


Figure 3.1: Problem definition flow-diagram. The unary potentials ($\psi_u(\cdot)$ in (3.1)) are computed based the user input (seed points) while the pairwise potentials ($\psi_p(\cdot)$ in (3.1)) are constructed based on the interactions among pixel intensities associated to the selected random variables in the clique structure and their corresponding states. As seen in the Figure, the image (a) is the input image which the user determined the regions corresponding to foreground and background with green and red colors. The Figure (b) shows the result of unary potential which here is a GMM model specifying each pixel belongs to background or foreground based the Gaussian models trained on seed points. Compared to the local random fields, here we optimize the model to satisfy two objectives: I) maximization of region F_1-score and II) maximization of boundary F_1-score. To achieve both objectives we introduce stochastic cliques which takes advantage of long-range connectivities while maintains the computational complexity by only using informative clique structures.

to be involved in the modeling and non-local interactions are incorporated in the random field modeling.

In an attempt to improve the modeling accuracy of the local MRF models, frameworks have been introduced that expand the clique structure to higher-order cliques [66, 83], as well as to introduce novel and effective penalty functions (i.e., the focus is on $\psi_p(\cdot)$ in (3.2)) in place of the standard Gibbs energy function [73] to better handle complex object structures. For example, in the work by Jegelka & Bilmes [73], the Gibbs energy was modified within the graph cuts optimization framework [42]. The smoothing issue was introduced as the short-boundary bias problem, resulting from the fact that penalizing the assignment of different labels to neighboring pixels leads to smooth segmentations in the standard pairwise models [15]. Although new potential functions address the short-boundary bias and overcome boundary smoothness, this approach has some drawbacks when dealing with situations characterized by background clutter.

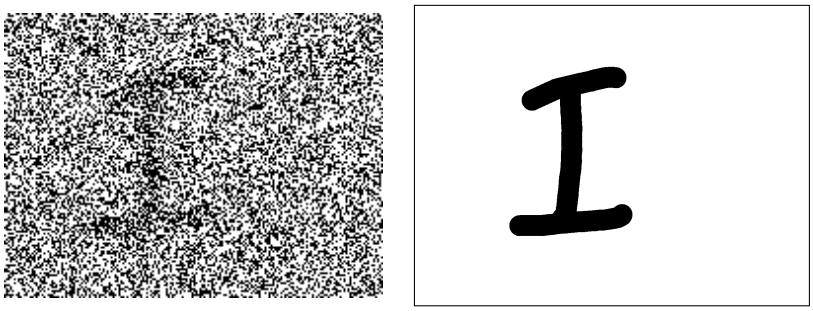
There are several inference frameworks that utilize fully-connected random fields for semantic image labeling [21, 131, 186]. By taking advantage of a large number of long-range connections (i.e., they focused on $\mathcal{C}$ in (3.2) however they extracted cliques deterministically), such methods have been shown to provide superior performance when compared to those with lower-order connectivities. However the complexity of inference in those initial inference frameworks using fully-connected conditional random field models limits their usage to only hundreds of nodes or fewer, as they become computationally intractable beyond such scenarios. To address the computational intractability issue of fully-connected conditional random fields, Krähenbühl and Koltun [89] proposed a tractable inference procedure by using specific potential functions. In their work, a fully-connected conditional random field (which we will refer to as FCRF) was presented to model the multi-class image segmentation problem, with the edge potentials obtained using Gaussian kernels (they focused on $\mathcal{C}$ while they reformulated $\psi_p(\cdot)$ in (3.2) in a form of Gaussian function). Based on these new feature functions, they formulated the inference problem as a filtering problem. However, the proposed methods are associated with a limitation that only specific potential functions can be used. This limitation restricts the effectiveness of CRFs

in modeling, as one of the key strengths of CRFs is the ability to utilize arbitrary feature functions.

Given the respective benefits and limitations of different streams of thought in addressing the smoothing issue in image modeling to better handle more complex objects, in this thesis we are motivated to investigate the marriage of random field theory and random graph theory to address the computational complexity of inference approximation on fully connected or long-range random field models. By combination of these two theories, the clique structure $\mathcal{C}$ in (3.1) is reconstructed with the proper number of cliques such that two mentioned objectives are optimized simultaneously. Before we introduce the proposed method, two intuitive demonstrations are provided to show the advantages and disadvantages of utilizing long-range interactions:

- As the first demonstration, it is shown in Figure 3.2 that fully connected interactions of all nodes (i.e., as discussed in Section 2.4.3 and (2.20)) during inference is not always useful via an experiment on a binary image classification (a binary classification is performed on every pixel in a noisy image using increasingly higher numbers of interactions). Figure 3.3 demonstrates the results based on F_1-score which is consistent with the above statement that long-range interactions are not always helpful. It is possible to draw a rough conclusion such that the effective range of interactions among random variables is highly dependent on the problem. However long-range connections and more information may be useful in modeling of several problems.

- Another simulation has been done to demonstrate the beneficial effects of long-range connections in certain situations. The problem is to classify each pixel of a noisy image to different classes. In this example, identical objects are distributed in the image. Since the identical objects are spatially distributed in the image, useful interactions corresponding to each node (pixels) are distributed in the image (i.e., the nodes with similar observation are distributed in the image). Valuable interactions are distributed in the whole image, therefore, long-range connectivities may be beneficial. Figure 3.4 shows the observation and the results of a CRF by incorporating different size of interaction. Figure 3.5 demonstrates the accuracy per number of



(A) (B)

Figure 3.2: Binary label classification image sample. The objective is to classify a noisy image to binary value (foreground and background) via (3.2) where the classification procedure results the maximum region F_1-Score while maintains the object boundary. (A) shows the observations X and (B) is the desirable result Y.

connections for this problem. As seen, the accuracy is increasing by adding more connections and it is consistent with the number of interactions.

As shown in Figures 3.3 and 3.5, problems are divided into two different categories:

1. The accuracy and effectiveness of the random field increases by adding more connections (e.g. Figure 3.5).

2. The long-range connections are not always effective (e.g. Figure 3.3) and there is a trade-off between the number of connections and the modeling accuracy.

Motivated by those examples and reported results in [21, 89], in this chapter we propose a new approach to take advantage of long-range interaction to some extent while maintains the computational complexity. As seen in Figure 3.1, the proposed framework assumes the random fields are fully connected random fields and all nodes are considered as neighboring nodes of other nodes in the random fields. However, a subset of cliques, $\mathcal{C}$, is considered in the inference step instead of all cliques in the fully connected random. Since the cliques are specified based on a distribution, this leads us to a new form of fully connected random fields where the generated cliques are stochastic in nature. To this end, we call this new

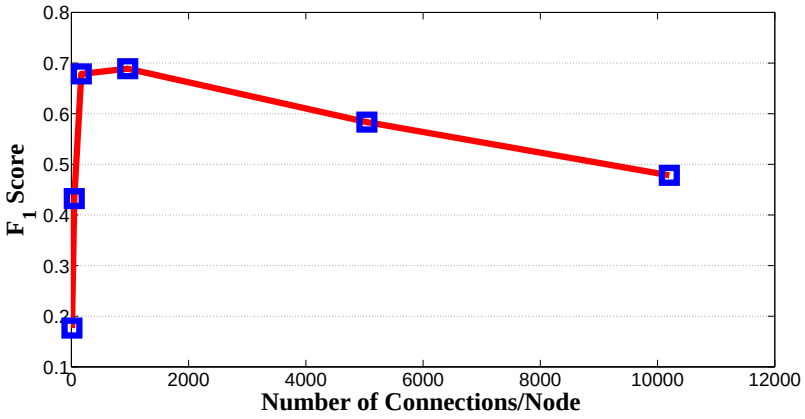


Figure 3.3: The region F_1-Score for binary image classification of noisy images (Figure 3.2) by CRFs with different interaction sizes. The curve shows that the accuracy increases by adding more interactions and peaking at a specific point that the number of useful interactions is smaller than the number of all connectivities in the graph. Increasing the number of connections more than the optimal value attenuates the model accuracy.

type of clique structure as stochastic clique. The proposed approach will be explained and analyzed in the following section.

3.3 Randomly-Connected Conditional Random Fields

In this section, we explain the concept of the proposed stochastic cliques in more details. The graph representation is demonstrated and the CRF model is derived via the novel concept of stochastic cliques structure. Considering the stochastic nature of the underlying clique structures in the new fully connected CRF model we coin the CRF model as randomly-connected conditional random field or short as RCRF.

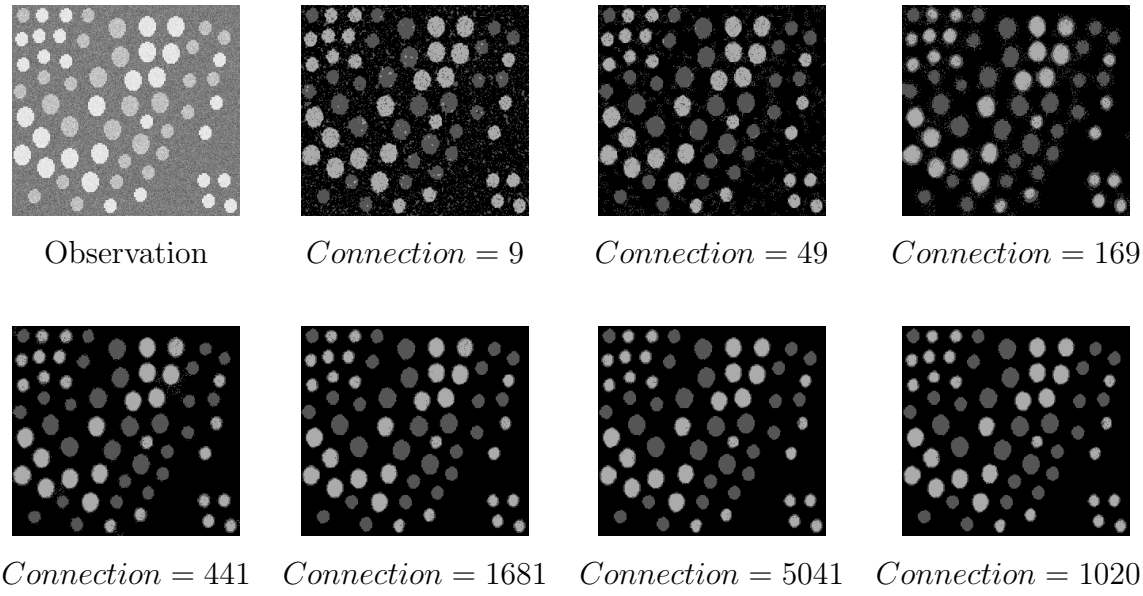


Figure 3.4: Multi-label image classification. The problem as discussed in Section 3.2 is to classify a noisy image to multiple labels. The upper left image shows the observation corrupted by noise and other images are the results of CRFs with different sizes of interactions. The inference step is applied with the same number of iterations in all CRFs. From left to right and top to bottom the number of interactions is increased in the CRFs model. As observed by the sequence of the results, the efficacy of the CRFs in classification is increased by incorporating more interactions in modeling since there are more useful information to incorporate into the model. It it worth to note that the similar objects in the provided observation (image) are distributed in the image such that long-range interactions most likely have positive effects on the modeling accuracy.

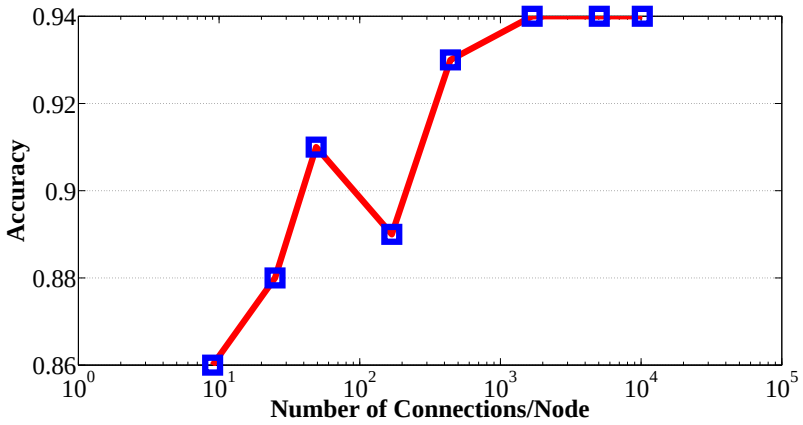


Figure 3.5: Accuracy of multi-label image classification. The accuracy is increased by adding more connections. Comparing the results with Figure 3.3 which is an example of local random fields structure, the performance of the CRF is increased by incorporating more interactions in the CRF model.

3.3.1 Stochastic Cliques

Randomly-connected conditional random fields (RCRF) are fully connected random fields in which cliques are defined stochastically. The term *fully connected* refers to the fact that each node in the graph can be connected to all other nodes of the graph, however the cliques for each node are determined based on distribution probabilities, so the number of pairwise cliques in the graph may not be the same as the number of neighborhood pairs.

The goal is to model $P(Y|X)$, the conditional probability of the state set Y given the measurement X. The conditional random field (CRF) approach to expressing $P(Y|X)$ is to write it as (3.1) where the potential function is the combination of unary and pairwise potential functions (3.1). The pairwise clique structure is the most regular cliques which is utilized to incorporate as the interactions among variables in the modeling; simplifying the formulations and relaxing the computational complexity are the main reasons of using

only pairwise cliques[1]:

$$\mathcal{C} = \Big\{C_p(i)\Big\}_{i=1}^{n} \tag{3.3}$$

$$C_p(i) = \Big\{(i,j)|j \in N(i), \mathbb{1}^S_{\{i,j\}} = 1\Big\}. \tag{3.4}$$

However, the formulation can be generalized for other clique structures. $C_p(i)$ for node i is determined based on a stochastic indicator neighbor function, $\mathbb{1}^S_{\{i,j\}} = 1$, to distinguish whether two nodes can construct a clique or not. This function itself is a combination of probability distributions. For image modeling, this function must consider the spatial relation among the nodes and must incorporate the observation information into the model; therefore, the proposed indicator function is the combination of spatially driven and data driven probabilities:

$$\mathbb{1}^S_{\{i,j\}} = \begin{cases} 1 & 1 - \Big(P^s_{i,j} \cdot Q^d_{i,j}\Big) \leq \gamma \cdot U(0,1) \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

$\mathbb{1}^S_{\{i,j\}} = 1$ has two responsibilities:

1. Incorporating the spatial information ($P^s_{i,j}$); The distance between two nodes in the underlying graph is utilized as one factor to determine the possibility of the connectivity between them. Here as we are dealing with spatial distances between pixels in the image, a simple Euclidean distance is utilized to model $P^s_{i,j}$:

$$P^s_{i,j} = \exp\Big(\frac{-\|u_i - u_j\|}{\sigma_p^2}\Big) \tag{3.6}$$

   where u_i and u_j determine the locations of nodes (random variables) i and j in the underlying graph and σ_p^2 is a controlling parameter.

2. Involving the data relation among the states ($Q^d_{i,j}$); The similarity of two nodes regarding to their associated observations is another factor which is used in the

[1]Readers are encouraged to study section 2.3.1 for more information about clique structures.

stochastic indicator neighbor function:

$$Q_{i,j}^{d} = \exp\Big(\frac{-\|x_i - x_j\|_2}{\sigma_q^2}\Big). \tag{3.7}$$

x_i encodes the color intensity of node i and σ_q^2 is a color controlling parameter which tunes the color similarity measure. It is worth to mention that L_2 norm between two pixels is a popular distance measure and also a simple one that has been utilized in the literature. This form of equation is selected here to have a simple formulation and to decrease the complexity of the problem.

The threshold γ in (3.5) determines the sparsity of the graph. The probability distributions $P_{i,j}^{s}$ and $Q_{i,j}^{d}$ are specified based on the problem.

3.3.2 Graph Representation

Graph $G(\mathcal{V}, \mathcal{E})$ is the realization of the RCRF where $\mathcal{V}$ is the set of nodes of the graph representing the states $Y = \{y_i\}_{i=1}^{n}$, and $\mathcal{E}$ is the set of edges of the graph where $|\mathcal{E}| \leq \frac{n(n+1)}{2}$. Corresponding to each vertex in the graph $G(\cdot)$, there is an observation $x_i \in X$. The edges in $G(\cdot)$ are randomly sampled, thus G is a realization of a random graph [24]. According to the Erdös-Rényi theorem [34] if the probability p' of the random graph $\hat{G}_{n,p'}$ is greater than $\frac{\log n}{n}$ the graph is connected with a high probability (Section 2.3.2). This property is important in the inference (i.e., Graph Cut) such that if the graph is not connected the inference result is trivial. As a result, the proposed graph $G(\cdot)$ is connected, has at least $n-1$ edges even for large values of γ, and satisfies Gibbs distribution [45] properties as well.

It is worth noting that the value of p_{ij} is very small if the random field is constructed for tackling problems where the number of random variables is large, such as the problem of image segmentation. As an example, for an image that is $n = 400 \times 300$, p_{ij} only needs to be greater than $\frac{\log n}{n} = 9.7460 \times 10^{-5}$ to satisfy the connectedness condition which corresponds to having 12 neighbors per pixel. As such, the connectedness condition is easily

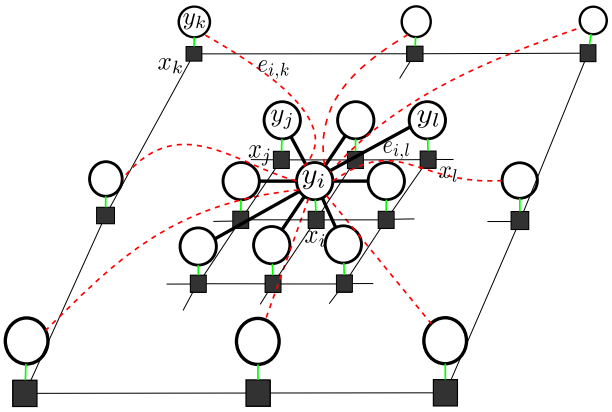


Figure 3.6: A visualization of a randomly-connected conditional random field graph. A connectivity between two nodes is determined based on a distribution; each two nodes in the graph can be connected according to a probability drawn from this distribution. There is a measurement x_i corresponding to each node y_i. The connectivity of each pair of two nodes y_i and y_k is distinguished by the edge $e_{i,k}$. Closer nodes are connected with a higher probability (black solid edges), whereas two nodes with a greater separation are less probable to be connected (red dashed edges).

satisfied for the purpose of image segmentation. Furthermore, it is possible to theoretically illustrate how many connectivities are needed to produce the optimal solution and to find the maximum number of connectivities which will be discussed in Section 6.2.1 as future work.

Figure 3.6 demonstrates an example of a RCRF. As seen, each node in the graph can be connected to all other nodes; However the connectivities of the centered node are highlighted to improve the visualization. The probability of connecting two nodes as a clique is different for each pair of nodes as shown in (3.5). Based on $P^s_{i,j}$, the connectedness probability of two nodes and the distance between them are inversely related. Nevertheless, there is a possibility for two distantly separated nodes y_i and y_k to be connected, as shown in Figure 3.6, which is how the RCRF takes advantage of the fully connected CRF.

By the amalgamation of random graph theory and random field theory, the proposed

RCRF might provide the merits of fully connected random fields by sampling the configuration of a fully connected random field which leads to a much smaller computational complexity than that of fully connected random fields. We will compared the proposed method with state-of-the-art fully connected random fields in Chapter 5.

3.4 Inference

There have been proposed several different approaches to compute inference over a random field models. Those frameworks can be divided into two distinct folds depends on the type of the outputs which are needed. In some situations, we need to evaluate the probability of a configuration of random variables with assigned values while in other situations we only need to find the best configuration (i.e., with highest probability) considering the underlying random field model. Finding the best configuration (the latter case) is the most popular problem [21, 83, 97, 134] such that we mostly care about the best result given the observation.

The inference methods can also be divided into I) Monte Carlo algorithms and II) variational algorithms [155]. Monte Carlo methods approximate the optimal result by sampling from the underlying distribution while variational inference approaches reformulate the problem to an optimization problem by changing the underlying distribution to a simple distribution that most closely matches to the original distribution.

Monte Carlo methods [135] iterate through random variables such that in each iteration they find the marginal distribution of a random variable considering others have fix states in the random field. This procedure is repeated until no significant change on states is observed and energy function reaches to optimal state. Gibbs sampling [45] is one of the most popular Monte Carlo methods known for inferencing on the random fields.

Variational inference [46, 85] utilizes a simpler yet similar distribution compared to the original one such that it makes the optimization procedure possible in terms of computing derivatives and also computational complexity. The variational Bayes inference [5] can be seen as an extension of the EM (expectation-maximization) algorithms [85] from maximum

a posteriori (MAP) estimation. Belief propagation (BP) [85, 114] is the most well-know variational inference approaches. The neighboring factors of each random variable makes a multiplicative contribution to the marginal of the interested random variable (known as a message) which each message is computed separately due to the tree structure constructed from the original graph [155]. Mean-field approximation [85], graph cut [71, 167] and beilef propagation [124, 154] are a subset of variational inference methods with different functionalities and properties.

In this thesis, we are more interested to take advantage of graph cut approach since our goal is image segmentation which the number of labels are limited and also finding the probability of the optimal solution is not important. We model the best configuration (optimal label set) as a maximum a posteriori problem and the inference of the MAP problem is solved by minimizing the energy function $\psi(\cdot)$ of the conditional probability distribution $P(Y|X)$:

$$\hat{Y} = \arg\min_{Y'} \psi(Y', X) \tag{3.8}$$

$\hat{Y}$ is the suboptimal configuration for states corresponding to observation X, and $\psi(\cdot)$ represents the energy function. The energy $\psi(\cdot)$ is minimized via formulating the underlying graph within a graph cut minimization.

3.4.1 Graph Cut

Graph cut [71, 167] formulates the energy minimization problems as finding the maximum-flow in a graph [19] such that the goal is to find those edge connectivities among nodes in the graph which can maximize the flow from the source node to the sink one. It is usually known as st-cut problem [86] which is the dual approach of the maximum flow problem.

A st-cut problem is to partition the vertices of V of graph $G(V, E)$ into two disjoint sets of S and T such that $s \in S$ and $t \in T$ and s and t are two terminal nodes. The cost of the cut is the summation of all edges that go from S to T:

$$c(S, T) = \sum_{i \in S, j \in T, (i,j) \in \mathcal{E}} c(i, j) \tag{3.9}$$

where i is a node in the set S, j is a node in set T and the pair (i, j) encodes an edge in the graph $G(\cdot)$. $c(i, j)$ is the cost of edge (i, j) in the graph $G(\cdot)$ which the goal of the st-cut problem is to find a cut with minimum cost. Ford and Fulkerson [41] proved that the st-cut algorithm is equivalent to maximum-flow from the source to the sink.

It is worth to note that the cut in a graph can be viewed as a binary partitioning of the graph where the nodes connected to the source are assigned with label "0" and those are connected to the sink are labeled by "1" (i.e., binary-valued labeling). Thanks to this intuitive similarity, it is very easy to formulate the image segmentation problem within a st-cut framework (graph cut).

Efficient global energy minimization algorithms for even the simplest class of discontinuity-preserving energy functions almost certainly do not exist and finding the global optimal solution is NP-Hard [20]. Due to this fact, there are several approximation algorithms to find the sub-optimal solution for the st-cut problem. The expansion move algorithms [86] are the well-known approximations for st-cut problems. The expansion move methods iterate through the possible labels α and change the label of random variables to find a lower energy. If the expansion move has lower energy than the current labeling, then it becomes the current labeling. The algorithm terminates with a labeling that is a local minimum of the energy with respect to expansion moves and there is no expansion move for any label with lower energy.

3.5 Example Problem

The proposed stochastic clique structure within a randomly-connected random field is examined with binary image classification as an example problem. However the comprehensive experiments will be conducted in Chapter 5.

3.5.1 Binary Image Classification

To demonstrate the power of the RCRF we performed experiments on binary classification datasets. It is worth to note that this experiment is for illustrative purposes and, therefore, the proposed method is not compared with other state-of-the-art algorithms. The first experiment studies the behaviors of local, long-range CRFs and the proposed RCRF frameworks on binary image classification. The utilized dataset is EnglishHnd, a set of handwritten characters [30], containing 3410 images grouped into 62 equally sized classes: 10 classes for digits, 26 classes for upper case letters, and 26 classes for lower case letters. We corrupted the given images with noise, and the goal is to classify the pixels of the noisy images as foreground and background. Salt & pepper noise at 80% and 90%, and Gaussian noise at 220% and 300%, where the Gaussian noise percentage is characterised by

$$\% = \left(\frac{\sigma}{\text{dynamic range}} \right) \times 100. \tag{3.10}$$

The images have a size of 480×360; therefore, the total number of pairwise connections of the fully connected graph is approximately 2.99×10^{10}. According to the random graph theory [34] mentioned earlier, the selection probability must be greater than 3.03×10^{-5} for the graph to be connected. The experiments were conducted with the selection probability 7×10^{-4}, leading to an expected number of pairwise cliques to be 2.09×10^{7} with an average of 121 pairwise cliques per node.

To test the effectiveness of the proposed framework, we compared our proposed RCRF against two other CRFs of different neighborhood sizes. Of the two compared approaches, the first is a regular CRF with adjacent neighbors (CRF-N3) where each node is connected to its closest eight neighbours (those within a 3×3 block); the second model, CRF-N11, has a larger neighborhood, where each node is connected to its closest 120 neighbors (those within an 11×11 block). The exact number of pairwise cliques of CRF-N3 is 1.38×10^{6} (8 pairwise cliques per node) and of CRF-N11 is 2.07×10^{7}(120 pairwise cliques per node).

All three methods (CRF-N3, CRF-N11 and RCRF) are implemented in Matlab, whereas the potential calculation was computed in C++ code integrated with Matlab through the

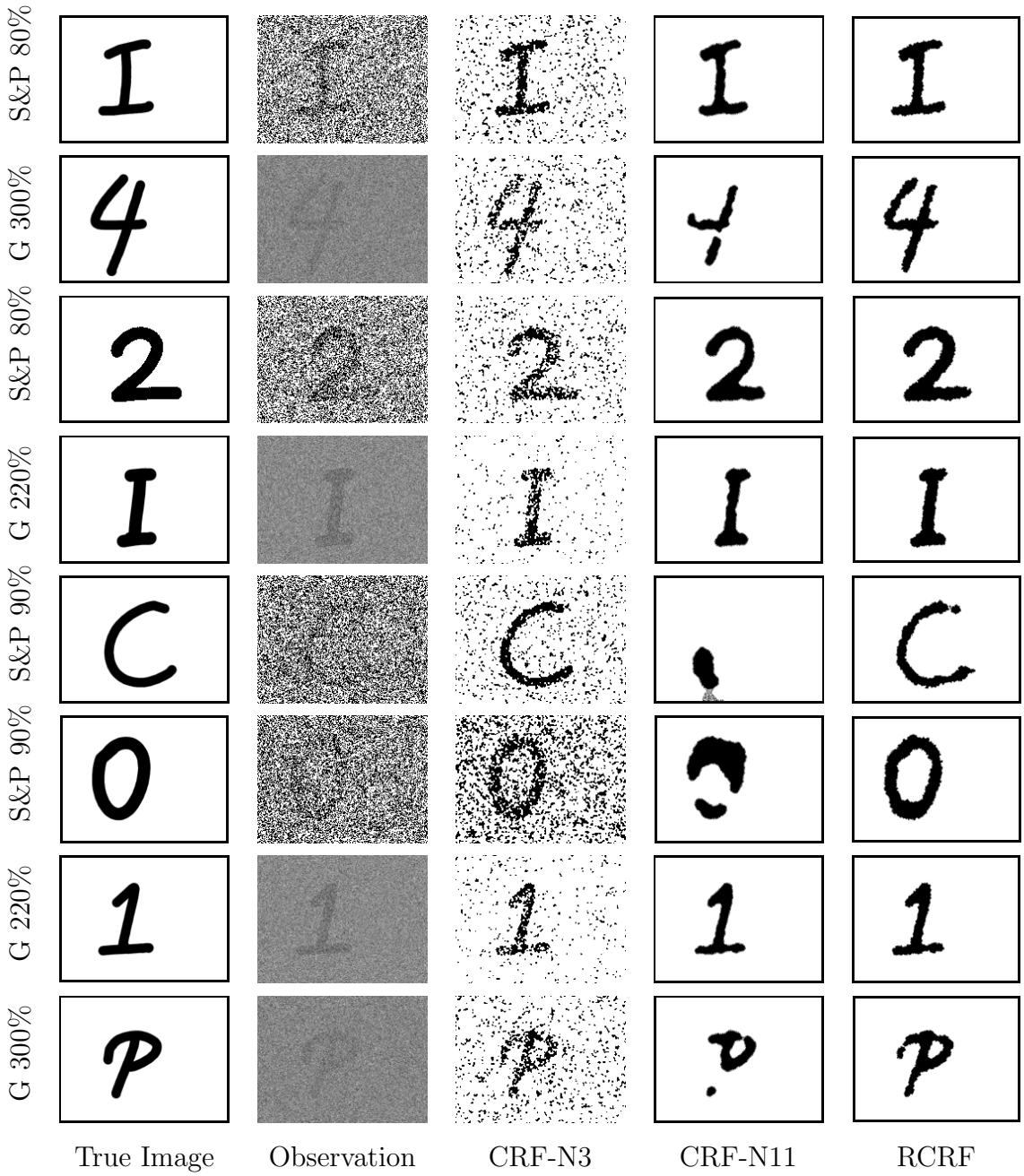


Figure 3.7: Qualitative results of RCRF; the proposed algorithm is examined based on two noise types with two strengths. The results clearly show how the RCRF outperforms both local and non-local CRFs.

Table 3.1: Quantitative results (F_1-score) based on the EnglishHnd dataset [30]. The proposed framework is examined by two noise types with two different levels. The RCRF is compared with the regular CRF (CRF-N3) and a CRF with a neighborhood size of 11 (CRF-N11). The per-iteration run time of each method is reported; all methods were run with an equal number of iterations.

	CRF-N3	CRF-N11	RCRF
Salt & Pepper (80%)	0.488	0.872	0.931
Salt & Pepper (90%)	0.235	0.313	0.859
Gaussian (220%)	0.566	0.818	0.895
Gaussian (300%)	0.391	0.646	0.842
Time per Iteration (s)	0.04	3.85	2.70

Mex interface. The average computational time for each iteration of the inference step is $0.04s$ for CRF-N3 and $3.85s$ for CRF-N11, a significant difference caused by the change in degree of connectivity between the two models. In contrast, the average runtime per iteration for the RCRF configuration is $2.7s$. Thus the inference time is decreased, while the flexibility in edge connectivity, in principle allowing arbitrarily distant connections, is increased based on this new clique structure.

Table 3.1 shows the F_1-score for the RCRF and other CRFs subject to the stated noise. The ground truth labels are obtained by binarizing the true images. The results show that the proposed RCRF outperforms the regular CRFs in all cases.

Figure 3.7 shows example results of the EnglishHnd dataset. As seen RCRF can classify the images even when they are distorted by the high level of the noise (i.e. 300%).

3.6 Summary

In this chapter we proposed a new clique structure, stochastic clique, which determines the set of cliques incorporated in the inference stochastically. The stochastic cliques are constructed via a stochastic process such that the useful connectivities in the inference step

are determined based on a probability distribution. Followed by that direction we propose a new random field structure which the clique connectivities from the underlying graph are extracted based on the stochastic cliques to address the computational complexity fo fully connected random fields. The proposed random field called randomly-connected random field (RCRF) is the marriage between random graph theory and random field theory such that the underlying graph of the random field is fully connected while useful cliques involved in the inference step are much less than the original fully connected random fields.

The performance of the proposed RCRF structure was examined on binary image classification as an example problem. The results showed the advantages of the proposed method compared to local random field models.

The inference of the proposed RCRF framework has been done by the graph cut minimization. However inference using graph-cuts is NP-hard thus the graph cut solution is necessarily approximate. Increasing the number of connectivities can make the problem more complex and might result to poor approximation. In the next chapter we aim to propose an extended framework that allows for the number of clique formations to scale without incurring an exponential increase in the computational complexity, while simultaneously supporting graph cut in producing solutions that are closer to the optimal solution than those achieved by RCRF.

Chapter 4

Deep Randomly-connected Conditional Random Field

"Look deep into nature, and then you will understand everything better."

Albert Einstein

4.1 Introduction

As discussed in chapters 2 and 3, structured inference where the goal is to infer a structured states output from a structured observation input, is a crucial component for a wide range of applications such as object recognition [127], image classification [119], natural language processing [180], gesture recognition [173], handwriting recognition [37], and bioinformatics. A powerful and commonly-used approach to structured inference is the use of Markov random field (MRF) and conditional random field (CRF) [98] models. A limitation of such graphical models is that they utilize unary and pairwise potentials on local neighborhoods only, and as such can result in smoothed state boundaries as well as prohibit long-range state boundaries given the limitations of constraint locality. This becomes particularly problematic in the presence of high observational uncertainties such as measurement noise and outliers.

Recently there has been significant interest in the application of two types of models for the purpose of structured inference that help address the issues associated with locally-connected graphical models: i) fully connected graphical models, and ii) deep-structured graphical models.

In two previous chapters we focus on fully connected random fields. We proposed the concept of stochastic clique structure which brings random graph theory and random field frameworks together to address the computational complexity of fully connected random fields. The proposed randomly-connected conditional random field benefits the long-range connectivities and interactions by use of the informative connectivities instead of all interactions in the graph. Utilizing this approach makes the computational complexity more feasible. Preliminary results demonstrates the potential of the proposed framework in structural modeling.

Although fully connected random fields address the local random fields' drawbacks in excessive smoothing of object boundaries, each fully connected approach (i.e., FCRF or RCRF) suffers by its own limitations. FCRF frameworks are limited to utilize Gaussian shape feature functions while RCRF has computational complexity issue when dealing with large random field yet. The graph cut inference framework is suboptimal and increasing

the number of connectivities makes it weaker in fining the best solution due the NP-hard property of the problem. In this chapter we extend the idea of RCRF framework to boost its modeling accuracy and computational complexity. We explain the notion of deep structures and propose a deep random fields to address image segmentation problem.

4.2 Deep Structures

Fully-connected graphical models address issues of locally-connected models by assuming full connectivity amongst all nodes in the graph, thus taking full advantage of long range relationships to improve inference accuracy. One of the main hurdles in utilizing fully-connected graphical models is the complexity of inference, which becomes computationally intractable as the size of the problem scales. Much of recent research in fully-connected graphical models have revolved around addressing the computational complexity of inference step. Krähenbühl and Koltun [89, 90] introduced an efficient inference procedure for fully-connected CRF based on specific potential functions which was previously explained in Section 2.5.

Nevertheless, while the aforementioned methods (FCRF method and its extensions as explained in Secion 2.5) significantly reduces the computational complexity of inference on fully-connected graphical models, they all address the problem similarly by defining specific potential functions to manage the inference as a filtering approach, thus limiting the effectiveness of such models as the one merit of such models is to allow for arbitrary feature function selection.

The proposed RCRF framework is associated with two limitations. First, while the computational complexity of the inference process is greatly reduced, the computational complexity of forming stochastic cliques is relatively high, thus resulting in a relatively higher overall computational complexity compared to the first direction. Second, to reduce the computational complexity of forming stochastic cliques, a weak spatial proximity is imposed when forming the stochastic cliques, in which nodal interactions are formed with decreasing probability with increasing spatial distance. As such RCRF does not

leverage long-range nodal interactions completely. Therefore, an approach allowing for high inference performance using fully-connected conditional random fields that relaxes restrictions on potential functions compared to the FCRF, without imposing spatial proximity is highly desired. In this chapter we investigate the feasibility of deep structure to address the mentioned drawbacks.

Deep-structured graphical models take a different approach to improve the inference performance by introducing intermediate state layers, where there is a dependency of each higher layer on its previously layer, and inference is carried out in a layer-by-layer manner from bottom to top. Prabhavalkar and Fosler-Lussier [125] and Peng *et al.* [123] both introduced multi-layer conditional random field models where the local factors in linear-chain conditional random fields are replaced by multi-layer neural networks and trained via back-propagation. Ratajczak *et al.* [134] introduced a context-specific deep conditional random field model by replacing the local factors in linear-chain conditional random fields with sum-product networks. Yu *et al.* [180, 179] proposed a deep-structured conditional random field model which consists of multiple layers of simple CRFs where each layer's input consists of the previous layer's input and the resulting marginal probabilities. While such deep-structured graphical models are good at handling high observational uncertainties such as measurement noise and outliers by characterizing different information at the different layers, they only implicitly take advantage of long range relationships and are more limited in this aspect when compared to fully-connected graphical models.

While fully-connected and deep-structured graphical models both have their own benefits and limitations, these two types of graphical models have been largely explored independently, leaving the unification of these two concepts ripe for exploration. Such a unified graphical model could yield significant benefits in improving state boundary preservation, better enable long-range state boundaries, and better handle high observational uncertainties such as measurement noise and outliers. A fundamental challenge with unifying these two types of graphical models is in dealing with computational complexity, as not only are all nodes fully-connected within a layer, there are also multiple layers to process due to the deep structure of the graphical model.

In this chapter, we investigate the feasibility of unifying fully-connected graphical models and deep-structured models in a computationally tractable manner for the purpose of statistical inference. To accomplish this, we propose a deep randomly-connected conditional random field (DRCRF) which extends upon RCRF through the introduction of a deep structure representation to address the sub-optimality of the utilized graph cut approach when applied to RCRF. Inference using the model proposed in [147] via graph-cuts is NP-hard, thus the graph cut solution is necessarily approximate. In contrast, the proposed DRCRF framework leverages a novel deep, multi-layer architecture that allows for the number of clique formations to scale without incurring an exponential increase in the computational complexity, while simultaneously supporting graph cut in producing solutions that are closer to the optimal solution than those achieved in [147]. By distributing the range of stochastic connectivities (i.e., stochastic cliques) in several layers, the computational complexity of graph cut method and as a result the sub-optimality is addressed to some extent and as a result the new model can lead to more accurate modeling.

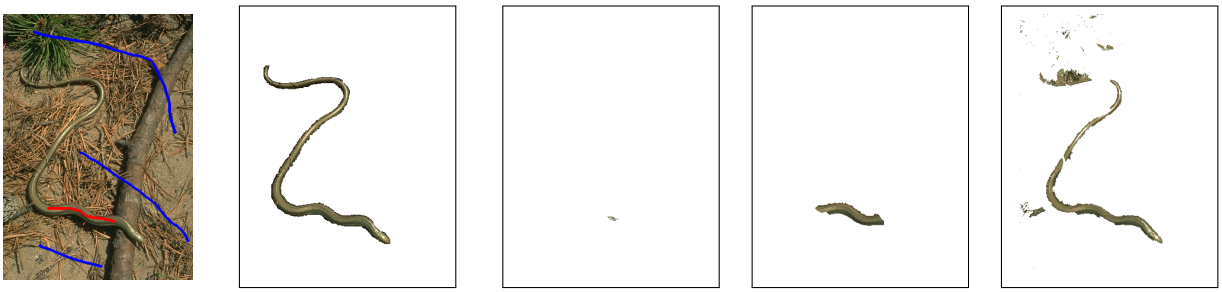
4.3 Deep Randomly-connected Conditional Random Fields

The main limitation of conventional random field models is in their ability to take advantage of long-range connections within the image and generally result in excessive smoothing of object boundaries [52, 168]. In an attempt to address this important issue, frameworks have been introduced that expand the clique structure to higher-order cliques [66, 83], as well as to introduce novel and effective penalty functions in place of the standard Gibbs energy function [73] to better handle complex object structures. For example, in the work by Jegelka & Bilmes [73], the Gibbs energy was modified within the graph cuts optimization framework [42]. The smoothing issue was introduced as the short-boundary bias problem, resulting from the fact that penalizing the assignment of different labels to neighboring pixels leads to smooth segmentations in the standard pairwise models [15]. Although new potential functions addresses the short-boundary bias and overcome the

boundary smoothness, this approach has some drawbacks when dealing with situations characterized by background clutter. Figure 4.1 demonstrates an example of such images. In these situations, the number of types of label discontinuities is more than usual and the penalty function has a negative effect. The aforementioned drawback is also encountered in corrupted and distorted images, as well as images having complex textures. In these aforementioned situations, there are varying types of discontinuities and there is a possibility that a background discontinuity and a boundary discontinuity may be grouped into the same cluster, leading to an incorrect labeling. As a result, image segmentation of complex object structures remains an open challenge.

More recently, there have been two main streams of approaches that aim at addressing the smoothing issue to better handle more complex object structures. One stream is to introduce inference frameworks that utilize fully-connected random fields for semantic image labeling. By taking advantage of a large number of long-range connections, such methods have been shown to provide superior performance when compared to those with lower-order connectivity. However the complexity of inference in those initial inference frameworks using fully-connected conditional random field models limits their usage to only hundreds of nodes or fewer, as they become computationally intractable beyond such scenarios. To address the computational intractability issue of fully-connected conditional random fields, Krähenbühl and Koltun [89] proposed a tractable inference procedure by using specific potential functions. In their work, a fully-connected conditional random field (which we will refer to as FCRF) was presented to model the multi-class image segmentation problem, with the edge potentials obtained using Gaussian kernels. Based on these new feature functions, they formulated the inference problem as a filtering problem.

Following [89], Zhang and Chen [185] relaxed the Gaussian assumption to any distribution by using a stationarity constraint. More statistical information was encoded by different distributions since they showed that the spatial potentials over two pixels depend only on their relative positions. Campbell *et al.* [21] generalized the pairwise potentials to a non-linear dissimilarity measure, such that the pairwise terms are encoded by the density estimates of the conditional probability, with the probabilities expressed by a dissimilarity measure. A continuous FCRF was proposed by Ristovski *et al.* [134], similar to [89], but



(a) Image (b) Ground Truth (c) FCRF (d) PD (e) DRCRF

Figure 4.1: Interactive image segmentation: (a) User-specified marked area; (b) Ground truth segmentation; Segmentation results of (c) FCRF [89], (d) principled deep random field [82], and (e) DRCRF. A GMM is trained by the user-specified areas corresponding to the object and background. The blue and the red areas show the background and the object seeds, respectively. It can be observed that the principled deep random field cannot find the optimal solution since the number of types of color dissimilarity is more than usual and the background is very complex. The FCRF cannot find the optimal boundary since the foreground and background have similar color distributions. The power of the DRCRF is to select informative clique connectivities in long-range distances. The image is selected from [3] for illustrative purposes.

targeting the regression problem with continuous outputs. By approximating the inference on a FCRF model using the aforementioned frameworks, the computational complexity of inference using a FCRF is reduced from $O(N^3)$ to near linear complexity, making it computationally tractable to solve, but with a corresponding limitation that only specific potential functions can be used. This limitation restricts the effectiveness of CRFs in modeling, as one of the key strengths of CRFs is the ability to utilize arbitrary feature functions.

The second stream introduces inference frameworks that utilize deep random fields for semantic image labeling [95, 97, 82]. Such approaches leverage deep, multi-layer graphical architectures to take advantage of higher-order potentials within the model. As such, a fundamental difference between fully-connected random fields and deep random fields is

that, while fully-connected random fields consider all possible pairwise connections in the inference process to enable higher-order connectivity, deep random fields consider only a subset of connections at each layer but rely on inter-layer connections within the deep graphical architecture to represent higher-order interactions. For example, Kohli *et al.* [82] proposed a deep, multi-layer pairwise model with hidden auxiliary random variables [70] for representing useful higher-order interactions, and derived an exact yet practical algorithm for MAP inference using this model. The advantage of this deep modeling approach is that it mitigates the computational tractability problem while allowing for the use of arbitrary feature functions which allows for greater modeling flexibility. While reported results [82] demonstrate the superiority of this new deep modeling approach over conventional approaches, and have been shown to provide state-of-the-art performance in image segmentation, they only implicitly take advantage of long range relationships and are more limited in this aspect when compared to fully-connected graphical models.

Given the respective benefits and limitations of two different streams of thought in addressing the smoothing issue to better handle complex objects, in this work we are motivated to investigate the marriage of fully-connected random fields and deep random field inference frameworks to achieve computationally tractable inference approaches that are well suited to the segmentation of complex objects. Here, we propose a new inference framework based on the concept of deep randomly-connected conditional random fields (DRCRF), which allows for computational tractability for inference without limiting the use of specific feature functions. Leveraging the idea of stochastic cliques, first introduced in [147], we introduce a deep graphical structure consisting of multiple layers of RCRFs where the clique connectivity is determined randomly to take better advantage of long-range interactions while maintaining computational tractability.

4.4 DRCRF Methodology

The main benchmark in this thesis is image segmentation, therefore, let us first formulate the image segmentation problem using a conditional random field (CRF) model. Let X

denote the set of pixels of image I being segmented and C denote the set of all pairwise interactions of the CRF model. The posterior probability of $P(Y|X) = \frac{1}{Z}\exp(-\psi(Y,X))$ is the factorization of unary and pairwise potential functions where Y is the set of random variables representing the label of each pixel $i \in X$:

$$\psi(Y,X) = \sum_{i \in X} \psi_u(y_i, X) + \sum_{c \in C} \psi_p(y_i, y_j, X) \tag{4.1}$$

where $\psi_u(\cdot)$ is the unary potential and $\psi_p(\cdot)$ represents the pairwise interaction between two nodes i and j. Since the model relies on pairwise interaction, $c = \{i, j\}$ represents a pairwise clique, $\psi_p(\cdot)$ a binary-variable function which is the contrast sensitive prior, and $\psi_u(y_i)$ the likelihood of pixel i to each class label. The pairwise potential $\psi_p(\cdot)$ plays the role of a penalizing function, the cost of assigning different labels to pixels i and j based on color similarity.

The inference of the Maximum a Posterior (MAP) solution of a CRF model $P(Y|X)$ can be formulated as minimizing the corresponding energy $\psi(\cdot)$. It has been shown [18] that the energy function can be minimized by graph cuts in polynomial time when the penalty function is non-negative for all configurations of i and j. However, the inference of a high-order random field is NP-hard [82]. As a result, each node i usually interacts with only 4 or 8 other neighboring nodes, depending on the order of the Markov assumption in the random field model. In other words, each node i is contributing to at most 8 different cliques c. Thus, the pairwise potential penalizes the assignment of different labels only to neighboring pixels, leading to smoothed segmentation results. There has been strong recent evidence [97, 83, 66] that increasing the number of model interactions can attenuate this smoothness effect, with the extreme case being fully-connected interactions [131], which are computationally intractable in general.

Although the framework proposed by Krähenbühl and Koltun [89], along with subsequent frameworks [185, 21] allow the inference of a fully-connected random field to be computationally tractable, they are limited to the use of specific potential functions which limit modeling flexibility. Furthermore, it was found in the experimental results conducted in [89] that classification accuracy peaked when not all possible interactions are involved in

the inference process. These issues become significantly amplified for deep, fully-connected conditional random fields, making the combination of fully-connected random fields and deep random fields intractable using conventional approaches.

We are motivated to tackle the challenge of achieving computational tractability inference from a different perspective, with the aim of retaining the powerful ability of CRFs within a deep, fully-connected graphical model to use arbitrary feature functions. We focus on the graphical structure of the deep, fully-connected conditional random field itself, in which all possible pairwise cliques between nodes exist. Accounting for all such interactions in a direct manner within a deep, fully-connected conditional random field leads to an inference problem that is computationally intractable, however studying the problem from this perspective leads to an interesting idea: What if the pairwise cliques take shape in a random manner, coming into and out of existence with a certain probability? This idea of a deep, randomly-connected conditional random field (DRCFR) relates directly to random graph theory [16], where graph nodes are connected based on some probability distribution. This new perspective allows us to retain the benefits of long range interactions and to account for them with arbitrary feature functions within a deep, fully-connected graphical model, while achieving computational tractability since the probability of clique formation can now be controlled via the choice of probability distribution.

Furthermore, the proposed deep, randomly-connected conditional random field approach has significant benefits over the preliminary work on stochastic cliques [147] by leveraging graph cut, one of the best–known frameworks used for random field inference. It was shown [54] that graph cut can compute the exact optimal solution in polynomial time within specific constraints:

1. The problem is a binary labeling problem,
2. The observation values are also binary.

These constraints can be relaxed [71], and it was proved that the exact minimum can also be found efficiently via graph cuts when $\psi(y_p, y_q) = |y_p - y_q|$ and Y is a finite 1D set where $\psi(\cdot, \cdot)$ is the pairwise potential and y_p is the state for node p. Nevertheless graph

cut minimization remains an NP-hard problem [167] in general frameworks. Specifically Zeng *et al.* [182] proved that on a planar 2D grid, when the foreground is "4-connected" and the background is "8-connected", the graph cut problem is NP-hard. Following upon that proof, it is clear that optimization with the model proposed in [147] by graph cuts is similarly NP-hard, since the connectivity for each node is varying in the graph and graph cut necessarily results in an approximation. It has furthermore been demonstrated by Juan and Boykov [75] that having a better initialization can help graph cut find the best solution faster and closer to the optimal solution. By introducing a deep fully-connected conditional random field where each layer acts as an initialization for its successor layer, such a model supports graph cut in producing solutions that are closer to the optimal solution than achieved in [147].

4.4.1 Graph Representation

Graph $G(\mathcal{V}, \mathcal{E})$ is the realization of the DRCRF where $\mathcal{V}$ is the set of nodes of the graph representing the states $Y = \{\bar{y_i}\}_{i=1}^{n}$, $\bar{y_i} = \{y_i^t | t = 1, ..., L\}$ and $\mathcal{E}$ is the set of edges of the graph with $|\mathcal{E}| \leq L \cdot \frac{n(n+1)}{2}$ for L graph layers and n nodes. Corresponding to each set of vertices $\bar{y_i}$ in the graph $G(\cdot)$ is an observation $x_i \in X$. The edges in $G(\cdot)$ are randomly realized via the stochastic indicator function, and thus $G(\cdot)$ is a realization of a random graph [34]. Based on the Erdös-Rényi theorem [34], if the probability p' of the random graph $\hat{G}_{n,p'}$ is greater than $\frac{\log n}{n}$ the graph is connected with a high probability. As a result, the proposed graph $G(\cdot)$ is connected, has at least $n-1$ edges in each layer, even for large values of γ, and satisfies a Gibbs distribution [45].

Figure 4.2 illustrates an example of a DRCRF, where the graph exhibits a deep, multi-layer structure composed of L layers, with each layer encoding a random field. As shown, each node in the graph is connected to all other nodes in one layer (the connections of the centered node in layer L are highlighted in Figure 4.2 to improve the visualization). The probability of two nodes forming a clique is different for each pair of nodes. According to $P_{i,j}^{s}$, the probability of two nodes forming a clique is inversely proportional to their distance from each other. However, there is a possibility for two distantly separated nodes

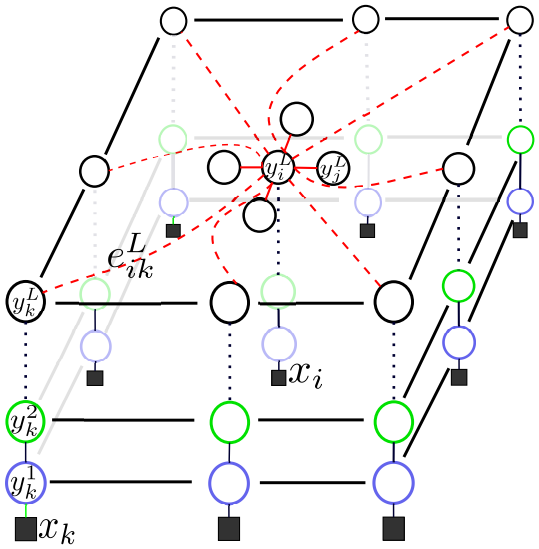


Figure 4.2: A realization of a deep randomly-connected conditional random field graph. Clique formation between two nodes in each layer is determined based on a stochastic clique indicator (i.e., the cliques can be different from one layer to another). There is a measurement x_i (pixel value) corresponding to each set of nodes $\bar{y}_i = \{y_i^t | t = 1, \ldots, L\}$. Nodes in closer spatial proximity form cliques with a higher probability (red solid edges, e.g., between y_i^L and y_j^L), whereas two nodes with a greater separation are less likely to form cliques (red dashed edges). The probability of clique formation varies in each layer, where the lower layers are designed to better explore the informative cliques in long-range interactions while the upper layers are design to exploit more local information. The clique formations for the center node in the last layer (L) are shown here for illustrative purposes only.

y_i^L and y_k^L to form cliques, as illustrated in Figure 4.2, and as such the DRCRF takes strong advantage of long-range interactions while significantly reducing inference complexity.

Each layer of $G(\cdot)$ encodes a separate random field where clique formations are determined by the stochastic clique indicator function. The clique formation probabilities varies in each layer of the graph: the lower layers are designed to better explore the informative cliques in long-range interactions, while the upper layers are design to exploit more lo-

cal information. Therefore, by combining random graph theory and random field theory within a deep, multi-layer structure, the proposed DRCRF provides the benefits of fully-connected conditional random fields and deep random fields while enabling computational tractability for inference.

4.4.2 MAP Inference

The inference of the MAP problem is solved by minimizing the energy function $\psi(\cdot)$ of the conditional probability of $P(Y|X)$. The goal here is to find the optimal solution Y^* which maximizes the conditional probability $P(Y|X)$. However designing a probability model to produce an optimal solution is not trivial, therefore, the goal here is to find the approximate solution $\hat{Y}$ instead of optimal solution Y^* by introducing long-range connectivities to the random fields:

$$\hat{Y} = \arg\max_{Y'} P(Y'|X) = \arg\max_{Y'} \Big[\frac{1}{Z}\exp\Big(-\psi(Y', X)\Big)\Big]. \tag{4.2}$$

Designing a random field with sufficient long-range connectivities to address the short-boundary bias problem has two issues: I) there is no known optimal energy function over the random field to address the short-boundary bias issue. II) As mentioned before increasing the number of connections in the graph increases the computational complexity which makes the inference more difficult. To relax these two problems, we proposed a deep structure of randomly connected random field (DRCRF) such that the goal is to find the solution $\hat{Y}$ for DRCRF which is close to the optimal solution Y^*. To this end, the DRCRF framework is minimized in a way that the result of the last layer of the deep structure model, $\hat{Y}^L$, is a suboptimal result close to Y^*:

$$min \left\| \hat{Y}^L - Y^* \right\| \tag{4.3}$$

such that

$$\hat{Y}^L = \arg\max_{Y'} P^L(Y'|X) \tag{4.4}$$

where $P^L(\cdot)$ is the conditional probability distribution over the random field of layer L.

The inference for each layer was conducted via a graph cut approach which is suboptimal. Due to this drawback, we applied a piecewise linear optimization to find the best approximate for the final segmentation result $\hat{Y}$. The energy function associated to each layer is minimized independently while in each layer the minimized result of the preceding layer is utilized as the initialization of graph cut approach in the present layer:

$$\hat{Y}^t = \mathcal{S}^t(Y^t, X; \hat{Y}^{t-1}) \tag{4.5}$$

where $\mathcal{S}^t(Y^t, X; \hat{Y}^{t-1})$ is the graph cut approach over random field Y^t given the graph structure at layer t, observation X and the initialization $\hat{Y}^{t-1}$. A different extent of connectivities is incorporated into each layer of the framework, such that graph cut can deal with only a limited number of pairwise connectivities each time.

The final result is approximated by finding the best solution of $\psi^L(\cdot)$ which is the result of the last layer:

$$\hat{Y} \equiv \hat{Y}^L = \mathcal{S}^L(Y^L, X; \hat{Y}^{L-1}). \tag{4.6}$$

In each layer t the energy function $\psi^t(\cdot)$ is minimized and the next factor, $\psi^{t+1}(\cdot)$, will be minimized by considering the optimal value of $\psi^t(\cdot)$. By using this approach a good approximation of $\psi(\cdot)$ is resulted by minimizing each factor of $\psi^t(\cdot)$ step-by-step.

4.5 DRCRF Layer-wise Analysis

In this section, we study the effects of introducing a deep structure representation into RCRF on overall modeling accuracy using an example, thus better illustrating the efficacy of the proposed DRCRF approach. It is worth to mention that this example is for illustrative purposes and comprehensive experiments are provided in Chapter 5. More specifically, we demonstrate the behaviour at different layers of a DRCRF model for the task of interactive image segmentation on noisy images. Figure 4.3(a) and (b) shows the true image and the noisy image corrupted with 25% Gaussian noise (as explained in Section 3.5.1).

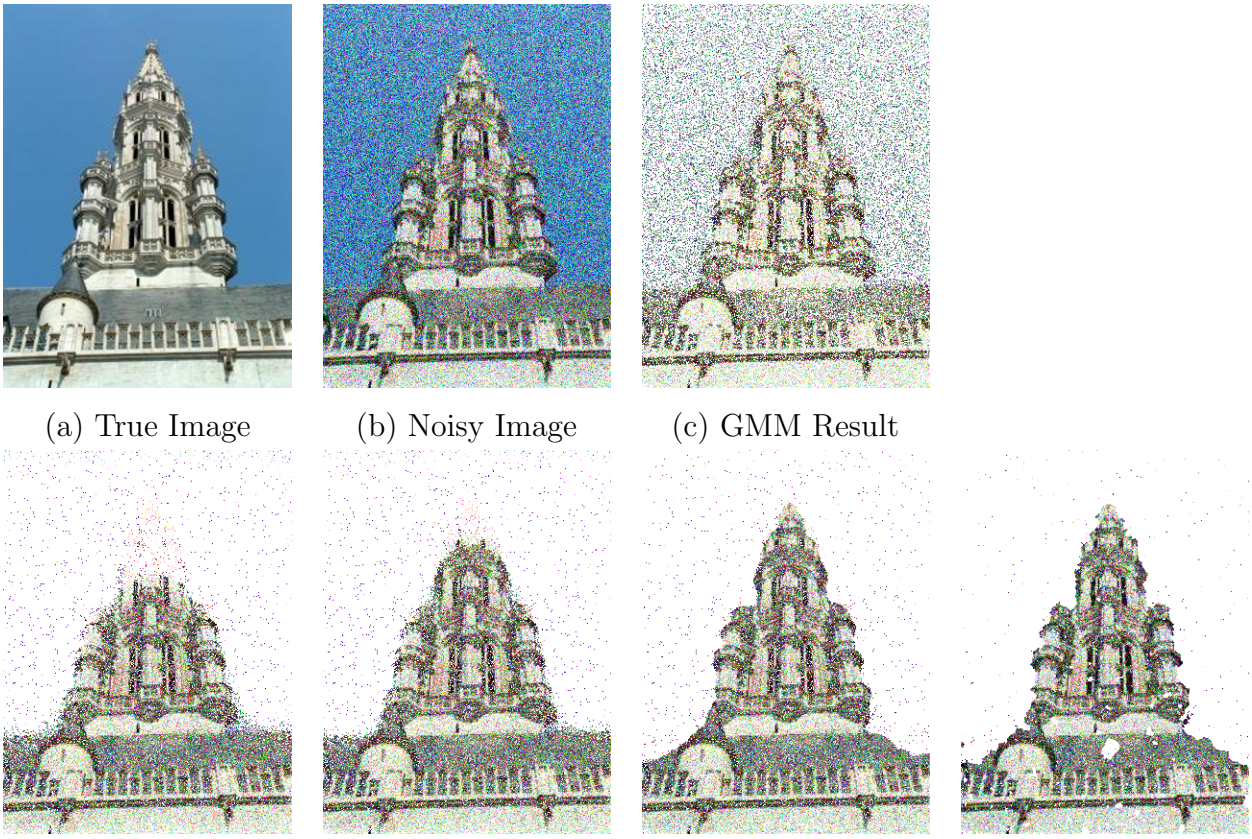
The unary potential is constructed via a GMM model and based on user's seed information within the context of interactive image segmentation[1] as shown in Figure 4.3(c). We classify the pixels of the image as foreground and background by formulating the problem via a deep structure randomly-connected random field framework. To show the effect of long-range connectivity and deep model approach, a 4-layer DRCRF model was utilized to segment the foreground from the background. Results of each layer of the deep model are demonstrated in Figure 4.3 (d)–(f). As seen by increasing the number of layers the framework can segment the foreground from the background more accurately. The combination of layers with extracting different ranges of information allows DRCRF to preserve object (foreground region) more accurately through layers.

As evident by the example, the first layer of DRCRF segments the main regions belong to foreground which can be considered as the result of a specific RCRF method. This result is an initial point for the second layer such that the second layer tries to optimize the energy from the initial point and gets a better sub-optimal result which preserves more areas of foreground. The initial point helps the graph cut framework to find a better direction to the optimal result. This process continues until the fourth layer which results the best possible distinction of the foreground and the background via DRCRF.

4.6 Summary

We addressed the drawbacks of RCRF method within a deep framework (DRCRF). There are two main drawbacks associated to the proposed RCRF framework including the exponential computational complexity of graph cut method as function of number of connectivities and the sub-optimality and limitedness of the neighborhood ranges in the RCRF approach. Here in this chapter, by proposing a new deep framework which applied a specific range of connectivities in each layer we relaxed the computational complexity of graph cut approach. By use of DRCRF method, it is possible to take advantage of more connectivities while it is possible to obtain the optimal result by a graph cut algorithm.

[1]For more information regarding the interactive image segmentation process, please refer to Chapter 5.



(a) True Image (b) Noisy Image (c) GMM Result

(d) First Layer (e) Second Layer (f) Third Layer (g) Fourth Layer

Figure 4.3: Interactive image classification result for noisy image. The image (a) is corrupted with 25% Gaussian noise (b). A GMM model is utilized as unary potential as shown in (c) The results of three different layers of a DCRF model are demonstrated. As seen by increasing the number of layers the model can segment the foreground from the background more accurately.

Results showed that having a deep structure can help the proposed RCRF model to capture the local and global connectivity interactions though several layers while maintains computational complexity and optimality at the same time.

Parallel to deep structures on the random field models, deep models specially deep learning approaches have attracted researchers in the past years for different applications from classification and segmentation to speech recognition and data analytics. The proposed methods were essentially a revisiting of neural networks with different approach. The proposed deep learning methods outperformed conventional algorithms in different applications of computer vision and machine learning.

A deep learning model can be seen as a generative model where it models the joint distribution of states and observations. This type of model also can be considered as a fully connected graph which nodes in each layer are connected to the nodes of other layers. Therefore, it is possible to expand the proposed method on deep learning framework. In Chapter 6, we discuss how the proposed approach can be applied on deep learning framework and motivates the future work of this thesis.

Chapter 5

Experimental Results

"Ideas do not always come in a flash but by diligent trial-and-error experiments that take time and thought."

Charles K. Kao

Figure 5.1: Examples images of the Weizmann dataset [148]. This dataset contains images with single and two objects.

5.1 Introduction

The performance of the proposed frameworks (i.e., RCRF and DRCRF) explained in Chapters 3 and 4 were compared with that of different state-of-the-art frameworks in the context of the interactive image segmentation problem within different situations. We consider noiseless and noisy situations as our main experimental results and we do a comprehensive analysis to explore the behavioral of the proposed methods in different conditions.

5.1.1 Dataset Description

Natural images from several datasets were used to evaluate the proposed methods including images from the Weizmann segmentation evaluation database [148] (examples seen in Figure 5.1), the complex scene saliency dataset (CSSD) [178] with some examples shown in Figure 5.2, and the Microsoft research interactive dataset (MRIS) [137] with examples visualized in Figure 5.3. The Weizmann database actually consists two different datasets: i) 100 images of single objects, and ii) 100 images of pairs of objects. The CSSD and MRIS datasets contain 200 and 50 images, respectively. To evaluate performance under noisy and corrupted situations, the images were also corrupted by white Gaussian noise with a standard deviation of 25% of the image dynamic range. The segmentation procedure is conducted based on user-specified areas corresponding to the object of interest and the background.

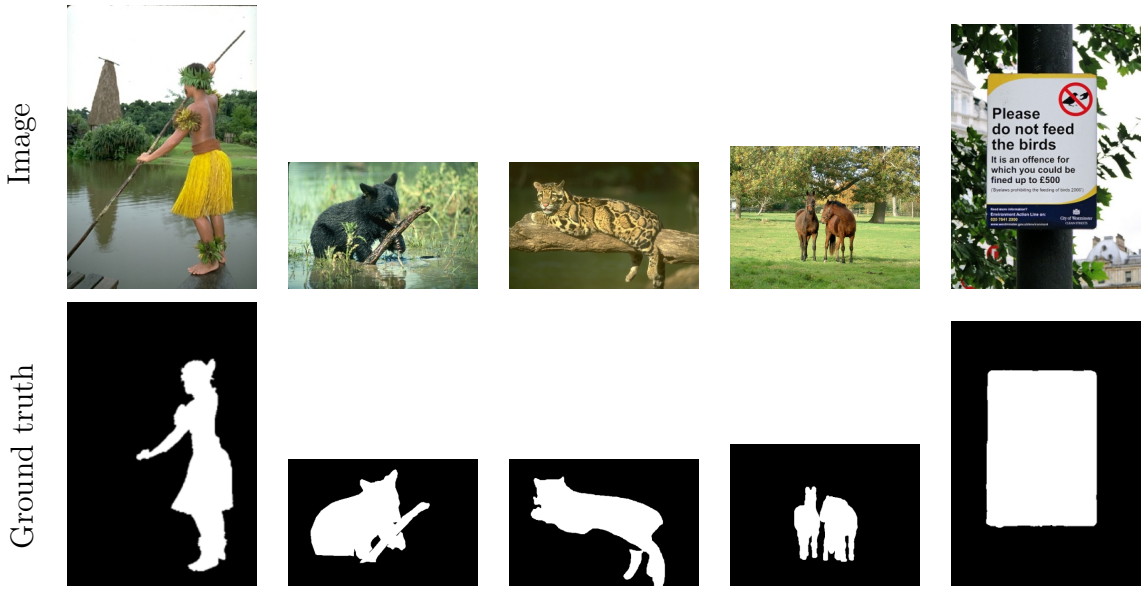


Figure 5.2: Image examples and corresponding ground truth for CSSD dataset [178].

Figure 5.3: Image examples and their ground truth of MRIS dataset [137].

The objects in the images are mostly natural objects with varied sizes in the images. The objects are with wide variation in shapes and complexity of the boundaries. For example, in Figure 5.3 the second image is with complex background while in the third image the object contains very fine structures and elongated boundaries.

The seed points to trained the unary potentials are annotated manually but they are the same for all competing algorithms.

5.1.2 Competing Algorithms

The proposed methods are compared against random field frameworks, including both deep as well as fully-connected random field approaches:

- **Pairwise graph cuts (P-GC)** [82] finds the MAP solution of a standard pairwise random field model by graph cuts.
- **Cooperative cut (Coop-cut)** [73] couples the edges based on the pairwise potential function and penalizes based on the number of types of label discontinuities.
- **Fully-connected CRF (FCRF)** [89] performs fast inference on a fully-connected CRF based on Gaussian potential functions and a permutohedral lattice [1].
- **Principled deep random field (PD)** [82] undertakes inference using a deep, multi-layer pairwise model with hidden auxiliary random variables [70] for representing useful higher-order interactions.

For more information about each method, readers are encouraged to study Sections 2.5 and 2.6

5.2 Model Configuration

5.2.1 Parameter Description

The parameters of two distributions $P^s_{i,j}$, $Q^d_{i,j}$ of the proposed methods and the number of layers for DRCRF were learned via a grid search on a holdout validation set. The 25 images from the Weizmann single object dataset were chosen as the validation set. The parameters for P-GC, Coop-Cut, PD, are selected as the optimal parameters reported by the authors [82], based on the validation set and the publicly available source code. The FCRF was based on the source code that the authors [82] provided publicly as well; however, their optimal parameters had not produced the best result, consequently the FCRF parameters were selected based on a grid search optimization using the validation set.

The same unary potentials are utilized in all methods, and are computed via a Gaussian mixture model with five components based on the color intensities of the pixels within the seed regions. The contrast-dependent Potts pairwise potential used is the same for all methods except FCRF (since it is implemented there via a different approach):

$$\psi_p(y_i, y_j, X) = \theta(x_i, x_j) \cdot \|y_i - y_j\| \tag{5.1}$$
$$\theta(x_i, x_j) = 0.05 + \frac{0.95 \exp(-0.5 \|x_i - x_j\|^2)}{\sigma}$$

where the value of σ is assigned by the mean of the color gradients of the image in P-GC, Coop-Cut, PD but selected as a constant of 0.2 for RCRF and DRCRF. Since FCRF has a different setup, based on a permutohedral lattice and fast implementation, the utilized potential function is a type of bilateral filter with different parameters. The standard deviations of the Gaussian pairwise potential functions (θ_γ) in (2.25) are set as $(3, 3)$ with a weight ($\omega^{(2)}$) of 5, while the standard deviations of the bilateral potential function (θ_α and θ_β) are $(20, 20, 0.08)$ with a weight ($\omega^{(1)}$) of 10, where the first two values show the spatial standard deviation and the last one is the color standard deviation, based on a normalized image dynamic range of $[0, 1]$.

Two-layer and three-layer DRCRF models (DRCRF(2L) and DRCRF(3L)) were utilized for the non-noisy case, where $P^s_{i,j} = \mathcal{N}(u_j|u_i, 15)$ in (3.5) and $\gamma = 5.5\%$ for the first layer, $P^s_{i,j} = \mathcal{N}(u_j|u_i, 10)$ and $\gamma = 10\%$ as the second layer, and $P^s_{i,j} = \mathcal{N}(u_j|u_i, 7)$ and $\gamma = 10\%$ as the third layer in the three-layer framework, where u_i denotes the location of node y_i in the random field, and $\mathcal{N}(u_j|u_i, \sigma_s)$ denotes a Gaussian function with a mean of u_i and a standard deviation of σ_s. For the noisy case, a four-layer DRCRF model (DRCRF(4L)) was utilized to better handle the effects of noise as we discuss in Section 5.3.5; the first two layers are the same as above, $P^s_{i,j} = \mathcal{N}(u_j|u_i, 3)$ and $\gamma = 80\%$ selected for the third layer, and a second-order Markov model chosen as the fourth layer. For both of these DRCRF model configurations, $Q^d_{i,j} = \mathcal{N}(x_j|x_i, 0.2)$ for all layers, where x_i is the color intensity of y_i, and $\mathcal{N}(x_j|x_i, \sigma_x)$ denotes a Gaussian function having a mean of x_i and a standard deviation of σ_x.

Based on (3.5), a larger γ gives those nodes having a lower probability a greater chance to be selected as neighbors. Since $P^s_{i,j}$ is derived via the spatial distance between two nodes i and j, the left side of the inequality in (3.5) is larger for the nodes which are far from node i. By increasing γ the likelihood of those nodes being selected is increased, meaning that longer-range information is incorporated into the model. Based on this explanation, a smaller γ is utilized in the lower layers to capture the most useful longer-range information, while a larger γ is selected to incorporate more local information in upper layers.

A graph cut approach was applied to minimize the potential function of the RCRF and DRCRF frameworks. The graph cut approach is initialized by maximizing the GMM model (unary potential) result.

5.2.2 Unary Potential

The same unary potential was utilized to evaluate the competing algorithms. A Gaussian mixture model (GMM) [109] with five components was trained via color intensity of pixels selected as seed points. To use a GMM approach as unary potential, for each label of foreground and background a Gaussian probability distribution with five components is trained based on the corresponding seed points of foreground or background.

After training the GMM models, the probabilities of a pixel to be assigned as foreground or background are computed by those trained GMM models for all pixels in the image. Two computed probabilities are utilized as unary potential in the random field model.

5.3 Quantitative Evaluation

Three different quantitative measures are utilized to analyze the behavior of the compared methods. The proposed methods were examined with noiseless and noisy images to explore their behaviors in different situations.

5.3.1 Quantitative Measures

Three different quantitative measures are utilized to analyze the behavior of the compared methods. Since the underlying problem being evaluated is the image segmentation problem, one can evaluate the performance of the tested method using the F_1-score:

$$F_1 = 2 \cdot \frac{Pr \cdot Re}{Pr + Re} \tag{5.2}$$

s.t.

$$\begin{aligned} Pr &= \frac{TP}{TP + FP} \\ Re &= \frac{TP}{TP + FN} \end{aligned} \tag{5.3}$$

where TP, FN and FP are the number of true positives, false negatives, and false positives, respectively. Here, two types of F_1-scores are evaluated, described below.

Region F_1-score: The conventional region F_1-score is evaluated based on the region-of-interest specified by the ground truth images. The foreground is chosen as the positive class label while the background is the negative class label:

$$\text{Region } F_1\text{-Score} = 2 \cdot \frac{Pr_r \cdot Re_r}{Pr_r + Re_r} \tag{5.4}$$

where Pr_r and Re_r are computed based on the pixel-wise accuracy over the whole target image. This measure measures the degree to which the evaluated method can distinguish the background and foreground in terms of their regions.

Boundary F_1-score: Boundary preservation accuracy is an important objective in image segmentation. Motivated by [3], the extracted boundary of the ground truth is taken as the positive class while the other pixels are specified as the negative class:

$$\text{Boundary } F_1\text{-Score} = 2 \cdot \frac{Pr_b \cdot Re_b}{Pr_b + Re_b} \tag{5.5}$$

here Pr_b and Re_b are computed based on the boundary of the foreground object in the image. This measure shows how much the boundary of the extracted object overlaps with that of the ground truth foreground. Similar to [3], a distance tolerance of 2 pixels is used in the calculation of this evaluation measure, meaning that the boundary detected by the algorithm is considered to be a true positive if it is within two pixels of a ground-truth boundary.

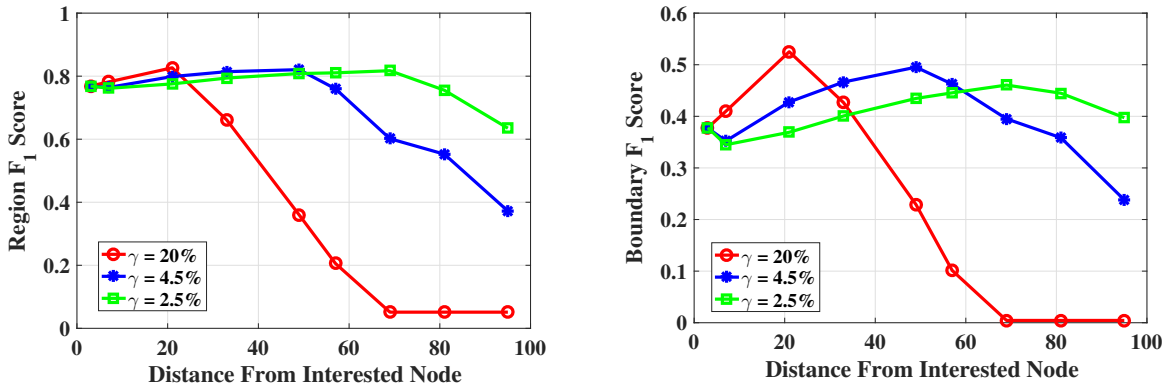
Another quantitative measure which is commonly utilized in image segmentation is **Intersection over union (IOU)** [35], also known as the Jaccard index. The intersection of the estimated segmentation result per class and the ground truth, divided by the union, is reported as a metric:

$$IOU = \frac{TP}{TP + FP + FN}. \tag{5.6}$$

The IOU measure demonstrates how much the segmented region overlaps with the ground truth, based on the number of pixels. Since our problem is a binary segmentation, the IOU metric is just reported based on the class corresponding to the object.

5.3.2 Connectivity Range Effect on RCRF

As the first experiment in this chapter we show the effect of the connectivity range and the distance from the interested node in the stochastic clique generation for the image



(a) (b)

Figure 5.4: The effect of γ on modeling accuracy. The value of γ determines the expected number of pairwise cliques such that one of the end-nodes is the interested node. The left plot, (a), shows the region F_1-score while the right plot (b) demonstrates the boundary F_1-score. As seen, the modeling accuracy is increasing when the range of long-distance pairwise cliques is being increased but to some extent and the range is bounded. However the number of pairwise connectivities involved in the inference affects the modeling accuracy observed by the varied accuracy as a function of γ.

segmentation problem. A subset of 30 images from the CSSD dataset [178] was selected to conduct the experiment. Here the goal is to show how γ can affect the modeling accuracy. Three different values of γ have been selected to create the underlying graph of the RCRF based on specific distances from the interested node. By increasing γ, the number of connectivities of the underlying graph is increased. In other words, the range of neighborhood structure in the random field is one of the variables in this experiment.

Figure 5.4 demonstrates the region F_1-score and boundary F_1-score corresponding to three different γ values with varied distance from the interested node. It is worth to mention that for distance of 3 (second-order Markov) we assumed that $\gamma = 100\%$ or all nodes are connected. The results shows three interesting observations:

- Increasing the range of pairwise cliques from local connections to long-range connec-

tions can improve the modeling accuracy. However there is a bound which determines the number of connections and the distance from the interested node, as evident by different ranges of γ.

- As seen, by increasing the range of connectivity from the interested node, γ should be decreased to maintain better modeling performance. The reasons that can be provided for that is, when the range of connectivity is increased, the possibility of utilizing wrong information (non-useful pairwise cliques) is increased, therefore decreasing γ helps to decrease the possibility of using them in the modeling and forces a negative effect on the modeling accuracy.

- As stated before, long-range connectivity can address the smoothness boundary issue (i.e., short boundary bias problem) of local random fields. This is evident by the result of $\gamma = 2.5\%$ (green curve) in Figure 5.4. As seen, even for distance ≈ 100 where the region F_1-score is worse than local connectivities, the boundaries have been preserved better than a random field with local connectivities as observed by plot (b) in Figure 5.4.

As expected, the conducted experiment illustrated that long range connectivities help to address the short boundary bias problem (i.e., smoothness issue) [82]. However it is important to consider that the range of connectivities must be bounded or the connectivities must be selected via a guided approach to obtain informative connections or cliques in the graph. It is worth to mention that the range of connectivities is a problem dependent factor.

5.3.3 Connectivity Range Effect on DRCRF

The effect of varied ranges of information on DRCRF are examined on 30 images from the CSSD dataset by a two-layer DRCRF as well. Figure 5.5 plots the boundary F_1-score and IOU as a function of the information ranges in the two layers. In this experiment, the effect of the selection probability is examined by varying the standard deviation of Gaussian function utilized as $P^s_{i,j}$ (3.6).

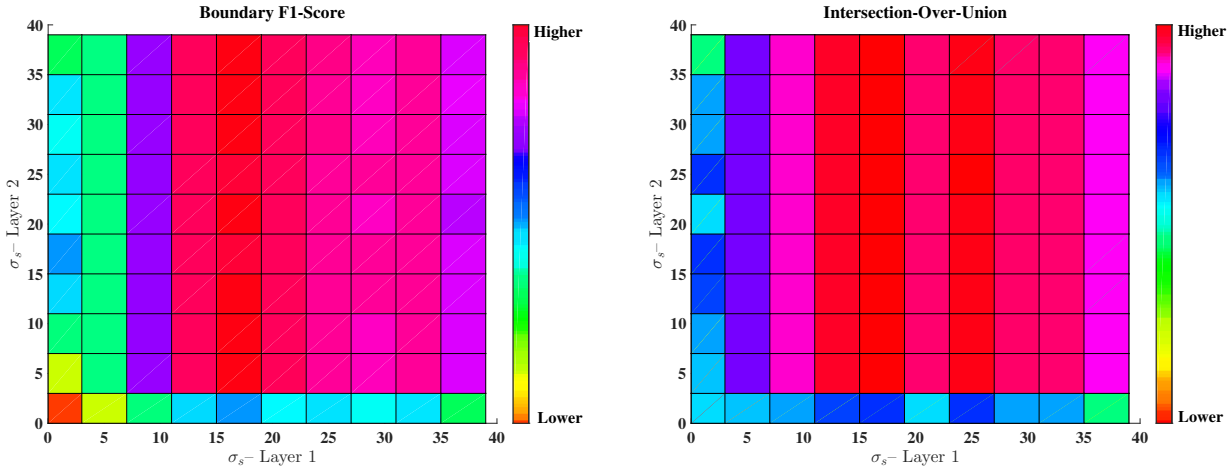


Figure 5.5: Quantitative analysis of σ on the performance of a two-layer DRCRF. The X and Y axes represent the σ value in layers one and two. The results are reported based on 30 images from the CSSD dataset. The plots show that to gain a strong segmentation result while preserving object boundaries, the first layer of the DRCRF needs to capture long-range information, while going up through layers, there is no such requirement, and the range of connectivity can be narrowed down to be limited to local information.

The key strength of DRCRF is that it is able to incorporate longer-range information more efficiently, leading to a better preservation of object boundaries, and more effectively addressing the short-boundary bias problem.

As observed in Figure 5.5, to effectively preserve object boundaries while maintaining object segmentation, it is essential that the first layer captures sufficiently long-range information, whereas the results are much less sensitive to non-local connectivities in the second layer.

5.3.4 Noiseless Images

Tables 5.1 and 5.2 show quantitative comparisons of the tested methods in terms of the region F_1-score and the boundary F_1-score. The proposed DRCRF is examined based

Table 5.1: Region F_1-score results. The performance of the comparison methods are demonstrated in a noiseless context for the Weizmann single-object, two-objects, CSSD and MRIS (Microsoft Research Interactive Segmentation) datasets. The time complexity is reported by averaging the running time (in seconds) of the methods.

	P-GC [82]	Coop-cut [73]	FCRF [89]	PD [82]	RCRF	DRCRF(2L)	DRCRF(3L)
Weizmann S-O	0.8511	0.8600	0.8655	0.8711	0.8651	**0.8755**	0.8578
Weizmann T-O	0.8515	0.8716	0.8397	**0.8840**	0.8404	0.8546	0.8711
CSSD	0.8219	0.8286	**0.8551**	0.8286	0.8268	0.8464	0.8425
MRIS	0.8929	0.8929	0.8717	**0.9032**	0.8756	0.8720	0.8933
Impl.	M-M	M-M	C++	M-M	M-M	M-M	M-M
Time Complexity (s)	15.51	16.58	0.48	106.86	34.41	77.32	94.37

Table 5.2: Boundary F_1-score results. The performance of the comparison methods is shown for the noise-free context for all datasets.

	P-GC [82]	Coop-cut [73]	FCRF [89]	PD [82]	RCRF	DRCRF(2L)	DRCRF(3L)
Weizmann S-O	0.5324	0.5573	0.5770	0.5782	0.5879	0.5857	**0.5942**
Weizmann T-O	0.7049	0.7408	0.6992	**0.7603**	0.7132	0.7499	0.7468
CSSD	0.5226	0.5333	0.5212	0.5349	0.5235	0.5232	**0.5625**
MRIS	**0.6441**	0.6393	0.5452	0.6389	0.6157	0.5949	0.6290

Table 5.3: Intersection Over Union (IOU) results. To ensure that the reported F_1-scores are consistent, all methods are evaluated by the IOU measure without noise.

	P-GC [82]	Coop-cut [73]	FCRF [89]	PD [82]	RCRF	DRCRF (2L)	DRCRF(3L)
Weizmann S-O	0.7579	0.7706	0.7850	0.7889	0.7850	**0.7999**	0.7829
Weizmann T-O	0.7606	0.7887	0.7637	**0.8083**	0.7601	0.7966	0.8009
CSSD	0.7212	0.7301	**0.7626**	0.7306	0.7328	0.7520	0.7573
MRIS	0.8182	0.8185	0.7912	**0.8308**	0.7601	0.7966	0.8222

on the DRCRF(2L) and DRCRF(3L) frameworks to obtain better insights about how the number of layers in the deep, multi-layer structure comprising the DRCRF model influences the performance of the method. It can be observed that DRCRF performs as well as or outperforms other competing methods, especially regarding the preservation of boundaries. Table 5.1 also demonstrates the average running time of each algorithm, from which we can see that although increasing the number of layers in DRCRF can lead to accuracy improvements, this comes at additional computational complexity.

It can be observed that DRCRF(2L) and DRCRF(3L) broadly perform similarly to or better than the competing approaches, typically achieving **1%** to **3%** improvement in F1-score, certainly when compared to RCRF and FCRF. For the boundary F_1-score, the proposed DRCRF performs strongly, with the three-layer DRCRF outperforming all compared approaches for the CSSD and single-object Weizmann databases, and fairly significantly outperforming FCRF and RCRF in all cases.

5.3.5 Noisy Images

The competing methods are also tested on images corrupted by 25% Gaussian noise, with results reported in Tables 5.4 and 5.5. Although DRCRF and PD exhibit comparable F_1 performance in the noiseless case, interestingly in the noisy context, particularly in the boundary assessment of Table 5.5, DRCRF was able to achieve significantly improved performance, by **10%** to **20%**, compared to PD and all other tested methods, indicating that it is able to better capture object boundaries in the presence of noise.

Table 5.4: Region F_1-score results for noisy images. The proposed method outperforms other methods in almost all datasets.

	P-GC [82]	Coop-cut [73]	FCRF [89]	PD [82]	RCRF	DRCRF(4L)
Weizmann S-O	0.6866	0.6914	0.6586	**0.7502**	0.6398	0.7203
Weizmann T-O	0.6510	0.6505	0.6718	0.7327	0.6078	0.7347
CSSD	0.5683	0.5683	0.5526	0.5919	0.5024	**0.6380**
MRIS	0.6034	0.6034	0.5424	0.6149	0.5380	**0.6399**

Table 5.5: Boundary F_1-score results in the presence of noise. The proposed DRCRF can preserve boundaries more effectively than competing methods.

	P-GC [82]	Coop-cut [73]	FCRF [89]	PD [82]	RCRF	DRCRF(4L)
Weizmann S-O	0.2299	0.2304	0.1797	0.3110	0.2211	**0.3910**
Weizmann T-O	0.3441	0.3404	0.2610	0.4152	0.2866	**0.5406**
CSSD	0.1618	0.1617	0.0941	0.1936	0.1738	**0.2820**
MRIS	0.1570	0.1560	0.0636	0.1694	0.1589	**0.2801**

To analyze the robustness of the proposed method, all competing algorithms are compared in the context of salt & pepper noise as well. Table 5.7 demonstrates the quantitative measures for MRIS dataset when corrupted with 10% of salt & pepper noise. The reported results show that the proposed DRCRF methods are robust to multiple noise types, with the DRCRF outperforming FCRF and PD in all three of region F_1, boundary F_1, and IOU, and outperforming RCRF in both region F_1 and boundary F_1.

Table 5.6: Intersection Over Union (IOU) results for noisy images.

	P-GC [82]	Coop-cut [73]	FCRF [89]	PD [82]	RCRF	DRCRF(4L)
Weizmann S-O	0.5389	0.5451	0.5170	**0.6234**	0.5124	0.5967
Weizmann T-O	0.5082	0.5075	0.5387	0.6073	0.4890	**0.6338**
CSSD	0.4174	0.4174	0.4020	0.4459	0.3850	**0.4996**
MRIS	0.4507	0.4507	0.4013	0.4696	0.4212	**0.5180**

Table 5.7: Quantitative segmentation results for MRIS dataset images corrupted with 10% salt & pepper noise. The proposed method preserves boundaries better than other competing algorithms.

	P-GC [82]	Coop-cut [73]	FCRF [89]	PD [82]	RCRF	DRCRF(4L)
Region F_1-score	0.8454	0.8470	0.8429	0.8732	0.8783	**0.8828**
Boundary F_1-score	0.5343	0.4996	0.3446	0.4391	0.6473	**0.6679**
Intersection-Over-Union	0.7459	0.7475	0.7493	0.7866	**0.8110**	0.8062

5.3.6 Performance Comparison on different Noise Powers

Since the databases contain natural RGB images, Gaussian noise is the most common type of noise utilized to examine the robustness of an algorithm. Gaussian is a good assumption for processes that are subject to Central Limit Theorem [120]. To allow a comprehensive analysis, all methods are tested as a function of noise level in the context of the Microsoft research interactive dataset (MRIS). Figure 5.6 demonstrates the performance of the competing algorithms, with the DRCRF performing very well, particularly under boundary F_1-score.

It worth to mention that PD and P-GC force the segmentation result on the selected seed pixels to be the same as the user selected labels and because of that they produced better results in terms of IOU measure after 35% noise.

5.4 Qualitative Evaluation

Example segmentation results for noise-free images are shown in Figure 5.7 (single-object and two-object Weizmann datasets) and in Figure 5.10 (CSSD and MRIS datasets). It can be seen that PD and Coop-Cut methods had difficulty in preserving boundaries in the test cases shown, with either the background being merged with the object or parts of the object being classified as background. The "reclining girl" and the "Man throws the ball" in Figure 5.10 (columns (e) and (f)) are good examples in which P-GC, Coop-Cut and

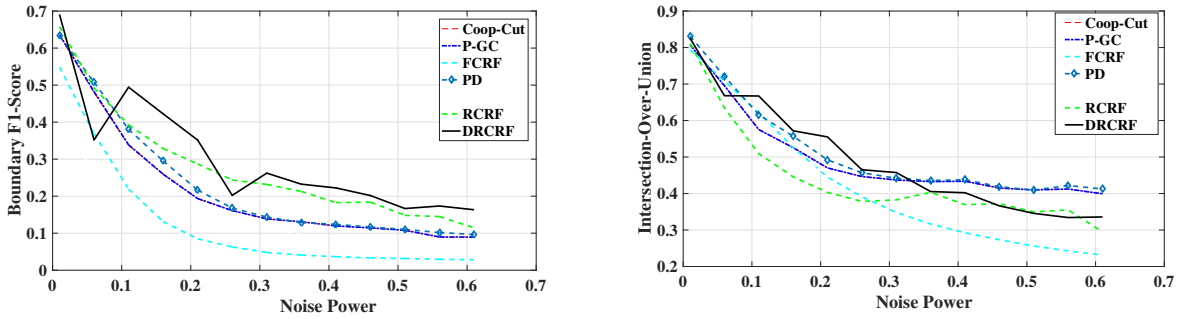


Figure 5.6: The performance of competing methods based on various noise power. As seen, the DRCRF outperforms other methods in preserving object boundaries. PD, Coop-Cut and P-GC perform slightly better in region F_1-score after 35% noise because they explicitly use the seed points label in the final result while others methods only utilize the seed points to train the unary potential (GMM).

PD could not separate the body from the grass completely (the hands in column (e) and the legs in column (f)) because of the way in which they formulated the pairwise potential in the random field. FCRF was able to preserve boundaries better than PD and Coop-Cut methods in the test cases shown, but also exhibited the introduction of segmentation artifacts (e.g., missing faces in Figure 5.10(a), large incorrectly segmented patch in the top right corner of Figure 5.10(f), etc.).

It can be observed that the proposed method is capable of preserving narrow and elongated boundaries, for example the preservation of the tree branches in Figure 5.7(e) and the airplane tail in Figure 5.7(a). Furthermore, it can be observed that the proposed method is capable of dealing with scenarios characterized by complex and cluttered backgrounds, in the starfish image of Figure 5.7(d).

Figure 5.8 demonstrates the airplane example with larger image resolution to see the results with more details. As seen, P-GC (i.e., which is a local random field) suffers from strong short boundary biased problem and could only capture the main area of the plane and assigned the tail as background. Although this problem is less in the Coop-cut's results, it also has problem with the regions which contain complex boundaries, as evident by the

head and tail of the plane. The aforementioned problem is resolved in FCRF as it is a fully connected random field. However as discussed before, because of long range connectivities it is prone to introduce segmentation artifacts to the foreground regions. Although the segmentation result produced by PD is much better than Coop-cut and P-GC results, as seen PD could not preserve the elongated regions of the plane and also it assigned a part of background in the tail as foreground. The segmentation results of RCRF and DRCRF(2L) outperform other results and it can be said that the DRCRF(2L) performed better than RCRF to preserve boundaries in the image. DRCRF(3L) preserved all parts of the plane as foreground, but as seen it assigned the part of background in the tail as foreground region. This can be explained by the fact that DRCRF(3L) uses more local interactions in the third layer which results to a more smoothness on the segmentation result.

Figure 5.9 is another example to compare different methods. As seen P-GC suffers from short boundary bias problem very baldly as evident by the bird's legs, beak and tail. This problem still presents with the result of Coop-cut and PD with small improvements in foreground segmentation. FCRF produced a good result however it could not preserve the bird's beak. RCRF segmented the bird from background however because of the long-range connectivities effect it introduced segmentation artifacts to the result. DRCRF(2L) resolves the artifacts seen in RCRF result and it is considered as the best result from the competing methods. DRCRF(3L) has the same problem as before as it used local connectivities in the last layer.

To have more illustrative comparisons, Figure 5.11 and 5.12 demonstrate the segmentation results of two images from CSSD datasets with higher resolution. As evident, the type of penalty function that Coop-cut and PD utilized enforced the segmentation procedure to add some parts of the background into the foreground region while those methods that utilizing long-range information such as FCRF or RCRF and DRCRF do not have this problem.

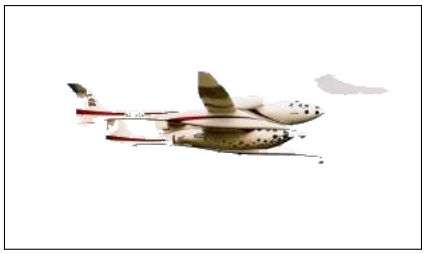
It is an interesting observation that DRCRF(2L) can preserve elongated boundaries better than DRCRF(3L), whereas DRCRF(3L) performs better when the object has a dense structure. A possible argument is that DRCRF(2L) has only two layers, but with higher

Figure 5.7: Example segmentation results produced by the tested methods on the Weizmann dataset [148] without noise. As seen, the proposed method can preserve elongated boundaries in complex images. Furthermore, the proposed method is capable of dealing with situations characterized by complex and cluttered backgrounds, as is evident in the starfish image (d).

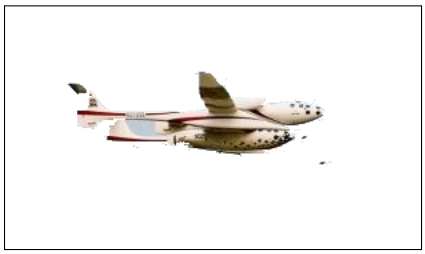
(a) Ground truth

(b) P-GC

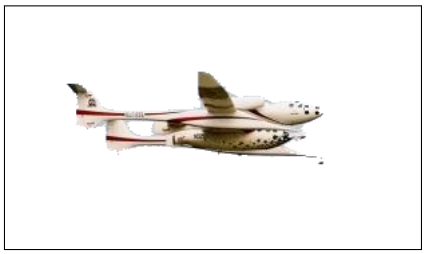
(c) Coop-cut



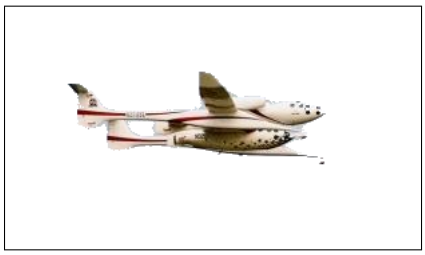
(d) FCRF



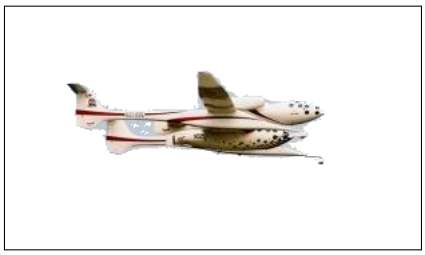
(e) PD



(f) RCRF

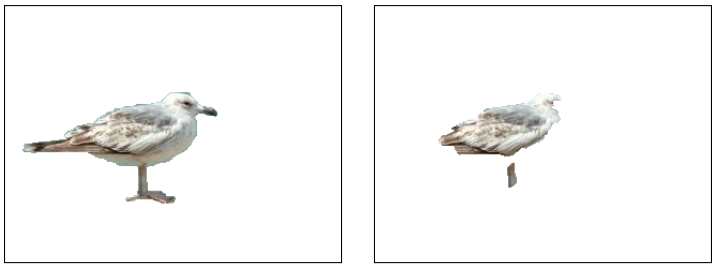


(g) DRCRF(2L)



(h) DRCRF(3L)

Figure 5.8: Airplane example; the segmentation results of different competing methods for airplane image are demonstrated with higher resolution. As seen each method has some issues to produce the desirable segmentation result while the proposed methods can result a good segmentation.



(a) Ground truth

(b) P-GC

(c) Coop-cut

(d) FCRF

(e) PD

(f) RCRF

(g) DRCRF(2L)

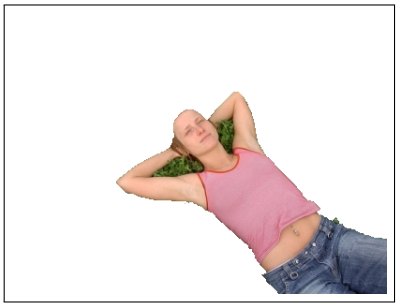
(h) DRCRF(3L)

Figure 5.9: The segmentation results of competing methods for the Bird image example. The subjects of comparison here is bird's legs, beak and tail. As seen, the long-range connectivities effect in RCRF introduced segmentation artifacts while DRCRF(2L) can preserve fine boundaries.

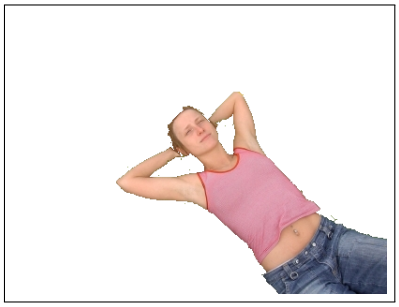
Figure 5.10: Example segmentation results produced on images from the CSSD and MRIS datasets without noise. Results demonstrate the effect of long-range connectivities and deep structure models on improving the modeling accuracy in random field modeling.

(a) Ground truth

(b) P-GC



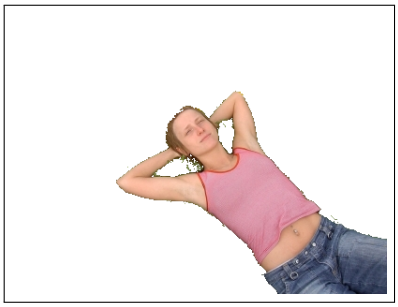
(c) Coop-cut



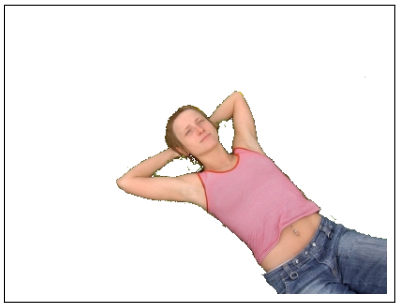
(d) FCRF

(e) PD

(f) RCRF



(g) DRCRF(2L)



(h) DRCRF(3L)

Figure 5.11: Segmentation results of different methods for reclining girl.

(a) Ground truth

(b) P-GC

(c) Coop-cut

(d) FCRF

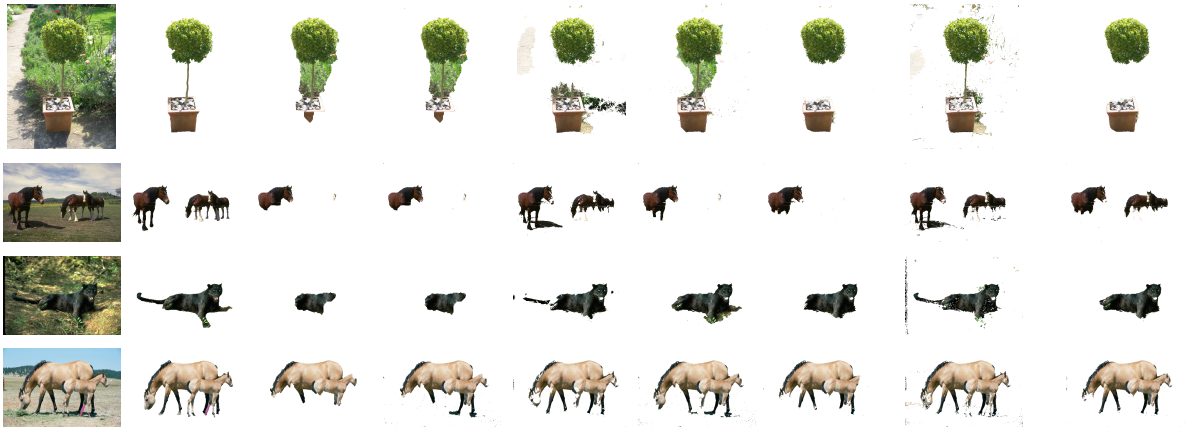
(e) PD

(f) RCRF

(g) DRCRF(2L)

(h) DRCRF(3L)

Figure 5.12: Segmentation results of different methods; As seen P-GC has short boundary bias problem andCoop-Cut and PD resulted bad segmentation because of the way they utilize pairwise connectivities. FCRF could not segment the faces correctly because of utilizing blind long-range connectivities and the similarity of face color with the background. RCRF and DRCRF(2L) produced the best results.

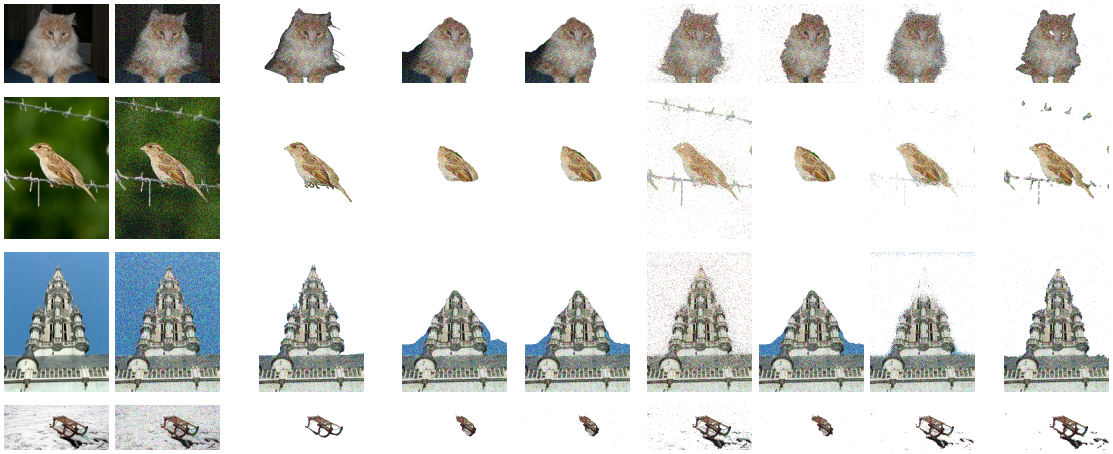


(a) **Image** (b) **GT** (c) **P-GC** (d) **Coop-Cut** (e) **FCRF** (f) **PD** (g) **RCRF** (h) **DRCRF(2L)** (i) **DRCRF(3L)**

Figure 5.13: Example segmentation results of objects with elongated boundaries. The DRCRF methods can preserve elongated boundaries more effectively compared to other methods. The "tree" image is a good example of the superiority of the proposed method.

σ values, whereas DRCRF(3L) has an extra layer configured with a lower σ. The higher σ prompts longer-range of connectivities resulting in a better preservation of elongated boundaries in DRCRF(2L), whereas having the final segmentation results produced based on a lower σ value, DRCRF(3L) is more powerful when facing objects with dense structure. In general, DRCRF(3L) results in a better quantitative boundary preservation since most of the database objects are dense, with relatively few elongated boundaries. Illustrations of the above explanation can be seen in "tree" (Figure 5.13), in which DRCRF(2L) performs better than DRCRF(3L) (i.e., due to the elongated boundary), whereas DRCRF(3L) better segments the dense structures in "snake" (Figure 5.7(c)), "plane" (Figure 5.10(b)), and "flower" (Figure 5.10(c)).

To provide a better visualization of the methods dealing with the object with elongated boundary, Figure 5.15 demonstrates the segmentation result for "tree" example by all competing methods. This example is a very complex image segmentation sample where there are complex background and elongated boundaries together. As seen based on a complexity of background Coop-cut and PD did not perform reasonably in finding the



(a) **Image** (b) **Noisy** (c) **Ground Truth** (d) **P-GC** (e) **Coop-Cut** (f) **FCRF** (g) **PD** (h) **RCRF** (i) **DRCRF(4L)**

Figure 5.14: Example segmentation results for images corrupted with 25% Gaussian noise. This level of noise is selected to examine the extreme robustness of competing methods dealing with noise. It can be observed that DRCRF is able to preserve boundaries in complex and noisy situations effectively.

interested object. FCRF could not preserve stem completely and also introduced segmentation artifact into the result. RCRF also could not preserve stem while it produced a good segmentation. DRCRF(2L) preserved all elongated boundaries, however it introduced artifacts in the results as well. The result of DRCRF(3L) is like RCRF with more smoothness in boundaries which is the effect of the last layer (i.e. local interactions).

The qualitative results of the algorithms when dealing with noisy images are demonstrated in Figure 5.14. It can be observed that DRCRF is able to achieve strong segmentation results in the presence of noise, where it is able to capture elongated boundaries ("bird on barbed wire"), fine boundary detail ("cat", where the cat's hairs are well captured), and boundaries around dense objects ("town hall", where the boundary around the top of the tower is well captured). The figure makes a compelling argument for the use of a multi-layer structure, in comparing the DRCRF (multi-layer) results against the single-layer RCRF model.

For better visualization of details, Figures 5.16 and 5.17 show the results for two example images. As seen in Figures 5.16 and 5.17 DRCRF(4L) outperforms other competing methods by far in preserving the foreground regions and objects boundaries which demonstrates the effectiveness of the proposed DRCRF method.

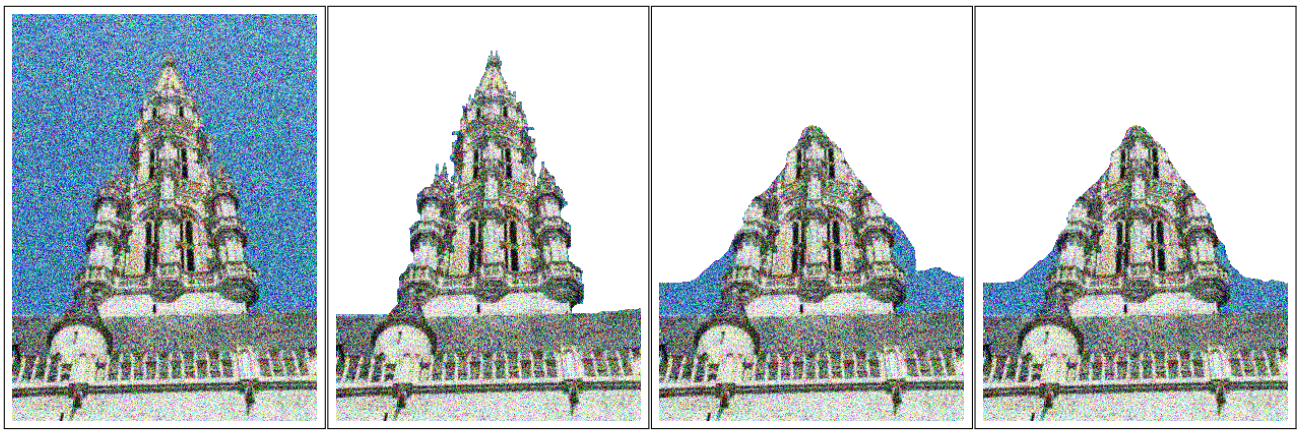
5.5 Summary

In this chapter, we evaluated the proposed method along with other competing algorithms on several image segmentation datasets. We compared different methods quantitatively and qualitatively based on region F_1-score, boundary F_1-score and intersection over union. All competing methods are evaluated on noisy situations as well. Results showed that the proposed RCRF and DRCRF methods can result as well as or outperform other state-of-the-art methods in the interactive segmentation problem. Experimental results also demonstrated that DRCRF outperforms other competing approaches in presence of noise on input images. To have a better understanding of the parameters of the proposed methods, we showed the accuracy as function of the variation of the parameters and the robustness of competing methods were measured by different noise power which showed that the proposed DRCRF method is the most robust approach in the presence of noise in preserving the object boundaries.

(a) Ground truth (b) P-GC (c) Coop-cut (d) FCRF

(e) PD (f) RCRF (g) DRCRF(2L) (h) DRCRF(3L)

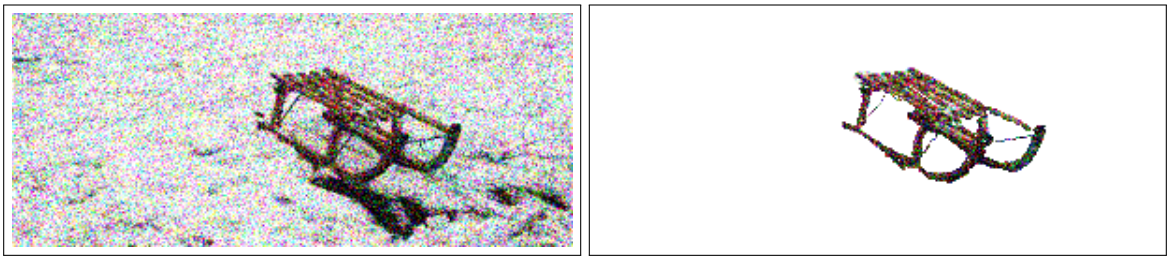
Figure 5.15: Example segmentation results for an object with elongated boundaries; as seen DRCRF can preserve elongated boundary more accurate than other methods.



(a) Noisy Image (b) Ground truth (c) P-GC (d) Coop-cut

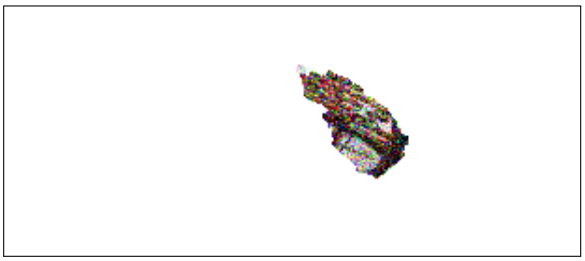
(e) FCRF (f) PD (g) RCRF (h) DRCRF(4L)

Figure 5.16: Example noisy segmentation of town hall via all competing algorithms. As seen DRCRF(4L) outperforms other method in noisy case.

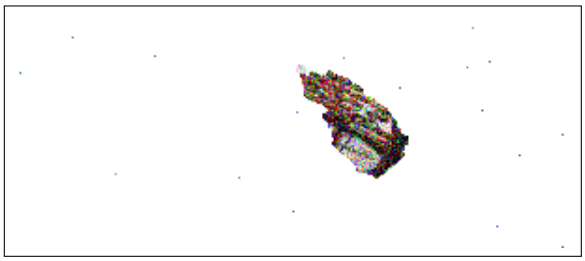


(a) Noisy Image

(b) Ground truth

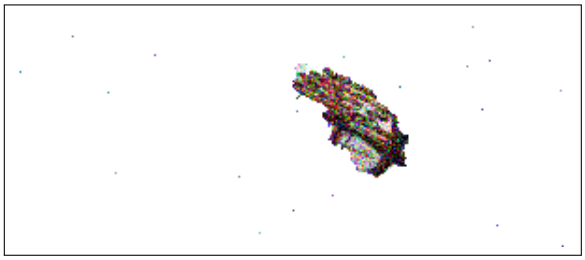


(c) P-GC

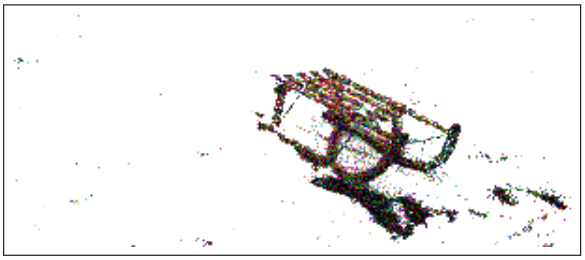


(d) Coop-cut

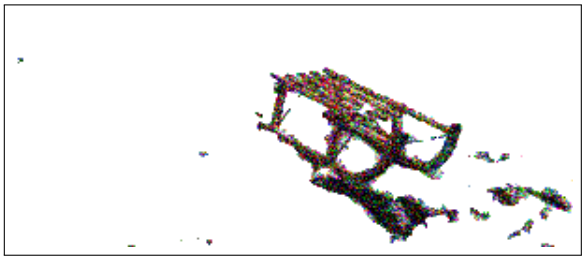
(e) FCRF



(f) PD



(g) RCRF



(h) DRCRF(4L)

Figure 5.17: Segmentation example of a noisy image; It is shown that DRCRF(4L) outperforms other competing methods in noisy situations.

Chapter 6

Conclusion & Future Work

"The more you observe life in relation to yourself the more you will see the fact that you are hardly ever correct when you think about something in the future. The future exists only in imagination; and that is why, no matter how hard you try to imagine it, you will not be able to predict the future with total certainty."

Barry Long

The goal of this research was to study the effectiveness of a stochastic approach to model dense random field structures via sparse graph generation to address the computational complexity of the inference process on dense and complex random fields. To address this issue we proposed a new clique structure in Chapter 3, stochastic clique, which is related to random graph theory. In Chapter 4, we showed that combining the proposed stochastic cliques with a deep structure model improves the modeling accuracy in the interactive image segmentation problem. This chapter summarizes the main contributions presented in this thesis and offers some promising directions for future research.

6.1 Thesis Contribution Highlights

Random fields have remained a topic of great interest over the past decades for the purpose of structured inference, especially for problems such as image segmentation. The local nodal interactions commonly used in such models often suffer the short-boundary bias problem [82], which are tackled primarily through the incorporation of long-range nodal interactions. However, the issue of computational tractability as discussed in Chapter 3 becomes a significant issue when incorporating such long-range nodal interactions, particularly when a large number of long-range nodal interactions (e.g., fully-connected random fields) are modeled.

Although recent work on fully connected random fields [21, 134, 186] and deep random fields [82] have been shown to be very promising in addressing these issues, both streams of approaches face certain limitations which could affect the performance of the inference and computational tractability. In this thesis we introduced the concept of stochastic cliques and proposed the randomly-connected conditional random fields (RCRFs), which fuse fully-connected random fields with the concept random graph theory. To obtain benefits from long-range interactions while allowing for efficient inference using arbitrary potential functions, we proposed a deep structure randomly-connected CRF which leverages random graph theory and the concept of stochastic cliques to incorporate into a deep conditional random field structure to take better advantage of long-range interactions while maintain-

ing computational tractability. The main contribution of this thesis can be summarized as follows:

- **Stochastic cliques (Chapter 3):** We proposed a new type of clique structure which specifies the set of cliques incorporated in the inference process stochastically. Intuitively, to formulate the energy function over the underlying conditional random field framework, the stochastic clique approach samples the most relevant pairwise clique structures such that minimizing the energy function based on this subset of pairwise cliques leads to the same result as a fully connected random field (i.e., all pairwise cliques are incorporated via the energy function). Based on this approach, a small subset of pairwise cliques are utilized in the energy function and, therefore, the energy function can be minimized by considering a lesser number of pairwise relations between nodes, making the optimization process much faster.

- **Randomly-connected conditional random fields (Chapter 3):** we explored tackling the problem of computational complexity by constructing a sparse graph representation stochastically from the fully-connected random field by randomly sampling the most informative nodal interactions and based on the concept of stochastic cliques. Inspired by random graph theory, active cliques are formed stochastically in the inference step to represent the fully-connected CRF with a sparse graph model that provides approximately the same results as the fully-connected CRF. By combining random graph theory with random field theory in such a way, the resulting sparse graph retains all of the properties of a CRF, and as such can be used in all of the same structured inference scenarios that CRFs are used for.

- **Deep randomly-connected conditional random fields (Chapter 4):** We focused on the graphical structure of the deep, fully-connected conditional random field itself, in which all possible pairwise cliques between nodes exist. Accounting for all such interactions in a direct manner within a deep, fully-connected conditional random field resulted to an inference problem that is computationally intractable. We studied the problem from a perspective such that what if the pairwise cliques take

shape in a random manner, coming into and out of existence with a certain probability? We proposed the idea of a deep, randomly-connected conditional random field (DRCFR) which relates directly to random graph theory, where nodes in the graph are connected based on some probability distribution. This new perspective allowed us to retain the benefits of long range interactions and to account for them with arbitrary feature functions within a deep, fully-connected graphical model, while achieving computational tractability since the probability of clique formation can be controlled via the choice of probability distribution.

6.1.1 Limitations

The experimental results demonstrated that the proposed methods produce comparable or even better results compared to state-of-the-art approaches. However there are some limitations accompanied with the methods which should be considered:

1. **Computational Complexity:** To generate the associated graph of the RCRF approach, it is needed to perform the stochastic indicator function for each pair of nodes in the random field. This process imposes an extra computational complexity to the model which limits the utilization of very large connections. However it is obvious that several pairs of connectivities have the same characteristics in applications such as image segmentation, which it is possible to leverage this property to decrease the computational complexity of the graph generation. This new idea will be explained in Section 6.2.2.

2. **Stochastic Clique Indicator:** In this thesis we utilized very simple functions to characterize the stochastic indicator function, (3.5), in the model. As explained in Section 3.3.1 the pixel intensities were utilized to characterized the relations of nodes (pixels) in the graph. Although the simplicity of this model helped us to analyze the model easier, utilizing a more sophisticated model to characterize which two nodes should be connected in the random field would boost the model accuracy of RCRF approach.

3. **Number of Layers:** The experimental results showed that by increasing the number of layers in DRCRF approach the modeling accuracy is increased for some problem. However it is obvious that by increasing the number of layers in the model, the computational complexity and the running-time is increased as result. Therefore, it needs to consider there is a trade of between the modeling accuracy and computational complexity for the proposed algorithms.

6.2 Future Work

The proposed methods in this thesis open several new directions as future work. Here we describe the main topics.

6.2.1 Mathematical Hypothesis

In this work, the inference framework was computed via a graph cuts approach (i.e., *s-t* min cut) [18]. Due to the randomness involved in representing the underlying graph of the fully-connected CRF, two factors should be considered:

1. The graph should be connected as discussed in Section 3.3.2. It is also showed that the underlying graph of the RCRF is connected with high probability.

2. It is important to show that the nodes in the sparse graph representation of the fully-connected CRF obtained via the proposed stochastic clique formation process can be partitioned into approximately the same sets of nodes as the original fully-connected graph of the fully-connected CRF by the use of *s-t* minimum cut approach with a limited variation range on the min cuts values (*Minimum Cut*), since the goal of the proposed framework is to address the computational complexity associated with structured inference using fully connected CRFs without impeding performance.

Karger [76] and Benczur and Karger [6] proposed random sampling techniques for approximating problems that involve cuts and flows in graphs. They proved that *given*

dense graph $\mathcal{H}$ and an error parameter $\epsilon \leq 1$, there is a sparse graph $\mathcal{G}$ which has $O(\frac{n \log n}{\epsilon^2})$ edges and the value of each cut in $\mathcal{G}$ is within $(1 \pm \epsilon)$ times the value of corresponding cut in $\mathcal{H}$.

As such, this theorem asserts that the upper bound of the sampling probability should be $p \approx \frac{n \log n}{n^2 \epsilon^2}$ to obtain a sparse graph with a bounded minimum cut error of ϵ. This theorem introduces a trade-off between the computational complexity of the graph and the minimum cut error, ϵ. Therefore, it is possible to sparsify a fully connected graph, by specifying a fixed error rate for the cut accuracy. Using the previous example mentioned of an image in Section 3.3.2 that is $n = 400 \times 300$, to represent a fully connected random field as a sparse representation via stochastic sparsification with an error parameter of $\epsilon = 0.1$, the number of edges in the underlying sparse graph should be less than or equal to $\frac{n \log n}{\epsilon^2} \approx 1.4034 \times 10^8$ (or alternatively a random graph generated with a selection probability of $p \leq 0.0097$) to satisfy the minimum condition.

The two aforementioned conditions determine the lower (connectedness condition $\frac{\log n}{n}$) and upper (minimum cut condition $\frac{n \log n}{n^2 \epsilon^2}$) bounds of the probability p considering a limited error for the result; within which the resulting sparse graph representation obtained via stochastic clique formation is a good approximation of the fully-connected CRF with a limited error bound. It should be noted that there is an adjustment between the accuracy and computational complexity of the sparse graph which should be optimized based on the application.

As one direction to future work, it is possible to investigate the connection between Karger's method and the proposed RCRF approach and to theoretically illustrate the lower and upper bounds for the number of connectivities in the graph to make sure with high probability that the result is the near optimum result for the problem.

6.2.2 Connectivity Computation via Abstraction

To construct the sparse graph representation of the fully-connected CRF based on the stochastic clique structure within the proposed framework, the one-to-one stochastic clique

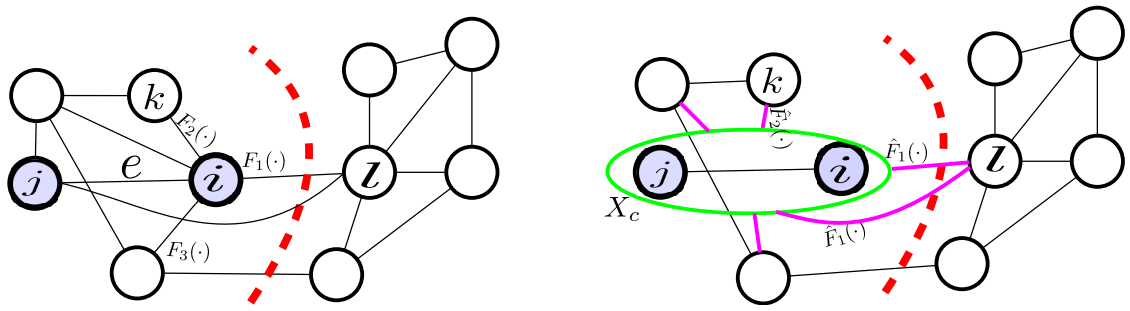


Figure 6.1: Nagamochi and Ibaraki theorem [116, 115]: If an edge in the graph is not in the minimum cut, then its corresponding nodes must be on the same side of the minimum cut result. It is assumed that the red dashed line is the minimum cut of the graph. In our example, the edge e is not crossed by the cut; therefore, two blue nodes corresponding to edge e are in the same side of the cut. As such, the connectivity measures between a node l and connected nodes that are similar to each other on the opposite side of the cut can be approximated as the same such that the resulting graph has the same minimum cut value as the original graph. The proposed abstraction strategy approximates the connectivity measure F between node l and node i as seen in left graph by the expected value of F between node l and the set of nodes $X_c = \{i, j\}$ (denoted by $E[F(x_l, X_c)]$) in the right graph. In this example after applying the abstraction strategy, $F_1(\cdot)$ and $F_3(\cdot)$ in the left graph are replaced by $\hat{F}_1(\cdot)$ in the right graph.

indicator function must be called for all nodes in the fully-connected CRF. The computational complexity of this procedure increases exponentially based on the number of random variables (e.g., number of pixels in the case of image modeling). However some of these similarity evaluations are redundant since there can be many similar nodes in the random field which have the same one-to-one similarity value with other nodes in the random field. To significantly reduce the computational complexity of computing connectivity measures, we are inspired by the work of Nagamochi and Ibaraki [116, 115], where it was shown that if an edge in the graph is not in the minimum cut, then its corresponding nodes must be on the same side of the minimum cut result. Figure 6.1 demonstrates the aforementioned the-

orem visually. As such, the connectivity measures between a node l and connected nodes that are similar to each other on the opposite side of the cut can be approximated as the same such that the resulting graph has the same minimum cut value as the original graph. Motivated by this, it is possible to propose an abstraction strategy where we approximate the one-to-one connectivity measures at significantly reduced computational complexity when compared to directly computing all connectivity measures.

Instead of computing the one-to-one connectivity measure between a node and all other nodes, the abstraction strategy computes the expected value of similarity, $F(\cdot)$, of the node and a group of nodes that are similar to each other:

$$F(x_l, x_i)|_{x_i \in X_c} \simeq E\Big[F(x_l, X_c)\Big] \tag{6.1}$$

$$F(x_l, X_c) = \Big\{F(x_l, x_i) | x_i \in X_c\Big\} \tag{6.2}$$

where X_c is the set of nodes in the graph, $x_i \in X_c$ is a particular node in the group of similar nodes X_c, and $E[\cdot]$ encodes the expectation function. The value of $E\Big[F(x_l, X_c)\Big]$ is approximately equal to the actual value of $F(x_l, x_i)$ since X_c is the combination of nodes that are similar to each other. Furthermore, even if this approximation does deviate from the actual value of $F(x_l, x_i)$, the nodes that are similar to each other are on the same side of the cut with high probability since they are grouped together as X_c and have zero value of $F(\cdot)$ between each other while result larger values (greater than zero or zero for exactly similar ones) of $F(\cdot)$ with outside nodes of X_c. As such computing the expected value instead of the actual value does not change the relationship amongst the nodes inside the set X_c and the outside nodes; therefore, the individual final cut edges are not changed based on the aforementioned theorem. It is worth noting that the intra-edges in the group of similar nodes have very large connectivity measures such that their corresponding edges have very low probability to be a cut edge. Therefore, the proposed abstraction strategy has a very low probability of changing the actual cut edges of the problem.

As shown in the right graph of Figure 6.1, instead of computing the connectivity measure $\Big\{F_1(\cdot), F_3(\cdot)\Big\}$ between node l and nodes i and j respectively, the abstraction strategy approximates these functions as $\hat{F}_1(\cdot)$, the expected value based on a set of the nodes

X_c which consists nodes i and j. Using this strategy, only one computation is done to approximate the connectivity measure between node l and all nodes in the set $X_c = \{i, j\}$.

As the second direction for the future work, we are proposing to design a new algorithm based on the aforementioned discussion to reduce the computational complexity of creating the underlying graph of RCRF. It will be a great study to analyze the behavior of the new algorithm and compare it with the proposed approach in running-time and modeling accuracy.

6.2.3 Graphical Models & Deep Learning Approaches

We proposed a stochastic approach in Chapter 3 to address the computational complexity of non-local random fields in the inference process. However it is possible to apply this approach on other dense graphical models to improve feasibility and usage of those methods in the training and inference steps.

Deep neural networks are the most interested methods in the past decade. It has been shown that they outperform conventional machine learning algorithms in several applications. They are a branch of machine learning that has seen a meteoric rise in popularity due to its powerful abilities to represent and model high-level abstractions in highly complex data. Deep neural networks have been shown to provide state-of-the-art performance for a number of complex tasks ranging from speech recognition [29, 61] and natural language processing [8, 26], to object recognition [63, 92, 102, 151]. However high-performance computing devices such as GPUs are the first and the foremost requirement to be able to take advantage of these types of algorithms.

There has been considerable focus in recent years on increasing the performance and capabilities of deep neural networks via strategies such as deeper architectures [151, 157, 184], network regularization [171, 181], and improvements in activation functions [48, 49, 64]. The reason behind this is that there is no high computational power system available in real-world applications. The available computing resources are practically limited to low-power, embedded GPUs and CPUs with limited memory and computing power in different

applications such as self-driving cars, surveillance cameras, and smartphone applications. However, the way that neural connections within deep neural networks are formed has not been an active area in deep neural network research in recent years. Therefore, deep exploration on different strategies for neural connectivity formation within a deep neural network may yield promising findings.

The idea of forming neural connections in deep neural networks as random graph realizations might help to address this issue. A deep neural network can be synthesized via a stochastic process with a possibility to specify the size of the network and also the sparsity of the network. Using the random graph modeling to characterize deep neural networks, one can then form the neural connections within a deep neural network as a realization of the random graph by initializing with a set of neurons, and randomly inserting neural connections between the set of neurons independently with a probability distribution.

Preliminary results [141, 142, 143, 144, 140] have demonstrated the effectiveness and great potential of this method in the deep learning community. For example, we showed in [140] that it is possible to take advantage of this approach to utilize a deep neural network as efficient feature extractor. We proposed an efficient learning and extraction of features [140] via this idea, where sparsely-connected deep neural networks can be formed via stochastic connectivity between neurons. Moving forward in this field of research, it is possible to utilize deep neural network (DNN) models in embedded systems [141]. We presented a new motion detection algorithm that leverages the power of DNNs while maintaining the low computational complexity needed for near real-time embedded performance without specialized hardware.

The sparse neural responses can be utilized in different applications such as image saliency detection [142]. In order to attain low computational complexity, random graph theory is leveraged to stochastically form the neural connectivity of deep neural networks such that the resulting deep neural networks are highly sparsely connected yet maintain the modeling capabilities of traditional densely connected deep neural networks. The neural responses are extracted from layers of a deep convolutional neural network architecture and are used to formulate the saliency detection problem.

Another interesting idea in this venue is to synthesize efficient deep neural networks in a iterative process [144]. A promising paradigm for achieving this is the concept of evolutionary deep intelligence, which attempts to mimic biological evolution processes to synthesize highly-efficient deep neural networks over successive generations. An important aspect of evolutionary deep intelligence is the genetic encoding scheme used to mimic heredity, which can have a significant impact on the quality of the offspring deep neural networks. It is possible to encode the heredity within the context of a stochastic process and random graph theory where a probabilistic model specifies the connectivity of a synapse in the network.

Based on the aforementioned ideas, continuing in this direction and applying random graph theory to model very large dense neural networks via a sparse and with very lesser number of parameters will help the methods to be applicable on practical real-world applications when there is no high-performance computing device. Analyzing and having a comprehensive study on the deep neural network and random graph theory is the third branch to be considered as future work for this thesis.

References

[1] A. Adams, J. Baek, and M. Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*. Wiley Online Library, 2010.

[2] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, and G. Diamos. Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR, abs/1512.02595*, 2015.

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2011.

[4] Y. Artan, M. Haider, D. Langer, T. van der Kwast, A. Evans, Y. Yang, M. Wernick, J. Trachtenberg, and I. Yetik. Prostate cancer localization with multispectral mri using cost-sensitive support vector machines and conditional random fields. *Transactions on Image Processing*, 2010.

[5] M. Beal. *Variational algorithms for approximate Bayesian inference*. University of London, 2003.

[6] A. Benczur and D. Karger. Randomized approximation schemes for cuts and flows in capacitated graphs. *SIAM Journal on Computing*, 2015.

[7] Y. Bengio. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2009.

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research (JMLR)*, 2003.

[9] A. Berger, V. Pietra, and S. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1996.

[10] C. Bishop. Pattern recognition and machine learning (information science and statistics), 2007.

[11] C. Bishop et al. *Pattern recognition and machine learning*. springer New York, 2006.

[12] C. Bishop, J. Lasserre, et al. Generative or discriminative? getting the best of both worlds. *Bayesian Statistics*, 2007.

[13] A. Blake, P. Kohli, and C. Rother. *Markov random fields for vision and image processing*. MIT Press, 2011.

[14] A. Blake, P. Kohli, and C. Rother. *Markov random fields for vision and image processing*. MIT Press, 2011.

[15] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *European Conference on Computer Vision (ECCV)*. Springer, 2004.

[16] B. Bollobás and F. Chung. *Probabilistic combinatorics and its applications*. American Mathematical Soc., 1991.

[17] B. Bollobás and Fan R. *Probabilistic combinatorics and its applications*. American Mathematical Soc., 1991.

[18] Y. Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *International Conference on Computer Vision (ICCV)*. IEEE, 2001.

[19] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2004.

[20] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2001.

[21] N. Campbell, K. Subr, and J. Kautz. Fully-connected crfs with non-parametric pairwise potential. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[22] K. Chan Kyo, P. Byung Kwan, and K. Bohyun. High-b-value diffusion-weighted imaging at 3 t to detect prostate cancer: comparisons between b values of 1,000 and 2,000 s/mm2. *American Journal of Roentgenology*, 2010.

[23] W. Chen, J. T Wilson, S. Tyree, K. Q Weinberger, and Y. Chen. Compressing neural networks with the hashing trick. *CoRR, abs/1504.04788*, 2015.

[24] F. Chung and L. Lu. *Complex graphs and networks*. American mathematical society Providence, 2006.

[25] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[26] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning (ICML)*. ACM, 2008.

[27] G. Cross and A.. Jain. Markov random field texture models. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1983.

[28] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. A framework and graphical development environment for robust nlp tools and applications. In *ACL*, 2002.

[29] G. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[30] T. de Campos, B. Babu, and M. Varma. Character recognition in natural images. In *International Conference on Computer Vision Theory and Applications*, 2009.

[31] N. Desouza, S. Reinsberg, E. Scurr, J. Brewster, and G. Payne. Magnetic resonance imaging in prostate cancer: the value of apparent diffusion coefficients for identifying malignant nodules. *The British Journal of Radiology*, 2007.

[32] P. Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 1968.

[33] J. Domke. Learning graphical model parameters with approximate marginal inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[34] P. Erdos and A. Renyi. On random graphs i. *Publ. Math. Debrecen*, 1959.

[35] M. Everingham, A. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 2014.

[36] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[37] S. Feng, R. Manmatha, and A. McCallum. Exploring the use of conditional random field models and hmms for historical handwritten document recognition. In *Second International Conference on Document Image Analysis for Libraries (DIAL'06)*. IEEE, 2006.

[38] P. Fieguth. Hierarchical posterior sampling for images and random fields. In *International Conference on Image Processing (ICIP)*. IEEE, 2003.

[39] P. Fieguth. Hierarchical mcmc sampling. In *International Conference Image Analysis and Recognition*. Springer, 2004.

[40] P. Fieguth. *Statistical image processing and multidimensional modeling*. Springer, 2010.

[41] L. Ford and D. Fulkerson. *Flows in networks*. Princeton university press, 2015.

[42] D. Freedman and P. Drineas. Energy minimization via graph cuts: Settling what is possible. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005.

[43] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *International Conference on Computer Vision (ICCV)*. IEEE, 2009.

[44] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *International Conference on Computer Vision (ICCV)*. IEEE, 2009.

[45] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1984.

[46] Z. Ghahramani and M. Beal. Variational inference for bayesian mixtures of factor analysers. In *Neural Information Processing Systems (NIPS)*, 1999.

[47] E. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 1959.

[48] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, 2010.

[49] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[50] M. Golumbic. *Algorithmic graph theory and perfect graphs*. Elsevier, 2004.

[51] Y. Gong, L. Liu, M. Yang, and L. Bourdev. Compressing deep convolutional networks using vector quantization. *CoRR, abs/1412.6115*, 2014.

[52] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision (IJCV)*, 2008.

[53] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE international Conference on Acoustics, Speech and Signal Processing*, 2013.

[54] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1989.

[55] J. Gross and J. Yellen. *Handbook of graph theory*. CRC press, 2004.

[56] M. Haider, T. van der Kwast, J. Tanguay, A. Evans, A. Hashmi1, G. Lockwood, and J. Trachtenberg. Combined t2-weighted and diffusion-weighted mri for localization of prostate cancer. *C American Journal of Roentgenology*, 2007.

[57] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971.

[58] S. Han, H. Mao, and W. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *CoRR, abs/1510.00149*, 2015.

[59] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Neural Information Processing Systems (NIPS)*, 2015.

[60] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, and A. Coates. Deep speech: Scaling up end-to-end speech recognition. *CoRR, abs/1412.5567*, 2014.

[61] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, and A. Coates. Deepspeech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[62] B. Hassibi and D. Stork. *Second order derivatives for network pruning: Optimal brain surgeon*. Morgan Kaufmann, 1993.

[63] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[64] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[65] X. He, R. Zemel, and MA Carreira-Perpindn. Multiscale conditional random fields for image labeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2004.

[66] X. He, R. Zemel, and MA Carreira-Perpi$\widetilde{n}$an. Multiscale conditional random fields for image labeling. In *Conference on Computer vision and pattern recognition (CVPR)*. IEEE, 2004.

[67] S. Hill, Y. Wang, I. Riachi, F. Schürmann, and H. Markram. Statistical connectivity provides a sufficient foundation for specific functional connectivity in neocortical neural microcircuits. *Proceedings of the National Academy of Sciences*, 2012.

[68] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 2012.

[69] M. Howard, E. Cox Pahnke, and W. Boeker. Understanding network formation in strategy research: Exponential random graph models. *Strategic Management Journal*, 2016.

[70] H. Ishikawa. Higher-order clique reduction in binary graph cut. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.

[71] H. Ishikawa and D. Geiger. Segmentation by grouping junctions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1998.

[72] E. Jaynes. Information theory and statistical mechanics. *Physical review*, 1957.

[73] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: coupling edges in graph cuts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011.

[74] M. Jordan. Learning in graphical models. 2004.

[75] O. Juan and Y. Boykov. Active graph cuts. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006.

[76] D. Karger. Random sampling in graph optimization problems. *Stanford University*, 1995.

[77] J. Kim and R. Zabih. Factorial markov random fields. In *European Conference of Computer Vision(ECCV)*. Springer, 2002.

[78] R. Kindermann, J. Snell, et al. *Markov random fields and their applications*. American Mathematical Society Providence, RI, 1980.

[79] k. Kitajima, Y. Kaji, K. Kuroda, and K. Sugimura. High b-value diffusion-weighted imaging in normal and malignant peripheral zone tissue of the prostate: effect of signal-to-noise ratio. *Magnetic Resonance in Medical Sciences*, 2008.

[80] R. Klinger and K. Tomanek. *Classical probabilistic models and conditional random fields*. TU, Algorithm Engineering, 2007.

[81] P. Kohli, M. Kumar, and P. Torr. P3 & beyond: Solving energies with higher order cliques. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007.

[82] P. Kohli, A. Osokin, and S. Jegelka. A principled deep random field model for image segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[83] P. Kohli, P. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision (IJCV)*, 2009.

[84] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[85] D. Koller, N. Friedman, L. Getoor, and B. Taskar. Graphical models in a nutshell. *Statistical Relational Learning*, 2007.

[86] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, 2004.

[87] I. Kovalenko. The structure of a random directed graph. *Theory of Probability and Mathematical Statistics*, 1975.

[88] I. N. Kovalenko. The structure of random directed graph. *Probab. Math. Statist.*, 1975.

[89] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Neural Information Processing Systems (NIPS)*, 2011.

[90] P. Krähenbühl and V. Koltun. Parameter learning and convergent inference for dense random fields. In *ICML (3)*, pages 513–521, 2013.

[91] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009.

[92] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*, 2012.

[93] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

[94] S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *Neural Information Processing Systems (NIPS)*, 2003.

[95] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *International Conference on Computer Vision (ICCV)*. IEEE, 2005.

[96] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. In *International Conference on Computer Vision (ICCV)*. IEEE, 2009.

[97] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. In *International Conference on Computer Vision (ICCV)*. IEEE, 2009.

[98] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, 2001.

[99] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 2015.

[100] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[101] Y. LeCun, J. Denker, S. A Solla, R. Howard, and L. Jackel. Optimal brain damage. In *Neural Information Processing Systems (NIPS)*, 1989.

[102] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2004.

[103] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[104] S. Li and S. Singh. *Markov random field modeling in image analysis*. Springer, 2009.

[105] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman. Sift flow: Dense correspondence across different scenes. In *European Conference of Computer Vision (ECCV)*. Springer, 2008.

[106] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2011.

[107] X. Liu, D. Langer, M. Haider, Y. Yang, M. Wernick, and I. Yetik. Prostate cancer segmentation with simultaneous estimation of markov random field parameters and class. *Transactions on Medical Imaging*, 2009.

[108] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference of Computer Vision (ICCV)*. IEEE, 2001.

[109] G. McLachlan and K. Basford. *Mixture models: Inference and applications to clustering*. Marcel Dekker, 1988.

[110] J. Moon and L. Moser. On cliques in graphs. *Israel journal of Mathematics*, 1965.

[111] D. Moran, R. Softley, and E. Warrant. The energetic cost of vision and the evolution of eyeless mexican cavefish. *Science advances*, 2015.

[112] K. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[113] K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 1999.

[114] K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1999.

[115] H. Nagamochi and T. Ibaraki. Computing edge-connectivity in multigraphs and capacitated graphs. *SIAM Journal on Discrete Mathematics*, 1992.

[116] H. Nagamochi and T. Ibaraki. A linear-time algorithm for finding a sparsek-connected spanning subgraph of ak-connected graph. *Algorithmica*, 1992.

[117] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems Workshop (NIPS)*, 2011.

[118] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Neural Information Processing Systems (NIPS)*, 2002.

[119] Y. Pan, X. Hou, and C. Liu. A hybrid approach to detect and localize texts in natural scene images. *IEEE Transactions on Image Processing (TIP)*, 2011.

[120] A. Papoulis and U. Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[121] S. Paris and F. Durand. A fast approximation of the bilateral filter using a signal processing approach. *International Journal of Computer Vision (IJCV)*, 2009.

[122] N. Payet and S. Todorovic. Rf2–random forest random field. In *Neural Information Processing Systems (NIPS)*, 2010.

[123] J. Peng, L. Bo, and J. Xu. Conditional neural fields. In *Neural Information Processing Systems (NIPS)*, pages 1419–1427, 2009.

[124] B. Potetz and T. Lee. Efficient belief propagation for higher-order cliques using linear constraint nodes. *Computer Vision and Image Understanding (CVIU)*, 2008.

[125] R. Prabhavalkar and E. Fosler-Lussier. Backpropagation training for multilayer conditional random field based phone recognition. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[126] Rohit Prabhavalkar and Eric Fosler-Lussier. Backpropagation training for multilayer conditional random field based phone recognition. In *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[127] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Neural Information Processing Systems*, 2004.

[128] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.

[129] L. Rabiner and B. Juang. Fundamentals of speech recognition. 1993.

[130] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *International Conference on Computer Vision (ICCV)*. IEEE, 2007.

[131] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *International Conference on Computer Vision (ICCV)*. IEEE, 2007.

[132] Martin Ratajczak, S Tschiatschek, and F Pernkopf. Sum-product networks for structured prediction: Context-specific deep conditional random fields. In *Proc Workshop on Learning Tractable Probabilistic Models*, 2014.

[133] J. Reynolds and K. Murphy. Figure-ground segmentation using a hierarchical conditional random field. In *Canadian Conference on Computer and Robot Vision*. IEEE, 2007.

[134] K. Ristovski, V. Radosavljevic, S. Vucetic, and Z. Obradovic. Continuous conditional random fields for efficient regression in large fully connected graphs. In *Conference on Artificial Intelligence (AAAI)*, 2013.

[135] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[136] A. Rosenkrantz, N. Hindman, H. Chandarana, F. Deng, S. Babb, S. Taneja, and C. Geppert. Computed diffusion-weighted imaging of the prostate at 3t: Impact on image quality and tumor detection. *Proc. Intl. Soc. Mag. Reson. Med.*, 2013.

[137] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 2004.

[138] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[139] M. J. Shafiee, A. G. Chung, A. Wong, and P. Fieguth. Improved fine structure modeling via guided stochastic clique formation in fully connected conditional random fields. In *International Conference on Image Processing (ICIP)*. IEEE, 2015.

[140] M. J. Shafiee, P. Siva, P. Fieguth, and A. Wong. Efficient deep feature learning and extraction via stochasticnets. 2015.

[141] M. J. Shafiee, P. Siva, P. Fieguth, and A. Wong. Embedded motion detection via neural response mixlture background modeling. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.

[142] M. J. Shafiee, P. Siva, C. Scharfenberger, P. Fieguth, and A. Wong. Nerd: A neural response divergence approach to visual saliency detection. *IEEE Signal Processing Letters*, 2016.

[143] M. J. Shafiee, P. Siva, and A. Wong. Stochasticnet: Forming deep neural networks via stochastic connectivity. *IEEE Access*, 2016.

[144] M. J. Shafiee and A. Wong. Evolutionary synthesis of deep neural networks via synaptic cluster-driven genetic encoding. *Neural Information Processing Systems Workshop (NIPS)*, 2016.

[145] M. J. Shafiee, A. Wong, and P. Fieguth. Deep randomly-connected conditional random fields for image segmentation. *IEEE Access*, 2016.

[146] M. J. Shafiee, A. Wong, P. Siva, and P. Fieguth. Efficient bayesian inference using fully connected conditional random fields with stochastic cliques. In *International Conference on Image Processing (ICIP)*. IEEE, 2014.

[147] M. J. Shafiee, A. Wong, P. Siva, and P. Fieguth. Efficient bayesian inference using fully connected conditional random fields with stochastic cliques. In *IEEE International Conference on Image Processing (ICIP)*, 2014.

[148] A. Sharon, G. Meirav, B. Ronen, and B. Achi. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007.

[149] A. Sharon, G. Meirav, B. Ronen, and B. Achi. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007.

[150] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision (IJCV)*, 2009.

[151] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[152] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR, abs/1409.1556*, 2014.

[153] R. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Neural Information Processing Systems (NIPS)*, pages 2368–2376, 2015.

[154] E. B Sudderth, Al. Ihler, M. Isard, W. Freeman, and A. Willsky. Nonparametric belief propagation. *Communications of the ACM*, 2010.

[155] C. Sutton and A. McCallum. *An introduction to conditional random fields for relational learning*. Introduction to statistical relational learning. MIT Press, 2006.

[156] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, 2007.

[157] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[158] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[159] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2008.

[160] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2002.

[161] Introduction to AI a modern approach. *Russell, S.J. and Norving, P.* Printece Hall, 2002.

[162] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Neural Information Processing Systems (NIPS)*, 2014.

[163] A. Torralba, K. Murphy, and W. Freeman. Sharing visual features for multiclass and multiview object detection. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2007.

[164] I. Ulusoy and C. Bishop. Generative versus discriminative methods for object recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005.

[165] V. Vapnik. Statistical learning theory. *Inc., New York*, 1998.

[166] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab.

[167] O. Veksler. *Efficient graph-based energy minimization methods in computer vision*. PhD thesis, Cornell University, 1999.

[168] J. Verbeek and W. Triggs. Scene segmentation with crfs learned from partially labeled images. In *Neural Information Processing Systems (NIPS)*, 2007.

[169] V. Vineet, J. Warrell, P. Sturgess, and P. Torr. Improved initialisation and gaussian mixture pairwise terms for dense random fields with mean-field inference. In *British Machine Vision Conference (BMVC)*, 2012.

[170] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008.

[171] L. Wan, M. Zeiler, S. Zhang, Y. Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning (ICML)*, 2013.

[172] C. Wang, N. Paragios, et al. Markov random fields in vision perception: a survey. 2012.

[173] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006.

[174] Y. Wang, K. Loe, and J. Wu. A dynamic conditional random field model for foreground and shadow segmentation. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2006.

[175] G. Welch and G. Bishop. An introduction to the kalman filter, 1995.

[176] J. Winn and C. Bishop. Variational message passing. In *Journal of Machine Learning Research (JMLR)*, 2005.

[177] A. Wong, M. J. Shafiee, P. Siva, and X. Wang. A deep-structured fully connected random field model for structured inference. *IEEE Access*, 2015.

[178] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[179] D. Yu, L. Deng, and S. Wang. Learning in the deep-structured conditional random fields. In *Neural Information Processing Systems Workshop (NIPS)*, 2009.

[180] D. Yu, S. Wang, and L. Deng. Sequential labeling using deep-structured conditional random fields. *IEEE Journal of Selected Topics in Signal Processing*, 2010.

[181] M. Zeiler and R. Fergus. Stochastic pooling for regularization of deep convolutional neural networks. *arXiv preprint arXiv:1301.3557*, 2013.

[182] Y. Zeng, D. Samaras, W. Chen, and Q. Peng. Topology cuts: A novel min-cut/max-flow algorithm for topology preserving segmentation in n–d images. *Computer Vision and Image Understanding (CVIU)*, 2008.

[183] X. Zhang, C. Moore, and M. Newman. Random graph models for dynamic networks. *arXiv preprint arXiv:1607.07570*, 2016.

[184] X. Zhang, J. Zou, K. He, and J. Sun. Accelerating very deep convolutional networks for classification and detection. *arXiv preprint arXiv:1505.06798*, 2015.

[185] Y. Zhang and T. Chen. Efficient inference for fully-connected crfs with stationarity. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[186] S. Zheng, M. Cheng, J. Warrell, P. Sturgess, V. Vineet, C. Rother, and P Torr. Dense semantic image segmentation with objects and attributes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.