Robust Image Labeling Using Conditional Random Fields

by

Maryam Nematollahi Arani

MSc, Tarbiat Modares University, Tehran, Iran, 2009

BSc, Shahrood University of Technology, Shahrood, Iran, 2005

A dissertation

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2017

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

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

Robust Image Labeling

Using Conditional Random Fields

Doctor of Philosophy 2017

Maryam Nematollahi Arani

Electrical and Computer Engineering

Ryerson University

Abstract

Object recognition has become a central topic in computer vision applications such as image search, robotics and vehicle safety systems. However, it is a challenging task due to the limited discriminative power of low-level visual features in describing the considerably diverse range of high-level visual semantics of objects. Semantic gap between low-level visual features and high-level concepts are a bottleneck in most systems. New content analysis models need to be developed to bridge the semantic gap. In this thesis, algorithms based on conditional random fields (CRF) from the class of probabilistic graphical models are developed to tackle the problem of multiclass image labeling for object recognition. Image labeling assigns a specific semantic category from a predefined set of object classes to each pixel in the image. By well capturing spatial interactions of visual concepts, CRF modeling has proved to be a successful tool for image labeling. This thesis proposes novel approaches to empowering the CRF modeling for robust image

labeling.

Our primary contributions are twofold. To better represent feature distributions of CRF potentials, new feature functions based on generalized Gaussian mixture models (GGMM) are designed and their efficacy is investigated. Due to its shape parameter, GGMM can provide a proper fit to multi-modal and skewed distribution of data in nature images. The new model proves more successful than Gaussian and Laplacian mixture models. It also outperforms a deep neural network model on Corel imageset by 1% accuracy. Further in this thesis, we apply scene level contextual information to integrate global visual semantics of the image with pixel-wise dense inference of fully-connected CRF to preserve small objects of foreground classes and to make dense inference robust to initial misclassifications of the unary classifier. Proposed inference algorithm factorizes the joint probability of labeling configuration and image scene type to obtain prediction update equations for labeling individual image pixels and also the overall scene type of the image. The proposed context-based dense CRF model outperforms conventional dense CRF model by about 2% in terms of labeling accuracy on MSRC imageset and by 4% on SIFT Flow imageset. Also, the proposed model obtains the highest scene classification rate of 86% on MSRC dataset.

Acknowledgements

It is my great pleasure to thank the many people whose support made this thesis possible.

Firstly, I wish to express my sincere gratitude to my Ph.D. supervisor, Prof. Xiao-Ping Zhang, for his great support, encouragement, and continual understanding. I learned a lot from him; many of his ideas became the very foundation of the present thesis.

I must thank the Department of Electrical and Computer Engineering of Ryerson University for giving me the opportunity to start this thesis and to conduct the necessary research work.

I wish to extend my thanks to my colleagues and friends whose warm support through the course of my Ph.D. helped me to complete my program. I am immensely grateful to Sheida Rasooli, Zahra Ahanchian, Nasim IranNejad, Jianan Han, Iris Choi, Farheen Fatima, Nastaran Rahnama, Yufang Hao, Behnaz PoorEbrahimi, Misagh Aghajani, Forough PoorHossein, Mahdi Takaffoli, Nikoo KuchakiPoor and Omid Alizadeh.

Last, but by no means least, my heartfelt thanks to my family for their never-faltering encouragement, support, and love.

Dedication

To my mother

Contents

List of Tables

List of Figures

Acronyms and Abbreviations

BN	Bayesian Network
BP	Belief Propagation
CBIR	Content Based Image Retrieval
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
cbDCRF	Context-Based Dense Conditional Random Field
DNN	Deep Neural Networks
DCRF	Dense Conditional Random Field
DRF	Discriminative Random Fields
DGM	Directed Graphical Models
EM	Expectation Maximization
FN	False negatives
FP	False positives
GMM	Gaussian Mixture Model
GGM	Generalized Gaussian Mixture
GGMM	Generalized Gaussian Mixture Modeling
HMM	Hidden Markov Model
KL	Kullback-Leibler

LMM	Laplacian Mixture Model
ML	Maximum Likelihood
MRF	Markov Random Field
NN	Nearest Neighbor
PGM	Probabilistic Graphical Models
PCA	Principal Component Analysis
PR	Precision-Recall
PCA	Principal Component Analysis
ROC	Receiver Operator Characteristic
SGD	Scholastic Gradient Descent
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
TP	True Positives
TN	True Negatives
TPR	True Positive Rate
FPR	False Positive Rate
UGM	Undirected Graphical Models

List of Important Symbols

a	Weight parameter
b	Bias parameter
c	Random variable representing scene type
$f_i(\cdot)$	Feature function corresponding to the association potential at the label site i
$f_{ij}(\cdot)$	Feature function corresponding to the interaction potential between the current site i and its neighboring site j
g	Gaussian function
i	Pixel/segment/node index
j	Pixel/segment/node index
k	Site/node/feature index
l	Label variable
$n(.)$	Number of
p	Probability of a random variable
$\widetilde{p}, \widetilde{P}$	Unnormalized probability distribution
q	Probability distribution
s	Clique or superpixel

w	Weight parameters
x	observation variable
y	random label variable
z	Unobserved variable
D	Number of observation data points
E	Energy field of a set of nodes in a graph
$\mathrm{E}, < \cdot >$	Expectation function
E_z	Expectation with respect to variable z
G	Graph nodes
J	Likelihood function
K	Number of features
L	Number of possible classes or labels, Likelihood function
M	Number of mixture components
N	Number of all pixel/segment nodes in the graph
P	Probability of a set of random variables
Q	Probability distribution of a set of random variables
S	Subset of connected nodes in a graph
X	Set of observation variables
Y	Set of random label variables
$Z(.), Z$	Normalizing partition function
$\mathcal{C}$	Set od possible scene types
$\mathcal{D}$	Entire dataset
$\mathcal{L}$	Predefined finite set of possible labels
$\mathcal{N}$	Set of neighborhood nodes

$\mathcal{R}$	Retrieval set of images
$\mathcal{S}$	Set of superpixels
α	Weight parameter
β	Shape parameter or weight parameter
δ	Delta function
η	Learning rate
θ	Set of all model parameters
μ	Mean variable, or compatibility function
π	Mixture coefficient
$\pi(.)$	Set of parent nodes
σ^2	Variance
ψ_i, ψ_u	Unary potential
ψ_c	Context-based unary potential
ψ_{ij}, ψ_p	Pairwise potential
$\psi^{(0)}$	First derivative of Gamma function
$\psi^{(1)}$	Second derivative of Gamma function
Ψ	Function defined over a subset of random variables
Γ	Gamma function

Chapter 1

Introduction

1.1 Objective

Scene understanding is one of the primary goals in the field of computer vision since it is a chief task in many applications of artificial intelligence. For example, in the field of robotic systems and autonomous vehicles, the high target is to autonomously plan and accomplish intended tasks by deliberately navigating through a typical environment. For successful navigation towards completion of any task, a detailed understanding of the target environment is necessary. Scene understanding covers a wide range of problems such as object detection, scene identification, image labeling and depth estimation, to name a few. Image labeling or semantic segmentation defined as simultaneous segmentation and recognition of objects in the image is an active research area [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] and the topic of concern in this thesis.

In image labeling the target is to categorize every pixel in the image to one of several predefined classes (figure 1.1). Given an image, the system should automatically partition it into semantically meaningful areas each labeled with a specific object class. In fact,

Figure 1.1: Goal of image labeling is to categorize each image pixel to one of several predefined classes.

primarily, identification of isolated objects and also object categories is a critical component of visual perception in human visual system [12]. Although, human visual system identifies and perceives the complex visual world comprehensively, visual perception or scene understanding in humanoid systems is a challenging problem.

1.2 Motivation

Image labeling is an intriguing research problem firstly because regarding low-level, mid-level and high-level computer vision, many problems could be formulated as an image labeling task. At each level, the semantic meaning of different label sets and label values differ based on the fact that which of the scene properties are captured by labels.

- *Edge detection*: In low-level vision where objects of interest are low-level image attributes such as edges, image labeling could be applied to label image pixels into edge or non-edge labels.

- *Noise removal and image restoration*: Image restoration compensates for or undoes defects which degrade an image. As a labeling problem, the label set contains the restored intensities.

- *Image segmentation*: In mid-level vision, the predefined labels do not carry any semantic meaning and arrangement of detected labels could vary based on different labeling criterion. Image segmentation, for example, partitions an image into non-overlapping segments that have a coherent appearance and the labels might take any order and might come in any number based on the segmentation criterion. As a labeling problem, region IDs constitute the label set.

- *Depth estimation*: Labeling might be applied to capture some continuous low-level attributes of the image. For example, a set of depth labels constitute the finite set of labels in depth estimation.

- *Stereo matching*: For every pixel in image 1, the target of stereo matching is to locate the corresponding pixel in image 2. As a labeling problem, the label set is the differences (disparities) between corresponding pixels.

In high level vision, objects are the attributes of interest in the image and the problem is to either separate a foreground object from the background in a binary labeling setting or partition the image into semantically meaningful regions such that each of them represents an object. Image labeling in this form is a chief task in many application of computer vision; such as:

- *Content-based image retrieval*: image labeling could be useful for image querying; For example, one might be interested to retrieve all images with animals in water from an image database as large as the World Wide Web. Image labeling could be applied to convert images to keywords using some statistical modeling approach that describe image components [13]; or develop a system that would also localize objects such as sky, trees, grass, and faces in images [14] and retrieve according images from the database. Labeling-based image retrieval methods are particularly necessary when the database carries no text annotations or incomplete annotations.

- *Inspection systems*: In industry, detection and recognition of machine appliances to assess safety conditions of different components is very important for insuring timely and safe operation of the system [15].

- *Humanoid robots and autonomous vehicles*: In order for robotic systems to be able to successfully navigate through a typical environment and accomplish tasks, it is necessary that they have a visual understanding of the environment that they explore in real-time [16, 17]. For example, if a robot knows what kind of object it is going to grasp, then it is easier for it to decide how to pick it up and hold [18]. Image labeling facilitates the process of identification of objects for the machine.

- *Medical image processing*: In the medical field, artificial intelligence has been frequently applied to highlight region of interest (ROI) in medical images [19]. Automatic ROI labeling has proved to help with surgical planning and improvement of diagnostic accuracy in automated diagnostic systems [20]. Automated ROI labeling also has the extra advantage of generating anatomical and functional atlases and also improving repeatability of the diagnostic studies and experiments [21] by substituting the rigorous practice of collecting training data through tedious manual

delineation of contours of organs, tumors and lesions or localization of ROIs.

- *Surveillance and security*: Automatic license plate recognition, automatic recognition of particular individuals and detection of suspicious behavior in sequential images are among other applications of image labeling which are useful in traffic surveillance and security systems [22, 23].

Image labeling could also be applied coupled with other computer vision systems. For example, knowing that a particular image belongs to the category of 'indoor' images, then the task of depth estimation for that image could be done more accurately [18].

1.3 Challenges

Due to the fact that computer vision systems lack the high power of human visual system for extracting high-level semantic information and also due to the loss of information in image formation, image labeling is not an easy task for computer-based systems. The extreme challenge stems from several factors. One source of difficulty is the great variation in types of objects. Some objects are structured with solid or deformable shape such as a 'horse' or a 'car'. On the other hand, other objects are formless and do not have a definite shape or structure, like 'sky' or 'tree'. Apart from variability of object types, different objects are characterized by different kinds of features in computer vision field. For example, bikes are best distinguishable with their shape or outline; however, 'grass' is identified with its color and texture; and 'sky' with its location extent in the image.

Another source of challenge is great within-class variability and also overlapping between-class characteristics. Regarding within-class variability, for example, although horses have distinctive shape but they show great variability in color. Also, although cars

Figure 1.2: Examples of sources of great within-class variability in objects.

are well-known with their rigid shape but they come with a different outline in different models. Besides, cars from same model might come in contrasting colors. Also, changes in object viewing angle, pose, image scale, lighting, partial occlusions, and environmental factors make computer-based image labeling difficult. As an example, a 'car' looks differently viewed from different angles; or a 'cow' might stand, sit or lie in different poses. Figure 1.2 shows several examples of the above mentioned challenges in computer-based image understanding. As an example of overlapping between-class characteristics, although 'sky' could be well distinguished by its geometric attribute but its color often is

Figure 1.3: Sky is often mistaken for water due to their reflection of each other.

Figure 1.4: Optical illusion and surreal art make image understanding challenging even for human eye.

mistaken for 'water' (figure 1.3). Also, local image patches might be ambiguous in terms of category labels. For example, a 'window' might be part of a 'car', a 'building' or an 'airplane'.

It is also notable that although understanding a complex image is an easy task for the human visual system, there are some image scenarios in which understanding the image might not be as straightforward and effortless not only for the computer but for the human eye too. Images of optical illusions and surrealist art are of the examples of these images (figure 1.4) [24].

In summary, segmentation and recognition of objects is a complicated task due to large variability in object types and great changes in imaging conditions and also presence of noise. The low-level local appearance information crudely available to computer vision systems such as color and texture are not enough to correctly identify objects. There is a lot of uncertainty not only in the information extracted from image but also in decision

making about class label of a patch of image.

1.4 Background

To harness the uncertainty inherent to the image labeling problem and to be able to develop robust automated recognition systems that scale well to large datasets, researchers have built successful vision systems based on probabilistic graphical models (PGM) [25, 1, 26, 2, 3, 27, 28, 29, 30, 31, 4, 32, 32, 33, 34, 11]. In the following, the logic behind success of graph-based image labeling is described.

1.4.1 Probabilistic Graphical modeling

Probabilistic graphical models bring probability theory and graph theory together [35]. Probability models capture the concealed orderly relations between image data and class labels; they exploit the informative prior knowledge about structures hidden in the data obtained from labeled images. Besides, since labels are dependent across pixels, graph theory is applied to take into account the long range spatial interactions within pixels, regions and objects. That is, probabilistic graphical models use a graph-based representation as the foundation for encoding a complete distribution over a multi-dimensional space and a graph that is a compact or factorized representation of a set of independences that hold in the specific distribution. Nodes in the graph represent the semantic label related to an image pixel or segment; and edges represent data/label dependencies (figure 1.5) [36].

Application of PGMs is promising because using contextual adjacancy information is necessary for successful image labeling. An image may provide information that could be utilized at several levels. Sometimes local information such as color and texture

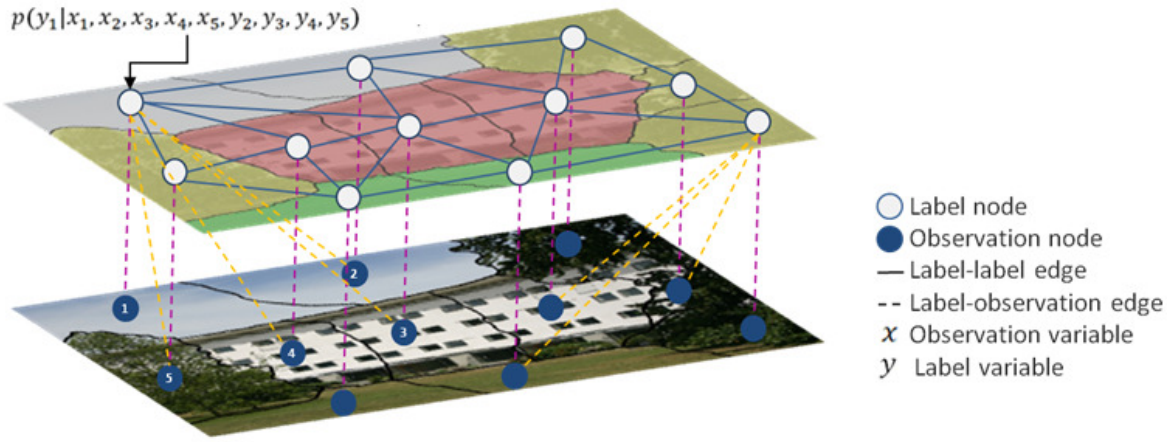


Figure 1.5: PGMs use a graph-based representation as the foundation for encoding a complete distribution over labels given observation data. Nodes in the graph represent the semantic label related to an image pixel or segment; and edges represent data/label dependencies. In this figure, magnet dash edges show dependency of current node to its local observation data. Yellow dash edges (shown only for two nodes for brevity) represent dependency of current label to contextual neighborhood observation data. Solid edges represent label compatibilities. The yellow edges are the surplus attribute of CRF graphical models in comparison with MRF models.

is enough to make a decision about the label of a patch of image. For instance, it is reasonable to label the green patches in the image as 'grass'. However, overlapping characteristics between classes might be confusing, such as the example in figure 1.3 where 'sky' and 'water' are two candidate labels for a blue or gray patch of image. Therefore, it is useful to take into account the spatial interaction of object classes in the image plane. In figure 1.3, the fact that 'sky' is usually located upper in the image whereas 'water' could be found at the bottom of the image, or in a similar scenario, the fact that airplanes fly in 'sky' not in 'water', help to disambiguate the decision-making for object identification [2]. Thus, information deployed from surrounding patches in the image could be helpful for successful image labeling. Employing probabilistic graphical models, internal configuration of objects and structural characteristics of their external

environment is extracted from labeled images and applied for accurate object recognition.

Markov random fields (MRF) [37, 38, 39, 40, 41] and Conditional random fields (CRF) [42, 29, 43] from the class of probabilistic graphical models have been frequently used to tackle the image labeling problem by capturing contextual information and spatial interaction of object classes at different granularity levels. Markov random fields are not the topic of discussion in this thesis because regarding image labeling task, they are not as successful as conditional random fields. The reason is that, in addition to interaction of labels, MRFs consider only local observation data for labeling the current pixel or segment of the image. However, labeling results of CRFs are furtherly a function of not only local, but also neighborhood or global observation data. Therefore, CRFs are more powerful in incorporating contextual information. Referring to graph in figure 1.5, the yellow edges which make the local label dependent on neighborhood observation are a characteristic of CRFs. The MRF graph will be a similar graph in which the yellow edges are excluded.

Moreover, MRF is a generative model in which the labeling problem is formulated as a joint distribution of the image observation data and corresponding labels. That is, for infering the label probabilities, two modeling steps are required. Generative models first require us to estimate the distribution for image features given the labels ($p(observation|label)$) and the object priors ($p(label)$); since they describe how image data has been generated regarding the labels and parameters of the model. Secondly, they use Bayes theorem to determine the posterior probability of labels given observations. Practically, generative models need a rich training set with many labeled images for estimating the parameters of the joint distribution of image data and labels. In contrast, CRFs [2, 3, 44, 28, 29, 31, 4, 33, 43] are a discriminative model which directly infer posterior probability of labels given observation data. Even when this posterior is

simple, the corresponding generative model might be complex and hard to train [45].

1.4.2 Conditional random fields

Basic CRF models are composed of unary (associative) and pairwise (interactive) potential terms (functions). Unary potentials are defined upon individual image pixels or segments (graph nodes) and capture the association between class labels and low-level image features such as color and texture; that is, they incorporate the image evidence to labeling task to obtain the probability of labels given the image low-level information. The pairwise potentials, however, are a function of all neighboring image pixels or segments and are meant to maximize local label agreement between neighboring pixels and to incorporate the compatibility between different object classes across the image. They are generally formulated as the probability of adjacency of class labels or occurrence of different object classes in certain spatial distances given the image data. Pairwise potentials assure data-dependent label smoothing and consistency.

Due to flexibility in the form of pairwise functions, contextual information has been investigated in CRF-based image labeling literature in different ways. Mostly, contextual information exploits relationship between objects in a scene in terms of semantic consistency, relative location and scale [46, 44, 2]. For example, the relative location potentials give higher probability to sky class than road class for pixels in the upper half of the image [44]. Also, inter-class relationship of objects may be considered so that identification of an object within the image would have an effect on the probability of finding others [1, 24, 46]. For example, polar bear and hippopotamus may never be seen together in one scene since one lives in the arctics and the other lives in tropical areas; or discovering a tree in the image suggests the label sky for the pixels above it and the label

grass for those lying underneath. CRF-based image labeling methodologies have proved to outperform other classification methods due to their capability to take into account different aspects of image context [47, 44, 30, 1, 31, 4, 48, 3, 43].

1.5 State of the art and proposed approches

As discussed above, developing an object recognition model which is not only accurate, but also efficient in terms of computational complexity and processing time is a challenging problem. The latter is particularly important when dealing with large image databases and for pixel-level identification of objects in high-resolution images. Regarding CRF-based image labeling, design and formulation of the potential terms and finding ways of applying contextual information are of major concern. Potential terms have to well represent the visual information; and contextual information have to be applied such that it ensures the self-consistency of the interpretation [49, 3, 4, 50, 33, 46, 6, 34, 7, 51, 52, 11, 9, 43]. Perspectives regarding implementation of both of these factors are disscussed in the following.

1.5.1 CRF potential functions

In image labeling and object recognition, conventional unary and pairwise potentials were primarily defined as summation of weighted feature functions. However, potentials of this form usually need an enormous number of features to render satisfactory results which makes their training and inference to be a difficult task. Besides, weighted feature functions are very sensitive to training initialization conditions and their training might

get stuck in local optima [42, 53]. To boost CRF performance, reserachers have tried to empower CRF potentials by high level prior modeling. Different studies use potential functions such as logistic [32], boosting [1, 5, 3], Neural Networks [2, 54], SVM classifiers [27], local support tensor machines [11], label transfer [7], mixture models [55] and combinations of them [49].

Recently, computer vision community has been dazzled with the advance of deep neural networks (DNN) due to its outstanding performance [56, 10, 49, 57]. In essence, deep networks transform the observation data into high-level abstract concepts in deeper layers [9]. They are able to learn compact, discriminative and high-level features [58, 59]. In fact, they render strikingly better results than systems with crafted hand-engineered features partially due to inherent invariance of deep DNN to local image transformations. Spatial shared weights derives deep networks to learn spatially shift-invariant features. This gives level to their ability of learning hierarchical abstractions of data. Also, pooling layers reduce sensitivity of the networks output to input shift and distortions. However, shift invariance hampers low-level vision tasks where precise localization is desirable; such as in semantic segmentation. Due to this invariance, deep DNNs are limited in refining local structures like pixels and patches [60].

Besides, the difficulty of solving non-convex optimization problems together with complexity of the prediction model can lead to overfitting phenomena or bad local optima in deep architectures. To moderate these problems, it is customary to initialize the supervised training with an unsupervised pre-training step. This procedure guides the optimization to a more reliable region in the weight space [61, 62]. However, it adds to computational cost of the model. It is also common to finetune one of the famous pretrained DNN architectures like AlexNet[56], GoogleNet [63] or VGG [64] which have been trained for an auxiliary task as scene (image) classification and re-purpose it for

semantic segmentation using a sizeable image dataset of annotated images. However, imagesets of at least order 10K number of images are required for finetuning of the network to achieve satisfactory results; which is not feasible particularly when training data is scarce such as in medical imagesets or satellite imagery; and also when having time constraints and when using hardware without GPUs (e.g. consumer laptops and smartphones).

Also, deep architectures struggle in discriminating visually similar observations since they have a limited context view and therefore, their results are spatially insensitive and very rough at object boundaries. To get around this problem, authors in [49] considered multi-scale context input to DNN and used the deep layer weights as features for training a contextual CRF model. Also, thanks to the factorized mean field approximation CRF inference model proposed by [3], in which message passing terms are redefined in terms of unary potentials, DNNs could be applied as the unary classifier in contextual CRF modeling. DNN and CRF parameters could be trained either jointly [65] or separately using a two step procedure [60]. However, this combination results in a computationally costly complex model in which parameter tuning is burdonsome.

Deep Neural networks and also boosting algorithms need many training data for obtaining good results; they are biased to vote for frequent classes and therefore produce poor results when applied to databases with imbalanced amount of training data in different classes. They are also computationally expensive and an inefficient method for ever-growing image databases with changing object variety. Label transfer image labeling methods are most successful when applied to large-scale databases with a rich variety of samples of different classes in different context [7, 8]. For example, they have proven high labeling performance in large imagesets such as SUN database [66] and SIFT Flow dataset [67]. However, they generate poor results when applied to small or noisy image

sets such as MSRC [1] and Corel [2, 68].

Mixture models are one of the high performance yet efficient approaches to image labeling [69, 70]. Mixture models capture well within class variability of objects (flowers come in different colors) [1, 71, 72]; Augmented by CRF modeling which well discriminate visually similar samples of different classes due to considering contextual information, mixture models have proved to achieve high labeling accuracy [73]. In [53], authors show that distributions of features in natural scene images are better approximated by a Laplacian distribution than a Gaussian. They show by experimentation that Laplacian feature functions outperform both conventional weighted feature functions, SVM classifiers and Gaussian mixture models. However, state of the art literature questions the ability of firmly-shaped distributions such as Gaussian or Laplacian densities to precisely approximate observation data of different object classes [70, 74]. Despite their efficiency and efficacy, rigidly-shaped distributions such as Laplacian and Gaussian fail to capture data characteristics where data fluctuations happen very smoothly; so that they even give rise to induction of atypical results due to erroneous modeling of data.

In this thesis, a new feature function for accurate segmentation and labeling of nature images by deploying generalized Gaussian mixture modeling (GGMM) of image features is studied. Having an additional shape manipulation parameter, GGMM can model data characteristics more accurately. We propose to bring the flexibility of GGMM for data modeling into the CRF framework to leverage the discrimination power of the feature functions while maintaining low complexity. We investigate the effectiveness of the proposed new feature functions in comparison with their Laplacian and Gaussian counterparts, conventional weighted feature functions, SVM and deep learning methods. The proposed GGMM-based CRF outperforms the other methodologies and produces less erronous and more consistent labeling particularly in even regions of background of

the image.

1.5.2 Wide extent context information

The success of CRF-based image labeling is very much restricted by the extent to which information is allowed to flow in the image via the pairwise potentials. However, pairwise potentials are limited in their ability to model long-range connections within the image and generally produce excessive smoothing of object boundaries. Higher-order potentials and hierarchical connectivity between image regions have proved to substantially improve labeling accuracy [4, 2, 27, 75]. Nonetheless, these approaches are restricted by the accuracy of unsupervised image segmentation, which is used to compute the regions on which the model operates.

To produce accurate label assignments around complex object boundaries, recent research on CRF-based image labeling has been devoted to development of efficient inference algorithms for fully-connected (dense) CRF models which connect each pixel to every other pixel in the image [3, 28, 33, 34]. They render iterative inference algorithms which are computationally efficient and refine object boundaries at a pixel level. However, dense random fields are confined to the success of the initial unary classifier. If the initial unary potentials fail to identify the objects in the image correctly, the iterative algorithm cannot revise the object labels and continues to refine boundaries of wrong labels. Dense random fields are also very prone to over-smoothing small objects from foreground (thing) classes in the large pool of pixels from background classes.

In this work, we proposed to integrate global semantics of the image with pixel-wise dense inference to make dense inference robust to initial misclassifications of the unary classifier and to preserve small foreground classes. We utilize global scene type of the

image to eliminate ambiguity of local context; certain types of objects are more likely to happen at specific scenes or object settings. For example, despite cars, boats are more probable to be seen at a sea shore scene than inside city; or mice are more known with office desks than kitchen tables. The new context-based inference corrects wrong predictions by the unary potentials in favor of objects coherent with the scene type.

The proposed model applies scene-object co-occurrence information in favor of object-object co-occurrence prediction. We factorize the joint probability of labeling configuration and image scene type and use the mean field approximation to obtain prediction update equations for labeling individual image pixels and the overall scene type of the image. We apply scene type context as a model selection cue in the mean field approximation inference to alleviate sensitivity to initialization and severe smoothing problem. We use whole image descriptors to discriminate distinct environmental categories using an SVM scene classifier and then define the CRF unary potentials conditioned on the overall scene type of the image. The CRF pairwise potentials connect each image pixel with all other pixels in the image to account for long-range interactions of objects. We derive the inference algorithm for the proposed context-based dense CRF model.

Contextual information in the form of prior knowledge about the whole scene and world regularities is informative to identification of individual objects. This phenomenon has been applied to facilitate object recognition in previous studies along with sparsely-connected random fields [11, 6, 76] and has been shown to be useful. Here, a new context-based dense CRF model is proposed which applies scene-level semantic information to improve pixel-level object recognition. A quantitative evaluation on the MSRC [1] and Sift flow [67] image sets show that the proposed model outperforms conventional dense CRF labeling. The proposed context-based dense CRF model competes with other state of the art context-based CRF approaches that apply scene level information and also

improves scene prediction results to render the highest scene detection results obtained on these datasets.

1.6 Main Contributions

As mentioned earlier, there are two important problems regarding CRF-based image labeling. The first one is proper design of the potential functions so that observation data of different semantic categories are effectively discriminated. The second problem is how to employ contextual information to boost labeling accuracy so that the CRF output is smooth and consistent. This thesis assigns two chapters to developing labeling models that each of them tackles one of the above mentioned problems. By addressing these problems, the proposed models leverage existing models in terms of labeling accuracy. Therefore, the main contributions of this thesis are:

1- New potential functions based on mixture of generalized Gaussian distribution are proposed. The wide variety of data distributions in natural images could be modeled with a generalized Gaussian distribution with proper value of the shape parameter. This is because in spite of Gaussian and Laplacian distributions which have fixed order, the adjustable shape parameter in generalized Gaussian distribution can capture data variations from a large range of sharp to very flat variations. Therefore, the proposed CRF model based on new feature functions outperforms its Laplacian and Gaussian counterparts in terms of labeling accuracy. The proposed model is particularly more accurate in even background of images where there are little variation in the image regions. EM algorithm and Newton-Raphson method are combined to estimate parameters of the proposed feature functions.

2- A new context-based dense CRF model (cbDCRF) is proposed. Model components are selected so that the new model yields optimum performance efficacy and computational efficiency. The proposed cbDCRF model outperforms the conventional dense CRF model in preserving small foreground objects in the large pool of pixels of image background classes. The new model also is able to correct wrong initializations of the unary classifier since it applies scene-object coherence to improve object-object consistency. Using the new model, we are able to increase both object recognition and scene classification rates. For the new model, mean field approximation from the class of variational inference algorithms is applied to write the inference equations in a format which makes them computationally efficient to calculate.

1.7 Outline

In chapter 2, the background work in image labeling and CRF-based object recognition are reviewed. Various types of design of graphs and structures and formulation of context in CRF model that has been applied to the image labeling task is discussed. Chapter 3 introduces the mathematics of conventional CRF-based image labeling and explain popular inference and training algorithms applied for performing image labeling.

Chapter 4 discusses the new GGMM feature functions proposed to improve the CRF-based labeling accuracy. To develop the GGMM-based feature functions, Expectation-maximization (EM) algorithm will be used [77, 78] to estimate mixture parameters. Belief propagation (BP) and stochastic gradient descent (SGD) algorithms will be utilized for CRF inference and training, respectively. We demonstrate that in comparison with their Laplacian and Gaussian counterparts, conventional weighted feature functions, SVM and

deep learning methods, the proposed GGM feature functions provide more accurate labeling of nature images.

Chapter 5 introduces the new context-based CRF model for image labeling. We define context-based unary and pair-wise potentials and also derive the inference algorithm for the context-based model based on the mean field approximation method. Quantitative evaluation of the new model on the benchmark database demonstrates that the proposed framework outperforms conventional dense CRF labeling in terms of object recognition accuracy. Qualitative comparisons show that where the conventional dense CRF adheres to the wrong initializations and fails, the new context-based model identifies the objects correctly. Chapter 6 discusses the main contributions of this thesis and elaborates on future work based on the current work and also state of the art literature.

Chapter 2

Literature Review

There are many ways to incorporate contextual information using PGM. Some studies have manipulated the unary and pairwise potentials to capture specific contextual characteristics; and some approaches have tried to add extra potential terms to traditional random fields to model different kinds and levels of contextual characteristics. That is, different methodologies vary in the contextual cues applied and their design of the potential function to capture them. Contextual reasoning is a critical piece of object recognition puzzle although benefit of context varies per object class so that for many object classes incorporation of context improves the detection accuracy whereas for other object classes it is not as effective. Ultimately, inclusion of context results in more reasonable detection errors. Most errors happen where classes share similar context; for example, in the confusion matrix analysis, airplanes are confused for birds or cats are confused with dogs [79]. This section reviews some of the prominent and current literature on the subject.

2.1 Unary Potentials

2.1.1 Parametric models

Weighted feature functions [42], logistic classifier [32], boosting [1, 5, 3], Neural Networks [2, 54], SVM classifiers [27], local support tensor machines [11], mixture models [55, 53, 73], and combinations of them [49] could be utilized as unary potentials of the labeling graphical model. Low-level image features such as color, texture, class-specific shape prior and location priors could be used to train these algorithms. Location priors are important because for example, sky tends to happen at the upper part of the image all the time whereas water appears in the lower part and windows are at the middle of the image. Shape priors work well particularly when dealing with rigidly shaped objects as 'cars' and 'faces' [6, 1].

Researchers in [80] show that for improving object recognition accuracy, extracting effective contextual features to represent image and object attributes are of more importance than developing complicated structures of recognition models. Regarding this fact, [1] proposes approaches for exploiting textural contextual information via definition of proper context-based features. Researchers in [1] define texture-layout filters that can capture contextual information in terms of texture appearance and their corresponding spatial layout within and between object classes. Texture-layout filters survey different types of textural information within specific regions relative to each pixel in the image. In [1], contextual information is deployed via computing features within different large neighborhoods of each pixel.

Also, output of hidden layers of DNN has become a popular source for rich visual features. Researchers in [9] apply the output of 6^{th}, 7^{th} and 8^{th} hidden layers of DNN

as hierarchical features for SVM classification. Nodes in hidden layers are aligned to concepts (classes to be learned). [81] also used convolutional networks to calculate zoom-out feature representation for each superpixel in the image obtained over a sequence of nested regions of increasing extent, and then, Caffe [82] was used to train DNNs for classification of each superpixel. Since training data is scarse, [9] and [10, 59] pre-train the network in a supervised way for an auxiliary correlated task such as scene (image) classification and then finetune it for image labeling. Some studies apply the available pre-trained DNN architectures like AlexNet[56], GoogleNet [63] or VGG [64].

Also, since due to shift-invariance property of DNN the output labeling of deep networks are spatially imprecise, [10] carries out a post-processing step which combines semantic features from a coarse deep layer with visuall features from a fine shallow layer to yeild visually accurate segmentation. [60, 83] also apply CRF modeling as a post-processing step to refine segmentation details at the output of the networks. In [65], DNN is trained jointly with CRF modeling to obtain precise segmentation. A deep feed-forward neural network is proposed by [84] that utilizes the contextual information from the entire image, through bottom-up followed by top-down context propagation via random binary parse trees. With this approach, feature representation of every super-pixel in the image is improved for better classification into semantic categories.

2.1.2 Nonparametric approach

As opposed to parametric approaches that are based on learning generative or discriminative algorithms, nonparametric definition of unary potentials rely on image retrieval and matching. In order to exploit contextual information, recent research looks for contextual cues beyond a single image. This approach is based on the fact that there are

images in the database that are very similar to the query image which share same spatial and semantic layout and for which annotations are available. The set of similar images and the query image are assumed to contain instances of the same object classes. They propose to transfer desired information from set of similar images to the query image and interpret its semantic configuration [7, 8, 85, 86, 87].

Researchers in [7] propose a labeling system based on recognition-by-matching approach. Using nearest neighbor approach, they first retrieve a set of images in the database that are similar to the query image in terms of scene configuration using GIST feature matching [88]. Then, they establish dense scene alignment between database images and the query image using SIFT flow matching [89]. Finally, they transfer annotations from most similar images onto the query by generating label probability maps.

Researchers in [8] propose an effective nonparametric approach to image labeling based on superpixel label transfer. They first retrieve the set of similar images and then, they perform a superpixel matching step. They perform feature extraction for all superpixels in the image and score each superpixel for the class labels that are present in the retrieved image set. They argue that, in comparison with image matching, super pixel matching allows for more variation between the layout of the test image and the images in the retrieval set. These studies show that the non-parametric approach to image labeling competes with training-based labeling methods in terms of performance with the extra advantage that their approach requires no training and scales well to large databases. A recent study integrates parametric and non-parametric models [54].

2.1.3 Object detectors

Window-wise object detectors are used as unary potentials within CRF model to combine object detection in semantic segmentation (image labeling) framework [90, 6, 91, 92].

2.2 Interaction Potentials

2.2.1 Interaction concepts

CRF is known to be an appropriate tool for image labeling because it can integrate many interactivity terms for modeling contextual relationships between different image sites and different object classes.

Label smoothing - Researchers in [29] apply CRF for the first time for using contextual information in the form of spatial dependencies to consider dependency of pixel intensity values in even areas or structures such as lines and edges. They define the interaction potentials as data dependent smoothing terms which are a function of characteristics of every two nodes that share an edge, similar to the Ising mode. They show that data interaction term in CRF reduces the false positives in addition to increasing the detection rate. In [29], both association potential and data-independent interaction potential are defined in a logistic regression form. Data-dependent term of the pairwise potential is written in the form of a Gaussian format similarity metric in [1, 3, 28].

Label Consistency - Some studies incorporate context via modeling statistics of object relations [93, 94]. They learn a joint distribution of different object classes and provide prior information on different combinations that objects appear in the world. Presence of certain types of objects in an image is often correlated with one another [46, 95]; for example, monitor, keyboard and mouse frequently appear close together. In [46] and [80], a new energy term is added to model object class co-occurrence statistics and it is shown that the co-occurrence potentials suppress uncommon combinations of classes such as boats on roads. [96] develops a data-dependent object relationship model using link propagation techniques. In [95], contextual relevance is applied to maximize label

agreement. They compare two sources of contextual relevance, one learned from training data and another queried from Google Sets. In [97], a sliding window method and unsupervised image region clustering are combined to leverage 'stuff' classes such as 'sea', 'sky' and 'road' to improve detection of objects of 'thing' classes such as 'cars'. In [5], easy objects such as 'monitor' are detected first and then contextual information is passed to detect more difficult objects such as 'mouse'.

Spatial relationships - Objects in images follow certain relative spatial configurations; for example, 'sky' appears to be above 'water' or above 'vegetation' [98]. Detection of one object anywhere in the image has effect on where other objects appear in the image. Researchers in [44] encode spatial offset preferences between objects by generating the relative location probability maps. Relative location feature maps are computed per class; for example, relative location features between chair and road class encourage placement of chair on top of road pixels. In [99], four prototypical spatial relationships - above, below, inside and around are quantized. Researchers in [100] estimate the horizon line in images and incorporate a measure of object positions relative to the horizon line. The viewpoint, defined by the horizon position in the image and the camera height, directly affects the position and size of the objects in the image.

Object presence - Some studies define a potential term which carries the information of presence or absence of each object in the image [101, 6]. Global image level features can be used to identify presence of objects [100, 102]. These studies are built on the hypothesis that there is a strong correlation between statistics of low-level features of the overall scene which describe structural scene properties and presence of objects in images. They used low-level feature statistics across the whole image to prime object detection.

Global context - There is a strong relationship between the environment and the objects in it [102, 103, 6, 76]. Psychological studies suggest that seeing a picture, humans analyze the overall scene presented in the image before scrutinizing individual objects and details [104, 105]. On the other hand, scenes are identified with the kind of objects that exist in the image. That is, scene classification and object detection are reciprocal tasks. Researchers in [106, 103, 11, 107, 108] propose to take advantage of scene type prediction for improving object detection accuracy; because it can reduce the number of classes to consider in an image. They use a global feature vector extracted from the whole image to represent and then classify the scene presented by the image [88, 109, 110, 27] and then combine this prediction with local object detectors in a discriminative classification framework to detect presence of objects in images. They argue that scene type classification is easier and much less time-consuming than labeling individual objects and it also improves object detection results. However, their method is not labeling the image at a pixel level.

2.2.2 Connectivity structure

Interaction order - In preliminary CRF work [29], the interaction potentials are assumed to have a 4-nearest neighbor grid structure. This structure fails to capture long range dependency priors between image sites and object classes. High-order connections between nodes that are spaced further apart in the image have been considered in response to the need for richer and more expressive prior information from the training data. Some approaches define hierarchical connectivity and high-order potentials over image regions [2, 94, 4, 46, 98, 80, 75, 111] whereas others consider fully-connected dense random fields defined over all image pixels [3, 28, 34, 33]. Comparing the two approaches,

accuracy of image labeling in the former approach is limited by the accuracy of image segmentation algorithm that is applied to partition the image into regions. The accuracy diminishes with the segmentation-based approaches particularly around object boundaries since the regions obtained via segmentation might not share exact boundaries with semantic objects. However, complexity of inference in dense random fields limits their application where real-time approaches are required for image labeling [95, 27]. Higher order connectivity in graph modeling results in large computational costs for different inference methods [80]. Recent research proposes the application of approximate inference models to overcome the limitations imposed by computational cost [3].

Scale - Researchers in [4, 112] combine multiple segmentations with various-scale granulaties and show that this approach results in more accurate segmentation of objects. In [4], authors define a criterion for evaluating consistency goodness of image regions obtained from image segmentation step and propose high-order variance-sensitive potentials that are sensitive to quality of segmented regions. They take account of the fact that one region might belong to several objects at the edges of object boundaries and therefore avoid over-smoothing and generate more delicate segmentation at object boundaries.

In [80], a segment-based CRF is proposed to incorporate multi-scale features of pixels and segments by directly considering the features of pixels that constitute a segment so that the unary potential associated with a segment is the sum of its pixels unary potentials. In [51] a fully connected CRF model is defined over overlapping image segments obtained from multiple segmentations of the image. Researchers in [2] formulate a multi-scale CRF model which combines information from three local, regional and global scales in a product of experts framework. Regional-scale feature functions are designed to capture within label regularities and cross-label boundary regularities in mid-scale; whereas global-scale feature functions capture coarser patterns in the entire label field.

Hierarchical and multi-scale models - Researchers in [98] propose a two-layer hierarchical CRF to develop a unified approach for incorporating local and global context in images. The first layer models short range interactions to ensure consistent labeling and the second layer models long-range interaction between objects and image regions. Researchers in [6] define a logistic-regression-based scene type unary potential in the CRF framework and train a one-vs-all SVM classifier over different scene types. In order to model the relationship between scene type and the objects that may appear in it, they define a scene-object compatibility potential to specifically narrow-down possible classes to those that are probable within a specific scene type. Researchers in [75] propose a hierarchical three-layer CRF (pixels, segments and super-segments) and further generalize the quality-sensitive potential terms in [4] to encourage consistency between neighboring segments within and between segments from different layers of hierarchical CRF. This framework integrates multi-scale features from multiple fine to course segmentations of the image. Also in [113], a joint-CRF on multiple levels of an image segmentation hierarchy is formulated. [6] also employs a higherarchical approach in a CRF framework to reason jointly about regions, location, class and spatial extent of objects, presence of a class in the image, as well as the scene type; the results reported are among highest accuracies obtained on MSRC dataset. Researchers in [18] argue that apart from scene type and object interactions there are more aspects to an image such as depth and objects saliency that could be applied to exploit contextual information. They combine classifiers of different image modalities such as scene categorization, depth estimation, saliency detection, event categorization and object detection to obtain an understanding of the whole image and objects in it. There are structured connections between different classifiers and outputs of later classifiers are available to earlier classifiers as a feedback signal to modify errors that could be corrected by the information from another modality.

Dense random fields - Researchers in [3] propose a highly efficient inference algorithm for pixel-wise fully-connected dense random fields. They define the interaction potentials to take the form of a mixture of Gaussian kernels multiplied by a Potts model with label consistency for similar colors. They apply mean field approximation algorithm and approximate high dimensional filtering to reduce computational complexity of message passing step in inference. Their experimentation shows that their fast approach enables greatly refined pixel-level object boundaries. Researchers in [34, 114] argue that the mean field inference method used in [3] is sensitive to unary initialization conditions and might get stuck in local minima. They propose a hierarchical mean-field approach to prepare a framework for providing good initial conditions. They apply SIFT-flow based label transfer method [7] to find good initial conditions. Then they transfer the initial labels from a coarse CRF to a finer grid CRF. They also generalize the zero-mean Gaussian pairwise potentials so that it could take non-zero mean values. Researchers in [28] relax the Gaussian assumption of pairwise potentials by [3] to be able to encode more statistics by arbitrary distributions. They propose an efficient inference algorithm with the help of convolution based on quadratic programming (QP) relaxation [115]. They show that gradient of QP relaxation can be efficiently computed using convolution. The interaction potential in their CRF model is defined as a combination of color contrast and spatial relationships between two classes. The color contrast term is the smoothing term and the spatial relation term models the probability that two categories co-occur at a relative distance. They argue that their potential terms can capture object size information as well. They show that their model captures detailed pixel-level spatial information and preserves object contours.

Arbitrary structure - Instead of assuming a fixed structured graph for images, some studies assume dynamic structure for graph connectivities [26]. Researchers in [5] exploit

contextual correlations of object classes by introducing boosted random fields. They apply boosting to learn the long range connectivities in the two dimensional CRF graph structure of an image. Node connectivities are chosen by a weak learner that has access to a dictionary of graph fragments. This dictionary of graph fragments implies the typical spatial arrangements of objects in images and is learnt during training. Connections from different locations in the image and between different object classes are added to the dictionary during training. The overall graph structure takes form by assembling graph fragments in an additive model.

Chapter 3

Preliminaries

In image labeling, a discrete value from a finite predefined set of label values is assigned to every pixel in the image so that recognition and segmentation of multiple object classes are performed concurrently. Mathematically, let $X = \{x_1, x_2, ..., x_N\}$ denote the set of observation vectors of image pixels/segments and $Y = \{y_1, y_2, ..., y_N\}$ the set of N random variables corresponding to pixel/segment labels, each of which may be a value from the finite set of labels $\mathcal{L} = \{1, 2, ..., L\}$. Image labeling can then be formulated as finding a mapping from X to Y. Mathematical algoritms usually model such a mapping through an optimization problem as:

$$Y^{opt} = \arg\min_{Y} E(Y, X; \theta) \tag{3.1}$$

where energy function $E(Y, X; \theta)$ is the cost function quantifying some quality measure of configuration Y in the solution space given observations X and model parameters θ. Finding this expert mapping involves three modeling, inference and learning tasks [116].

The modeling step includes: (i) the choice of discrminative visual features X; (ii)

appropriate representation of the solution space of variables Y; and (iii) designing the energy function $E(Y, X; \theta)$. Observations X could be low-level image features such as pixel intensity values, color distributions, texture features, shape features, output of multi-scale bandpass filters, descriptors in the frequency domain, etc. The necessity of steps (ii) and (iii) comes from the fact that pure object-centered representations such as local color and texture are not sufficient to find the correct mapping to the label space particularly in poor quality images due to degraded imaging conditions such as large distance, noise or occlusion. A good labeling approach takes advantage of all three levels of object characteristics; low-level representations such as color and texture; mid-level cues of region continuity and shape; and high-level semantics considering object co-occurrences and inter-object relationships such as relative location, scale or compatibility.

In other words, finding the true mapping in (3.1) involves performing complex visual and also contextual reasoning. The image structure and prior knowledge about world regularities, in other words contextual information, are essential to reliable object recognition when local low-level evidence of objects are not enough to identify them. For example, a small isolated black blob-like object in a slightly blurred image might not be easily identifiable. However, in a context of a computer desk, a screen and a keyboard the same vague blob-like object is most likely a computer mouse. Importance of employing contextual reasoning has been recognized in early research in this domain [117, 118].

Contextual reasoning calls for a compact structure to model the inter- and intra-relationship between variables $\{Y, X\}$. Most successful approaches formulate labeling problem in the framework of a probabilistic graphical model (PGM); each image site and random variables associated with their semantic labels are represented with the nodes in a graph [103, 5, 4, 3, 60]. Dependencies between neighboring image sites at different scales and adjacency orders are represented with edges between these nodes. The model is then

solved as a discrete energy minimization task performed over the entire graph to find out the data likelihood and the dependency properties of graph variables. PGMs provide a flexible modular way to combine regularization, data likelihood, and prior terms with other contextual cues in a single formulation. It also provides a means for visualizing the model structure and therefore facilitates the design and definition of different terms.

The inference step for finding the optimum mapping in (3.1) has to search the entire solution space including all configurations of variables Y to find the optimum configuration that minimizes the energy function $E(Y, X; \theta)$. The computational demands of inference could grow largely depending on the number of variables and edge orders. Thanks to the factorization property of PGMs , a graph structure helps to develop efficient inference algorithms. An Example of inference algorithms solved for models proposed in this thesis (section 5.2) is elaborated in appendix A.

The learning step aims to select the optimal model parameters based on the training data (section 4.3). The factorization property of PGMs also provide a way for piecewise training of the structured models [119]. Mathematical demands of using graph modeling to incorporate context information will be discussed in this chapter.

3.1 Probabilistic graphical modeling

Probabilistic graphical models (PGM) provide a flexible and consistent framework for employing contextual information to label image regions. Labeling approach based on PGM treats image components (pixels or patches of regular (windows) or irregular (superpixels) shape) as random variables and applies parametric probability distributions to model the regulation and interaction among this random variables. A structured graph G is comprised of a number of nodes which represent random variables, and edges which in

the form of an inter-node line connection represent the interaction between these random variables. An edge between two nodes represents dependency between their corresponding random variables. Two random variables with no edge in between are conditionally independent. The joint probability distribution of all random variables in the graph is formulated as products of functions defined on connected subsets of nodes [25]:

$$P(X) \propto \prod_{S \subset G} \Psi_S(x_S) \tag{3.2}$$

where S is a subset of connected nodes in graph G, x_S denotes the set of random variables represented by node S, and Ψ_S is a function defined over random variables x_S which is not necessarily a probability distribution but models the inner workings of variables x_S. Definition of Ψ_S depends on the type of variables x_S (observation/label) and the type of graph which may be directed or undirected.

3.1.1 Directed Graphical Models

If edges in the graph show a direction from one node (parent) to another (child) indicated with an arrow, then we call this graph a directed graph and interpret that there is a causal relationship between the nodes; although causality is not an inherent attribute of directed graphs. Directed graphical models (DGM) have a topological ordering such that parent nodes come before children nodes. A graph has Markov property if we assume that a node only depends on its immediate parents. Then, equation (3.2) can be written as:

$$P(X) \propto \prod_{S \subset G} p(x_S | x_{\pi(S)}) \tag{3.3}$$

where $\pi(S)$ is the set of parent nodes of node S. One example of a DGM is the Naive Bayes Classifier in which it is assumed that given the class labels y, the observation

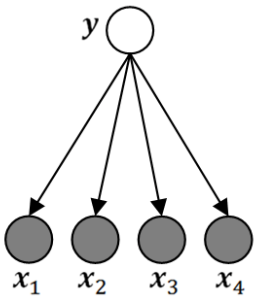


Figure 3.1: A Naive Bayes classifier represented as a DGM. Each feature type has been represented with a node in DGM. Shaded nodes are observed and the unshaded node is hidden (a random variable).

features x are conditionally independent (figure 3.1), so that:

$$p(y, X) = p(y) \prod_{i=1}^{D=4} p(x_i|y) \tag{3.4}$$

where D is the number of observations. Hidden Markov model (HMM) [120] for analyzing sequence data such as speech language is a major application of directed graphical models.

3.1.2 Undirected Graphical Models

Undirected graphical models (UGM) do not require us to specify a direction for edges in the graph. They are more natural for applications such as the spatial lattice of an image in which the intensity values of neighboring pixels are correlated to each other. However, the UGM parameters are less interpretable and less modular than DGM and also their estimation is more computationally expensive [35]. Undirected graphs have no topological ordering and chain rule can not be used to represent the joint probability in UGM. In an undirected model, instead of using conditional probability distributions to associate with each node, potential functions are associated with each maximal clique in the graph. A clique is defined as a set of nodes which are all connected to each other,

and a maximal clique is the clique that cannot be made larger without losing the clique property. A potential function is a non-negative function of its arguments. Therefore, the joint distribution is proportional to a product of potential functions of maximal cliques. The following theorem explains this statement formally.

Theorem 1. ***Hammersley_Clifford theorem****: If $\mathcal{S}$ is the set of maximal cliques of undirected graph G and P is a positive definite distribution which satisfies the conditional independence properties of G, then P can be represented as a product of factors over maximal cliques of G:*

$$P(Y|X,\theta) = \frac{1}{Z(\theta)} \prod_{s \subset \mathcal{S}} \Psi_s(y_s|x_s,\theta_s) \tag{3.5}$$

where θ is the model parameter set; and $Z(\theta)$ is the normalizing partition function:

$$Z(\theta) \triangleq \sum_{X} \prod_{s \subset \mathcal{S}} \Psi_s(y_s|x_s,\theta_s) \tag{3.6}$$

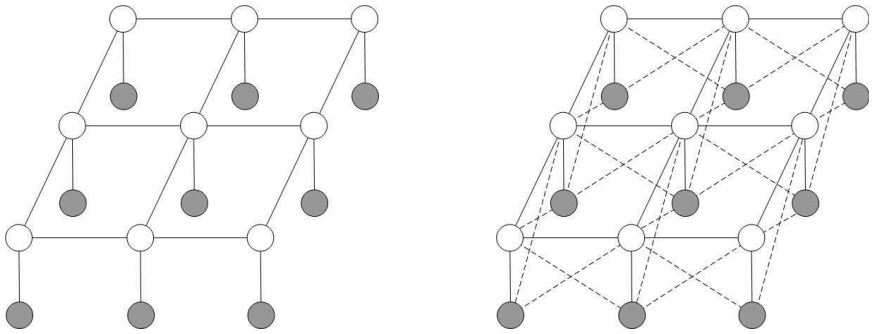
Based on the Gibbs distribution, if strictly positive function $E(.)$ is the energy associated with the variables in clique s, then equation (3.5) can be rewritten as:

$$P(Y|X,\theta) = \frac{1}{Z(\theta)} \exp\Big(- \sum_{s} E(y_s|x_s,\theta_s)\Big) \tag{3.7}$$

that is, $\Psi_s(y_s|x_s,\theta_s) = \exp\big(- E(y_s|x_s,\theta_s)\big)$; which means that low energy configurations of variables correspond to high probability states. The potential or energy functions are not probabilities; they represent the relative compatibility between their arguments. Potential functions may generally be defined as a linear function of the parameters θ_s:

$$\log \Psi_s(y_s) \triangleq \phi_s(y_s,x_s)^T \theta_s \tag{3.8}$$

where $\phi_s(y_s,x_s)$ is the feature vector derived from observation variables.



(a) (b)

Figure 3.2: (a) An example of two grid MRF model, (b) an example of a 2D grid CRF model. In CRF, local labels depend on the local observation as well as neighborhood observation. Shaded nodes are observation nodes and unshaded nodes are random variables.

3.1.3 Markov Random Fields

Markov Random Fields (MRF) [35, 36, 76, 26] are one of the most popular undirected graphical models in physics and machine learning. An example of a 2D grid MRF model is shown in figure 3.2-(a). The local label at each site is dependent on the local observation and 4 adjacent labels in the 4-neighborhood grid. The posterior probability over this graph can be written as:

$$\begin{aligned} P(Y|X) &= \frac{P(X,Y)}{P(X)} \propto P(X|Y)P(Y) \\ &= \prod_{i=1}^{N} p(x_i|y_i).\frac{1}{Z}\prod_{s\subset\mathcal{S}}\Psi_s(y_s) \\ &= \frac{1}{Z}\prod_{i=1}^{N} p(x_i|y_i)\prod_{j=1}^{4} p(y_i,y_j) \end{aligned} \tag{3.9}$$

In equation (3.9), i is the node index. The term $P(x_i|y_i)$ indicates the prior knowledge, and the term $\Psi_s(y_s)$ (or $p(y_i,y_j)$) accounts for local spatial interactions in the graph.

3.1.4 Generative models versus discriminative models

Generative models such as Markov random fields model the joint probability of observations and corresponding labels as $P(X,Y) = P(Y)P(X|Y)$ based on the Bayes rule. Therefore, not only the label patterns but also observation model under each class needs to be encoded. The observation likelihood model is assumed to have a factorized form over all nodes (image sites); that is, MRF assume that given the labels, observation nodes are conditionally independent of each other. However, this independency assumption is too restrictive due to the fact that data of a class at a local site is dependent on its neighbors. For example in image labeling, it is assumed that areas bounded to objects are smooth and structures of objects in terms of shape, texture and configuration follow some underlying organization rules within image pixels rather than being random [29].

The independency assumption simplifies the model but ignores regularities and structures within neighboring observations. Visual features in adjacent image sites are very much correlated such as even backgrounds or lines and curves in images. The assumption that observation features are conditionally independet given the labels, ignores this correlation structure; this is not favorable because the independency assumption allowes the $P(Y|X)$ model to count the same feature again and again (figure 3.3). For example, if we have five copies of the same feature, that is, five very correlated features that effectively measure the same thing, they will be counted five times and make the model too confident because of that one feature type. If we had 100 copies of that feature, it will be counted and relied on 100 times and will push the model towards a very skewed probability distribution that are not good reflections of the true probability because of incorrect independecy assumptions. It is reasonable to add edges to the network to capture feature dependencies and harness the problem of correlated observation features.

Figure 3.3: Visual features in adjacent image sites are very much correlated for example in even backgrounds of images. The assumption that observation features are conditionally independet given the labels (such as in Naive Bayes classifier), ignores correlation among features and counts the same feature again and again.

This makes the problem more complex since it is hard to figure out the correlation structure unless densely connected models are used. A completely different solution lies in the fact that the purpose of labeling is not to predict the distribution of features and their structure; but to use the known features to predict the labels or model the image synthesis [36].

Discriminative models such as CRF model the conditional probability of the labels given the observations, $P(Y|X)$, which is what interests us in image labeling task (instead of modeling the joint distribution $P(X,Y)$). A CRF model looks like a Gibbs distribution in which, the potential functions over cliques get multiplied and an unnormalized measure $\widetilde{P}(X,Y)$ is obtained. To have a conditional distribution $P(Y|X)$, CRF has a normalization constant or partition function that is a function of X: $Z(X) = \sum_{Y} \widetilde{P}(X,Y)$. This normalizing constant resolves the feature correlation problem since for any given x, the partition function has a sum of all the y's that correspond to that x; and the distribution is normalized by it. By conditioning on the observations as opposed to generating them, we can incorporate arbitrary and overlapping (correlated or from neighborhood

area) features, without the need to make strong independence assumptions [103].

Where generative models could become very complex, their corresponding class posterior model might be quite simple. That is, generative approaches apply a lot of resources on modeling the observation space which is not particularly helpful to solve the main labeling problem, $P(Y|X)$. Also, learning the class density model, $P(X|Y)$, becomes difficult when there is not enough training data. Therefore, in comparison with generative models, discriminative models solve the labeling task by using fewer resources and less complex models. Also, these models facilitate training discriminative random fields (DRF) [29] which are a special type of CRF in which potential functions are designed using local discriminative classifiers such as mixture models, neural networks or boosting.

3.1.5 Conditional Random Fields

Conditional random field (CRF) [42, 29] are the most popular form of a Markov random field in which the potentials of cliques are conditioned on input features:

$$P(Y|X,\theta) = \frac{1}{Z(X,\theta)} \prod_{s} \Psi_s(y_s|x,\theta) \tag{3.10}$$

Note that observations x are not restricted to clique s, and it might refer to all or a subset of neighboring cliques. CRF model is an extension of the logistic regression and the potentials are usually in the form of a log-linear function:

$$\Psi_s(y_s|x,\theta) = \exp\left(\theta_s^T \phi(x,y_s)\right) \tag{3.11}$$

where the $\phi(x,y_s)$ is the feature vector derived from input x under label set y_s and θ is the vector of weight coefficients.

Figure 3.2-(b) shows an example of a 2D CRF model. Note that there is interaction between observations at nodes with their neighboring labels (shown with the dashed lines) and despite MRF, the assumption of independency of observations is relaxed. The CRF-based posterior probability of the graph in figure 3.2-(b) is defined as:

$$P(Y|X,\theta) = \frac{1}{Z(X,\theta)} \prod_{i=1}^{N} \Psi_u(y_i|x_i) \prod_{i=1}^{N} \prod_{j \subset \mathcal{N}_i} \Psi_p(y_i, y_j|x_i, x_j) \tag{3.12}$$

where i and j are node indecies; $\mathcal{N}_i$ is the set of neighboring nodes of node i. $\Psi_u(y_i|x_i)$ is called unary or associative potential and $\Psi_p(y_i, y_j|x_i, x_j)$ is called the pairwise or interaction potential assuring label consistency and smoothing; both unary and pairwise potentials can be expressed by (3.11).

To emphasize, in a CRF model, the unary potential at node i could be a function of label y_i and all or a subset of neighboring observations $\{x_1, x_2, ..., x_N\}$; that is, it is possible to incorporate global features obtained from the whole image; whereas in a MRF, the unary potential is a function of y_i and local x_i only. Also, the pairwise potentials are independent of observations and are a function of labels only. However, in CRF model, the pairwise potentials are a function of both label and observations.

Advantages of CRF model over MRF model are analogous to merits of discriminative classifiers over generative models. In CRF, we do not waste resources to analyze observations or seen data. But we focus on modeling what interests us, which is the probability of labels given observation data. Another advantage of CRF in comparison with MRF is that the potential terms are data-dependent. For example, in image labeling application, if there is an intensity discontinuity in between two pixel observations, we may turn off the interaction potential. However, CRFs are slower to train than MRFs [35, 121].

Chapter 4

Generalized Gaussian mixture CRF

In this chapter, new feature functions based on mixtures of generalized Gaussian distribution are proposed to improve accuracy of multi-class image labeling using conditional random fields (CRF). As discussed in previous chapter, CRF modeling has proved to be a successful approach to image labeling task. In this approach, a probabilistic graphical model (PGM) is defined over the image grid. Image units, pixels or patches of regular or irregular shape, constitute the nodes in the graph. In a primary step, color, texture and shape features are extracted from each unit in the image to represent the appearance attributes across the image. CRF models the relationship between image attributes and the class labels via unary potentials. Unary or association potentials represent the log-likelihood of a class label y_i, for i-th image site given the observation x_i of that site:

$$\psi_i(y_i|x) = \log P(y_i|x_i) \tag{4.1}$$

The interaction between two neighboring image sites i and j is modeled with CRF pairwise or interaction potentials which represent the log-likelihood of neighboring labels y_i and y_j given the observation vectors x_i and x_j:

$$\psi_{ij}(y_i, y_j|x) = \log P(y_i, y_j|x_i, x_j) \tag{4.2}$$

Given the feature vectors, both unary and pairwise potentials can be formulated as weighted feature functions. For each site i, if $f(x)$ is a function that maps the observation x on a feature vector such that $f : x \to R^l$, then both potentials can be written as $\sum_k w_k f_k(x)$, where k ranges over arbitrary feature functions and the weights w_k are estimated during CRF training for different feature functions. CRF modeling with potentials of this form requires a large number of features to achieve satisfactory results; particularly, with the growth of number of classes they generate poor results and a large number of features makes their training and inference cumbersome and therefore their application is limited. Potentials of this form are also sensitive to parameter initialization and parameter estimation might get stuck in local minima. Besides, using potentials in the form of weighted features makes the choice of right discriminative features critical.

As discussed in section 1.5.1, potential functions can take forms such as logistic [32], boosting [1, 5, 3], Neural Networks [2, 54], SVM classifiers [27], local support tensor machines [11], label transfer [7], mixture models [55] and combinations of them [49]. Mixture models are one of the high performance yet efficient approaches to image labeling [69, 70]. Mixture models capture within-class variability of objects well (flowers come in different colors) [1, 71, 72]; Augmented by CRF modeling which well discriminate visually similar samples of different classes due to considering contextual information, mixture models have proved to achieve high labeling accuracy [73].

The feature functions might also take the form of a mixture distribution of class labels given the observations [55]. In [1], color-based unary potentials are defined as Gaussian mixtures. In [53] and [73], both unary and pairwise potentials over color and texture features are modeled as mixture of Laplacian and Gaussian distributions, respectively.

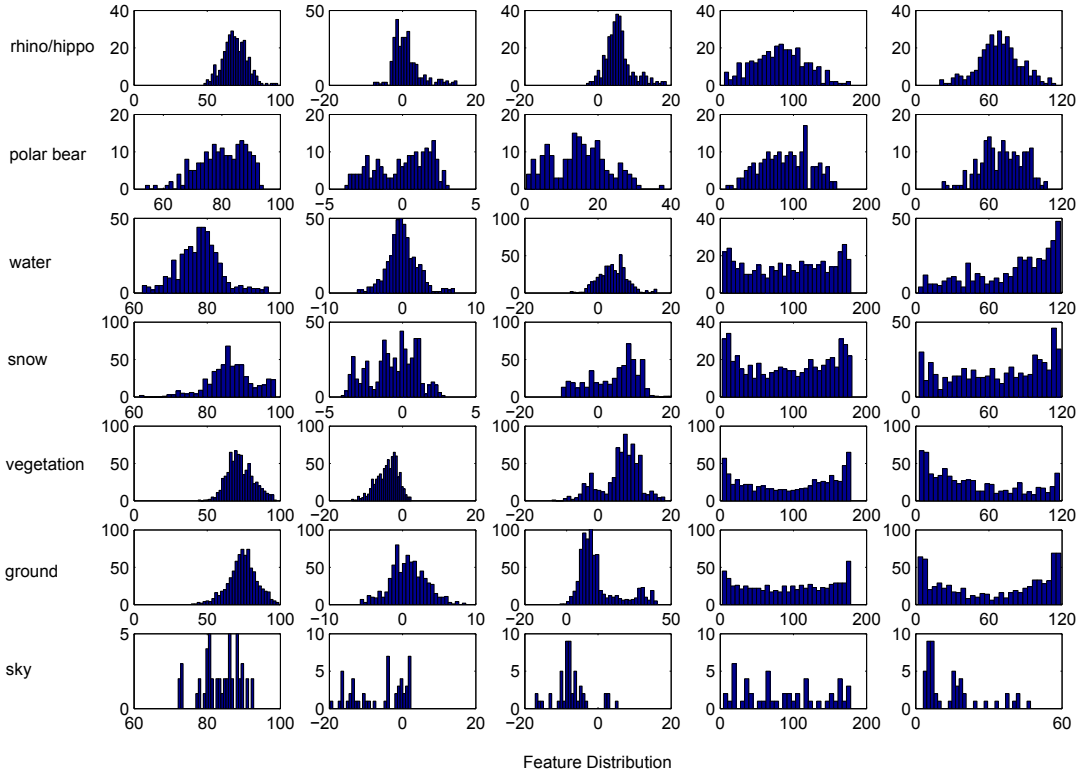


Figure 4.1: Feature distribution of 7 classes of Corel image database. The columns correspond to five different features: Lab colors (L: lightness, A,B: color-opponent dimensions) and positions (horizontal and vertical offset from the image center), from left to right.

Features in nature images follow certain statistical distributions. An example is shown in figure 4.1. The rows are 7 classes in the Corel image database and the columns are five different features: 3 Lab colors and 2 positions (horizontal and vertical offset from the image center). The use of mixture distributions for representing feature distributions in a natural system may also be motivated by the intuitive notion that the individual component densities may model some underlying set of hidden classes (cars come in different shape outlines).

In [53], researchers show that distributions of features in natural scene images are

better approximated by a Laplacian distribution than a Gaussian. They argue that selecting feature functions that better reflect the distribution of the dominant features could reduce the need for more features and increase the convergence speed. Their experiments show that CRF with Laplacian feature functions outperform CRF models with Gaussian feature functions. The advantage of applying mixture distributions as feature function, particularly for labeling nature scene images, is that a lower number of descriptive features will be needed to accomplish successful labeling. However, state of the art literature questions the ability of firmly-shaped distributions such as Gaussian or Laplacian densities to approximate the data precisely [70, 74].

In this chapter, we propose a new feature function based on generalized Gaussian mixture (GGM) modeling of image features. We investigate the effectiveness of the proposed GGMM feature functions to improve the labeling and segmentation accuracy in comparison with their Laplacian counterparts as proposed in [53]. It will be argued that distribution of features can be better approximated with a generalized Gaussian mixture feature function than a Laplacian mixture distribution. Rigidly-shaped Laplacian potentials fail to capture data characteristics where data fluctuations happen very smoothly such as in plain even backgrounds of natural images; so that they even give rise to induction of atypical results due to erroneous modeling of data. Having an additional shape manipulation parameter, generalized Gaussian mixtures can model data characteristics more precisely.

To compare the fitting accuracy of Gaussian, Laplacian and generalized Gaussian mixtures, the goodness-of-fit statistic value χ^2 [74] is used, which is defined as:

$$\chi^2 = \sum_x \frac{(\boldsymbol{H}(x) - \boldsymbol{p}(x))^2}{\boldsymbol{p}(x)} \tag{4.3}$$

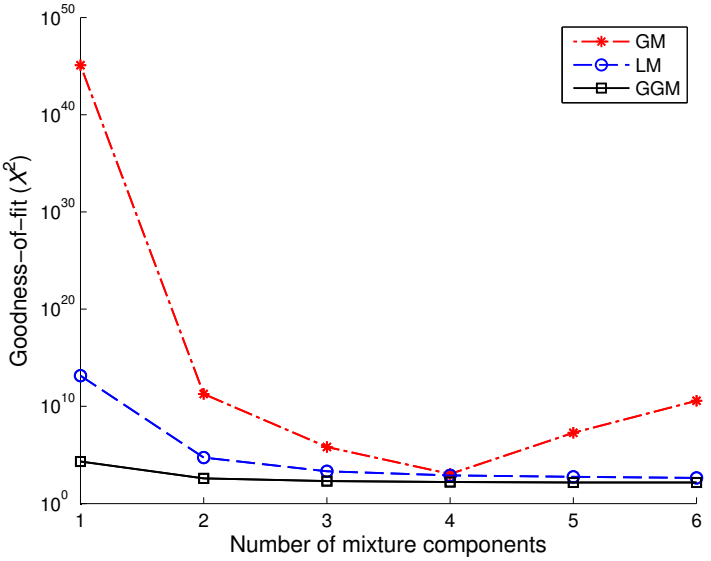


Figure 4.2: Comparison of average χ^2 statistic values versus different number of mixture components for different mixture types GM, LM and GGM.

where $\boldsymbol{H}(x)$ and $\boldsymbol{p}(x)$ are the empirical and expected feature distributions for feature x.

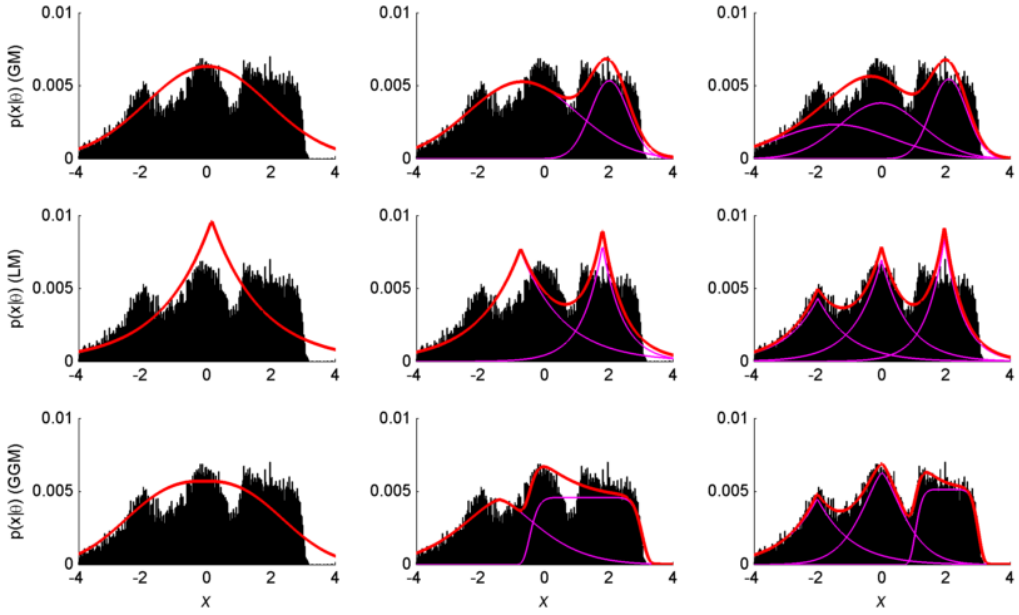


Figure 4.3: An example of a qualitative comparison of different mixture types using different number of mixture components (1, 2 and 3 components from left to right).

Figure 4.2 shows a comparison of χ^2 value versus number of mixture components for different mixture types, Gaussian mixture (GM), Laplacian mixture (LM) and generalized Gaussian mixture (GGM). Figure 4.2 is obtained by averaging χ^2 value over distributions of 53 feature types in all 7 classes in figure 4.1. The GGM model requires lesser number of components to reach a high level of histogram fitting accuracy. That is, GGM method uses less complex models to fit the feature distribution. In figure 4.2, the average χ^2 value for the Gaussian mixture model rises after the 5-th component due to poor convergence. An example of qualitative comparison of different mixture types using different number of mixture components is given in figure 4.3.

In this chapter, the performance of the proposed GGM-based CRF will be evaluated on the commonly used 7 class Corel database. We investigate the performance of the proposed new GGMM feature functions in comparison with their Laplacian and Gaussian counterparts, conventional weighted feature functions, SVM and deep learning methods. In the following, we first bring problem formulation, introduce notation and explain the general mixture CRF framework. In section 4.2, we introduce the new feature functions based on the generalized Gaussian distribution. In section 4.3, we discuss the training and inference for the proposed model. The new image labeling model is applied to 7 class Corel database in section 4.4 and the simulation results are shown. We close this chapter with discussions regarding the proposed approach.

4.1 Problem Formulation

CRF is used to model the conditional distribution of class labels given image observations. Observations could be a set of image measurements such as pixel intensity values, color distributions, texture features, shape features, bag-of-words features, output of multi-

scale band-pass filters or descriptors in the frequency domain, etc. Let $X = \{x_1, ..., x_N\}$ denote the observation data from input image. N is the total number of image sites including pixels or patches of regular or irregular shape. x_i is the feature vector from site i and $Y = \{y_1, ..., y_N\}$ is the set of all image labels. Corresponding to each image site, there is a label $y_i \in \mathcal{L}$ where $\mathcal{L} = \{1, 2, ..., L\}$ is the set of all possible labels. CRF equation for image labeling models the conditional distribution over labels Y given the observations X:

$$P(Y|X) = \frac{1}{Z(X)} \exp\{\sum_{i=1}^{N} \psi_i(y_i|x) + \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \psi_{ij}(y_i, y_j|x)\} \tag{4.4}$$

where ψ_i is the association potential between the observation data x_i and the label y_i of site i. Pairwise potential ψ_{ij} models the interaction between current site i and its neighboring site j given the observed features x_i and x_j. The set $\mathcal{N}_i$ refers to all neighboring sites of site i. Normalizing factor $Z(X)$ ensures that $\sum p(y_i|x_i) = 1$. Potential functions ψ_i and ψ_{ij} are primarily defined as summation of weighted feature functions such that:

$$\begin{aligned} \psi_i(y_i|x_i) &= \sum_{k \in K_u} w_{uk} f_{ik}(y_i|x_i) \\ \psi_{ij}(y_i, y_j|x_i, x_j) &= \sum_{k \in K_p} w_{pk} f_{ijk}(y_i, y_j|x_i, x_j) \end{aligned} \tag{4.5}$$

where k is feature index; f_{ik} indicates the k-th appearance feature at site i; and K_u and K_p are respectively the total number of unary and pairwise features extracted. Parameters w_{uk} and w_{pk} are weights for k-th feature of unary and pairwise feature functions and will be computed during training phase. The task of CRF image labeling is to infer labels Y with the maximum likelihood given data of an input image X and parameters w of the CRF model.

4.2 New GGMM CRF model

We define new feature functions based on generalized Gaussian mixture (GGM) modeling of image features. The new feature functions f_{ik} and f_{ijk} will be formulated by log-likelihood functions, such that:

$$\begin{aligned} f_{ik}(y_i|x_i) &= \sum_{l\in\mathcal{L}} \delta(y_i - l) \log \sum_{m\in M} \pi_{y_i m} P_i(x_{ik}|y_i, m) \\ f_{ijk}(y_i, y_j|x_i, x_j) &= \sum_{l\in\mathcal{L}} \sum_{l'\in\mathcal{L}} \delta(y_i - l)\delta(y_j - l') \log \sum_{m\in M} \pi_{y_i y_j m} P_{ij}(x_{ik}, x_{jk}|y_i, y_j, m) \end{aligned} \tag{4.6}$$

where m is the index of the mixture component and M is the total number of components per mixture. Here $\pi_{y_i m}$ and $\pi_{y_i y_j m}$ are mixture coefficients. Note that $\sum_{m\in M} \pi_{m y_i} = 1$ and $\sum_{m\in M} \pi_{m y_i y_j} = 1$. The function $\delta(y - l) = [y = l]$, where $[.]$ is the indicator function and $l \in \mathcal{L}$ is the index of image classes. The conditional probabilities P_i and P_{ij} will be defined as generalized Gaussian distributions defined as:

$$p(x|\mu, \sigma, \beta) = \frac{\beta \sqrt{\frac{\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})}}}{2\sigma\Gamma(\frac{1}{\beta})} \exp\left(- \left(\frac{\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})} \right)^{(\frac{\beta}{2})} \left| \frac{x - \mu}{\sigma} \right|^{\beta} \right) \tag{4.7}$$

where $\Gamma(.)$ denotes the gamma function, and μ and σ are mean and standard deviation of the distribution. Shape parameter $\beta > 0$ determines the peakedness of the distribution. The smaller the value of β, the more peaked the distribution is around its mean and as β grows larger, the distribution becomes flatter. The shape parameter β makes the distribution flexible to fit the data properly [74, 122]. Since the GGM formulation for each feature will be the same, we drop the image site index i and feature index k for simplicity. However, it should be noted that above equation has to be considered to represent the distribution of each k-th feature in the data collection.

To compute the feature functions in equation (4.6), we need to estimate parameters π, μ, σ and β for each component of the mixture. $\theta = \{\pi_m, \mu_m, \sigma_m, \beta_m, m = 1, ..., M\}$ is the set of model parameters to be estimated for each unary and pairwise feature in each class. Parameters θ are tied across image sites i and j. With known class labels for each site of the training images, one can group the features of the same label and use the EM algorithm to calculate parameters of the model. When class labels are known for one site and its neighboring site, the parameters can be learned for their label interaction. Knowing these parameters, feature function of the model can be calculated using (4.6).

Given a data sample $x = \{x_1, x_2, ..., x_n\}$, which in our case represents a specific feature, e.g. red color, from all available samples of a particular class, we intend to estimate the generalized Gaussian mixture distribution parameters so that it fits the distribution of the feature in a class. We use the maximum likelihood ML method such that:

$$\theta = \arg\max_{\theta}\{p(x|\theta)\} \tag{4.8}$$

Given a predefined number of mixture components, M, we use EM algorithm [77, 123] to solve this optimization problem. EM algorithm is an iterative process with two steps in each iteration: expectation step (E-step) and maximization step (M-step). We formulate EM algorithm to estimate GGM distribution parameters.

E-step: Find the expected value of the complete data log-likelihood $\log(p(x; z))$ with respect to unobserved data z given the observed data x:

$$\begin{aligned} Q(\theta, \theta^{t-1}) &= \mathbb{E}_z[\log(p(x, z|\theta))|x, \theta^{t-1}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{M} \mathbb{E}_z[z_{ij}|x_i, \theta^{t-1}].\log(\pi_j^{t-1} p(x_i|j, \theta^{t-1})) \end{aligned} \tag{4.9}$$

where t denotes iteration number, $p(x_i|j,\theta^{t-1})$ is calculated from the generalized Gaussian distribution formulation in (4.7); and:

$$\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}] = \frac{\pi_j^{t-1}p(x_i|j,\theta^{t-1})}{\sum_{k=1}^{M}\pi_k^{t-1}p(x_i|k,\theta^{t-1})} \tag{4.10}$$

M-step: Obtain the following updating equations at each iteration t:

$$\begin{aligned}
&\frac{\partial Q(\theta,\theta^{t-1})}{\partial \pi_j} = 0 \Rightarrow \pi_j^t = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}] = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_j^{t-1}p(x_i|j,\theta^{t-1})}{\sum_{k=1}^{M}\pi_k^{t-1}p(x_i|k,\theta^{t-1})}\\
&\frac{\partial Q(\theta,\theta^{t-1})}{\partial \mu_j} = 0 \Rightarrow \mu_j^t = \frac{\sum_{i=1}^{n}x_i\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}]\left|x_i-\mu_j^{t-1}\right|^{\beta_j-2}}{\sum_{i=1}^{n}\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}]\left|x_i-\mu_j^{t-1}\right|^{\beta_j-2}}\\
&\frac{\partial Q(\theta,\theta^{t-1})}{\partial \sigma_j} = 0 \Rightarrow \sigma_j^t = \left(\frac{\Gamma(\frac{3}{\beta_j^{t-1}})}{\Gamma(\frac{1}{\beta_j^{t-1}})}\right)^{\frac{\beta_j^{t-1}}{2}}\frac{\sum_{i=1}^{n}\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}]\left|x_i-\mu_j^{t-1}\right|^{\beta_j^{t-1}}}{\frac{1}{\beta_j^{t-1}}\sum_{i=1}^{n}\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}]}
\end{aligned} \tag{4.11}$$

Finally, Newton-Raphson method is applied to estimate the value for parameter β in each iteration:

$$\beta_j^t = \beta_j^{t-1} - \frac{\frac{\partial Q(\theta,\theta^{t-1})}{\partial \beta_j}}{\frac{\partial^2 Q(\theta,\theta^{t-1})}{\partial \beta_j^2}} \tag{4.12}$$

Formulation of numerator and denominator terms are:

$$\begin{aligned}
\frac{\partial Q(\theta,\theta^{t-1})}{\partial \beta_j} = &\sum_{i=1}^{N}\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}]\Bigg(\frac{\Gamma(\frac{3}{\beta_j})\left(2\beta_j + 3\psi^{(0)}(\frac{1}{\beta_j}) - 3\psi^{(0)}(\frac{3}{\beta_j})\right)}{8\sigma^2\Gamma(\frac{1}{\beta_j})^3}\\
&- \frac{1}{2\beta_j}\left(\frac{\Gamma(\frac{3}{\beta_j})}{\Gamma(\frac{1}{\beta_j})}\right)^{\frac{\beta_j}{2}}\left|\frac{x-\mu}{\sigma}\right|^{\beta_j}\left(2b\log\left(\left|\frac{x-\mu}{\sigma}\right|\right) + \psi^{(0)}(\frac{1}{\beta_j})\right.\\
&\left.- 3\psi^{(0)}(\frac{3}{\beta_j}) + b\log\left(\frac{\Gamma(\frac{3}{\beta_j})}{\Gamma(\frac{1}{\beta_j})}\right)\right)\Bigg)
\end{aligned} \tag{4.13}$$

$$
\begin{aligned}
\frac{\partial^2 Q(\theta,\theta^{t-1})}{\partial \beta_j^2} =& \sum_{i=1}^{N} \frac{\mathbb{E}_z[z_{ij}|x_i,\theta^{t-1}]\Gamma(\frac{3}{\beta_j})}{8\beta_j^2\sigma^2\Gamma(\frac{1}{\beta_j})^3}\Bigg(2\beta_j^2 - 6\beta_j\psi^{(0)}(\frac{3}{\beta_j}) + 9\psi^{(0)}(\frac{1}{\beta_j})^2 \\
&+ 9\psi^{(0)}(\frac{3}{\beta_j})^2 + 6\psi^{(0)}(\frac{1}{\beta_j})\Bigg(b - 3\psi^{(0)}(\frac{3}{\beta_j})\Bigg) - 3\psi^{(1)}(\frac{1}{\beta_j}) + 9\psi^{(1)}(\frac{3}{\beta_j})\Bigg) \\
&- \Bigg(\frac{\left|\frac{x-\mu}{\sigma}\right|^b \Big(\frac{\Gamma(\frac{3}{\beta_j})}{\Gamma(\frac{1}{\beta_j})}\Big)^{\frac{\beta_j}{2}}}{4\beta_j^2}\Bigg)\Bigg(4\beta_j^2\log^2\Big(\Big|\frac{x-\mu}{\sigma}\Big|\Big) \\
&+ 4\beta_j^2\log\Big(\Big|\frac{x-\mu}{\sigma}\Big|\Big)\log\Big(\frac{\Gamma(\frac{3}{\beta_j})}{\Gamma(\frac{1}{\beta_j})}\Big) + 8\beta_j^2\log\Big(\Big|\frac{x-\mu}{\sigma}\Big|\Big) \\
&+ \Bigg(2\beta_j\psi^{(0)}(\frac{1}{\beta_j}) - 6\beta_j\psi^{(0)}(\frac{3}{\beta_j})\Bigg)\Bigg(\frac{2}{\beta_j} + 2\log\Big(\Big|\frac{x-\mu}{\sigma}\Big|\Big) + \log\Big(\frac{\Gamma(\frac{3}{\beta_j})}{\Gamma(\frac{1}{\beta_j})}\Big)\Bigg) \\
&+ \beta_j^2\log^2\Big(\frac{\Gamma(\frac{3}{\beta_j})}{\Gamma(\frac{1}{\beta_j})}\Big) + \psi^{(0)}(\frac{1}{\beta_j})^2 + 9\psi^{(0)}(\frac{3}{\beta_j})^2 \\
&- 6\psi^{(0)}(\frac{1}{\beta_j})\psi^{(0)}(\frac{3}{\beta_j}) - \frac{2}{\beta_j}\psi^{(1)}(\frac{1}{\beta_j}) + \frac{18}{\beta_j}\psi^{(1)}(\frac{3}{\beta_j})\Bigg)
\end{aligned}
\tag{4.14}
$$

where $\psi^{(0)}$ is the first derivative of Gamma function and $\psi^{(1)}$ is the second derivative of Gamma function. Careful initialization of the parameters is very critical to the success of the EM algorithm. In our experiments, we employed the overall mean (μ_{pop}) and variance (σ^2_{pop}) of population x to initialize mean and variance parameters for each of the mixture components. For example, for $M = 2$, we set: $\mu_1^0 = \mu_{pop} - \sigma_{pop}$, $\mu_2^0 = \mu_{pop} - \sigma_{pop}$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2_{pop}$. In case $M = 3$, then a good initialization for parameter μ would be: $\mu_1^0 = \mu_{pop} - \sigma_{pop}, \mu_2 = \mu_{pop}$ and $\mu_1^0 = \mu_{pop} + \sigma_{pop}$. For parameter π, we set $\pi_j^0 = \frac{1}{M}$ for all components. Regarding parameter β, we set the initial value as $\beta_j^0 = \frac{m_1}{\sqrt{m_2}}$ for all components; where $m_1 = \frac{1}{n}\sum_{i=1}^{n}|x_i|$ is first statistical moment of the absolute values and m_2 is the second statistical moment.

Once the functions in equation (4.6) are known, stochastic gradient descent (SGD) and belief propagation (BP) [124] algorithms can be applied to learn the weight parameters

w using training samples and to infer the test sample labels, respectively.

4.3 Training and Inference

Stochastic gradient descent (SGD) is used for training of the new CRF model. The mixture parameters are known using EM algorithm before SGD training. Weight parameters are tied across image sites i and j. Given a set of training examples, the goal is to choose parameter values w that maximize the conditional probability of the training examples. In other words, the objective function for training is the conditional log-likelihood of the set of training examples:

$$\frac{\partial}{\partial w_k} \log \prod_{n=1}^{N} P(y^{(n)}|x^{(n)}) \tag{4.15}$$

The parameters are updated based on a batch of training examples each time. In our experiment, the number of training images in a batch is set to be 3. There is one weight for each mixture feature function in the new CRF model. The partial derivative of the conditional log-likelihood $\log P(y^{(n)}|x^{(n)}; w)$ with respect to the weight w_k (that could be w_{ik} or w_{ijk}) is calculated as follows [125]:

$$\begin{aligned}
&\frac{\partial}{\partial w_k} \log P(y^{(n)}|x^{(n)}; w) \\
&= f_k(x^{(n)}, y^{(n)}) - \frac{\partial}{\partial w_k} \log Z(x^{(n)}, w) \\
&= f_k(x^{(n)}, y^{(n)}) \\
&\quad - \tfrac{1}{Z(x^{(n)}, w)} \textstyle\sum_{y^{(n)\prime}} \frac{\partial}{\partial w_{k'}} \exp \sum_{k'} w_{k'} f_{k'}(x^{(n)}, y^{(n)\prime}) \\
&= f_k(x^{(n)}, y^{(n)}) \\
&\quad - \textstyle\sum_{y^{(n)\prime}} f_k(x^{(n)}, y^{(n)\prime}) \frac{\exp \sum_{k'} w_{k'} f_{k'}(x^{(n)}, y^{(n)\prime})}{\sum_{y^{(n)\prime\prime}} \exp \sum_{k''} w_{k''} f_{k''}(x^{(n)}, y^{(n)\prime\prime})} \\
&= f_k(x^{(n)}, y^{(n)}) - \langle f_k(x^{(n)}, y^{(n)\prime}) \rangle_{P(y^{(n)\prime}|x^{(n)}; w)}.
\end{aligned} \tag{4.16}$$

Here n is the current training example and both $y^{(n)'}$ and $y^{(n)''}$ represents the possible labels. The $f_k(\cdot)$ ($f_{ik}(\cdot)$ or $f_{ijk}(\cdot)$) are the feature functions in the equation (4.6) and $P(y^{(n)'}|x^{(n)}; w)$ is the conditional probability of label $y^{(n)'}$ given the weights w and features $x^{(n)}$. According to this partial derivative, the weights w_k are updated iteratively as:

$$w_k^t = w_k^{t-1} - \eta(f_k(x^{(n)}, y^{(n)}) - \langle f_k(x^{(n)}, y^{(n)})\rangle), \tag{4.17}$$

where t is the iteration number and η is the learning rate. Therefore, the weight change is proportional to the value of the feature function for the known label $y^{(n)}$ minus the expected value of the feature function for all possible labels $y^{(n)'}$.

Once the weights $w = \{w_{ik}, w_{ijk}\}$ are calculated during training, the beliefs for labels of each image site could be inferred by using the belief propagation (BP) algorithm.

4.4 Experimentation

We apply the CRF model with the new generalized Gaussian based feature function to the image labeling task. The task of image labeling is to find an appropriate content label for every image pixel. Since the CRF is a computationally costly model, we apply the new model in a superpixel level than pixel level. This approach is feasibly reasonable since most likely a pixel belongs to the same object category as the neighboring pixel. Therefore, images are first oversegmented to superpixels.

Superpixels are small homogenous regions composed of a group of adjacent similar pixels and they are the result of oversegmentation of the image. With a large number of small regions, the potential error induced by such an oversegmentation at object boundaries is relatively small. Superpixels constitute the nodes in the CRF graph. Unary

potential will be defined as the label of a superpixel given the features extracted from that superpixel. Also, interaction potentials in the CRF modeling will be defined on adjacent superpixels. Using superpixels instead of pixels, the number of nodes in the CRF graphical model used is greatly reduced and so the computational cost of the CRF training and inference.

The basic steps of image labeling using a mixture model are listed as follows:

Step 1 - Feature extraction of training images: Training images are oversegmented to superpixels and features of each superpixel are generated.

Step 2 - Learning the mixture parameters: Superpixel features for each class are grouped. EM algorithm is used to compute the parameters of the mixture distribution of features in each class.

Step 3 - CRF training: Using the parameters calculated in step two, unary and pairwise feature functions are computed. Then, stochastic gradient decent training is performed iteratively to estimate the weight parameters for potential feature functions.

Step 4 - Feature extraction of test images: Test images are oversegmented to superpixels and features of each superpixel are generated.

Step 5 - Feature function computation: The potentials are computed using mixture parameters learned from Step 2 to perform inference over the test image.

Step 6 - CRF Inference: Inference predicts the testing image superpixel labels using BP algorithm.

4.4.1 Image database

To evaluate the performance of this new CRF model with generalized Gaussian mixtures, image labeling experiments were conducted on the commonly-used Corel image database [2]. There are seven classes in this dataset, rhino/hippo, polar bear, water, snow, vegetation, ground, and sky. The task is to recognize and segment these 7 classes. The database has 100 images and the size of images is 180x120 pixels. In the experiment, the database is divided randomly to 50 training images and 50 test images.

4.4.2 Superpixels

Due to the fact that the pixel-based CRF is computationally intensive, the new mixture CRF is built on superpixels, similar to [2, 53, 73]. Each image is oversegmented using mean shift segmentation algorithm [126]. Mean shift segmentation method belongs to the class of unsupervised segmentation algorithms which work by clustering pixels on the basis of low level image features. Mean shift algorithm is defined as the product of spatial and range kernels. The spatial domain contains the (i, j) coordinates, while the range domain contains pixel colour information in LUV color space. The number of superpixels per image is roughly 60 superpixels. This number might affect the accuracy of image labeling at image boundaries [38] but it does not affect the comparative results of our experiments.

4.4.3 Feature extraction

We use the set of low level features including superpixel location, mean Lab color values and texture features returned by Leung-Malik (LM) filter bank [127] of size 49x49x48 over the image. The LM set is a multi-scale, multi orientation filter bank with 48 filters.

It consists of first and second derivatives of Gaussians at 6 orientations and 3 scales, 8 Laplacian of Gaussian (LOG) filters, and 4 Gaussians. The exact feature value of a superpixel is the mean over all pixel feature values of this superpixel. Therefore, including color and location features each superpixel is represented by a 53 dimensional feature vector. A bias term 1 is always added to the feature vector. The feature vector for pairwise potential is calculated as the absolute difference of the features of two neighboring superpixels.

4.4.4 Performance analysis

We investigate the effectiveness of the proposed GGMM feature functions to improve CRF-based labeling accuracy in comparison with their Laplacian and Gaussian counterparts, conventional weighted feature functions, SVM and deep learning methods. Particularly in comparison with Laplacian and Gaussian mixture-based CRF modeling, we use different number of features and different number of components to show that GGM modeling outperforms Laplacian and Gaussian CRF modeling. We show that generalized Gaussian mixture feature functions outperform other methods in terms of recall and precision defined as:

$$\begin{aligned} \text{recall} &= \frac{\sum\limits_{\text{all classes}} \text{true positive}}{\sum\limits_{\text{all classes}} \text{condition positive}} = \frac{TP}{TP+FN} \\ \text{precision} &= \frac{\sum\limits_{\text{all classes}} \text{true positive}}{\sum\limits_{\text{all classes}} \text{test outcome positive}} = \frac{TP}{TP+FP} \end{aligned} \tag{4.18}$$

where true possitives (TP) are examples correctly labeled as positives. False positives (FP) refer to negative examples incorrectly labeled as positive. True negatives (TN)

correspond to negatives correctly labeled as negative. False negatives (FN) refer to positive examples incorrectly labeled as negative. Based on these definitions, true positive rate (TPR) and false positive rate (FPR) are also defined as:

$$\begin{aligned} \text{True positive rate} &= \frac{TP}{TP+FN} \\ \text{False positive rate} &= \frac{FP}{FP+TN} \end{aligned} \tag{4.19}$$

The FPR measures the fraction of negative examples that are incorrectly labeled as positive; and TPR is the same as recall and measures the fraction of positive examples that are labeled correctly. Precision measures the fraction of examples classified as positive that are truely positive.

We use the 50 training images $D = \{(x^{(n)}, y^{(n)}), n = 1, ..., 50\}$ to estimate the parameters of GGMM, LM and GM-based CRF models using stochastic gradient descent algorithm. The learning rate η is fixed to be 0.0001. Starting with random weights, the stochastic gradient descent algorithm converges after about 10 iterations for all three models. The same training images are used to train the SVM and DNN models and the conventional CRF with weighted feature functions.

Figure 4.4 compares results of GGMM, LM and GM-based CRF labeling using different number of mixture components versus different number of features. Considering different number of mixture components from 1 to 5, the proposed GGMM-based CRF labeling generates the best performance in terms of recall criteria. Maximum recall performance of 74.56 is obtained via GGMM-based CRF labeling using 2 number of mixture components and 42 features; the next best performance is also obtained via GGMM-based CRF labeling using 1 mixture component and 30 features. Also, figure 4.5 elaborates the comparison of recall performance of the three mixture types for differ-

ent number of features using different number of components. As illustrated, for most feature combinations the proposed GGMM-based CRF modeling obtains the maximum recall performance over different number of components.

Figure 4.6-(a) and figure 4.6-(b) illustrate the average recall versus different number of features and different number of components, respectively. Graph in figure 4.6-(a) is obtained by averaging all 5 graphs in figure 4.4 for different number of components; and figure 4.6-(b) is obtained by averaging all 9 graphs in figure 4.5 for different number of features. Figure 4.6 shows that best recall performances are resulted from proposed GGMM-based model over different number of features and different number of components. Note that higher recall means higher true positive rate. Figure 4.7 shows the corresponding average precision graphs obtained over different number of features and different number of components. As illustrated, best precision performances are resulted from proposed GGMM-based model. Higher precision means lower false positive rate.

Figure 4.8 shows a comparison of average receiver operator characteristic (ROC) and Precision-Recall (PR) curves of CRF labeling using three mixture types. In the ROC space, a curve closer to upper-left-corner indicates better performance and in the PR space a curve close to the upper-right-corner is an indication of better performance. PR is an alternative to ROC performance measure. The performances of the three algorithms appear to be comparable in ROC space; however, the new GGMM-based CRF modeling is outperforming the two other mixture types in PR space. Since our labeling problem is not a binary classification problem and we are categorizing 7 different classes, the number of negative examples in each class greatly exceeds the number of positive examples. That is, a large number of false positives make a small change in the false positive rate in ROC space. However, precision measure compares false positives to true positives rather than true negatives and therefore, captures the effects of large number of negative examples

on performance of the algorithms [128].

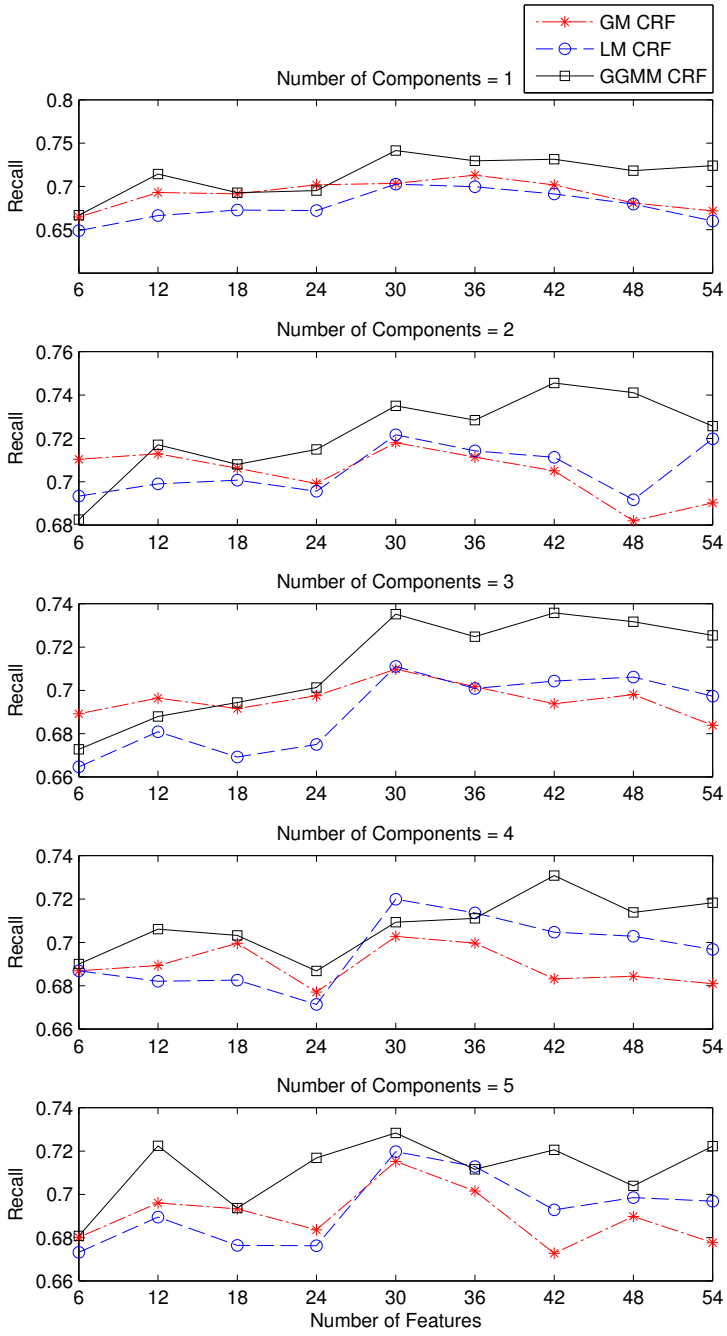


Figure 4.4: Using different number of components, best performances are resulted from proposed GGMM-based model.

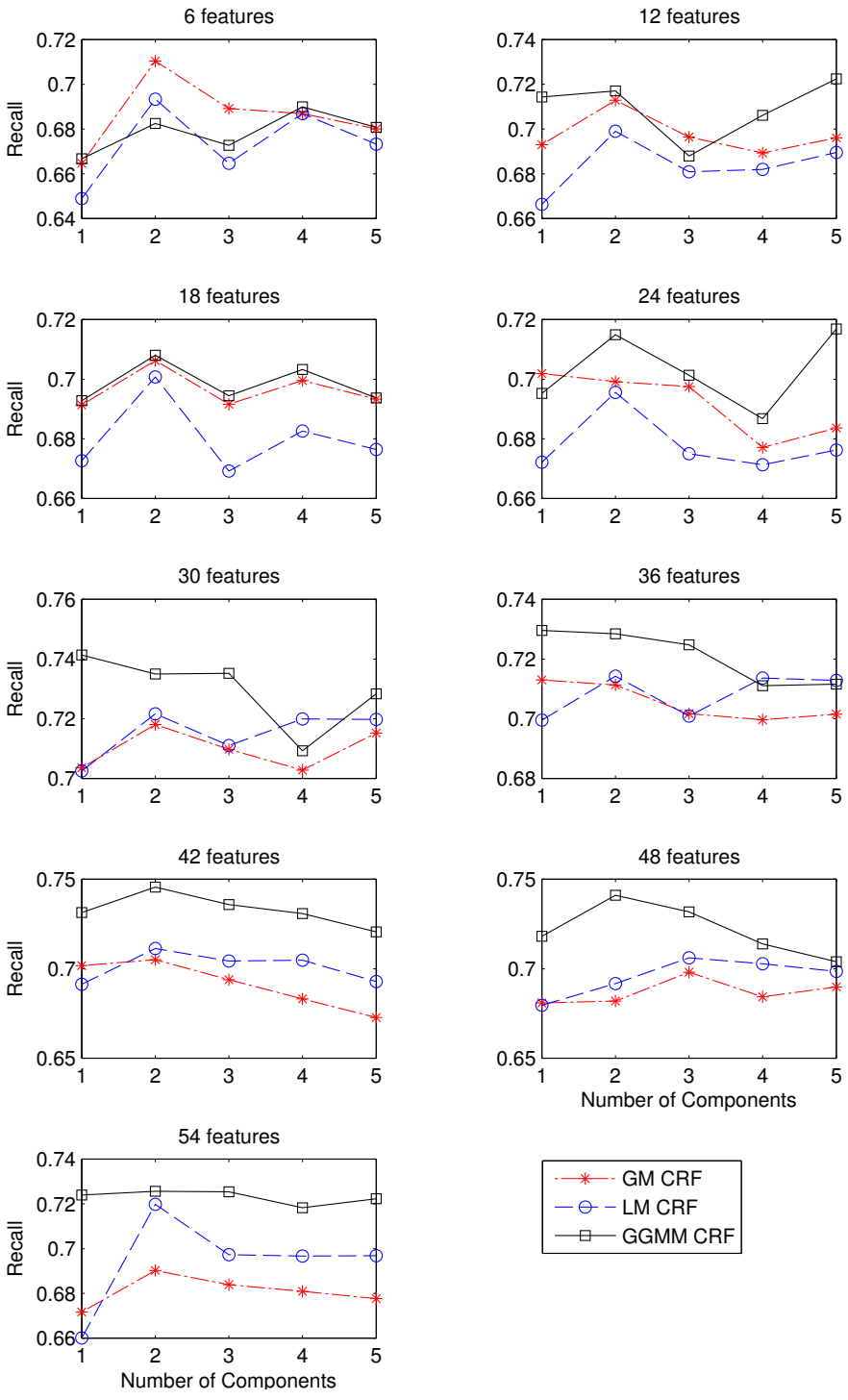
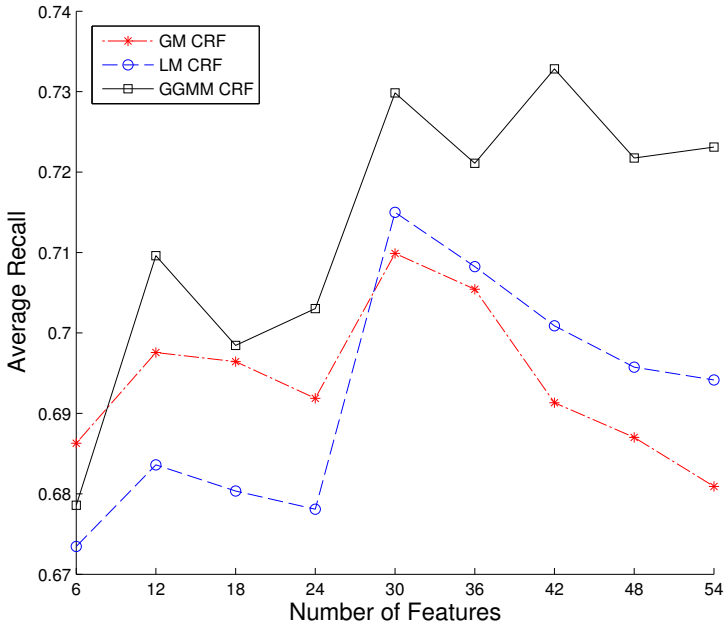
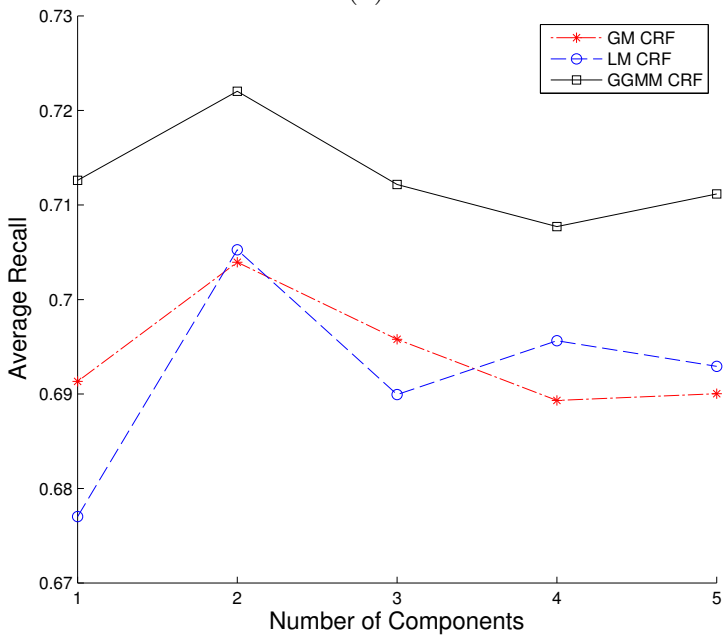


Figure 4.5: For most feature combinations the proposed GGMM-based CRF modeling obtains the maximum recall performance over different number of components.

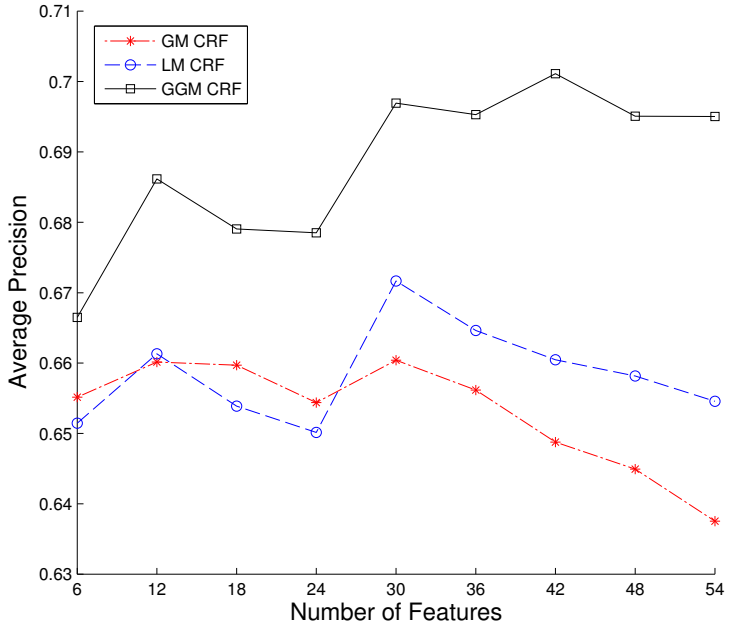


(a)

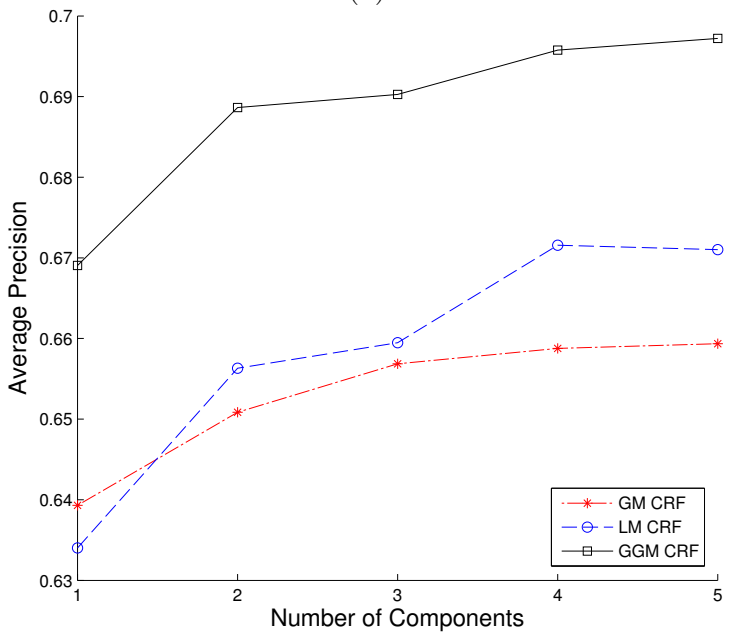


(b)

Figure 4.6: Average recall versus (a) different number of features and (b) different number of components.

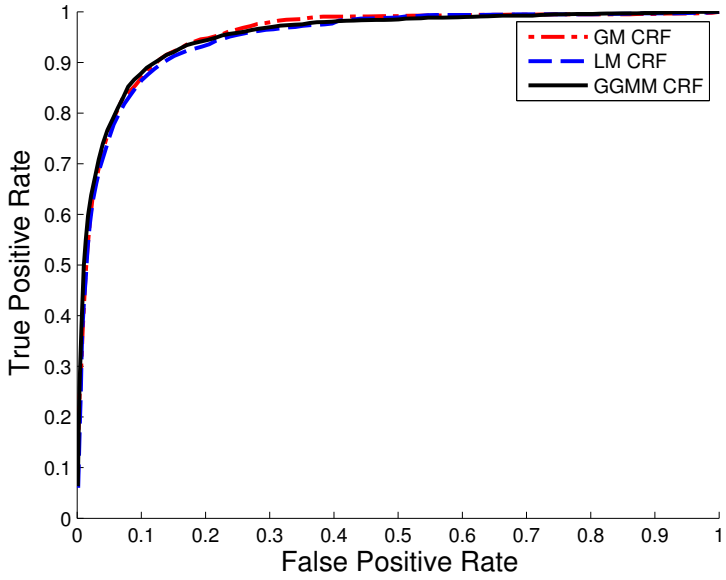


(a)

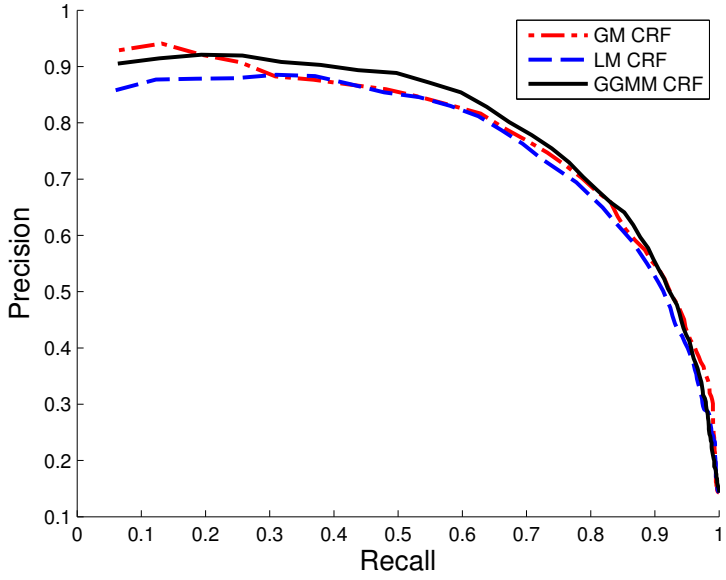


(b)

Figure 4.7: Average precision versus (a) different number of features and (b) different number of components.



(a)



(b)

Figure 4.8: (a) Comparison of ROC curves of CRF labeling using three mixture types, (b) Comparison of precision vs recall curves of CRF labeling using three mixture types

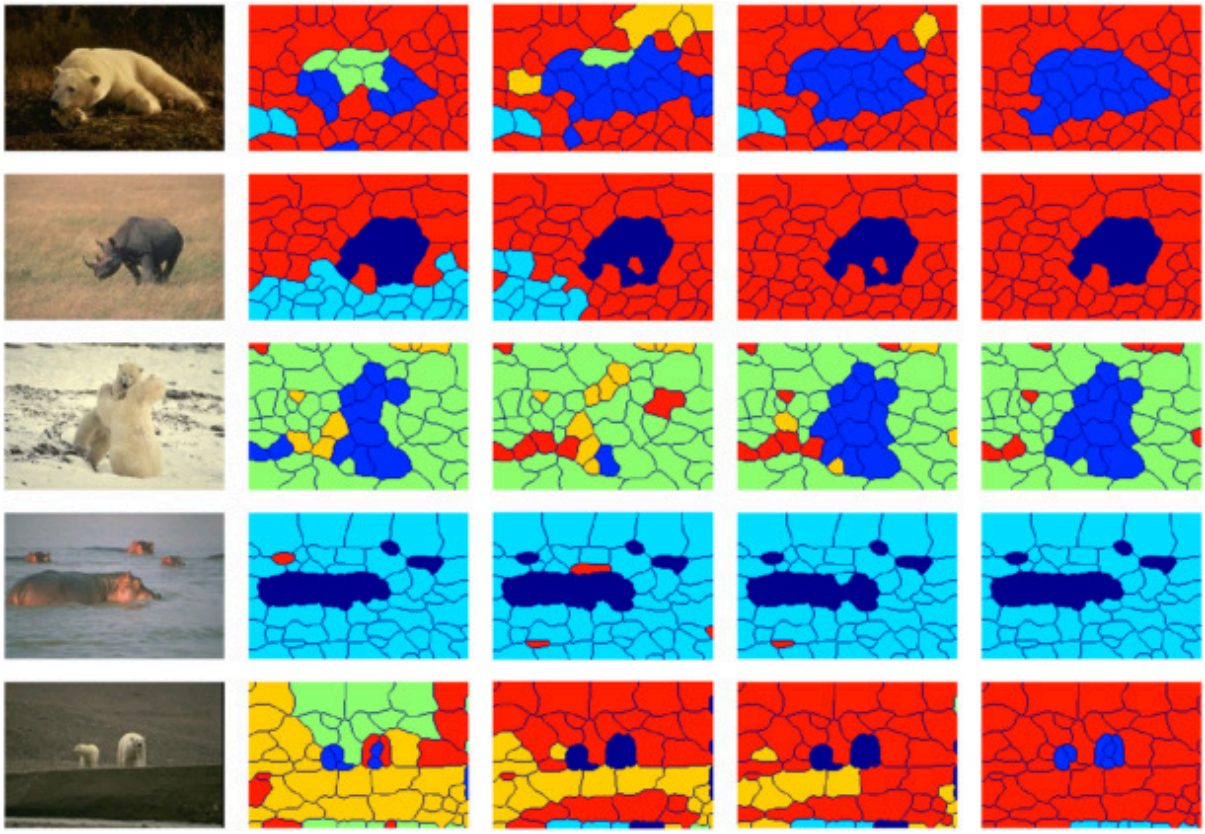


Figure 4.9: Examples of labeling images from Corel dataset using GM, LM and GGM CRF modeling.

Some examples of mixture-based CRF labeling are illustrated in figure 4.9. The proposed GGMM-based CRF produces less erronous and more consistent labeling particularly in even areas of the image background.

For comparison of the proposed mixture-based CRF model with SVM, deep neural networks (DNN) and baseline CRF using logistic potential feature functions, we fix the number of mixture components and number of features to values for which all mixture types render the best performance; that is, $M = 2$ and 30 features.

We trained a DNN with two hidden layers with different number of input features and

Table 4.1: Recall performance of the GGMM-based CRF model comparing to other methods.

	Recall Performance
GGMM CRF	**73.5**
LM CRF	72.1
GM CRF	71.8
CRF	64.57
DNN+MRF	65.32
DNN+CRF	72.32
SVM+MRF	64.08
SVM+CRF	69.48

different hidden layer sizes. The best performance for DNN was obtained by utilizing all 54 features; the size of the first and second hidden layers were set to 35 and 25, respectively. The size of the input layer was set to the number of features and the size of the output layer was set to the number of classes. Furthermore, we trained an SVM classifier with a radial basis function for 54 features [129]. The SVM parameters were selected using cross-validation.

For fairness of comparisons, we enforce contextual constraints on labeling results of DNN and SVM classifiers using MRF modeling and CRF modeling on the probability images obtained by these classifiers. The probability images were used as unary potentials of the random fields model. For CRF model, we used the dense random field in [3] with color variance $\sigma_r = \sigma_g = \sigma_b = 7$, location variance of $\sigma_x = \sigma_y = 3$ pixels, bilateral filter of width 40 pixels, and weight coefficients $w = 1$ for both location and bilateral kernels. The MRF model was defined with energy function $J(Y) = \sum_{i=1}^{N} E_{data}(y_i, x_i) + \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} E_{smooth}(y_i, y_j)$; where $E_{data}(y_i, x_i)$ is the sigmoid function of class likelihood scores from DNN and SVM classifiers. The smoothing term $E_{smooth}(y_i, y_j)$ was defined based on probabilities of label cooccurrence as:

$$E_{smooth}(y_i, y_j) = -\log[(p(y_i|y_j) + p(y_j|y_i))/2] \times \delta[y_i \neq y_j]$$

where $p(c|c')$ is the conditional probability of one superpixel having label c given that its neighbor has label c', estimated by counts from the training set. The term $\delta[y_i \neq y_j]$ ensures that this energy term is semi-metric as required by the graph cut algorithm applied for MRF inference [46].

Table 4.4.4 shows the recall performance of the GGMM-based CRF model comparing to the baseline CRF, LM- and GM-based CRF and SVM and DNN with CRF smoothing. Note that in the baseline CRF the quadratic expansion of the features is used. Performance of the proposed model is higher than other methods.

4.5 Discussion

A new image labeling model based on mixture CRFs is introduced in this chapter. We apply generalized Gaussian mixture distributions as CRF unary and pairwise potentials. That is, the proposed model takes advantage of both the unstructured generalized Gaussian feature distribution and structural discriminative CRF model. The combination of the two provides a successful system for nature image labeling. The training of the new CRF is performed by stochastic gradient descent. To infer the most probable labels, Belief propagation inference is used. Performance analysis was carried out on Corel image dataset with 7 different categories of natural environment including animals and vegetation. The results showed the prominence of proposed GGMM-based CRF over baseline CRF, LM-based CRF, GM-based CRF, SVM and DNN methods. Rigidly-shaped potentials such as Gaussian and Laplacian fail to capture data characteristics where data fluctuations happen very smoothly such as in plain even backgrounds of natural images; so that they give rise to induction of atypical results due to erroneous modeling of data

and therefore, reduce labeling accuracy. The shape parameter of generalized Gaussian mixture makes the distribution flexible to fit the data properly and therefore leads to more accurate labeling. The new mixture CRF model is a general framework with the advantage of high classification rate and low computational training, which can be applied to other applications related to multimedia content analysis. Future works include improving the performance by incorporating more relevant features, testing the method for other more complex databases.

Chapter 5

Context-based dense CRF

Apart from the design of CRF potential functions, another important aspect of CRF modeling is how to perform contextual reasoning. As elaborated in section 2, the range of connectivity of nodes in the graph determines the extent to which contextual information such as class co-occurrences can be exploited. High-order connectivity among image regions boost labeling performance. However, accuracy of these methods is depending on the accuracy of the image segmentation algorithm applied to generate the regions.

Pixel-wise CRF labeling can produce accurate label assignments around complex object boundaries. However, its computational cost hindered its application until recently Krähenbühl and Koltun [3] proposed an efficient inference algorithm for pixel-wise dense conditional random fields (DCRF) model which connects each pixel to every other pixel in the image. They applied mean field approximation to write inference equations in a factorized form so that they were able to use high-dimensional filtering to approximate and speed computations. However, although fully-connected DCRF generates precise object boundaries at the pixel level (figure 1.2), it is prone to over-smoothing small objects from thing classes (foreground objects) in the large pool of pixels from background

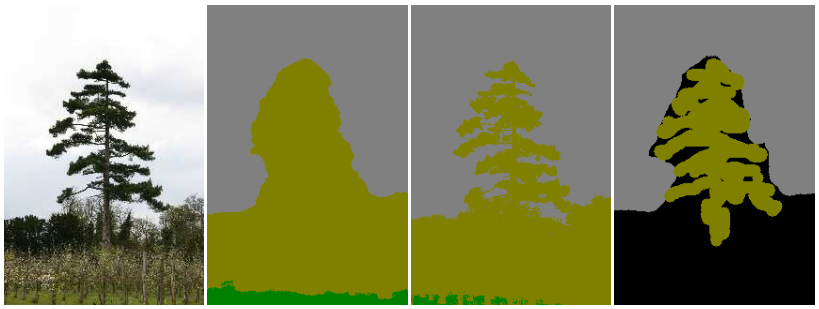


Figure 5.1: DCRF generates precise object boundaries at the pixel level.

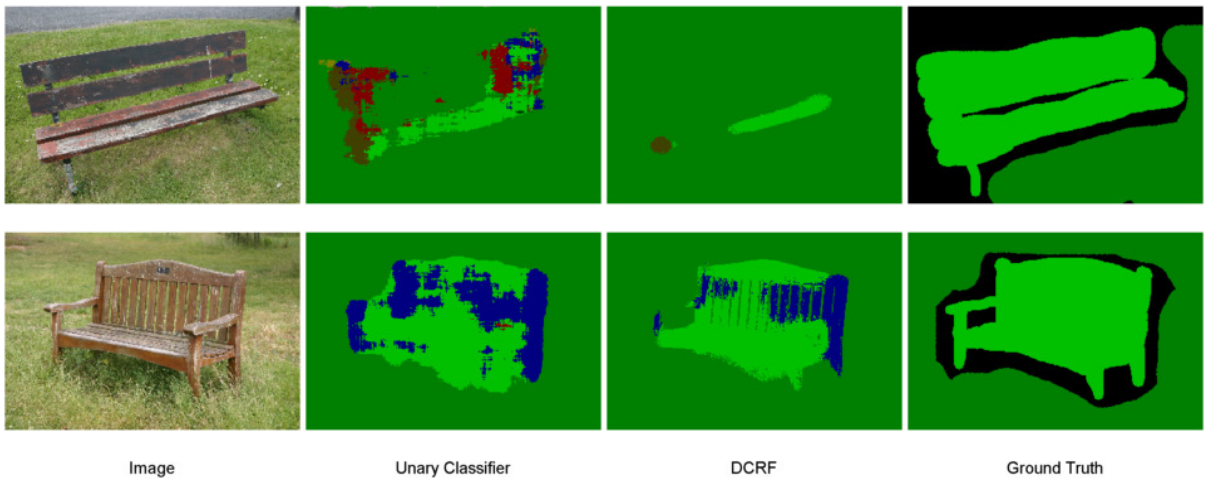


Figure 5.2: DCRF is prone to over-smoothing small objects from thing classes (upper image); moreover, dense random fields are confined to the success of the initial unary classifier (lower image).

classes (figure 5.2, upper image). Moreover, success of dense random fields are restricted to correctness of the initial unary classifier. If the initial unary potentials fails to identify the objects in the image correctly, DCRF does not revise the object labels and continues to refine boundaries of wrong labels (figure 5.2, lower image).

In this chapter, a new context-based dense conditional random field (cbDCRF) model is proposed which integrates global semantics of the image with pixel-wise dense inference to preserve small thing classes and to make dense inference robust to initial misclassifi-

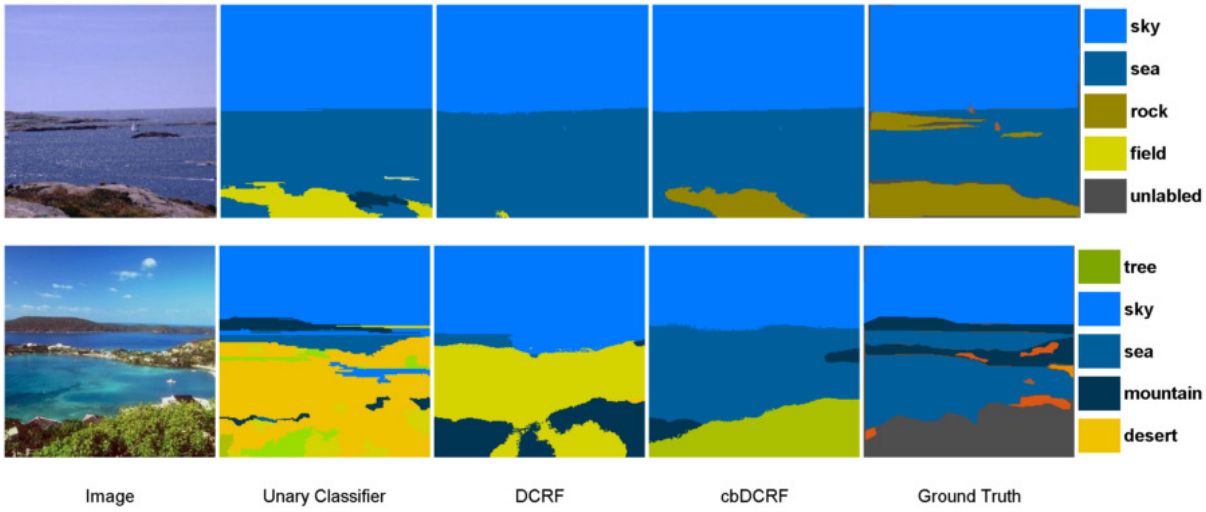


Figure 5.3: Knowing that an image is picturing a coastal area based on the global visual characteristics of the image, the probability of rock and sea labels are increased over desert and field classes.

cations of the unary classifier. We propose to utilize global scene type of the image to strengthen probability of objects coherent with the global scene and eliminate ambiguity due to between-class visual similarity. Distribution of each object varies significantly through different scene types. Even frequency of occurence of a common object such as 'tree' is greatly changing from 289 samples under 'forest' category to 30 samples in 'coastal' scenes; or class 'desert' occurs only in 'open country' scenes. We show that prior knowledge about the image scene type can improve the performance of dense CRF labeling. As an example in figure 5.3, knowing that an image is picturing a 'coastal' area based on the global visual characteristics of the image, the probability of 'rock' and 'sea' labels are increased over 'desert' and 'field' classes.

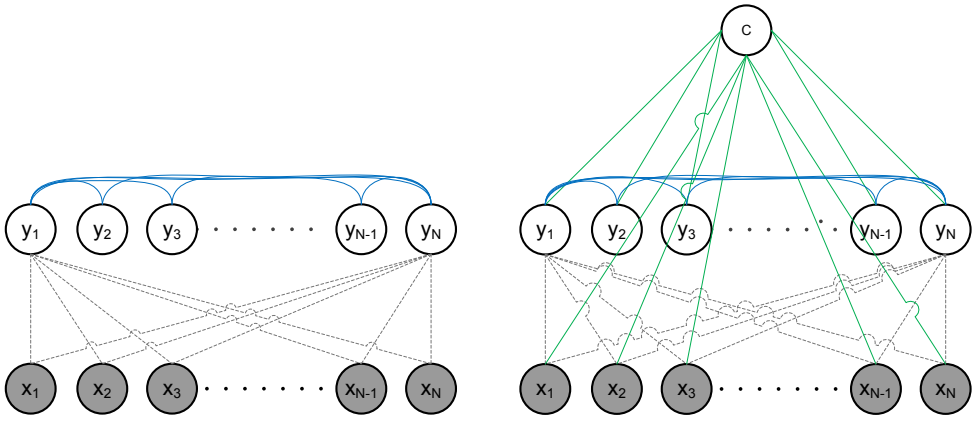
The new model forces scene-object co-occurrence restrictions to improve object-object cooccurrence prediction. Joint probability of labeling configuration and image scene type is factorized using the mean field approximation method to obtain prediction update

equations for labeling individual image pixels and predicting overall scene type of the image. CRF pairwise potentials connect each image pixel with all other pixels in the image to account for long-range interactions of objects. Scene type context is integrated as a model selection cue in the new model to alleviate sensitivity to unary initialization and severe smoothing problem by elevating scene-object and object-object co-occurrence prediction. Whole image descriptors are used to discriminate distinct environmental categories using an SVM scene classifier; the CRF unary potentials are then conditioned on the overall scene type of the image to impose scene-object and object-object consistency. We derive the inference algorithm for the proposed context-based dense CRF model which enhances both scene and object prediction.

In the following, the new cbDCRF model is elaborated. In section 5.2, the related inference algorithm is derived. Section 5.1 explains the individual components of the model for implementation purposes. Experiment results are nailed down in section 5.4.

5.1 Context-based dense CRF model

The new context-based dense CRF model is shown in figure 5.4-(b) in comparison with conventional dense CRF in 5.4-(a). The following notation is used to describe the model. N is the total number of pixels in the image I; $X = \{x_1, x_2, ..., x_N\}$ is the set of feature vectors of pixels 1 to N, where $x_i \in \Re^d$ is the d-dimensional observation feature vector obtained at pixel i. $Y = \{y_1, y_2, ..., y_N\}$ is the set of random variables representing labels of corresponding pixels where y_i can be any label l from the set of all possible labels $\mathcal{L} = \{1, 2, ..., L\}$; where $L = |\mathcal{L}|$ indicates the size of set $\mathcal{L}$. c is the random variable representing the image scene type and can take a value from a set $\mathcal{C} = \{1, \ldots, C\}$ of C possible scene types. In the cbDCRF model, every pixel is connected to every other pixel;



(a) DCRF (b) cbDCRF

Figure 5.4: In the cbDCRF model, every pixel is connected to every other pixel. The green edges represent the dependency of scene type C on observations X and the inter-dependency with the pixel labels Y. Full connectivity of the dense CRF is shown with blue edges (edges are shown only for two y_1 and y_N nodes). The gray dash edges illustrate the dependency of local labels to neighborhood observations.

the full connectivity of the dense CRF is shown in figure 5.4 with blue edges (edges are shown only for two y_1 and y_N nodes). The gray dash edges illustrate the dependency of local labels to neighborhood observations as previously explained as the merit of general CRF modeling. The green edges represent the dependency of scene type c on global observation X and inter-dependency with the pixel labels Y.

Let $P(Y|X,c)$ to be probability of labeling configuration Y under context c given image observations X, and $P(c|X)$ to be the probability of infering context c given image observations X, then the formulation for proposed context-based dense CRF is:

$$P(Y,c|X) = P(Y|X,c)P(c|X) = \frac{1}{Z}P(c|X)\exp\left(-E(Y|c,X)\right) \tag{5.1}$$

where Z is the normalizing partition function that ensures $\sum_i P(y_i|x) = 1$. Moreover, $E(Y|X,c)$ is the Gibbs energy defined as:

$$E(Y|X,c) = \alpha \sum_{i=1}^{N} \psi_u(y_i|x) + \beta \sum_{i=1}^{N} \psi_c(y_i|x) + \sum_{i<j} \psi_p(y_i, y_j|x_i, x_j) \tag{5.2}$$

where $\psi_u(y_i|x)$ is the association (unary) potential between label y_i of site i and observation data x of site i and its neighborhood. $\psi_p(y_i, y_j|x_i, x_j)$ is the interaction potential between current pixel i and pixel j given the observed features x_i and x_j. $\psi_c(y_i|x)$ represents the context-based association potential defined as the negative log-likelihood of label y_i given scene type c and observations x. The unary or associative potentials are defined upon individual image pixels and model the association between class labels and low-level image features such as color and texture; that is, they incorporate the image evidence to labeling task in the format of probability of a class label given the image low-level informarion at each pixel. Both ψ_u and ψ_c could be the output of a classifier for probability of label y_i given observations x [42]. We describe the detailed formulation of the conventional and proposed context-based unary potentials in section 5.3. α and β parameters control the degree of employment of contextual scene-based unary potentials.

$\psi_p(y_i, y_j|x_i, x_j)$ is the interaction potential between current pixel i and pixel j given the observed features x_i and x_j. The pairwise potentials are defined on labels of neighboring image pixels and are meant to maximize local label agreement between neighboring pixels given the degree to which their apprearance is similar. It is a data dependent term whose aim is to have similar labels at a pair of sites for which the observed data is alike. A data dependent pairwise term can compensate for the errors and noise in modeling the unary potential. This is one of the advantages of CRF modeling over MRF in which the pairwise term smoothes labels independent of observation features. The data dependent pairwise potential ψ_p is formulated as:

$$\psi_p(y_i, y_j|x_i, x_j) = \mu(y_i, y_j) g(x_i, x_j) \tag{5.3}$$

where function $g(x_i, x_j)$ measures the similarity of two pixels i and j in terms of appearance and relative location. Note that the pairwise term is the most costly term in CRF models with a dense field since in such model every pixel is connected to every other pixel in the image making message passing to have quadratic complexity in the number of pixel variables N. Following Krähenbühl and Koltun [3] and Zhang et. al. [28], we apply a Gaussian function in the form of:

$$g(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right) \tag{5.4}$$

Intuitively, nearby pixels with similar features are very probable to have the same label. Therefore, if nearby similar pixels take different labels a penalty (μ) is considered which could be a Potts model in the form of $\mu(x_i, x_j) = [x_i \neq x_j]$ (where [.] is the zero-one indicator function).

Moreover, penalty of taking two different labels should be further reduced when the two pixels are less similar; in (5.4), $g(x_i, x_j)$ inflicts this notion. σ characterizes the extent to which we regard the neighborhood of a pixel whithin the image. Images with patchy and uneven objects such as 'trees' need to consider a larger neighborhood of pixels than those with even solid objects. That is because looking at trees you might see patches of sky through it. Therefore, if there are blue pixels within green pixels that do not conform to their surrounding, they might actually belong to some other object such as sky somewhat farther in the image. Larger values of σ consider a larger neighborhood as valid to explore for similarities.

Furthermore, from a computational point of view, Gaussian pairwise potentials facilitate application of a high dimensional filtering approach for efficient inference. Utilizing mean field approximation for inference, the permutohedral lattice [130], a highly efficient

convolution data structure could be applied to compute an approximation of the message passing by truncating the Gaussian kernels and making the complexity of message passing to reduce from quadratic to linear [3].

5.2 Inference

Efficient inference is critical in development of fully-connected CRF labeling. We develop an inference algorithm which is not only computationally efficient but also robust to the unary initialization and over-smoothing issue. Applying the mean field approximation [35], the inference update equations are derived for both pixel labels and scene label as described in the following. To employ mean field approximation method, the context-based joint posterior $P(Y, c|X)$ is assumed to have a fully factorized approximation of the form:

$$Q(Y, c) = q(c) \prod_{i=1}^{N} q_i(y_i) \tag{5.5}$$

To derive $q(c)$ and $q_i(y_i)$, mean field approximation minimizes the KL divergence:

$$KL(\widetilde{P}||Q) = \sum_{c,Y} Q(Y, c) \log \frac{Q(Y, c)}{\widetilde{P}(Y, c)} \tag{5.6}$$

where $\widetilde{P}(Y, c)$ is the unnormalized true distribution so that $\widetilde{P}(Y, c) = P(Y, c)Z$. Expanding the KL divergence expression the closed form inference update equations are derived as:

$$\begin{aligned} q(c) &= \frac{1}{Z_c} \exp\left(\mathrm{E}_{allq_i}\left[\log \widetilde{P}(Y, c)\right]\right) \\ q_i(y_i) &= \frac{1}{Z_i} \exp\left(\mathrm{E}_c \mathrm{E}_{-q_i}\left[\log \widetilde{P}(Y, c)\right]\right) \end{aligned} \tag{5.7}$$

where $\widetilde{P}(Y,c)$ is the term in equation (5.1) when not normalized by $Z(X,c)$. E_{allq_j} means expected value under all distributions q_i; E_c refers to expected value under the distribution $q(c)$ and E_{-q_i} means expectation with respect to all distributions q_j excluding the distribution for node i. Z_c and Z_i are normalizing factors enforcing $q(c)$ and $q_i(y_i)$ to be probability values:

$$\begin{aligned} Z_c &= \sum_{c=1}^{C} \exp\Big(\mathrm{E}_{allq_i}\Big[\log \widetilde{P}(Y,c)\Big]\Big) \\ Z_i &= \sum_{y_i=1}^{L} \exp\Big(\mathrm{E}_c \mathrm{E}_{-q_i}\Big[\log \widetilde{P}(Y,c)\Big]\Big) \end{aligned} \tag{5.8}$$

Regarding formula (5.1) and dropping conditionality on y_i and y_j for simplified notation, the equations in (5.7) are expanded as (appendix A derives these update equations):

$$\begin{aligned} q(c) &= \frac{1}{Z_c} \exp\Big(\log p(c|I) + \sum_{i=1}^{N}\sum_{y_i=1}^{L} q_i(y_i)\mathrm{E}_{-q_i}\left[-E(Y|c,I)\right]\Big) \\ q_i(y_i) &= \frac{1}{Z_i} \exp\Big(\sum_{c=1}^{C} q(c)\mathrm{E}_{-q_i}\left[-E(Y|c,I)\right]\Big) \end{aligned} \tag{5.9}$$

where the constant term $\sum_c q(c)\log P(c|I)$ is removed from $q_i(y_i)$ expression and:

$$\mathrm{E}_{-q_i}\left[-E(Y|X,c)\right] = -\alpha\psi_u(y_i) - \beta\psi_c(y_i) - \sum_{y_j=1}^{L}\sum_{j\neq i} q_j(y_j)\psi_p(y_i,y_j)$$

Rewriting this formulation for $q_i(y_i = l)$ and replacing $\psi_p(y_i,y_j)$ from (5.3), then:

$$\begin{aligned} E_{-q_i}\left[-E(Y|X,c)\right] = &-\alpha\psi_u(y_i=l) - \beta\psi_c(y_i=l) \\ &- \sum_{y_j=1}^{L}\mu(l,y_j)\sum_{j\neq i} g(x_i,x_j)q_j(y_j) \end{aligned} \tag{5.10}$$

Algorithm 1 cbDCRF Inference

- Obtain $\psi_c(y_i)$ from (5.14)
- Initialize $q(c)$ with $p(c|X)$ from (5.11)
- Initialize $q(y_i)$ with $\psi_u(y_i)$ from (5.12)
- **while** not converged **do**
  Compute $q(c)$ from (5.9)
  while not converged **do**
  Compute $q(y_i)$ from (5.9) and (5.10)
  end while
  end while
- $y_i = \arg\max_{l \in \mathcal{L}} \{q(y_i)\}$

where $Q(y_j) = \sum_{j \neq i} g(x_i, x_j) q_j(y_j)$ is a message passing term which is the computational bottleneck of the inference algorithm since for each variable y_i corresponding to each pixel, it requires a sum over all other variables y_j. This implies equation (5.10) has quadratic complexity in the number of pixels in the image. To reduce the computational complexity of message passing, we follow Krähenbühl and Koltun [3] to express the $Q(y_j)$ term as a convolution with a Gaussian kernel and approximating the Gaussian kernel by setting all values beyond two standard deviations to zero. Then the convolution at each pixel is computed approximately by aggregating values from only a limited number of neighboring variables such that message passing term $Q(y_j)$ can be roughly computed at a linear complexity $\mathcal{O}(N)$.

The Gibbs energy in the argument of exponential for infering $q_i(y_i)$ in (5.9) can be viewed as an extension of a mixture of experts model over C different scene types [131]. The $q(c)$ function works as a gating function for influence of labeling under each of the scene types over final labeling output. Regarding the scene-based labeling, the complexity of infering class labels of pixels is actually of order $\mathcal{O}(CN)$ which is still linear and could be compensated for by using parallelisam and GPU based computing. Algorithm 1 is

showing an overview of the inference procedure.

5.3 Model Learning

Since exact maximum-likelihood training is intractable for large undirected graphical models, our training algorithm is based on piecewise training method [119]. We take a modular approach to implementation of the proposed model which requires finding scene category likelihood $p(c)$, learning of general ($\psi_u(y_i)$) and scene-based ($\psi_c(y_i)$) unary potentials, and adjustment of model parameters. Each part is described in the following subsections. Algorithm 2 lists required implementation steps.

5.3.1 Scene classification

To do the initial scene classification, we train a multiclass SVM classifier [129] with global feature observations X_g for each image I. For observation features, we used a standard bag-of-words spatial pyramid with 1, 2 and 4 levels over a 1024 sparse coding dictionary of SIFT features, colorSIFT, RGB histograms and color moment invariants

Algorithm 2 Implementation Steps

- Select train imageset $\mathcal{D}$ with true scene ($c \in \mathcal{C} = \{1, \ldots, C\}$) and object ($l \in \mathcal{L} = \{1, \ldots, L\}$) labels
- Compute global feature vectors X_g for each image $I \in \mathcal{D}$
- Train SVM classifier for $\{X_g\} \in \mathcal{D}$ and $c \in \mathcal{C}$ using [129]
- Train unary classifier $\psi_u^{par}(y_i|x)$ on $\mathcal{D}$ using [1]
- Train unary classifier $\psi_u^{npar}(y_i|x)$ on $\mathcal{D}$ using [8]
- Train scene-based unary classifier $\psi_c^{par}(y_i|x)$ using [1]
- Train scene-based unary classifier $\psi_c^{npar}(y_i|x)$ using [8]
- Train calibration parameters a_l and b_l, $l \in \mathcal{L}$ by minimizing (5.15)
- To Label a new query, go to Algorithm 1

[6]. All the positive and negative examples in the train set of the databases are used for training the classifier. We used a validation set of images to tune the SVM parameters, the penalty parameter and kernel parameters. For image I, probability of being of scene type $c \in \mathcal{C} = \{1, \ldots, C\}$ is:

$$p(c) = \frac{1}{1 + \exp(w_c X_g + b_c)} \tag{5.11}$$

where w_c is the trained weight vector for scene c and b_c is the corresponding bias vector.

5.3.2 Unary potentials

We utilize the combination of a parametric (ψ_u^{par}) and a non-parametric (ψ_u^{npar}) object classifier as unary potential $\psi_u(y_i|x)$ for discriminating different classes based on observation features. That is:

$$\psi_u(y_i|x) = \psi_u^{par}(y_i|x) + \psi_u^{npar}(y_i|x) \tag{5.12}$$

Due to the fact that different datasets have varying characteristics in terms of object distributions and structure, we employ the combination of a parametric [1] and a non-parametric [8] object classifier. Distribution of objects in some imagesets are even so that there are roughly the same number of samples from each class (although the frequency of background classes such as 'sky' are inevitably slightly dominant). This characteristic facilitates the employment of parametric classifiers such as boosting [1] and neural networks [2] to produce satisfactory recognition results. However, objects in larger imagesets have a power-law distribution so that objects have imbalanced counts and there are very few samples for some classes. Due to this class imbalance problem, parametric classifiers

such as boosting do not produce satisfactory results; whereas non-parametric discriminative algorithms [7, 8] based on nearest-neighbor (NN) classifiers and content-based image retrieval (CBIR) provide higher labeling accuracy.

Parametric classifier: TextonBoost

For the parametric $\psi_u^{par}(y_i|x)$ classifier, we apply the TextonBoost model in [1] to train $\psi_u^{par}(y_i|x)$ over the entire training dataset. TextonBoost fuses appearance and contextual features in a grid-structure CRF model. In [1], authors propose new texton-shape features [132] which are capable of modeling object shape, appearance and context by capturing the relative texton locations for certain classes. They train texton-shape features using boosting to produce a multi-class logistic classifier. Pixel color information and a prior on class locations in the image are added in the form of extra unary potentials to the overal CRF model for improved performance.

Non-parametric classifier: SuperParsing

For the non-parametric $\psi_u^{npar}(y_i|x)$ classifier, we apply the the SuperParsing method in [8]. For each new test image, non-parametric models retrieve the most similar training images and transfer their appropriate label information onto the label space of the query image; this approach moderates the severe effect of biased vote of parametric methods for frequent classes. Another advantage of non-parametric models over competent classifiers is that parametric models are most suitable for discrimination of a fixed number of object categories and as the database or number of object categories grows, they become inefficient since they need to be trained anew. Tighe et. al. in [8] have proposed a successful non-parametric appraoch which does not need training and are suitable for large and heterogeneous datasets.

5.3.3 Context-based unary potentials

To compute the scene-based unary classifier, $\psi_c(y_i|x)$ in (5.2), we also utilize the combination of the parametric TextonBoost model (ψ_c^{par}) and non-parametric SuperParsing model (ψ_c^{npar}) to discriminate different classes under each scene type given observation features x:

$$p(y_i|x,c) = \psi_c^{par}(y_i|x) + \psi_c^{npar}(y_i|x) \tag{5.13}$$

For each $c \in \mathcal{C} = \{1,\ldots,C\}$, ψ_c^{par} is trained using all training images with scene label c. That is, TextonBoost is separately run $|\mathcal{C}|$ times to train each $\psi_c^{par}(y_i|x)$ under each scene type. Moreover, non-parametric scene-based unary potentials $\psi_c^{npar}(y_i|c)$ are obtained by performing the image retrieval phase of SuperParsing over images with scene label c only. The scene-based unary potentials are then determined using the sigmoid function as:

$$\psi_c(y_i|x,c) = -\log\left(\frac{1}{1+\exp\big(-(a_l \times p(y_i|x,c) + b_l)\big)}\right) \tag{5.14}$$

The coefficient a_l and bias parameters b_l are defined to moderate the class imbalance problem. Class frequencies usually follow a power-law distribution particularly in large databases. Positive samples of classes 'building' and 'tree' are widely available across different kinds of images; but samples for 'boats', 'street lights' and 'awnings' are rarely photographed even in large heterogeneous databases. Apart from class imbalance, faulty segmentation also intensifies the misclassification problem between frequent classes and rare objects because many times the rare objects are of smaller size too. This poses a challenge to the recognition system by biasing the output probabilities with more frequent classes. In equation (5.14), a_l and b_l work as a calibration parameter to compensate for this class imbalance. In the following we explain how to learn these parameters.

Train calibration parameters

To train the calibration parameters a_l and b_l, we apply the method by [133]. Coordinate descent is applied to minimize the following cost function over the calibration parameters defined as pixel labeling accuracy averaged over all classes:

$$\Gamma(y,t) = 1 - \frac{1}{L}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{i;t_i=l}[t_i = y_i] \tag{5.15}$$

where the second summation is over all pixels in all training images. y_i is the output pixel label; where i indexes over all pixel in the imageset. n_l is the number of pixels with ground truth label l. L is the number of classes. [.] is 1 if the internal condition in true and 0 otherwise. The inverse class counting factor $\frac{1}{n_l}$ is considered to deal with the class imbalance problem. The reason for using coordinate descent instead of gradient descent is the fact that the cost function in (5.15) is not differentiable due to maximization in y_i. Coordinate descent iteratively applies line search to optimize the loss over a single parameter at a time, keeping all others fixed. This process cycles through all parameters until convergence [133].

5.4 Performance Analysis

In this section, we study the performance of proposed context-based dense CRF and compare it with some of the state-of-the-art labeling algorithms. Both per-pixel and per-class classification rates have been reported as quantitative measures for fairness in evaluation of recognition rates of objects from 'thing' and 'stuff' classes. Consideration of both of these measures is particularly critical for large datasets with imbalanced class distributions. Per-pixel rate is obtained by deviding the number of correctly classified

pixels by the total number of labeled test pixels; it gives the proportion of the correctly labeled. Therefore, it is majorly biased with recognition rate of common and large background objects in 'stuff' classes. On the other hand, per-class rate is obtained by averaging the recall rates ($\frac{TP}{TP+FN}$) of individual classes and it gives a better measure of performance over 'thing' objects.

The proposed model is tested on MSRC imageset and a subset of SIFT Flow dataset. In this work, we investigated that for MSRC dataset, the best results could be obtained by $\psi_u^{npar}(y_i|x) = 0$, that is, $\psi_u(y_i|x) = \psi_u^{par}(y_i|x)$. This is due to the fact that MSRC has a roughly even object distribution. However, there are only limited samples (20 to 40 images) from all classes in MSRC imageset which makes application of non-parametric classifiers such as Nearest Neighbors (NN) incompetent. SIFT Flow objects have imbalanced counts such that there are very few samples for classes such as boat, bus, or bird. Therefore, parametric classifiers perform poorly for these classes. Therefore, for SIFT Flow imageset, we set $\psi_u^{par}(y_i|x) = 0$, meaning $\psi_u(y_i|x) = \psi_u^{npar}(y_i|x)$.

5.4.1 MSRC imageset

We run the first experiment on the frequently-investigated MSRC-21 dataset [1] which has 591 images manually labeled in 21 different classes of objects from both 'thing' and 'stuff' categories. For scene labels, we used the annotations provided by [6] where each image is given the label of the salient foreground object as scene type. There are 21 different scene categories obtained. To train our model, we used the train/validation/test split by [3].

Table 5.1 reports the quantitative performance of the proposed system in comparison with other models in the literature. We compare against grid-structure CRF [1] (used

Table 5.1: Compariative quantitative analysis of performance of proposed context-based dense CRF on MSRC imageset.

numbers in %	Harmony potential [107]	Unary classifier [1]	DCRF μ: Potts [3]	CB DCRF μ: Potts (ours)
Per Pixel Recall	77	84	81.6	83.26
Per Class Recall	75	76.6	70.8	75.41
building	60	71.9	67	69.58
grass	78	98.1	98.3	97.59
tree	77	89.7	84.6	82.51
cow	91	84.3	73.2	74.99
sheep	68	80.5	68.6	82.21
sky	88	93.3	94.6	92.87
plane	87	82.4	60.3	65.18
water	76	67.5	71.9	75.77
face	73	88.1	76	89.13
car	77	84.2	82.5	80.85
bike	93	91.1	81.7	91.35
flower	97	90.7	95.2	97.57
sign	73	70	79.4	69.84
bird	57	47.6	38.4	50.73
book	95	94.1	95.5	96.93
chair	81	59.3	47.7	83.54
road	76	88.8	87.6	83.62
cat	81	75.7	62.4	68.21
dog	46	46	42.1	46.30
body	56	79.9	67.3	75.18
boat	46	25.1	13.7	9.60

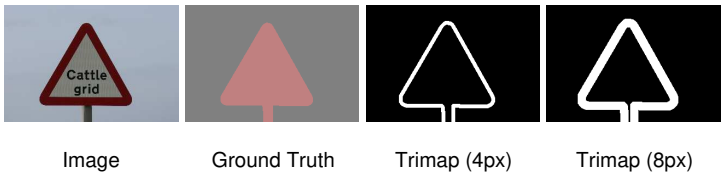
as unary classifier), context-based Harmony potential [107] and conventional dense CRF [3]. Our algorithm as well as [3] use the classifier by [1] for CRF unary potentials. [107] applys similar scene level information as ours to their CRF model. The proposed model outperforms the two other methods in terms of per-pixel and per-class accuracy conveying it can well detect objects from the 'thing' classes as well as 'stuff' classes.

Our context-based dense CRF model (cbDCRF) is built on the dense CRF (DCRF) work of [3]. The conventional dense CRF model in [3] with the Potts compatibility function severely damages recognition rate of small 'thing' objects (such as 'bird', 'cat',

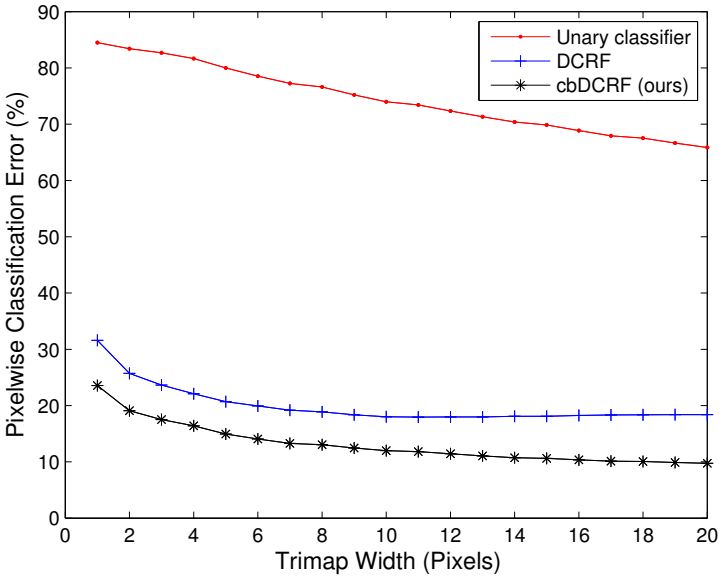
'dog' and 'chair') which are usually objects of interest in images. Krähenbühl and Koltun discuss reasonably in [3] that a Potts model for the compatibility function μ in (5.3) has the shortcoming to penalize an incompatible label pair like 'sky' and 'cat' to the same extent as a certainly compatible label pair 'sky' and 'bird'; and they instead train a general symmetric compatibility function using L-BFGS to maximize the log-likelihood of CRF model for a validation set of images. However, learning the compatibility function requires computation of the gradient of the dense CRF which is very computationally expensive and becomes intractable with growth of the number of classes. The proposed context-based model with an undemanding Potts compatibility function alleviates the severe smoothing effect of DCRF model and outperforms dense CRF.

It is noteable that, although these dense models seem to be behind the unary classifier on the MSRC dataset, boundaries of objects and things are clearer, finer and more precise in dense models than grid-structure unary in [1]. Note that there are many objects of 'thing' class in this dataset which are vulnerable to drowning in the large pool of pixels from the background class. Furthermore, dense models deliver highly better segmentation results as stated above. To illustrate this fact, we use 'trimap' measure [4] to compute segmentation errors of each method. Trimap measure of segmentation error counts the number of misclassified pixels within a narrow band (trimap) surrounding actual object boundaries which is obtained from accurate ground truth object boundaries [3]. Figure 5.5-(a) showes the visualization of trimap method and resulting error percent computed by it using trimaps of width 1 to 20 pixels in Fig. 5.5-(b). As illustrated, proposed cbDCRF model generates the least segmentation error. To obtain this figure, the test was run on the accurate ground truth images from Krähenbühl in [3].

α and β adjust the degree to which contextual scene-based potentials are employed. Gradient-based optimization or grid search on a holdout validation set can be applied to



(a)



(b)

Figure 5.5: (a) Visualization of trimaps of different width (b) Percent of misclassified pixels within trimaps of different width

obtain best values for α and β. In Fig. 5.6 graph, it is shown that by changing the ratio of α and β, the optimum performance is obtained when α and β are equal. This is due to the fact that in the proposed method, we are optimizing the label of the scene and the object labels reciprocally. Therefore, equal contributions of the general and context-based unary potentials results in the best performance. Note that the proposed model is able to improve scene classication performance up to 85.95% which to the best of our knowledge is the highest score reported on MSRC so far.

Figure 5.7 showes the effect of increasing/reducing the number of gating functions. To

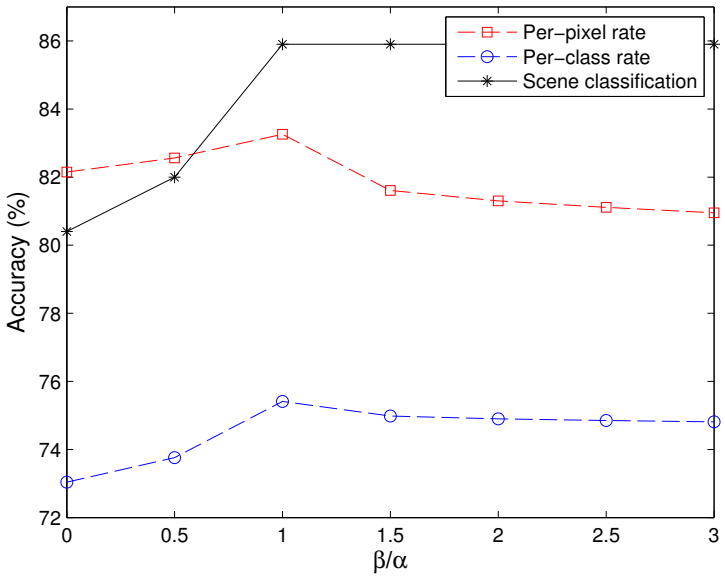


Figure 5.6: $\beta = 1$ gives the maximum per-pixel accuracy, per-class accuracy and sccene classification rate.

obtain this figure, for each image, all possible scene types are arranged in the descending order of value of $p(c)$; then, $|C| = 1$ implies applying only the scene type with maximum value of $p(c)$. $|C| = 2$ implies the first and second scene type with maximum ranking of $p(c)$. As illustrated in Fig. 5.7, gating function of the scene type with maximum $p(c)$ gives the best results in terms of per-pixel accuracy and per-class accuracy. In fact, adding more number of gating functions loosens the advantage of applying prior knowledge about the overal scene type of the image and this reduces the accuracy to a small extent. Figure 5.7 is obtained by fixing the parameters α and β to one. Figure 5.6 is obtained by fixing the number of scene-based gating functions ($q(c)$) of $q_i(y_i)$ formulation in (5.9) to only one.

Figure 5.8 shows some examples of results obtained by porposed cbDCRF model. In this figure, first and last column from left are the original image and ground truth labeling. Second column from left is the output of the original unary classifier without

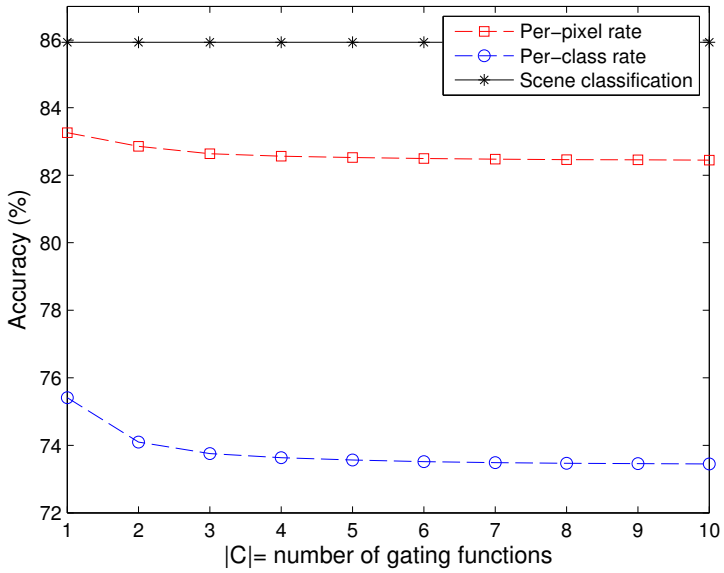


Figure 5.7: The strongest gating function $q(c)$ gives the maximum per-pixel accuracy and per-class accuracy.

the help of CRF. Third column from left is the result of conventional dense CRF model as proposed by Krähenbühl and Koltun [3]; and forth column is the output by the proposed context-based (scene-based) dense CRF (cbDCRF) model. First and forth figures from the top are illustrating a bench which are initially identified partially as a chair and partly a cow due to brown color surrounded with green grass color and texture. The DCRF model refines the chair boundaries delicately; however, it does not correct the pixels misclassified as cow. The proposed cbDCRF model modifies the misclassified pixels and labels all the area of the benches correctly. This is due to the fact that the integrated scene classifier identifies the whole structure of the image to belong to 'chair' scene type (context). In the third row, the DCRF has refined the boundaries of the 'bird' area as 'dog'; but our cbDCRF model refines the the boundaries of the 'bird' area correctly as 'bird'. Similarly for other illustrated examples in Fig. 5.8, accessing prior information in the form of context of the image, the proposed model corrects mis-labeling of the unary

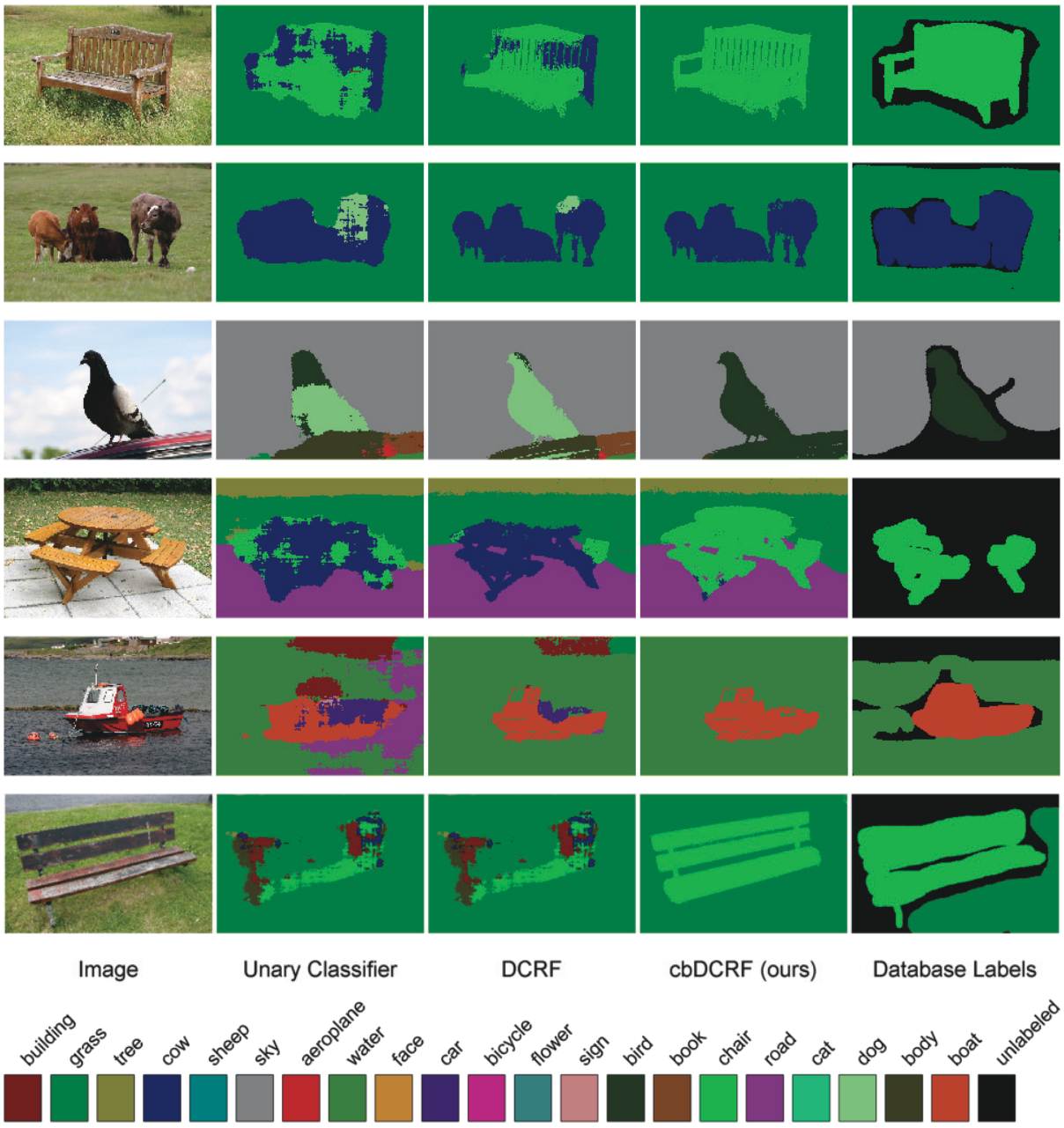


Figure 5.8: Examples from MSRC imageset: Accessing prior information in the form of scene type of the image, the proposed cbDCRF model corrects mis-labeling of the unary classifier; whereas the conventional DCRF keeps refining the wrong labels.

classifier; whereas the conventional DCRF keeps refining the wrong labels.

5.4.2 SIFT Flow database

We run experiments on a subset of the SIFT Flow database from [7] which contains 1692 fully annotated images with 33 different classes of objects. The imageset is randomly split into 1556 images for training and 136 test images. Images are organized in 5 scene categories including 'Coast', 'Forest', 'Mountain', 'Open country' and 'Tall building'. Distribution of objects varies significantly under different scene types. We believe that prior knowledge about scene type can improve the performance of dense CRF over labeling in a diversive imageset as SIFT Flow.

For SIFT Flow database, we compared different set of global features (X_g) to train the SVM for scene classification. We split the train set to 1333 images for SVM training and 223 validation images for kernel selection and parameter adjustment. Gist features (512 dimensional) obtained the highest scene classification rate of 90% with a linear kernel.

Table 5.2: Quantitative analysis of performance of two implementations of proposed context-based dense CRF (cbDCRF with/without rare class calibration) against original unary classifier, Grid CRF and conventional dense CRF (DCRF) on SIFT Flow imageset.

numbers in %	Unary Classifier ([8])	DCRF μ : Potts [3]	cbDCRF no calibration (ours)	cbDCRF calibration (ours)
Per Pixel Recall (all data)	73.00	75.45	79.39	74.61
Mean Top 10 Class (>90%)	50.61	52.71	60.84	58.02
Per Class Recall (all data)	25.24	22.92	26.71	27.09
sky	90.55	86.31	92.50	81.28
building	85.50	90.50	87.15	84.92
tree	86.07	82.90	82.34	74.90
mountain	64.76	79.38	82.35	82.37
sea	70.37	69.98	88.68	88.70
field	42.62	69.90	52.35	44.90
sand	12.55	14.39	31.19	31.19
river	10.24	9.87	41.98	42.00
plant	5.35	3.59	7.33	7.39
grass	38.11	20.34	42.55	42.56
rock	6.2721	0.1063	6.0759	42.8395

The classisication rate was calculated by deviding the number of correctly classified images by the total number of test images (136 for SIFT Flow database). We use the $p(c|X_g)$ in (5.11) which the LIBSVM library [129] provides for multi-class SVM.

In table 5.2, we compare performance of two implementations of the proposed context-based dense CRF (cbDCRF) with the original unary classifier in [8] and fully-connected dense CRF (DCRF) in [3]. Proposed model without rare class calibration outperforms the baseline unary classifier and DCRF, both in terms of per-pixel and per-class accuracy. The full proposed cbDCRF model with rare class calibration outperforms the mean per-class detection rate of both the baseline unary classifier and DCRF model in terms of per-class accuracy and in top 10 most frequent classes which constitute more than 90% of all test data. The listed objects in table 5.2 are sorted in descending order of number of pixels in the test set. Figure 5.9 shows some examples of quality of results obtained by porposed cbDCRF model on SIFT Flow imageset.

Since an analysis of the errors of the proposed model helps to identify the shortcomings of our method and can suggest directions for future research, examples of erronous cases for MSRC database and SIFT Flow database are presented in Fig. 5.10. The overal scene type of these images has failed to be identified correctly. The cat has mistakenly been identified as bird; the dog as human face and body; the boat as bike; and lastly, open country area and vegetation has wrongly been identified as mountain. That is, SVM has identified the image as a mountain scene but the image is an open country scenery. Therefore, both conventional DCRF and the proposed model failed to refine the results for correct labels. This issue can be moderated by integrating object detectors [92] with the proposed model.

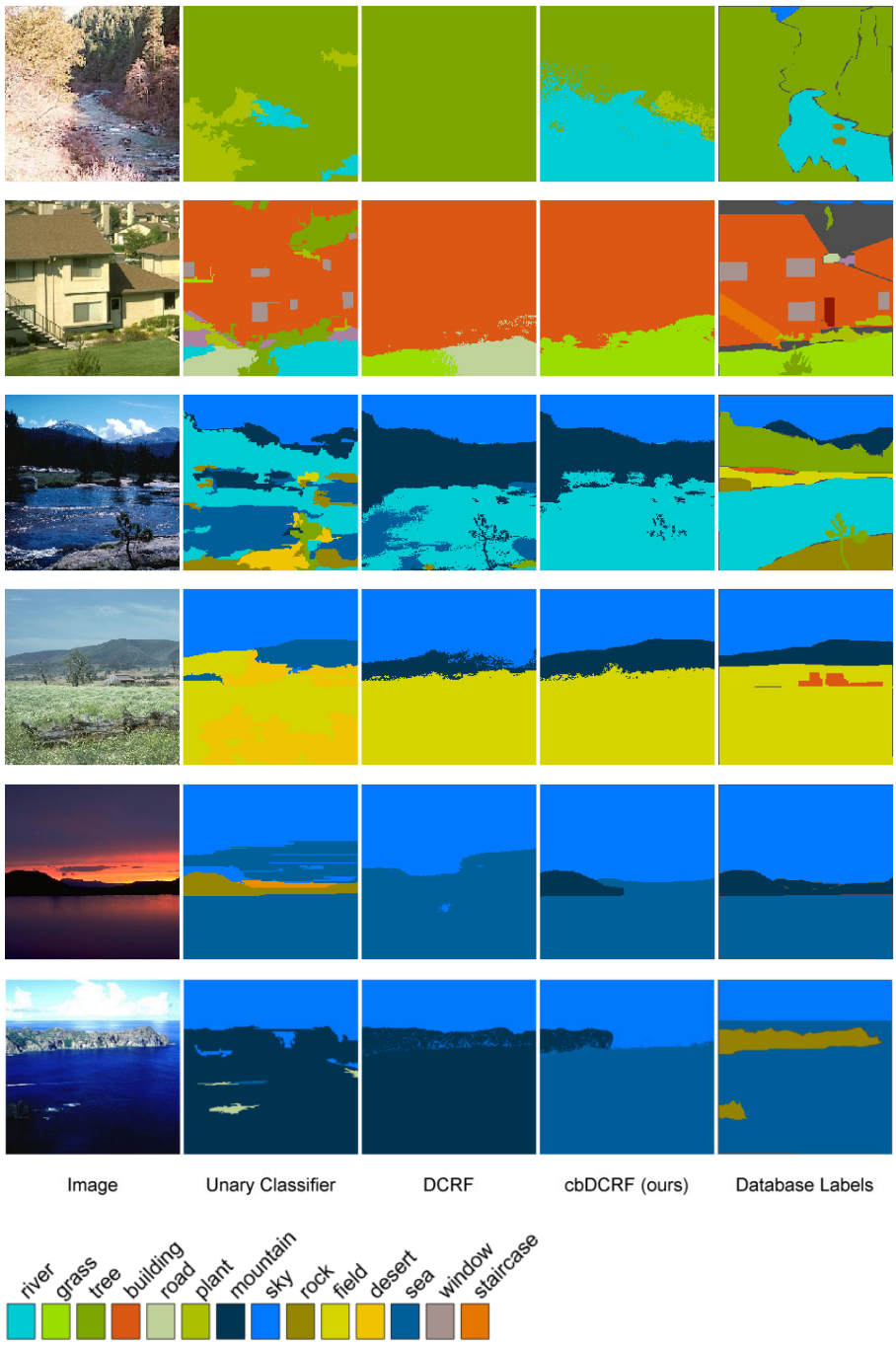


Figure 5.9: Examples from SIFT Flow imageset: Accessing prior information in the form of scene type of the image, the proposed cbDCRF model corrects mis-labeling of the unary classifier; whereas the conventional DCRF keeps refining the wrong labels.

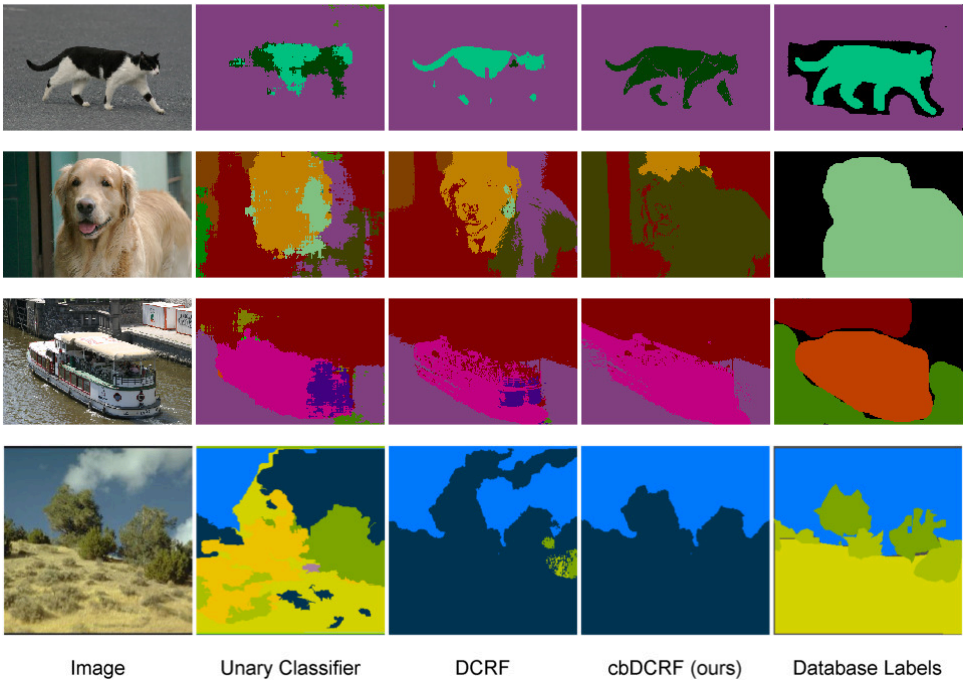


Figure 5.10: The cat has mistakenly been identified as bird; the dog as human face and body; the boat as bike; and lastly, open country area and vegetation has wrongly been identified as mountain. Therefore, both conventional DCRF and the proposed model failed to refine the results for correct labels.

5.5 Comparison to GGM-based CRF

In chapter 4, we introduced a CRF model based on the generalized Gaussian mixture distribution. We compare performance of the context-based dense CRF model proposed in current chapter with performance of the GGM-based CRF over the Corel imageset. Table 5.3 summarizes the results of our experimentation. The first two columns of table 5.3 compare the performance of the GGM-based CRF with the performance of the parametric TextonBoost classifier in [1] which was applied in section 5.3.2. That is, we

Table 5.3: Comparison of performance of the cbDCRF model with GGM-based CRF over the Corel imageset.

numbers in %	TextonBoost Classifier [1]	GGM CRF (Chapter 4)	cbDCRF TextonBoost Unary	cbDCRF GGM CRF Unary
Per Pixel Recall	69.16	68.02	74.55	71.24
Per Class Recall	63.79	69.54	66.91	69.82
Ground	76.38	71.87	73.23	80.95
Vegetation	62.86	59.58	73.14	60.09
Water	76.85	73.68	95.94	85.95
Snow	68.10	68.35	90.69	80.21
Hippo	68.23	80.96	62.78	66.27
Bear	56.03	47.06	41.51	34.73
Sky	38.06	85.27	31.06	80.58

first compare the performance of unary classifiers before enforcing scene-based analysis and context-based constraints. TextonBoost classifier is ahead of proposed GGM-based CRF model by about 1% in terms of per-pixel accuracy; however, proposed GGM-based CRF model is ahead of TextonBoost classifier in terms of per-class accuracy by about 5%. That is, proposed GGM-based CRF model has better performance across different classes. Furthermore, proposed context-based dense CRF model introduced in current chapter enhances the performance of both TextonBoost classifier and proposed GGM-based classifier introduced in chapter 4.

5.6 Discussion

This chapter proposes a method for alleviating the severe smoothin effect of pixel-level dense CRF (DCRF) model by employing the image level scene type contextual information. The proposed context-based dense CRF model (cbDCRF) also makes DCRF robust to misclassifications initially imposed by the integrated unary classifier. Global

scene type of the image is utilized to eliminate ambiguity due to between-class visual similarity by strengthening probability of objects coherent with the global scene. That is, scene-object co-occurrence restrictions are applied to improve object-object cooccurrence prediction. Joint probability of labeling configuration and image scene type is factorized using the mean field approximation method to obtain prediction update equations for labeling individual image pixels and predicting overall scene type of the image. We derive the inference algorithm for the proposed context-based dense CRF model which enhances both scene and object prediction. The proposed model is able to improve scene classification performance as well as accuracy of semantic segmentation.

Chapter 6

Conclusion

6.1 Summary

This thesis studies development of robust image labeling systems for the purpose of semantic segmentation of images. There are two primary problems ragarding successful image labeling. The first one is reliable representation of visual features and finding their proper mapping onto the possible label space. The second problem of great importance in image labeling is proper usage and formulation of contextual information to leverage labeling accuracy using algorithms which have low computaional cost.

Conditional random fields (CRF) from the class of probabilistic graphical models provide a good framework for studying both of these problems and have proven to deliver prominent results on various benchmark imagesets. Ragarding each of the above labeling problems, this thesis proposes a solution in the CRF framework whose performances competes with or outperforms state of the art literature.

The first approach proposes novel feature functions based on generalized Gaussian mixture (GGM) distribution to be utilized as CRF potential functions. The shape pa-

rameter in GGM distribution proves its efficacy to deliver more accurate data fitting and therefore, more accurate labeling of data, particularly in smooth even image regions. The proposed feature functions deliver more consistent semantic segmentation. In comparison with their Laplacian and Gaussian conterparts, the proposed GGM-based feature functions generated higher performmance in terms of both recall and precision criteria. That is, the new approach obtains higher accuracy with less type-I and type-II errors.

Performance of the proposed feature functions was also compared with support vector machines (SVM) and powerful deep neural networks. The results showed the higher performance of the new model. We deduct that in spite of great capabilities of DNNs, it is behind mixture modeling in performance where availability of training data is limited and where precise localization of labels is required.

Secondly, this thesis proposes a new context-based dense CRF model (cbDCRF) which takes advantage of scene type information of the image to make dense CRF model robust to initialization condition of the unary potentials. The new model also proves to generate more accurate labeling of small foreground objects in the large drowning pool of labels of objects of background classes.

In development of the new cbDCRF model, we propose to apply both parametric and non-parametric discriminative methods for the unary potentials. Parametric models produce good results when applied to controlled and structured imagesets with roughly same number of samples of all classes; however, non-parametric models such as Superparsing [8] deliver better results when applied to large heterogeneous imagesets.

Since our cbDCRF model has a pixel-wise fully-connected graph structure, an inference algorithm based on mean field approximation is applied to lower the computational cost of high order message passing. Moreover, the calibration parameters in the proposed non-parametric context-based unary potentials compensate for the class imbalance prob-

lem in large imagesets.

6.2 Future Work

Based on the current study, the following problems are open for investigation and further progress of robust labeling systems.

- Applying feature functions of objectness [134] and saliency [135] measures as another unary cue for boosting rare class and foreground object recognition.

- Scene-based contextual information reduces the space of possible labels to a few type of objects under a typical scene. This reduced label set could be used as image tags for refining the retrieval set to obtain more relevant image matchings. Also, the available reduced tags facilitate the application of tag-based models such as [136] to improve semantic segmentation.

- The proposed context-based dense CRF model assumes that the benchmark imageset is not only labeled at the object level, but also at the scene level. It is useful to investigate the performance of the proposed model if scene level information is not available. The proposed model may or may not require a clustering step based on the goodness of the initial unary labeling. In case of clustering, the proposed model could be built in combination with [112, 69].

- Combining object detectors with region labeling also can improve labeling accuracy particularly for rare objects and objects of 'thing' classes. In this approach, a likelihood map is obtained for each object class by projecting the object detector mask at the location that a detection has been fired. Then, class likelihoods from

the region-based labeling for each pixel/patch is combined (added/concatenated) with the detector-based labeling maps [92].

Generally, computer vision systems aim to determine the full 3 dimensional structure of the scene and inter-relations of the objects and their components in it to create artificial intelligence; to provide the means for machines to understand sceneries. I am excited to continue to explore these possibilities in my research.

Appendix A

Variational Inference: Mean Field Approximation

The basic idea of variational inference is to approximate the true but intractable distribution $P(Y)$ with a simple distribution $Q(Y)$ which is from a family of tractable distributions such as multivariate Gaussian or factored distribution. The assumption is that $Q(Y)$ has some free parameters that could be optimized to make it as close as possible to $P(Y)$. To do so, one cost function to minimize is the KL divergence defined as [35]:

$$\mathbb{KL}(Q||\widetilde{P}) = \sum_{Y} Q(Y) \log \frac{Q(Y)}{\widetilde{P}(Y)} \tag{A.1}$$

where $\widetilde{P}(Y)$ is the unnormalized true distribution so that $\widetilde{P}(Y) = P(Y)Z$. Alternatively, we can try to maximize the following quantity which is a lower bound on the log-likelihood of the data ($\mathcal{D}$). That is, variational algorithms reduce the inference problem to an optimization problem. By minimizing the $\mathbb{KL}(Q||\widetilde{P})$, we are actually maximizing a lower bound on the log-likelihood of the data since $\mathbb{KL}(Q||P) > 0$.

$$J(Q) = -\mathbb{KL}(Q||P) + \log Z \leqslant \log Z = \log P(\mathcal{D}) \tag{A.2}$$

Mean field approximation is one of the most popular forms of variational inference. Regarding cbDCRF model (P(Y,c)), in the mean field approach, the approximation $Q(Y, c)$ is assumed to have a fully factorized form: $Q(Y, c) = q(c) \prod_{i=1}^{N} q(y_i)$; and the goal is to minimize $\mathbb{KL}(Q||P)$.

$$\begin{aligned}
J(q_c) &= \sum_c \sum_Y Q(Y,c) \log \frac{\widetilde{P}(Y,c)}{Q(Y,c)} \\
&= \sum_c \sum_Y q(c) \prod_i q(y_i) \Big[\log \widetilde{P}(Y,c) - \sum_k \log q_k(y_k) - \log q(c) \Big] \\
&= \sum_c q(c) \underbrace{\sum_Y \prod_i q(y_i)}_{\mathrm{E}_{all q_i}} \log \widetilde{P}(Y,c) \\
&- \sum_c q(c) \underbrace{\sum_Y \prod_i q(y_i)}_{1} \log q(c) \\
&- \underbrace{\sum_c q(c)}_{1} \underbrace{\sum_Y \prod_i q(y_i) \sum_k \log q_k(y_k)}_{const} \\
&= \sum_c q(c) \mathrm{E}_{all q_i} \big[\log \widetilde{P}(Y,c) \big] - \sum_c q(c) \log q(c) + const \\
&= -\mathbb{KL}\Big(q(c) \Big|\Big| \mathrm{E}_{all q_i} \big[\log \widetilde{P}(Y,c) \big] \Big)
\end{aligned} \tag{A.3}$$

Assume $\log h_c(c) = \mathrm{E}_{all q_i} \big[\log \widetilde{P}(Y,c) \big]$; then, we maximize $J(q_c)$ by minimizing this KL, which we can do by setting $q_c = h_c$, as follows:

$$q(c) = \frac{1}{Z_c} \exp \Big(\mathrm{E}_{all q_i} \big[\log \widetilde{P}(Y,c) \big] \Big) \tag{A.4}$$

Similarly:

$$
\begin{aligned}
J(q_j) &= \sum_c \sum_Y Q(Y,c) \log \frac{\widetilde{P}(Y,c)}{Q(Y,c)} \\
&= \sum_c \sum_Y q(c) \prod_i q(y_i) \Big[\log \widetilde{P}(Y,c) - \sum_k \log q_k(y_k) - \log q(c) \Big] \\
&= \sum_c q(c) \sum_{y_j} \sum_{Y_{-j}} q_j(y_j) \prod_{i \neq j} q_i(y_i) \Big[\log \widetilde{P}(Y,c) - \sum_k \log q_k(y_k) - \log q(c) \Big] \\
&= \sum_c q(c) \sum_{y_j} q_j(y_j) \sum_{Y_{-j}} \prod_{i \neq j} q_i(y_i) \log \widetilde{P}(Y,c) \\
&- \sum_c q(c) \sum_{y_j} q_j(y_j) \sum_{Y_{-j}} \prod_{i \neq j} q_i(y_i) \Big[\prod_{k \neq j} \log q_k(y_k) + \log q_j(y_j) + \log q(c) \Big] \\
&= \sum_{y_j} q_j(y_j) \underbrace{\sum_c q(c)}_{\mathrm{E}_c} \underbrace{\sum_{Y_{-j}} \prod_{i \neq j} q_i(y_i)}_{\mathrm{E}_{-q_j}} \log \widetilde{P}(Y,c) \\
&- \sum_{y_j} q_j(y_j) \underbrace{\sum_c q(c) \sum_{Y_{-j}} \prod_{i \neq j} q_i(y_i)}_{1} \log q_j(y_j) \\
&- \underbrace{\sum_{y_j} q_j(y_j) \sum_c q(c)}_{1} \underbrace{\sum_{X_{-j}} \prod_{i \neq j} q_i(y_i) \sum_{k \neq j} \log q_k(y_k)}_{const} \\
&- \underbrace{\sum_{y_j} q_j(y_j) \sum_{Y_{-j}} \prod_{i \neq j} q_i(y_i)}_{1} \underbrace{\sum_c q(c) \log q(c)}_{const} \\
&= \sum_{y_j} q_j(y_j) \mathrm{E}_c \mathrm{E}_{-q_j} \big[\log \widetilde{P}(Y,c) \big] - \sum_{y_j} q_j(y_j) \log q_j(y_j) + const \\
&= -\mathbb{KL}\Big(q_j \Big|\Big| \mathrm{E}_c \mathrm{E}_{-q_j} \big[\log \widetilde{P}(Y,c) \big] \Big)
\end{aligned}
\tag{A.5}
$$

Assume $\log h_j(y_j) = \mathrm{E}_c \mathrm{E}_{-q_j} \big[\log \widetilde{P}(Y,c) \big]$; then, we maximize $J(q_j)$ by minimizing this KL, which we can do by setting $q_j = h_j$, as follows:

$$
q_j(y_j) = \frac{1}{Z_j} \exp \Big(\mathrm{E}_c \mathrm{E}_{-q_j} \big[\log \widetilde{P}(Y,c) \big] \Big) \tag{A.6}
$$

References

[1] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[2] Xuming He, Richard S Zemel, and Miguel A Carreira-Perpinán. Multiscale conditional random fields for image labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–695. IEEE, 2004.

[3] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems (NIPS)*, pages 109–117, 2011.

[4] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[5] Antonio Torralba, Kevin P Murphy, and William T Freeman. Contextual models for object detection using boosted random fields. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1401–1408, 2004.

[6] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–709. IEEE, 2012.

[7] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2368–2382, 2011.

[8] Joseph Tighe and Svetlana Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *Proceedings of the Europian Conference on Computer Vision (ECCV)*, pages 352–365. Springer, 2010.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[11] Zhiming Qian, Ping Zhong, and Jia Chen. Integrating global and local visual features with semantic hierarchies for two-level image annotation. *Neurocomputing*, 171:1167–1174, 2016.

[12] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10, 2007.

[13] Jia Li and James Z Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

[14] Michael S Lew. Next-generation web searches for visual content. *Computer*, 33(11):46–53, 2000.

[15] Kambiz Nayebi, H. Braren, and S. Williams. Application of machine vision technology in coupler securement inspection. In *Proceedings of the International Heavy Haul Conference*, pages 1–8, 2013.

[16] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[17] Stephen Se, David G Lowe, and James J Little. Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics*, 21(3):364–375, 2005.

[18] Congcong Li, Adarsh Kowdle, Ashutosh Saxena, and Tsuhan Chen. Towards holistic scene understanding: Feedback enabled cascaded classification models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1351–1359, 2010.

[19] François Chabat, David M Hansell, and Guang-Zhong Yang. Computerized decision support in medical imaging. *Engineering in Medicine and Biology Magazine*, 19(5):89–96, 2000.

[20] Sigurd Angenent, Eric Pichon, and Allen Tannenbaum. Mathematical methods in medical image processing. *Bulletin of the American Mathematical Society*, 43(3):365–396, 2006.

[21] Mert R Sabuncu, BT Thomas Yeo, Koen Van Leemput, Bruce Fischl, and Polina Golland. A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714–1729, 2010.

[22] Shan Du, Mohammad Ibrahim, Mohamed Shehata, and Wael Badawy. Automatic license plate recognition (alpr): A state-of-the-art review. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(2):311–325, 2013.

[23] Daniel Barbará, Carlotta Domeniconi, Zoran Durić, Maurizio Filippone, Richard Mansfield, and Edgard Lawson. Detecting suspicious behavior in surveillance images. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 891–900. IEEE, 2008.

[24] Jamie Shotton. *Contour and texture for visual recognition of object categories*. PhD thesis, Citeseer, 2007.

[25] Michael I Jordan. Graphical models. *Statistical Science*, pages 140–155, 2004.

[26] Quan Zhou, Jun Zhu, and Wenyu Liu. Learning dynamic hybrid markov random field for image labeling. *IEEE Transactions on Image Processing*, 22(6):2219–2232, 2013.

[27] Takahiro Toyoda and Osamu Hasegawa. Random field model for integration of local information and global information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1483–1489, 2008.

[28] Yimeng Zhang and Tsuhan Chen. Efficient inference for fully-connected crfs with stationarity. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 582–589. IEEE, 2012.

[29] Sanjiv Kumar and Martial Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proceedings of Ninth IEEE International Conference on Computer Vision*, pages 1150–1157. IEEE, 2003.

[30] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 670–677. IEEE, 2009.

[31] Bill Triggs and Jakob J Verbeek. Scene segmentation with crfs learned from partially labeled images. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1553–1560, 2007.

[32] Sanjiv Kumar and Martial Hebert. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, 2006.

[33] Neill DF Campbell, Kartic Subr, and Jan Kautz. Fully-connected crfs with non-parametric pairwise potential. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1658–1665. IEEE, 2013.

[34] Vibhav Vineet, Jonathan Warrell, Paul Sturgess, and Philip Torr. Improved initialisation and gaussian mixture pairwise terms for dense random fields with mean-field inference. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 73.1–73.11. BMVA Press, 2012.

[35] Kevin P Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[36] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[37] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[38] Fahmi Khalifa, Garth M Beache, Georgy Gimel Farb, Guruprasad Giridharan, Ayman El-Baz, et al. Accurate automatic analysis of cardiac cine images. *IEEE Transactions on Biomedical Engineering*, 59(2):445–455, 2012.

[39] Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[40] Charles Bouman, Michael Shapiro, et al. A multiscale random field model for bayesian image segmentation. *IEEE Transactions on Image Processing*, 3(2):162–177, 1994.

[41] Jean-Marc Laferte, Fabrice Heitz, Patrick Perez, and Eric Fabre. Hierarchical statistical models for the fusion of multiresolution image data. In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pages 42–53. International Society for Optics and Photonics, 1995.

[42] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceed-*

ings of the International Conference on Machine Learning (ICML), pages 282–289, 2001.

[43] Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellappa. Gaussian conditional random field network for semantic segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. CVPR, 2016.

[44] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008.

[45] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[46] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip H.S. Torr. Graph cut based inference with co-occurrence statistics. In *Proceedings of Europian Conference on Computer Vision (ECCV)*, pages 239–253. Springer, 2010.

[47] Yihong Gong and Wei Xu. *Machine learning for multimedia content analysis*, volume 30. Springer Science & Business Media, 2007.

[48] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *Proceedings of the IEEE Conference on Computer vision and pattern recognition (CVPR)*, pages 1–8. IEEE, 2008.

[49] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[50] Thomas Mensink, Jakob Verbeek, and Gabriela Csurka. Tree-structured crf models for interactive image labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):476–489, 2013.

[51] Wei Xia, Csaba Domokos, Loong-Fah Cheong, and Shuicheng Yan. Background context augmented hypothesis graph for object segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(4):582–594, 2015.

[52] Peng Wang, Chunhua Shen, and Anton van den Hengel. Efficient sdp inference for fully-connected crfs based on low-rank decomposition. *arXiv preprint arXiv:1504.01492*, 2015.

[53] Xiaofeng Wang and Xiao-Ping Zhang. A new laplacian mixture conditional random field model for image labeling. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Procrssing (ICASSP)*, pages 2118–2121, 2010.

[54] Bing Shuai, Gang Wang, Zhen Zuo, Bing Wang, and Lifan Zhao. Integrating parametric and non-parametric models for scene labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (cvpr)*, pages 4249–4258, 2015.

[55] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006.

[56] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[57] Pedro HO Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene parsing. *arXiv preprint arXiv:1306.2795*, 2013.

[58] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Europian Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[59] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[60] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[61] Taygun Kekeç, Rémi Emonet, Elisa Fromont, Alain Trémeau, and Christian Wolf. Contextually constrained deep networks for scene labeling. In *British Machine Vision Conference (BMVC)*, 2014.

[62] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660, 2010.

[63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[65] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015.

[66] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE, 2010.

[67] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *Europian Conference on Computer Vision (ECCV)*, pages 28–42. Springer, 2008.

[68] Xuming He. *Learning structured prediction models for image labeling*. PhD thesis, Department of Computer Science, University of Toronto, 2008.

[69] Daniel M Steinberg, Oscar Pizarro, and Stefan B Williams. Hierarchical bayesian models for unsupervised scene understanding. *Computer Vision and Image Understanding*, 131:128–144, 2015.

[70] Mohand Said Allili, Djemel Ziou, Nizar Bouguila, and Sabri Boutemedjet. Image and video segmentation by combining unsupervised generalized gaussian mixture modeling and feature selection. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(10):1373–1377, 2010.

[71] Chad Carson, Megan Thomas, Serge Belongie, Joseph M Hellerstein, and Jitendra Malik. Blobworld: A system for region-based image indexing and retrieval. In *Visual Information and Information Systems*, pages 509–517. Springer, 1999.

[72] Jan Puzicha, Thomas Hofmann, and Joachim M Buhmann. Discrete mixture models for unsupervised image segmentation. In *Mustererkennung 1998*, pages 135–142. Springer, 1998.

[73] Xiaofeng Wang, Xiao-Ping Zhang, Ian Clarke, and Yury Yakubovich. A new gaussian mixture conditional random field model for indoor image labeling. In *Proceedings of the first international workshop on Interactive Multimedia for Consumer Electronics*, pages 51–56. ACM, 2009.

[74] Mohand Saïd Allili. Wavelet modeling using finite mixtures of generalized gaussian distributions: application to texture discrimination and retrieval. *IEEE Transactions on Image Processing*, 21(4):1452–1464, 2012.

[75] Chris Russell, Pushmeet Kohli, Philip HS Torr, et al. Associative hierarchical crfs for object class image segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 739–746. IEEE, 2009.

[76] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2036–2043. IEEE, 2009.

[77] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[78] Moritz Blume. Expectation maximization: A gentle introduction. *Technical University of Munich Institute for Computer Science*, 2002.

[79] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei Efros, Martial Hebert, et al. An empirical study of context in object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1271–1278. IEEE, 2009.

[80] L Yu, J Xie, and Songcan Chen. Conditional random field-based image labelling combining features of pixels, segments and regions. *IET computer vision*, 6(5):459–467, 2012.

[81] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385, 2015.

[82] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[83] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[84] Abhishek Sharma, Oncel Tuzel, and Ming-Yu Liu. Recursive context propagation network for semantic scene labeling. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2447–2455, 2014.

[85] Jimei Yang, Bob Price, Sholom Cohen, and Ming-Hsuan Yang. Context driven scene parsing with attention to rare classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3294–3301. IEEE, 2014.

[86] Xiaofeng Wang, Xiao-Ping Zhang, Ian Clarke, and Yury Yakubovich. A new image labeling method based on content-based image retrieval and conditional random field. In *Proceedings of the International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 221–226. IEEE, 2009.

[87] Xi Chen, Abhishek Jain, and Larry S Davis. Object co-labeling in multiple images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 721–728. IEEE, 2014.

[88] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[89] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *European Conference on Computer Vision (ECCV)*, pages 28–42. Springer, 2008.

[90] L'ubor Ladickỳ, Paul Sturgess, Karteek Alahari, Chris Russell, and Philip HS Torr. What, where and how many? combining object detectors and crfs. In *Europian Conference on Computer Vision (ECCV)*, pages 424–437. Springer, 2010.

[91] Stephen Gould, Tianshi Gao, and Daphne Koller. Region-based segmentation and object detection. In *Advances in Neural Information Processing Systems (NIPS)*, pages 655–663, 2009.

[92] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3008. IEEE, 2013.

[93] Xubo B Song, Joseph Sill, Yaser S Abu-Mostafa, and Harvey Kasdan. Image recognition in context: Application to microscopic urinalysis. In *Advances in Neural Information Processing Systems (NIPS)*, pages 963–969. Citeseer, 1999.

[94] Sanjiv Kumar and Martial Hebert. A hierarchical field framework for unified context-based classification. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1284–1291. IEEE, 2005.

[95] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[96] Heesoo Myeong, Ju Yong Chang, and Kyoung Mu Lee. Learning object relationships via graph-based context model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2727–2734. IEEE, 2012.

[97] Byung-soo Kim, Min Sun, Pushmeet Kohli, and Silvio Savarese. Relating things and stuff by high-order potential modeling. In *European Conference on Computer Vision (ECCV)*, pages 293–304. Springer, 2012.

[98] Sanjiv Kumar and Martial Hebert. A hierarchical field framework for unified context-based classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1284–1291. IEEE, 2005.

[99] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-occurrence, location and appearance. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[100] Derek Hoiem, Alexei A Efros, and Martial Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[101] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *International Conference on Computer Vision (ICCV)*, pages 643–650. IEEE, 2011.

[102] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003.

[103] Kevin Murphy, Antonio Torralba, William Freeman, et al. Using the forest to see the trees: a graphical model relating features, objects and scenes. *Advances in Neural Information Processing Systems (NIPS)*, 16:1499–1506, 2003.

[104] Irving Biederman. *On the semantics of a glance at a scene*. 1981.

[105] David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383, 1977.

[106] Gabriela Csurka and Florent Perronnin. A simple high performance approach to semantic segmentation. In *British Machine Vision Conference (BMVC)*, pages 1–10, 2008.

[107] Josep M Gonfaus, Xavier Boix, Joost Van de Weijer, Andrew D Bagdanov, Joan Serrat, and Jordi Gonzalez. Harmony potentials for joint classification and segmen-

tation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3280–3287. IEEE, 2010.

[108] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 129–136. IEEE, 2010.

[109] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531. IEEE, 2005.

[110] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE, 2006.

[111] Tien-Vu Nguyen, Nghia Pham, Trung Tran, and Bac Le. Higher order conditional random field for multi-label interactive image segmentation. In *Proceedings of the International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, pages 1–4. IEEE, 2012.

[112] Bryan C Russell, William T Freeman, Alexei A Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1605–1614. IEEE, 2006.

[113] Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. Pylon model for semantic segmentation. In *Advances in neural information processing systems (NIPS)*, pages 1485–1493, 2011.

[114] Vibhav Vineet, Jonathan Warrell, and Philip HS Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *International Journal of Computer Vision*, 110(3):290–307, 2014.

[115] Pradeep Ravikumar and John Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. In *Proceedings of the International Conference on Machine learning (ICML)*, pages 737–744. ACM, 2006.

[116] Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627, 2013.

[117] Josef Kittler and John Illingworth. Relaxation labelling algorithmsa review. *Image and Vision Computing*, 3(4):206–216, 1985.

[118] Jay M Tenenbaum and Harry G. Barrow. Experiments in interpretation-guided segmentation. *Artificial Intelligence*, 8(3):241–274, 1977.

[119] Charles Sutton and Andrew McCallum. Piecewise training for undirected models. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, pages 568–575. AUAI Press, 2005.

[120] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[121] Roman Klinger and Katrin Tomanek. *Classical probabilistic models and conditional random fields*. TU, Algorithm Engineering, 2007.

[122] Mahesh K Varanasi and Behnaam Aazhang. Parametric generalized gaussian density estimation. *The Journal of the Acoustical Society of America*, 86(4):1404–1415, 1989.

[123] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, Series B (methodological)*, pages 1–38, 1977.

[124] Jonathan S Yedidia, William T Freeman, Yair Weiss, et al. Generalized belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 689–695, 2000.

[125] Charles Elkan. Log-linear models and conditional random fields. *Tutorial notes at CIKM*, 8, 2008.

[126] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[127] Jan-Mark Geusebroek, Arnold WM Smeulders, and Joost Van de Weijer. Fast anisotropic gauss filtering. *IEEE Transactions on Image Processing*, 12(8):938–943, 2003.

[128] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 233–240. ACM, 2006.

[129] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[130] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, volume 29, pages 753–762. Wiley Online Library, 2010.

[131] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[132] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *International journal of computer vision*, 43(1):7–27, 2001.

[133] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Joint calibration for semantic segmentation. *arXiv preprint arXiv:1507.01581*, 2015.

[134] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 73–80. IEEE, 2010.

[135] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–740. IEEE, 2012.

[136] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Tell me what you see and i will show you where it is. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3197, 2014.