Institut für Geodäsie und Geoinformation
Bereich Photogrammetrie

Hierarchical and Spatial Structures for Interpreting Images of Man-made Scenes Using Graphical Models

Inaugural-Dissertation

zur

Erlangung des Grades

Doktor-Ingenieur

(Dr.-Ing.)

der

Hohen Landwirtschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

vorgelegt am 12. Oktober 2011 von

Michael Ying Yang

aus Linhai, China

Referent:	Prof. Dr.-Ing. Dr. h. c. mult. Wolfgang Förstner
Korreferent:	Prof. Dr. rer. nat. Lutz Plümer
	Prof. Dr. Stefan Wrobel
Tag der mündlichen Prüfung:	16. 12. 2011
Erscheinungsjahr:	2012
	Diese Dissertation ist auf dem Hochschulschriften-server der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

Zusammenfassung

Hierarchische und räumliche Strukturen zur Interpretation von Bildern anthropogener Szenen unter Nutzung graphischer Modelle

Ziel der semantischen Bildinterpretation ist es, Bildregionen und ihre gegenseitigen Beziehungen zu kennzeichnen und in sinnvolle Klassen einzuteilen. Dies ist eine der Hauptaufgabe in vielen Bereichen des maschinellen Sehens, wie zum Beispiel der Objekterkennung, 3D Rekonstruktion oder der Wahrnehmung von Robotern. Insbesondere Bilder anthropogener Szenen, wie z.B. Fassadenaufnahmen, sind durch starke räumliche und hierarchische Strukturen gekennzeichnet. Diese Strukturen zu modellieren ist zentrale Teil der Interpretation, für deren statistische Modellierung graphische Modelle ein geeignetes konsistentes Werkzeug darstellen. Bayes Netze und Zufallsfelder sind zwei bekannte und häufig genutzte Beispiele für graphische Modelle zur Erfassung kontextabhängiger Informationen. Die Motivation dieser Arbeit liegt in der Überzeugung, dass wir eine generische Formulierung der Bildinterpretation mit klarer semantischer Bedeutung finden können, die die Vorteile von Bayes Netzen und Zufallsfeldern verbindet.

Der Hauptbeitrag der vorliegenden Arbeit liegt daher in der Entwicklung eines generischen statistischen graphischen Modells zur Bildinterpretation, welches unterschiedlichste Typen von Bildmerkmalen und die räumlichen sowie hierarchischen Strukturinformationen über eine multiskalen Bildsegmentierung integriert. Das Modell vereinheitlicht die existierender Arbeiten zugrunde liegenden Ideen, wie bedingter Zufallsfelder (conditional random field (CRF)) und Bayesnetze (Bayesian network (BN)). Dieses Modell hat eine klare statistische Interpretation als Maximum *a posteriori* (MAP) Schätzer eines mehrklassen Zuordnungsproblems. Gegeben die Struktur des graphischen Modells und den dadurch definierten Faktorisierungseigenschaften leiten wir die Wahrscheinlichkeitsverteilung des Modells ab. Dies führt zu einer Energiefunktion, die näherungsweise optimiert werden kann. Der jeweilige Typ der Bildmerkmale, die räumliche sowie hierarchische Struktur ist von dieser Formulierung unabhängig.

Wir zeigen die Anwendung des vorgeschlagenen graphischen Modells anhand der mehrklassen Zuordnung von Bildregionen in Fassadenaufnahmen. Wir demonstrieren, dass das vorgeschlagene Verfahren zur Bildinterpretation, durch die Berücksichtigung räumlicher sowie hierarchischer Strukturen, signifikant bessere Klassifikationsergebnisse zeigt, als klassische lokale Klassifikationsverfahren. Die Leistungsfähigkeit des vorgeschlagenen Verfahrens wird anhand eines öffentlich verfügbarer Datensatzes evaluiert. Zur Klassifikation der Bildregionen nutzen wir ein Verfahren basierend auf einem effizienten Random Forest Klassifikator. Aus dem vorgeschlagenen allgemeinen graphischen Modell werden konkret zwei spezielle Modelle abgeleitet, ein hierarchisches bedingtes Zufallsfeld (hierarchical CRF) sowie ein hierarchisches gemischtes graphisches Modell. Wir zeigen, dass beide Modelle bessere Klassifikationsergebnisse erzeugen als die zugrunde liegenden lokalen Klassifikatoren oder die einfachen bedingten Zufallsfelder.

Abstract

Hierarchical and Spatial Structures for Interpreting Images of Man-made Scenes Using Graphical Models

The task of semantic scene interpretation is to label the regions of an image and their relations into meaningful classes. Such task is a key ingredient to many computer vision applications, including object recognition, 3D reconstruction and robotic perception. It is challenging partially due to the ambiguities inherent to the image data. The images of man-made scenes, e. g. the building facade images, exhibit strong contextual dependencies in the form of the spatial and hierarchical structures. Modelling these structures is central for such interpretation task. Graphical models provide a consistent framework for the statistical modelling. Bayesian networks and random fields are two popular types of the graphical models, which are frequently used for capturing such contextual information. The motivation for our work comes from the belief that we can find a generic formulation for scene interpretation that having both the benefits from random fields and Bayesian networks. It should have clear semantic interpretability.

Therefore our key contribution is the development of a generic statistical graphical model for scene interpretation, which seamlessly integrates different types of the image features, and the spatial structural information and the hierarchical structural information defined over the multi-scale image segmentation. It unifies the ideas of existing approaches, e. g. conditional random field (CRF) and Bayesian network (BN), which has a clear statistical interpretation as the maximum *a posteriori* (MAP) estimate of a multi-class labelling problem. Given the graphical model structure, we derive the probability distribution of the model based on the factorization property implied in the model structure. The statistical model leads to an energy function that can be optimized approximately by either loopy belief propagation or graph cut based move making algorithm. The particular type of the features, the spatial structure, and the hierarchical structure however is not prescribed.

In the experiments, we concentrate on terrestrial man-made scenes as a specifically difficult problem. We demonstrate the application of the proposed graphical model on the task of multi-class classification of building facade image regions. The framework for scene interpretation allows for significantly better classification results than the standard classical local classification approach on man-made scenes by incorporating the spatial and hierarchical structures. We investigate the performance of the algorithms on a public dataset to show the relative importance of the information from the spatial structure and the hierarchical structure. As a baseline for the region classification, we use an efficient randomized decision forest classifier. Two specific models are derived from the proposed graphical model, namely the hierarchical CRF and the hierarchical mixed graphical model. We show that these two models produce better classification results than both the baseline region classifier and the flat CRF.

To
my parents & my wife

Acknowledgements

This dissertation would not have been possible without the help and encouragement of a number of people. I start by thanking my advisor, Prof. Wolfgang Förstner, who made it possible for me to come to Germany to pursue my PhD studies. His passion for research is infectious, and has helped me immensely in my research. I cannot thank him enough for his time and support. It was the best decision in my life to join his research group. I am also grateful to Prof. Lutz Plümer for agreeing to review my work and for his continuing support. I also thank him for his help and encouragement during the joint project under the Sino-German bundle. I thank Prof. Stefan Wrobel for agreeing to review my work.

A special thanks goes to my colleagues at the Department of Photogrammetry for all the productive discussions. The open work atmosphere was one of the reasons that made this thesis a success. I thank Susanne Wenzel for translating the German version of Abstract, and Heidi Hollander for checking English spelling. I have enjoyed collaborating with Martin Drauschke, Filip Korč, Falko Schindler, Jan Siegemund and Ribana Roscher. I thank them for the many enlightening discussions we have had in the last few years. I would also like to thank Lutz Plümer, Helmut Mayer, Liqiu Meng, Sven Behnke, Uwe Stilla, Christian Heipke, Olaf Hellwich, Claus Brenner, Monika Sester, Yanpeng Cao, Liangpei Zhang, Xianfeng Huang, Fan Zhang, Huijing Zhao and many others for conversations which have influenced my research.

My stay in Bonn was made pleasurable by numerous friends and colleagues who I would like to thank for their company. These include Barbara Förstner, Heidi Hollander, Lihua Li, Susanne Wenzel, Filip Korč, Timo Dickscheid, Richard Steffen, Jörg Schmittwilken, Thomas Läbe, Birgit Klein, Monika Tüttenberg, Elke Grub, Udo Grub and others. Most important of all, I thank my wife Dandan Chai. Her love, encouragement and tolerance have made this work possible. Finally, I am indebted to my parents who have supported me in all my endeavours.

Contents

List of Figures

List of Tables

LIST OF TABLES

Chapter 1

Introduction

Everything you can imagine is real.

-*Pablo Picasso* (1881 - 1973)

1.1 Motivation

The problem of scene interpretation in terms of classifying various image components, say pixels, regions, or objects, in the images is a challenging task partially due to the ambiguities in the appearance of the image data (Tsotsos, 1988). These ambiguities may arise either due to the physical conditions such as the illumination and the pose of the scene components with respect to the camera, or due to the intrinsic nature of the data itself. Images of man-made scenes, e. g. building facade images, exhibit strong contextual dependencies in the form of spatial interactions among the components. Neighbouring pixels tend to have similar class labels, and different regions appear in restricted spatial configurations. Modelling these spatial structures is crucial to achieve good classification accuracy, and help alleviate the ambiguities. For example, as shown in Fig. 1.1 on page 2, one region from a chimney may locally appear very similar to another region from a building facade. With the help of neighbouring spatial context, it is more likely that the object between the roof and the sky is a chimney.

Graphical models, either directed models or undirected models, provide consistent frameworks for the statistical modelling. Two types of graphical models are frequently used for capturing such contextual information, i. e. Bayesian networks (BNs) (Sarkar & Boyer, 1993) and random fields (RFs) (Besag, 1974), corresponding to directed and undirected graphs. RFs mainly capture the mutually dependent relationships such as the spatial correlation. Attempts were made to exploit the spatial structure for semantic image interpretation by using RFs. Early since nineties, Markov random fields (MRFs) have been used for image interpretation (Modestino & Zhang, 1992); the limiting factor that MRFs only allow for local features has been overcome by conditional random

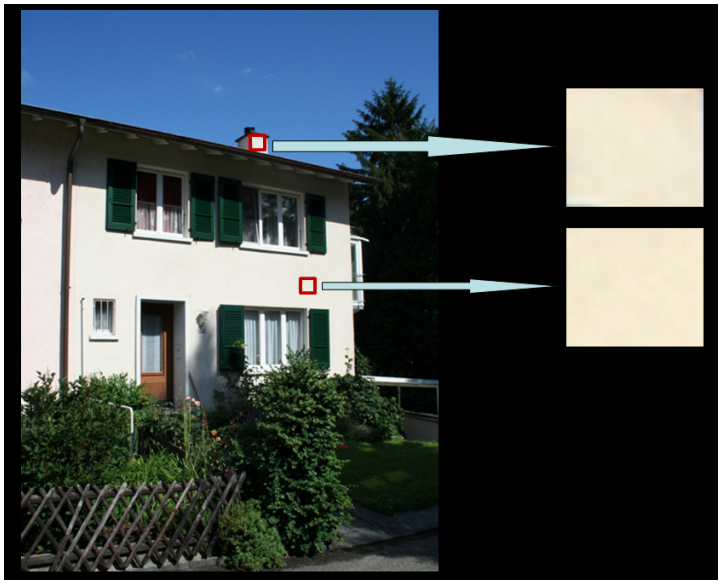


Figure 1.1: Classification of image regions is difficult due to the ambiguities in their appearance. The chimney region (upper red square patch) and the facade region (lower red square patch) look very similar. Neighbouring spatial context, such as the object between the roof and the sky more likely to be a chimney region than a building region, can help resolve these ambiguities. (Best view in colour.)

fields (CRFs) (Lafferty *et al.*, 2001; Kumar & Hebert, 2003a), where arbitrary features can be used for classification, at the expense of a purely discriminative approach. On the other side, BNs usually model the causal relationships among random variables. Early in nineties, Sarkar & Boyer (1993) have proposed the perceptual inference network with the formalism based on Bayesian networks for geometric knowledge-base representation. Both have been used to solve computer vision problems, yet they have their own limitations in representing the relationships between random variables. BNs are not suitable to represent symmetric relationships that mutually relate random variables. RFs are natural methods to model symmetric relationships, though not restricted to symmetric relations (cf. Korč 2011), but they are not suitable to model causal or part-of relationships.

Furthermore, for the real world vision problems, there are often complex relationships among the image entities. Fig. 1.2 on page 3 shows a synthetic example of image classification to illustrate this situation. Two layers are connected via overlap of the regions from the multi-scale segmentation. The hierarchical part-of relations can be captured by the directed edges. In the meantime, neighbouring region relationships representing the interactions between the spatial regions, can be captured by the undirected edges. Capturing and exploiting these spatial and hierarchical relationships are very important in solving some difficult computer vision problems. The aim of the thesis is to develop a consistent graphical model framework, which generalizes RFs and BNs, and apply this framework to scene interpretation to demonstrate its potential.

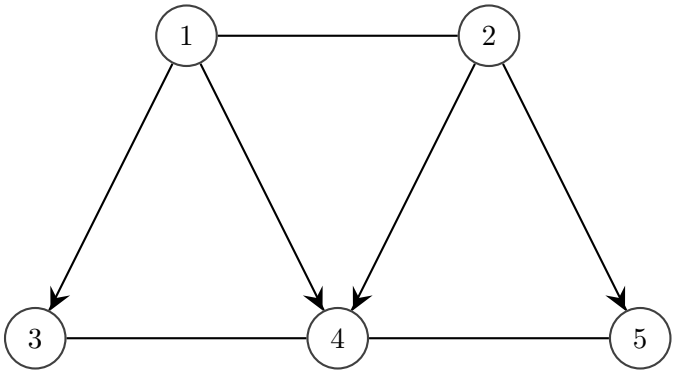


Figure 1.2: A synthetic example of image classification to illustrate the complex relationships among the image entities. Each number represents one image region. The spatial neighbouring region relationships are modelled by the undirected edges, while the hierarchical part-of relations are modelled by the directed edges.

1.2 Goal and achievements of the thesis

The goal of this work is to perform the semantic scene interpretation task, which is to label regions of an image and their relations into meaningful classes. Such task is a key ingredient to many computer vision applications, including object recognition, 3D reconstruction and robotic perception. The key achievement is a sound consistent probabilistic graphical model framework for the classification problem, which unifies conditional random fields and Bayesian networks by incorporating the spatial structure and the hierarchical structure. The key idea for integrating the spatial and the hierarchical structural information into the interpretation process is to combine them with the low-level region class probabilities in a classification process by constructing the graphical model on the multi-scale image regions.

1.3 Application domain

Applications of graphical models are numerous, including information extraction, speech recognition, computer vision, medical disease diagnosis, and protein structure classification. Although our method is applicable to each of these problems, we will focus on semantic scene interpretation, where the goal is the interpretation of the scene contained in an image as a collection of meaningful regions. As a specifically difficult problem, we direct our attention to terrestrial man-made scenes, i. e. building facade images. Building facades may appear as a narrow domain, yet facades comprise a multitude of object structures in terms of varying configurations of storeys, window arrays, balconies, entrance ensembles, and simultaneously a multitude of object appearances. Fig. 1.3 on page 4 shows a selection of some facades with moderate variability. There are single windows, but simultaneously window arrays, balcony windows and entrance windows. Windows constitute more than 50% of all facade objects but are almost in-

Figure 1.3: Some example images of terrestrial man-made scenes: a selection of some building facade images. From these images, we see the facades comprise a multitude of object structures in terms of varying configurations of window arrays, entrance ensembles, and simultaneously a multitude of object appearances.

conclusive regarding possible aggregates of which they might be a part. The structural variability has the natural consequence for probabilistic models. Both, the probabilities for the existence of aggregates given certain parts, and the probabilities for particular spatial relations between parts are not very decisive.

1.4 Challenges in image interpretation

In this section, we highlight the challenge issues that image interpretation faces.

Many satisfactory studies on image interpretation have been presented since the nineties (Modestino & Zhang, 1992; Kumar & Hebert, 2003a; Dick *et al.*, 2004), yet it remains an unsolved problem, because possibly it is one of the most challenging and ambitious problems in computer vision. Humans are able to recognize a tree even if it is far away from a building, or if it is very close to a building. The same tree has different appearances depending on the season of the year: it has no leaves in winter, brown leaves in autumn, green leaves in spring etc., which humans can recognize in all these situations. Humans can recognize and interpret objects in many different scenes, but for machines this is far from an easy task. Here are the major aspects we have to take into account to perform an image interpretation task.

Illumination change in the images is critical for image interpretation. For example, if

Figure 1.4: Illumination challenge: three building scenes affected by different illumination conditions. *Left*: a snowy day scene. *Middle*: a cloudy day scene. *Right*: a night scene.

Figure 1.5: Intra-class & Inter-class variation problem. *Left*: different windows present high intra-class variation, and there are windows with different sizes, windows with rolling shutter. *Right*: the pavement looks very similar to the road on the ground level, and there is no clear border between road and pavement.

we look at Fig. 1.4, we can recognize three building scenes even though the illumination in all images is rather different. So we have to consider that it must also be able to recognize objects and scenes under different illumination conditions.

Intra-class variability is also one reason. Identifying instances of general scene classes is an extremely difficult problem, partly because of the variations among instances of many common object classes, many of which do not afford precise definitions. For example, a window can appear in different positions, in different shapes, with or without rolling shutter, as shown in Fig. 1.5 *Left*. This means we need an approach that can generalize across all possible instances of a certain class.

Inter-class variability within the model is another major difficulty. We do not want to confuse between scenes of different classes that are quite similar. For example, the pavement and road are not labelled as the same class and we can see in Fig. 1.5 *Right* that would easily be confused.

Figure 1.6: Appearance variation problem. *Left*: flowers in front of windows as decorative objects. *Middle*: tree branches occluding the building and the sky. *Right*: windows reflecting tree branches, which are not even seen in this image.

Variability of appearances also exists in most of the vision tasks. For the scene interpretation task, the following three appearance variation problems exist extensively: decorative objects, occluded objects, and reflective objects. Three examples are given in Fig. 1.6.

Scale invariance is also important to take into account for the scene interpretation problem. We can have images with a balcony in front of us, or images with a balcony far away and in both cases it is a balcony class that the system must classify. We can also have some objects (e. g. a building) which appear at different scales in the images.

Furthermore, for the scene interpretation task there are other factors related to the human perception on which we would like to comment: the ambiguities and the subjectivity of the viewer. The obtainable classification accuracies depend strongly on the consistency and accuracy of the manual annotations, and sometimes annotation ambiguities are unavoidable.

Apart from the above mentioned problems, different approaches (Feng *et al.*, 2002; Kumar & Hebert, 2003a; Mortensen & Jia, 2006; Toyoda & Hasegawa, 2008) have been developed for capturing the probabilistic nature of structural information. In one class of approaches, the spatial structures of man-made scenes are modelled by means of Markov random fields and conditional random fields. In another class of approaches, the probabilistic structures of aggregates are modelled by Bayesian networks. Providing an unified probabilistic framework integrating both random fields and Bayesian networks will be a key challenge.

We try to address and resolve these challenges using a generic graphical model framework, by exploiting spatial and hierarchical structures in the images.

1.5 Outline

This thesis is organized as follows:

Previous work In Chapter 2, we start by introducing some previous work on interpreting images of man-made scenes, and work on the approaches for facade inter-

pretation. Then, we review some classification methods based on Bayesian networks, Markov random fields, and conditional random fields. At the end, we discuss some techniques concerning integration of random fields and Bayesian networks. The review will show the strengths and weaknesses of previous attempts to solve the interpretation problem.

Theoretical basis In Chapter 3, we present a theoretical basis needed for this thesis. First, we survey some of the basic notations in graph theory. Then, we introduce two graphical frameworks for representing probability distributions, i. e. Bayesian networks and random fields, corresponding to directed and undirected graphs. In addition, we introduce two approaches to build relations between them: a moral graph, which converts a directed graph to an undirected graph; a factor graph, which could represent both directed and undirected graphical models.

A generic framework for image interpretation of man-made scenes In Chapter 4, we develop a generic graphical model framework for scene interpretation that includes both information about the spatial structure and the hierarchical structure. We start by constructing the graphical model. The graphical model could consist of either the directed edges or the undirected edges. We can parametrize the directed edges by conditional probabilities, and the undirected edges by potential functions. Then, the statistical model is formulated as a multi-class labelling problem, where we derive the corresponding energy function. We compare our model with the previous models and show that at certain choices of the parameters of our model, these methods fall out as special cases. We also derive particular models for the energy potentials and the conditional probability energy that are suited well for scene interpretation. We derive the features from each region obtained from the unsupervised segmentation algorithm, and employ a classifier to calculate the label distribution for the local unary potential. We give one particular formulation for each of the pairwise potentials and the conditional probability energy. Finally, we discuss the learning and the inference issues of this graphical model.

Experimental results In Chapter 5, we present a number of experimental results that characterize the performance of the proposed model, and demonstrate the application of the proposed model on building facade image classification.

Conclusion and future work In Chapter 6, we give the concluding remarks and discuss the limitations and some potential future directions.

1.6 Notation

A list of frequently used mathematical symbols is given in Table 1.1. It covers the major part of symbols occurring in this thesis.

With a few exceptions, we will denote sets by calligraphic uppercase letters, vectors by bold lowercase letters, and matrices by bold uppercase letters. Elements of a set are either represented by their index, or the same letter as the set itself and carry their index as a lower right subscript. The first element in a set has index 1. For example, the set $\mathcal{V}$ representing a set of nodes in a graph is $\{1, \cdots, i, \cdots, n\}$.

Finally, we denote the discrete probability of a random variable $\underline{\boldsymbol{x}}$ by $P(\underline{\boldsymbol{x}} = \boldsymbol{x})$, abbreviated as $P(\boldsymbol{x})$.

Table 1.1: List of mathematical symbols and notation.

symbol	meaning
$\mathcal{G}$	graph
$\mathcal{V}$	set of nodes
$\mathcal{A}$	set of directed edges
$\mathcal{E}$	set of undirected edges
$\mathcal{D}$	directed graph
$\mathcal{H}$	undirected graph
Pa_i	parents of the node i
Ch_i	children of the node i
N_i	neighbours of the node i
$\mathcal{N}$	neighbourhood system of the random field
(i, j)	node j is the child of node i and i is the parent of j
$\{i, j\}$	nodes i, j are neighbours
$\langle i, j \rangle$	nodes i, j are adjacent
$\underline{\boldsymbol{x}}_1, \cdots, \underline{\boldsymbol{x}}_n$	random variables (vectors)
$\{\underline{\boldsymbol{x}}_i, i \in \mathcal{V}\}$	a set of variables, defined over a graph
$\underline{\boldsymbol{x}}$	compound random vector containing all the random vectors
$\mathrm{Pa}(\underline{\boldsymbol{x}}_i)$	the random variable, associated with the parent of the node i
$E(\cdot)$	Gibbs energy function
$\boldsymbol{c}$	clique
$\mathcal{C}$	the set of cliques
$\phi(\cdot)$	potential function
Z	partition function (normalization constant)
$\mathcal{F}$	factor graph
$f_s(\cdot)$	a factor function
$\boldsymbol{h}$	feature sets

Chapter 2

Previous Work

The stones of those hills, May be made into grind-stones.

The stones of those hills, May be used to polish gems.

-He Ming, Minor odes of the kingdom
The Book of Odes (1100 B.C. - 600 B.C.)

In this chapter we will review the most recent and significant work in the fields of image interpretation of man-made scenes, Markov random fields, conditional random fields, and Bayesian networks. The review will show the strengths and weaknesses of previous attempts to solve the interpretation problem. We start by introducing some previous work on interpreting images of man-made scenes, and work on the approaches for facade interpretation. Then, we review some classification methods based on Bayesian networks, Markov random fields, and conditional random fields. At the end of this chapter, we discuss some work concerning integration of random fields and Bayesian networks.

2.1 Interpreting images of man-made scenes

Automatic interpretation of man-made scenes and particularly building facades has been a consistent interest early since eighties. As an often cited early approach for the extraction of buildings, Herman & Kanade (1984) uses AI-focused 3D-reasoning and heuristics about the vertical and horizontal directions of lines to extract buildings as rectangular prisms. Comprehensive study and comparison of automatic building extraction can be found in Mayer (1999).

Early attempts to 3D city modelling are based on sets of prototypes or parametrized geometrical models (Fischer *et al.*, 1997) with the possibility of aggregation (Fischer *et al.*, 1999), on the restriction to roof structures (Brenner *et al.*, 2001) made possible by using the ground planes of the buildings from a 2D GIS. Practical approaches

are clearly interactive, e. g. InJect (Gülch *et al.*, 1998), CyberCityModeler (Gruen & Wang, 1999), with some support by automatic procedures. Modelling the architecture of complete building blocks by using generative models (Dick *et al.*, 2004) pushes theoretical research onto a new level. Dick *et al.* (2004) describe the automatic acquisition of 3D architectural models for reconstruction from images, which introduces reversible jump Markov Chain Monte Carlo (MCMC) techniques for estimation. A building is described as a set of walls together with a 'Lego' kit of parameterised primitives, such as doors or windows. A prior on wall layout, and a prior on the parameters of each primitive are defined. Part of this prior is learned from training data and part comes from expert architects. Their model, however, only consists of walls and primitives. Mayer & Reznik (2006, 2007) use image data. They get special information using implicit shape models by means of MCMC and plane sweeping for the reconstruction of windows in a building facade. But, MCMC based techniques are quite slow for convergence in general. Frahm *et al.* (2010) present a system approaching fully automatic 3D modelling of large-scale environments. The system achieves high computational performance through algorithmic optimizations for efficient robust estimation, the use of image-based recognition for efficient grouping of similar images, and two-stage stereo estimation for video streams that reduces the computational cost while maintaining competitive modelling results. All the aforementioned approaches only exploit a coarse scale of level of detail (LOD) in building modelling. They fall into geometric modelling category, not semantic modelling. In the similar spirit of the methods discussed above, but being closer to ours, there is a work of Micusik & Kosecka (2010), which presents an approach utilizing properties of piecewise planarity and restricted number of plane orientations to suppress reconstruction and matching ambiguities. The problem of the 3D reconstruction is formulated as an MRF framework. Similar to our work where we choose image regions as an image representation, they choose superpixels as an image representation. Our work, focusing on semantic image classification, could be an important pre-step for 3D city modelling, where the resulting 3D model has semantic meanings for each element.

Facade classification is an important subtask for scene interpretation and automatically building large 3D city models. Despite the substantial improvements during the past decade, the classification of building facade images remains a challenging problem, which receives a great deal of attention in the photogrammetry community (Rottensteiner *et al.*, 2007; Korč & Förstner, 2008; Micusik & Kosecka, 2009; Fröhlich *et al.*, 2010; Kluckner & Bischof, 2010; Teboul *et al.*, 2010). Micusik & Kosecka (2009) present an approach for image semantic segmentation of street scenes into coherent regions. They introduce an explicit model of spatial co-occurrence of visual words associated with superpixels and utilization of appearance, geometry and contextual cues in a probabilistic framework yielding a second-order MRF with unary and binary functions. The weighting parameters of the unary and binary terms are set manually, while in our setting, these parameters are learned from training images automatically. They use image sequences and employ 3D geometric information from Structure-from-Motion estimation to improve the recognition accuracy. In our experiments, we only have single

images, no image sequences. Multi-class facade segmentation by combining a machine learning approach with procedural modelling as a shape prior is presented by Teboul *et al.* (2010). Generic shape grammars are constrained so as to express buildings only. Randomized forests are used to determine a relationship between the semantic elements of the grammar and the observed image support. Fröhlich *et al.* (2010) also show a pixelwise labelling method of facade images using an efficient randomized decision forest classifier and the robust local opponent-SIFT features (van de Sande *et al.*, 2010). Both Teboul *et al.* (2010) and Fröhlich *et al.* (2010) show that a randomized decision forest is a good local classifier for image classification, therefore, we also employ a randomized decision forest as the local classifier for our graphical model. However, Fröhlich *et al.* (2010) only exploit local features, no spatial neighbourhood information is considered. While Teboul *et al.* (2010) use shape grammars to impose global constraints, the grammars lack flexibility compared to the pairwise potential functions in Markov random fields. Drauschke & Mayer (2010) evaluate the potential of seven texture filter banks for the pixel-based classification of terrestrial facade images. They provide some useful features for our scene interpretation task.

In recent years, mobile mapping systems increasingly provide terrestrial data, which changes the focus on facades. Due to their specific structure models based on grammatical rules have been developed, exploiting the long tradition in natural language understanding. Stochastic attribute grammars (Abney, 1997) have evolved and today appear as generalizations of Markov random fields and Bayesian networks, cf. (Liang *et al.*, 2009). Müller *et al.* (2006) introduce split grammars in order to model the structure of 2D facades and 3D buildings by irregular tessellations and hierarchical volumetric models. Becker (2009) adapts and extends this approach for the reconstruction of facades from terrestrial images and 3D point clouds, and learns context-free production rules. Ripperda & Brenner (2009) use formal grammars and a reversible jump Markov chain Monte Carlo approach to estimate the building model parameters. Integrating graphical models and the grammar is an ongoing research direction. Liang *et al.* (2009) present a nonparametric Bayesian generalization of the probabilistic context-free grammars based on the hierarchical Dirichlet process. Schmittwilken *et al.* (2009) propose a concept for integration of low- and high- level reasoning for the interpretation of images of man-made objects including a one-layer-graphical model for mid level reasoning integrated with a stochastic grammar for simple aggregates of facade objects. A single image reconstruction of building scenes is promised in Koutsourakis *et al.* (2009). The authors use a special shape grammar which translates to a tree-based MRF. For the work of this thesis, we will not address the problem of integrating graphical models and the grammar. We put this as a future work.

Many man-made and natural structures consist of similar elements arranged in regular patterns. Hartz & Neumann (2007) show that ontological concept descriptions for spatially related objects and aggregates can be learned from positive and negative examples. Using examples from the buildings domain, the authors show that learned aggregate concepts for window arrays, balconies and other structures can be successfully applied to discover repetitive patterns of objects. Hartz *et al.* (2009) introduce an

automatic way of incremental model learning for the interpretation of complex scenes by using annotated examples. The authors present a learning, interpretation, and evaluation cycle to deal with repetitive patterns of objects. Spinello *et al.* (2010) present an unsupervised approach for discovering and reasoning on repetitive patterns of objects in a single image. CRFs are used as a formalism to predict the location of elements at places where they are partially occluded or detected with very low confidence. Wu *et al.* (2010) present a robust framework to analyse large repetitive structures in urban scenes, which finds the salient boundaries of the repeating elements even when the repetition exists along only one direction. Wendel *et al.* (2010) introduce an approach for segmenting individual facades from streetside images, which incorporates prior knowledge about arbitrarily shaped repetitive regions. These repetitive regions are detected using intensity profile descriptors and a voting-based matcher. In Yang *et al.* (2010b); Yang *et al.* (2011), the authors present a general scheme for automatically aligning two widely separated 3D scenes via the use of the viewpoint invariant features. The viewpoint invariant features provide robust local feature information including patch scale and dominant orientation for effective repetitive structure matching in man-made environments. Our work focus on probabilistic graphical modelling. So, we do not have to deal with repetitive structures in the scene. However, if repetitive structures are detected (e. g. a window detector (Wenzel & Förstner, 2008)) and serve as priors, better classification results will surely be achieved.

The cited works, which are far from complete, show the progress regarding the particular methods which contribute to the overall problem of interpreting man-made scenes. For a long time, the difficulty of interpreting man-made scenes has been underestimated. The main reason is the high variability of man-made structures and their appearance, and the resulting complexity of the acquired data. In this thesis, we try to address these challenges by exploiting spatial and hierarchical structures in the images of man-made scenes. We focus on probabilistic graphical models, e. g. Markov random fields (MRFs) and Bayesian networks (BNs), which can be employed for modelling the spatial structures and the partonomies.

2.2 Previous work on Markov and conditional random fields

Markov random fields (MRFs) are the most commonly used undirected graphical models in computer vision, which allow one to incorporate local contextual information in a principled manner. MRFs have been made popular in computer vision by the early work of Besag (1974); Geman & Geman (1984); Besag (1986). Their limiting factor that they only allow for local image features has been overcome by conditional random fields (CRFs) (Lafferty *et al.*, 2001; Kumar & Hebert, 2003a), where arbitrary features can be used for classification, at the expense of a purely discriminative approach. In this section, we review most recent works on MRFs and CRFs that address the spatial neighbourhood relationships, the combination of global and local features, the higher

order potentials, and the hierarchical relationships.

There are many recent works on contextual models that exploit the spatial dependencies between the objects. For this, several authors explore MRFs and CRFs for the probabilistic modelling of local dependencies, e. g. (Modestino & Zhang, 1992; Barnard & Forsyth, 2001; Kumar & Hebert, 2003a; He *et al.*, 2006; Shotton *et al.*, 2006). The goal of these works is to label every pixel in the image with a single class label. Typically, these algorithms construct (conditional) Markov random fields over the pixels with a unary term based on pixel appearance and a pairwise smoothness term to encourage neighboring pixels to take the same label. The works differ in the details of the energy functions and the inference algorithms used. Kumar & Hebert (2003a) present a discriminative conditional random field framework for the classification of image regions by incorporating neighbourhood interactions in the labels as well as the observed data. The advantage of this model is its flexibility in using any type of class relevant observations, especially such which allow to discriminate between classes. This in general leads to much better classification results than achievable with MRFs. The disadvantage is, common with all discriminative models, that incremental learning is at least difficult, if not impossible. Shotton *et al.* (2006) propose an approach for learning a discriminative model of object classes, incorporating texture, layout, and contextual information. Unary classification and feature selection is achieved using a boosting scheme. Image segmentation is achieved by incorporating the unary classifier in a CRF, which captures the spatial interactions between class labels of neighboring pixels. They use an absolute location prior as a feature in their probabilistic construction, which we also adopt this idea. They only use local features, while we use both local and global features in our approaches. Levin & Weiss (2006) propose an approach that learns a CRF to combine bottom-up and top-down cues for class specific object segmentation. A similar purpose serves the harmony potentials, proposed by Gonfaus *et al.* (2010). They impose global shapes as a top-down cue, however, generalizing their binary classification formulation to a multi-class classification task is not straightforward.

A number of CRF models for image interpretation address the combination of global and local features (Brunn & Weidner, 1997; He *et al.*, 2004; Yang *et al.*, 2007; Reynolds & Murphy, 2007; Gould *et al.*, 2008; Toyoda & Hasegawa, 2008; Plath *et al.*, 2009; Schnitzspan *et al.*, 2009). They showed promising results and specifically improved performance compared with making use of only one type of feature - either local or global. He *et al.* (2004) propose a multi-layer CRF to account for global consistency, which shows improved performance. The authors introduce a global scene potential to assert consistency of local regions. Thereby, they are able to benefit from integrating the context of a given scene. This method infers a single scene context and do not allow the discovery of one class to influence the probability of finding others. Yang *et al.* (2007) propose a model that combines appearance over large contiguous regions with spatial information and a global shape prior. The shape prior provides local context for certain types of objects (e. g. cars and airplanes), but not for regions representing general objects (e. g. animals, buildings, sky and grass). Gould *et al.* (2008) propose a method for capturing global information from inter-class spatial relationships and

encoding it as a local feature. Toyoda & Hasegawa (2008) present a proposal of a general framework that explicitly models local and global information in a CRF. Their method resolves local ambiguities from a global perspective using the global image information. It enables locally and globally consistent image recognition. But their model needs to train on the whole training data simultaneously to obtain the global potentials, which results in high computational time.

Besides the above approaches, there are more popular methods to solve multi-class classification problems using higher order conditional random fields (Kohli *et al.*, 2007, 2009; Ladicky *et al.*, 2009). Kohli *et al.* (2007) introduce a class of higher order clique potentials called P^n Potts model. The higher order potential functions proposed in Kohli *et al.* (2009) take the form of the Robust P^n model, which is more general than the P^n Potts model. The higher order potentials, motivated by overcoming the smoothing properties of the CRFs with pairwise potentials, have been used to integrate results from multiple segmentations, to obtain crisper boundaries, and to improve the error due to an incorrect initial segmentation. Ladicky *et al.* (2009) generalize the Robust P^n model to P^n based hierarchical CRF model. Inference in these models can be performed efficiently using graph cut based move making algorithms. However, the work on solving higher order potentials using move making algorithms has targeted particular classes of potential functions. Developing efficient large move making for exact and approximate minimization of general higher order energy functions is a difficult problem. Parameter learning for a higher order CRF is also a challenging problem. Delong *et al.* (2010) propose the use of a soft cost over the number of labels present in an image for clustering. Their work extends α-expansion so that it can simultaneously optimize label costs as well. Ladicky *et al.* (2010) consider a class of global potentials defined over all variables in the CRF model. They add one cue called global object co-occurrence statistics, a measure of which classes (such as chair or motorbike) are likely to occur in the same image together. These approaches for capturing global contextual information about spatial co-occurrence of different class label are meaningful when the number of classes per image and the change of the viewpoint are relatively small as in the MSRC dataset (Shotton *et al.*, 2006). There, the cows typically appear next to grass and below the sky. In the man-made scenes with the larger number of object class appearing in the same image, these types of contextual relationships are no longer so persistent (Micusik & Kosecka, 2009) (cf. Fig. 1.3 on page 4).

The use of multiple different over-segmented images as a preprocessing step is not new to computer vision. For example, Russell *et al.* (2006) use multiple over-segmentations for finding objects in the images, and many of the depth reconstruction methods, e. g. (Hoiem *et al.*, 2007), make use of over-segmentations for computing feature statistics. In the context of multi-class image classification, the work of Plath *et al.* (2009) comprises two aspects for coupling local and global evidences both by constructing a tree-structured CRF on image regions on multiple scales, which largely follows the approach of Reynolds & Murphy (2007), and using global image classification information. Thereby, Plath *et al.* (2009) neglect direct local neighbourhood dependencies. The work of Schnitzspan *et al.* (2008) explicitly attempts to combine

the power of global feature-based approaches with the flexibility of local feature-based methods in one consistent framework. Briefly, Schnitzspan *et al.* (2008) extend classical one-layer CRF to a multi-layer CRF by restricting the pairwise potentials to a regular 4-neighbourhood model and introducing higher-order potentials between different layers. Yang *et al.* (2010a) present a concept of a hierarchical CRF that models region adjacency graph and region hierarchy graph structure of an image. Yang & Förstner (2011b) realize this concept in the application of classifying the images of man-made scenes. Rather than 4-neighbourhood graph model in Schnitzspan *et al.* (2008), Yang *et al.* (2010a); Yang & Förstner (2011b) build region adjacency graph based on unsupervised image segmentation, which leads to a irregular graph structure. Also, they apply an irregular pyramid to represent different layers, while Schnitzspan *et al.* (2008) use a regular pyramid structure. Third, their model only exploits up to second-order cliques, which makes learning and inference much easier.

2.3 Previous work on Bayesian networks

Although not as popular as random fields (MRFs and CRFs), Bayesian networks (BNs) have also been used to solve computer vision problems (Sarkar & Boyer, 1993; Feng *et al.*, 2002; Mortensen & Jia, 2006; Zhang & Ji, 2011). BNs provide a systematic way to model the causal relationships among the entities. By explicitly exploiting the conditional independence relationships (known as prior knowledge) encoded in the structure, BNs could simplify the modelling of joint probability distributions. Based on the BN structure, the joint probability is decomposed into the product of a set of local conditional probabilities, which is much easier to specify because of their semantic meanings (Zhang & Ji, 2010; Zhang *et al.*, 2011).

Early in nineties, Sarkar & Boyer (1993) have proposed the perceptual inference network with the formalism based on BNs for the geometric knowledge-base representation. The network provides a scheme to combine the bottom-up process of recognizing the regular components in the images and the top-down process of inferring the geometric structures from multiple cues and the knowledge of Euclidean geometric structures. This is the first application of BNs to low-level vision. Feng *et al.* (2002) integrates BNs with neural networks for scene segmentation. The BN models the prior distribution of the label fields. Neural networks are used to make local predictions given the pixel features. The predictions can be combined with the prior in a principled manner using the scaled-likelihood method. This model has a fixed structure and good initialization is required for the variational inference approach. Mortensen & Jia (2006) present a semi-automatic segmentation technique called Bayesian cut that formulates object boundary detection as the most probable explanation of a BN's joint probability distribution. A two-layer BN structure is formulated from a planar graph representing a watershed segmentation of an image. The network's prior probabilities encode the confidence that an edge in the planar graph belongs to an object boundary while the conditional probability tables (CPTs) enforce the global contour properties of closure and simplicity. Although these works have successfully applied BN in their specific

problems, most of them only use a simple BN structure (typically a naive BN). For complex problems, these models may not be expressive enough to model many different kinds of image entities and their relationships. How to effectively capture these relationships using a BN is crucial to solving these difficult problems. In Zhang & Ji (2011), the authors propose a BN model for both automatic and interactive image segmentation. A multilayer BN is constructed from an over-segmentation to model the statistical dependencies among regions, edge segments, vertices and their measurements. The BN also incorporates various local constraints to further restrain the relationships among these image entities. Given the BN model and various image measurements, belief propagation is performed to update the probability of each node. Image segmentation is generated by the most probable explanation inference of the true states of both region and edge nodes from the updated BN. Although their model improves segmentation results on the Weizmann horse dataset (Borenstein *et al.*, 2004), they need a lot of domain expert knowledge to design the local constraints. Their BN model is focused on the figure\ground segmentation problem, generalizing to multi-class segmentation faces the difficulty of designing and changing local constraints due to the complex boundaries in a multi-class segmentation.

2.4 Integration of random fields and Bayesian networks

From the last two sections, we see graphical models, underlying undirected and directed graphs, have reached a state where both, hierarchical and spatial neighbourhood structures can be efficiently handled. The concept of factor graphs allows integrating Bayesian networks (BNs) which are efficient for modelling partonomies, and random fields (RFs) which are standard for modelling spatial neighbourhoods in a common Markov field (Zhang & Ji, 2010). RFs and BNs are suitable for representing different types of statistical relationships among the random variables. RFs mainly capture the mutually dependent relationships such as the spatial correlation, while BNs usually model the causal relationships among random variables. Their combination can create a more powerful and flexible probabilistic graphical model. Yet only a few previous works focus on integrating RFs with BNs.

Kumar & Hebert (2003b) present a generative model based approach to man-made structure detection in 2D natural images. They use a causal multiscale random field as a prior model on the class labels. Labels over an image are generated using Markov chains defined over coarse to fine scales. Instead of assuming the conditional independence of the observed data, they propose to capture the local dependencies in the data using a multiscale feature vector. However, the spatial neighbourhood relationships are only considered at the bottom scale. So, essentially, this model is a tree-structured belief network (Feng *et al.*, 2002) plus a flat Markov random field. Kumar *et al.* (2005) propose a combination of an MRF with a layered pictorial structure model for object detection and segmentation. The layered pictorial structure model represents the global shape of the object and restrains the relative location of different parts of the object. They formulate the layered pictorial structure model using a fully connected MRF.

Therefore, the whole model is essentially an extended MRF model. Liu *et al.* (2006) propose an integration of a BN with an MRF for image segmentation. A naive Bayes model is used to transform the image features into a probability map in the image domain. The MRF enforces the spatial relationships of the labels. The use of a naive Bayes model greatly limits the capability of this method because it is hard to model the complex relationships between the random variables using a naive Bayes model.

Hinton *et al.* (2005) present a learning procedure for a chain graphical model that contains both directed and undirected connections. Their model is constructed by connecting several MRFs at different layers using the directed edges. In Hinton *et al.* (2005), they show that combining multiple MRFs into causal hierarchies as a chain graphical model has a major advantage over combining them into one big MRF by using the undirected connections. The causal connections between layers act as insulators that prevent the partition functions of the individual MRF from combining together into one large partition function. This also gives us motivation to build our graphical model. However, compared to Hinton *et al.*'s, our model has two major differences. In their model, the configuration of a top-level MRF provides the biases that influence the configuration of the next level MRF through the directed edges. While, in our model, the directed edges capture the causalities among the image regions and the undirected edges capture the spatial neighbourhood relationships conditioned on the observation. Their model exploits an approximation of the true posterior probability distribution of the hidden nodes by implicitly assuming the posterior of each hidden node is independent of each other. In contrast, we derive the factored probability distribution based on the graphical model structure, and therefore, do not have such an assumption.

Zhang & Ji (2010) propose a unified graphical model that can represent both the causal and noncausal relationships among the random variables and apply it to the image segmentation problem. They first employ a CRF to model the spatial relationships among the image regions and their measurements. Then, they introduce a multilayer BN to model the causal dependencies. The CRF model and the BN model are then combined through the theories of the factor graphs to form a unified probabilistic graphical model. Their graphical model is too complex in general. While the CRF part performs region-based image segmentation, the BN part performs edge-based segmentation, which is constructed to capture the causalities among the regions, edges, vertices (or junctions), and their measurements. The two parts are connected through the region nodes. The region nodes act as the parents of an edge node. The parents of the edge node correspond to the two regions that intersect to form this edge. Although their model improves state of the art results on the Weizmann horse dataset (Borenstein *et al.*, 2004) and the MSRC dataset (Shotton *et al.*, 2006), they need a lot of domain expert knowledge to design the local constraints. Also, they use a combination of supervised parameter learning and manual parameter setting for the model parameterization. Simultaneously learn the BN and CRF parameters automatically from the training data is not a trivial task. In Zhang *et al.* (2011), the authors apply a similar strategy to extend the conventional chain-like chain graphical model to a chain

graphical model with more general topology, which essentially appears to be a restrict version of their unified graphical model in Zhang & Ji (2010). There, they apply an approximate learning approach called the contrastive divergence learning, where the distribution over the n-step reconstruction of the sampled data are generated by n full-step Markov Chain Monte Carlo sampling via Gibbs sampling. This procedure produces better local minimum but rather slow. This kind of parameter learning remains a difficult problem and is also the most time-consuming part (Alahari *et al.*, 2010).

Compared to the graphical models in Kumar & Hebert (2003b) and Liu *et al.* (2006), which are too simple, the graphical models in Zhang & Ji (2010) and Zhang *et al.* (2011) are too complex in general. Our graphical model lies in between (cf. Fig. 4.2 on page 36). We try to construct our graphical model that is not too simple in order to model the rich relationships among the neighbourhood of pixels and image regions in the scene, yet not too complex in order to make parameter learning and probabilistic inference efficiently. Furthermore, our model underlies a clear semantic meaning. If the undirected edges are ignored, meaning no spatial relationships are considered, the graph is a tree representing the hierarchy of the partonomy among the scales. Within each scale, the spatial regions are connected by the pairwise edges.

In this chapter we have surveyed the work in the field of scene interpretation mainly using the graphical models. These models include Markov random fields, conditional random fields, Bayesian networks, and integration of random fields and Bayesian networks. It can be observed that the existing approaches score well in some scenarios. However, performing semantic scene interpretation in general still seems to be very challenging.

Chapter 3

Theoretical Basis

Everything should be made as simple as possible, but not simpler.

-*Albert Einstein* (1879-1955)

3.1 Overview

Graphical models are a marriage between probability theory and graph theory (Jordan, 1998). As a modelling and inference tool, graphical models use intuitive, powerful, and flexible graph structures to represent the probability distributions of the random variables. The graph structures encode the conditional dependency and independency among the random variables. The nodes in the graph are identified with the random variables, the edges linking the nodes represent the statistical relationships between the random variables, and the joint probability distributions are defined as the products over the functions of the connected subsets of the nodes.

In this chapter, we first introduce basic notations in graph theory. We then present two types of graphical models for representing the probability distributions: one with the directed graphs and one with the undirected graphs. Then we discuss the relations between directed and undirected graphical models in terms of the moral graphs and the factor graphs.

3.2 Basic notations in graph theory

In this section we survey some of the basic notations in graph theory used in the thesis. We will briefly describe graph, directed graph, undirected graph, path, trail, and directed acyclic graph (cf. Bang-Jensen & Gutin, 2008; Koller & Friedman, 2009).

Definition 3.1 *Graph. A graph is a structure consisting of a non-empty finite set of the nodes and a set of the edges connecting pairs of the nodes.*

In the following we denote the graph with $\mathcal{G}$. A pair of the nodes can be connected by a *directed edge* or an *undirected edge*. We will often write $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, which means that $\mathcal{V}$, $\mathcal{E}$, and $\mathcal{A}$ are the set of the nodes $\mathcal{V} = \{1, \cdots, i, \cdots, n\}$, the set of the undirected edges $\mathcal{E} = \{\{i, j\} \mid i, j \in \mathcal{V}\}$, and the set of the directed edges $\mathcal{A} = \{(i, j) \mid i, j \in \mathcal{V}\}$, respectively. We denote the directed edge as (i, j) and the undirected edge as $\{i, j\}$. An example of a graph $\mathcal{G}$ with the directed and undirected edges is given in Fig. 3.1.

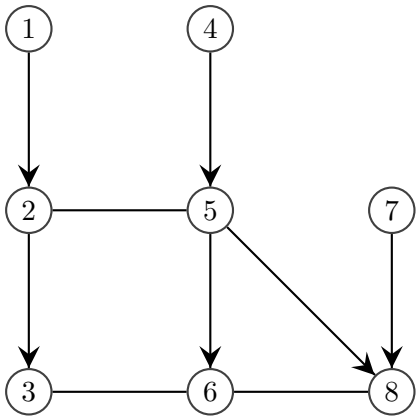


Figure 3.1: An example of a graph $\mathcal{G}$ with the directed and undirected edges.

In many cases, we want to define the graphs that contain only edges of one kind or another.

Definition 3.2 *Directed graph. A graph is directed if all edges are directed.*

Definition 3.3 *Undirected graph. A graph is undirected if all edges are undirected.*

A directed graph means $\mathcal{E} = \varnothing$ in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$. An undirected graph means $\mathcal{A} = \varnothing$ in a graph $\mathcal{G}$. In the following we denote a directed graph with $\mathcal{D} = (\mathcal{V}, \mathcal{A})$, and an undirected graphs with $\mathcal{H} = (\mathcal{V}, \mathcal{E})$. Examples of a directed graph and an undirected graph are given in Fig. 3.2.

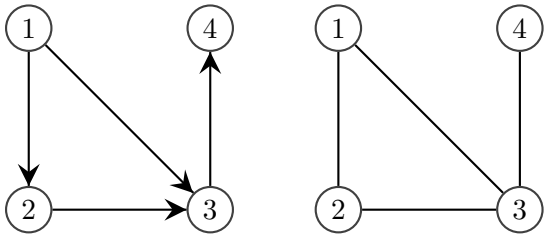


Figure 3.2: Examples of a directed graph $\mathcal{D}$ and an undirected graph $\mathcal{H}$. *Left*: all the edges are directed. *Right*: all the edges are undirected.

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, when we have that (i, j), we say that j is the *child* of i in $\mathcal{G}$, and i is the *parent* of j in $\mathcal{G}$. When we have $\{i, j\}$, we say that i, j are *neighbours*

in $\mathcal{G}$. We say that i, j are adjacent whenever i and j are connected via some edge, whether directed or undirected, denoted as $\langle i, j \rangle$. We use Pa_i to denote the parents of the node i, Ch_i to denote its children, and N_i to denote its neighbours. For example, in Fig. 3.1, node 1 is the only parent of node 2, and node 3 is the child of node 2. The only neighbour of node 2 is node 5, but its adjacent nodes are $1, 3, 5$.

In many cases, we want to consider only the part of the graph that is associated with a particular subsets of nodes. A subgraph is complete if every two nodes in this subgraph are connected by some edge. This kind of set is called a clique.

Using the basic notation of edges, we can define different types of connections in the graph.

Definition 3.4 *Path. We say that $s_1, \cdots, s_k$ form a path in the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, $\mathcal{S} = \{s_1, \cdots, s_k\} \subseteq \mathcal{V}$, if we have that either (s_i, s_{i+1}) or $\{s_i, s_{i+1}\}$, for every $i = 1, \cdots, k-1$. A path is directed if we have (s_i, s_{i+1}), for at least one i.*

Definition 3.5 *Trail. We say that $s_1, \cdots, s_k$ form a trail in the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, $\mathcal{S} = \{s_1, \cdots, s_k\} \subseteq \mathcal{V}$, if s_i, s_{i+1} are adjacent, for every $i = 1, \cdots, k-1$.*

In Fig. 3.1 on page 20, nodes $1, 2, 5, 6, 8$ form a path, and hence also a trail. On the other hand, nodes $1, 2, 3, 6, 5$ form a trail, which does not form a path.

Definition 3.6 *Cycle. A cycle in $\mathcal{G}$ is a directed path $s_1, \cdots, s_k$ where $s_1 = s_k$. A graph is acyclic if it contains no cycles.*

Definition 3.7 *Loop. A loop in $\mathcal{G}$ is a trail $s_1, \cdots, s_k$ where $s_1 = s_k$.*

The graph $\mathcal{G}$ of Fig. 3.1 on page 20 is acyclic. However, if we add the undirected edge $\{1, 5\}$ to $\mathcal{G}$, we have a path $1, 2, 5, 1$ from node 1 to itself. Clearly, adding a directed edge $(5, 1)$ would also lead to a cycle.

Definition 3.8 *DAG: directed acyclic graph. A DAG is a directed graph with no directed cycles.*

DAGs are the basic graphical representation that underlies Bayesian networks (cf. Section 3.3). An example of a DAG is given in Fig. 3.3.

We sometimes convert a graph to an undirected graph by ignoring the directions on the edges (Koller & Friedman, 2009).

Definition 3.9 *Graph's undirected version. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, its undirected version is a graph $\mathcal{H} = (\mathcal{V}, \mathcal{E}')$, where every directed edge is replaced by an undirected edge.*

Undirected version $\mathcal{H}$ of $\mathcal{G}$ in Fig. 3.1 on page 20 is given by Fig. 3.4.

The different types of graphs used in this thesis and their characteristic property are listed in Table 3.1 on page 22. We see the following relations among these different graphs: DAG $\subseteq \mathcal{D} \subseteq \mathcal{G}$ and $\mathcal{H} \subseteq \mathcal{G}$.

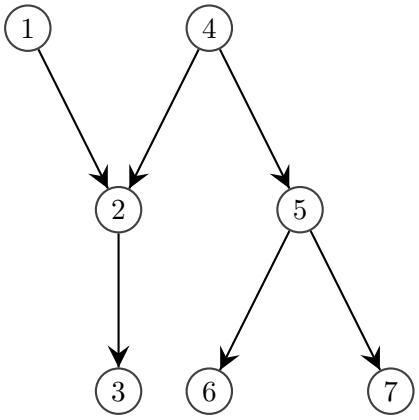


Figure 3.3: An example of a DAG. There is no directed cycle in this graph.

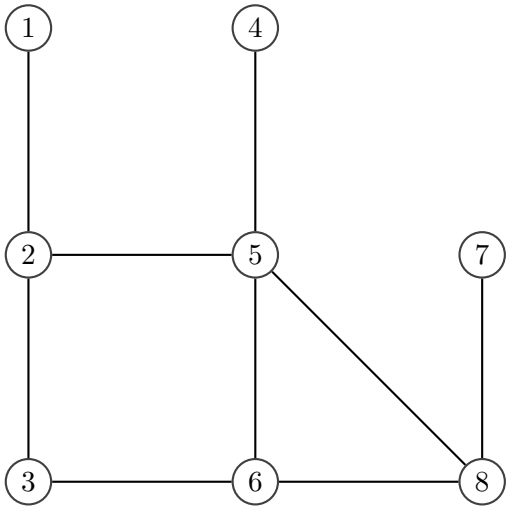


Figure 3.4: Undirected version of the graph in Fig. 3.1 on page 20.

Table 3.1: List of the graph types.

name	symbol	characteristic
Graph	$\mathcal{G}$	structure with a set of nodes and a set of edges
Directed graph	$\mathcal{D}$	all edges are directed
Undirected graph	$\mathcal{H}$	all edges are undirected
Directed acyclic graph	DAG	directed graph with no directed cycles

3.3 Directed graphical models - Bayesian networks

Directed graphical models use the directed edges to link the nodes in the graph. These directed edges encode the casual relationships among the random variables. Here, we introduce one type of directed graphical models, Bayesian networks (BNs). A Bayesian

network is a probabilistic graphical model that represents a set of the random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a BN could represent the probabilistic relationships between labels and observations in image classification. Given the observations, the network can be used to compute the probabilities of the presence of different labels.

3.3.1 Bayesian networks

Consider a set of the random variables $\{\underline{\boldsymbol{x}}_i, i \in \mathcal{V}\}$ defined over a DAG $\mathcal{D} = (\mathcal{V}, \mathcal{A})$. Each random variable $\underline{\boldsymbol{x}}_i$ is associated with a node $i \in \mathcal{V} = \{1, \cdots, i, \cdots, n\}$. The random variable, associated with the parents of the node i, is denoted as $\mathrm{Pa}(\underline{\boldsymbol{x}}_i)$. In this thesis, we restrict the random variable $\underline{\boldsymbol{x}}_i$ to random vectors, then all the random vectors could be put into a large compound vector $\underline{\boldsymbol{x}} = [\underline{\boldsymbol{x}}_1; \cdots; \underline{\boldsymbol{x}}_i; \cdots; \underline{\boldsymbol{x}}_n]$.

Definition 3.10 *Bayesian network.* $\underline{\boldsymbol{x}}$ *is a Bayesian network with respect to* $\mathcal{D}$ *if its joint distribution* P *can be expressed as a product*

$$P(\boldsymbol{x}) = \prod_{i \in \mathcal{V}} P(\boldsymbol{x}_i \mid Pa(\boldsymbol{x}_i)) \tag{3.1}$$

If $\underline{\boldsymbol{x}}_i$ does not have a parent, the conditional probability $P(\boldsymbol{x}_i \mid \mathrm{Pa}(\boldsymbol{x}_i))$ becomes the prior probability of $\underline{\boldsymbol{x}}_i$. Eq. (3.1) is called the chain rule for Bayesian networks. This key equation expresses the *factorization properties* of the joint distribution for a directed graphical model. The individual factor $P(\boldsymbol{x}_i \mid \mathrm{Pa}(\boldsymbol{x}_i))$ is the conditional probability distribution. For the DAG in Fig. 3.3 on page 22,

$$P(\boldsymbol{x}) = P(\boldsymbol{x}_1)P(\boldsymbol{x}_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_4)P(\boldsymbol{x}_3 \mid \boldsymbol{x}_2)P(\boldsymbol{x}_4)P(\boldsymbol{x}_5 \mid \boldsymbol{x}_4)P(\boldsymbol{x}_6 \mid \boldsymbol{x}_5)P(\boldsymbol{x}_7 \mid \boldsymbol{x}_5).$$

If the joint distribution over a set of the random variables is given as a product of the conditional distributions, i. e. (3.1), then we could test whether any potential conditional independence property holds in principle. In practice, such test would be time consuming. A decent feature of the graphical models is that the conditional independence properties of the joint distribution can be read directly from the graph without having to perform any analytical manipulations. The general framework for achieving this is called *d-separation* (Pearl, 1988). For detail description, we refer the reader to Bishop (2006); Koller & Friedman (2009).

A conditional Bayesian network is a BN conditioned on the observed data. Each random variable $\underline{\boldsymbol{x}}_i$ representing the class membership of the corresponding region node i is modelled in condition of the observed features in the image. In the tree-structured conditional Bayesian network (Drauschke & Förstner, 2011), the classification of a region is based on the unary features derived from the region and the binary features derived from the relations of the region hierarchy graph.

3.3.2 Inference in Bayesian networks

The inference problem for a BN aims at calculating the marginal probability. The task is to infer the most probable or maximum *a posteriori* (MAP) labelling $\boldsymbol{x}^*$ of the BN, which is defined as follows

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} P(\boldsymbol{x}) \tag{3.2}$$

The inference algorithms can be roughly divided into exact inference methods, such as belief propagation algorithm (Pearl, 1988), junction tree algorithm (Lauritzen & Spiegelhalter, 1988), and approximate inference methods, such as loopy belief propagation, variational algorithms (Jordan *et al.*, 1999) and Monte Carlo algorithms (MacKay, 2002).

For tree-structured BNs, belief propagation (BP) (Pearl, 1988; Yedidia *et al.*, 2000) can find the exact solution based on the local message-passing principle. Loopy belief propagation (LBP) is a widely used approximate inference method. The LBP directly applies the BP principle to a graphical model with loops. It can produce an approximate solution and may not guarantee the convergence of the message-passing process in general. However, the LBP works surprisingly well in many applications involving networks with loops (Murphy *et al.*, 1999; Yedidia *et al.*, 2000).

3.4 Undirected graphical models - random fields

Undirected graphical models use undirected edges to link the nodes in the graph. These undirected edges encode the mutual dependency relationships among the random variables. In this section, we introduce two types of undirected graphical models, Markov random fields (MRFs) and conditional random fields (CRFs). MRFs are appropriate in situations when associations between the random variables are considered to be more correlational than causal. The CRFs are the discriminative models that directly model the conditional distribution over the labels. This approach allows one to capture arbitrary dependencies between the observations without resorting to any model approximations. Both MRFs and CRFs are undirected graphical models.

3.4.1 Random field models

Consider a set of the random variables $\{\underline{\boldsymbol{x}}_i, i \in \mathcal{V}\}$ defined over a undirected graph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$. Each random variable $\underline{\boldsymbol{x}}_i$ is associated with a node $i \in \mathcal{V} = \{1, \cdots, n\}$ and takes a vector value from the label set $\mathcal{L} = \{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_C\}$. $\underline{\boldsymbol{x}} = [\underline{\boldsymbol{x}}_1; \cdots; \underline{\boldsymbol{x}}_i; \cdots; \underline{\boldsymbol{x}}_n]$ is called a random field. Any possible assignment of the labels to the random variables is called a *labelling* or *configuration*, which is denoted by the vector $\boldsymbol{x}$ and takes values from the set $\mathcal{L}^n$. The neighbourhood system $\mathcal{N}$ of the random field is defined by the sets $\{\mathrm{N}_i, i \in \mathcal{V}\}$, where N_i denotes the set of all neighbours of the node i. Three typical neighbourhood graphs (Pérez, 1998) used in image interpretation, i. e. a rectangular lattice grid, an irregular graph associated to an image partition, and a pyramid for hierarchical models, are shown in Fig. 3.5. For each graph, the blue nodes

are the neighbours of the white node. A rectangular lattice grid (Fig. 3.5 *Left*) is used to build the conditional random field model for the image region classification by Kumar & Hebert (2003a), an irregular graph (Fig. 3.5 *Middle*) for building facade image classification by Yang & Förstner (2011c), and a tree-structure as a simplified version of a pyramid (Fig. 3.5 *Right*) is used to build the hierarchical random field model for scene classification by Yang & Förstner (2011b). A clique $\boldsymbol{c}$ is as a subset of the nodes in a graph such that there exists an edge between all pairs of nodes in the subset. In the following, we give a formal definition of Markov random fields.

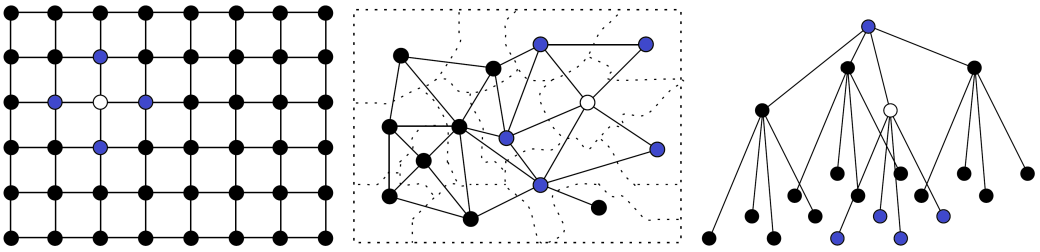


Figure 3.5: Three typical graphs supporting MRF-based models for image interpretation: *Left* a rectangular lattice grid; *Middle* an irregular graph associated to an image partition; *Right* a pyramid for hierarchical models. For each graph, the blue nodes are the neighbours of the white one. The rectangular lattice grid (*Left*) is used to build the conditional random field model for image region classification by Kumar & Hebert (2003a), the irregular graph (*Middle*) for building facade image classification by Yang & Förstner (2011c), and a tree-structure as a simplified version of the pyramid (*Right*) is used to build the hierarchical random field model for scene classification by Yang & Förstner (2011b). (Figure courtesy of Patrick Pérez (Pérez, 1998).)

A Markov random field (MRF) models the probability of the labelling $\boldsymbol{x}$, denoted by $P(\boldsymbol{x})$. According to the Bayes' rule, the posterior probability is proportional to the product of the likelihood and the prior probabilities as follows

$$P(\boldsymbol{x} \mid \boldsymbol{d}) \propto P(\boldsymbol{d} \mid \boldsymbol{x})P(\boldsymbol{x}) \tag{3.3}$$

where $P(\boldsymbol{d} \mid \boldsymbol{x})$ is the likelihood, $\boldsymbol{d}$ is the data, and $P(\boldsymbol{x})$ is known as the prior.

Definition 3.11 *Markov random field. A random field $\underline{\boldsymbol{x}}$ is said to be a Markov random field (MRF) with respect to a neighbourhood system $\mathcal{N} = \{N_i, i \in \mathcal{V}\}$ if and only if it satisfies the positivity property: $P(\boldsymbol{x}) > 0$, and the Markov property*

$$P(\boldsymbol{x}_i \mid \boldsymbol{x}_{\mathcal{V}-\{i\}}) = P(\boldsymbol{x}_i \mid \boldsymbol{x}_{N_i}) \tag{3.4}$$

The Markov property (3.4) implies that the prior probability of the assignment $\underline{\boldsymbol{x}}_i = \boldsymbol{x}_i$ depends only on the labelling of its neighbouring random variables given by N_i.

Using the Hammersley-Clifford theorem [1] (Hammersley & Clifford, 1971), the dis-

[1] A probability distribution that has a positive distribution satisfies the pairwise Markov property

tribution $P(\boldsymbol{x})$ over the labellings of the MRF is a *Gibbs* distribution [1] and can be written in the form

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left(-E(\boldsymbol{x})\right) = \frac{1}{Z}\exp\left(-\sum_{c\in\mathcal{C}}\phi_c(\boldsymbol{x}_c)\right) \tag{3.5}$$

where $\mathcal{C}$ is the set of cliques formed by the neighbourhood system $\mathcal{N}$, and $Z = \sum_{\boldsymbol{x}}\exp\left(-E(\boldsymbol{x})\right)$ is a normalization constant called the partition function. The term $\phi_c(\boldsymbol{x}_c)$ is known as potential function of the clique c, where $\boldsymbol{x}_c = \{\boldsymbol{x}_i, i \in c\}$. The term $E(\boldsymbol{x})$ is the so-called Gibbs energy function.

For a pairwise MRF, by assuming only up to pairwise clique potentials to be nonzero, the energy function E can be written as

$$E(\boldsymbol{x}) = \sum_{i\in\mathcal{V}} E_1(\boldsymbol{x}_i) + \sum_{\{i,j\}\in\mathcal{N}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{3.6}$$

where the set $\mathcal{N}$ is the set of *unordered* pairs of the neighbouring nodes. E_1 is called as the unary potential, which models the likelihood of the label assignment $\underline{\boldsymbol{x}}_i = \boldsymbol{x}_i$. E_2 is called as the pairwise potential, which models the cost of the assignment $\underline{\boldsymbol{x}}_i = \boldsymbol{x}_i$ and $\underline{\boldsymbol{x}}_j = \boldsymbol{x}_j$. While E_1 depends on the data, E_2 is independent of the data. In computer vision, a pairwise potential commonly takes the form of the Potts model (Potts, 1952), which gives a low energy value when $\underline{\boldsymbol{x}}_i = \underline{\boldsymbol{x}}_j$, and penalizes with a high energy values otherwise.

A conditional random field (CRF) may be viewed as an MRF globally conditioned on the observed data $\boldsymbol{d}$. The conditional distribution $P(\boldsymbol{x} \mid \boldsymbol{d})$ (Lafferty *et al.*, 2001) over the labellings of the CRF is a *Gibbs* distribution and can be written in the form

$$P(\boldsymbol{x} \mid \boldsymbol{d}) = \frac{1}{Z}\exp\left(-E(\boldsymbol{x} \mid \boldsymbol{d})\right) = \frac{1}{Z}\exp\left(-\sum_{c}\phi_c(\boldsymbol{x}_c \mid \boldsymbol{d})\right) \tag{3.7}$$

where $\boldsymbol{x}_c$ is the set of the nodes in a clique c, the term $\phi_c(\boldsymbol{x}_c \mid \boldsymbol{d})$ is the potential function of the clique c, and $Z = \sum_{\boldsymbol{x}}\exp\left(-E(\boldsymbol{x} \mid \boldsymbol{d})\right)$ is a normalization constant. The term $E(\boldsymbol{x} \mid \boldsymbol{d})$ is the Gibbs energy function.

For a pairwise CRF, by assuming only up to pairwise clique potentials to be nonzero,

(Koller & Friedman, 2009) with respect to an undirected graph $\mathcal{H}$ if and only if it is a Gibbs random field, that is, its distribution can be factorized over the cliques of the graph. The pairwise Markov property says any two non-adjacent variables are conditionally independent given all other variables.

[1] A distribution is a Gibbs distribution (Geman & Geman, 1984) if the joint distribution can be written as a product of the potential functions over the maximal cliques of the graph.

we can express the energy function E [1] more specifically as

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \sum_{\{i,j\} \in \mathcal{N}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) \tag{3.8}$$

where the set $\mathcal{N}$ is the set of *unordered* pairs of the neighbouring nodes. E_1 and E_2 are the unary and pairwise potentials respectively, which both depend on the observed data $\boldsymbol{d}$.

The most probable or maximum *a posteriori* (MAP) labelling $\boldsymbol{x}^*$ of the random field [2]is defined as

$$\boldsymbol{x}^* = \arg \max_{\boldsymbol{x} \in \mathcal{L}^n} P(\boldsymbol{x} \mid \boldsymbol{d}) \tag{3.9}$$

and can be found by minimizing the energy function E.

3.4.2 Inference in random field models

The task is to infer the most probable or MAP labelling $\boldsymbol{x}^*$ of the random field, which is defined as (3.9), and can be found by minimizing the energy function E. In general, minimizing the energy function E is NP-hard. But, there exist a number of algorithms which compute the exact solution for a particular family of the energy functions in polynomial time. For example, max-product belief propagation exactly minimizes the energy functions defined over the graphs with no loops (Yedidia *et al.*, 2000). And, some submodular energy functions (Fujishige, 1990) can be minimized by solving an st-MINCUT problem (Greig *et al.*, 1989; Kolmogorov & Zabih, 2004). However, many energy functions encountered in MRF and CRF models do not fall under the above classes, and are NP-hard to minimize (Kolmogorov & Rother, 2007). Most multi-label energy functions are non-submodular. For example, the Potts model potential (Potts, 1952) is a non-submodular function. They are instead solved using the approximate algorithms. These algorithms belong to two categories: message passing algorithms, such as sum-product algorithm, belief propagation (Yedidia *et al.*, 2000), tree-reweighted message passing (Wainwright *et al.*, 2005; Kolmogorov, 2006), and move making algorithms, such as Iterated Conditional Modes (Besag, 1986), $\alpha\beta$-swap, and α-expansion (Boykov *et al.*, 2001).

As will be seen in Chapter 4, the inference of the hierarchical CRF model is carried out with the multi-label graph optimization library of Boykov *et al.* (2001); Kolmogorov & Zabih (2004); Boykov & Kolmogorov (2004) using $\alpha\beta$-swap and α-expansion. Therefore, in the following part, we will provide an overview of $\alpha\beta$-swap and α-expansion algorithms.

$\alpha\beta$-swap and α-expansion are the two most popular graph cut algorithms, which are widely used to minimize the energy functions involving multi-valued discrete variables. Both algorithms work by repeatedly computing the global minimum of a binary

[1] Note that the CRF model with this specific energy function is denoted as the flat CRF in Chapter 5, to distinguish it from the hierarchical CRF.

[2] Note that the posterior probability distribution in the case of an MRF is proportional to the joint distribution.

labelling problem in their inner loops. This process converges to a local minimum. For a pair of labels α, β, a swap move takes some subset of the nodes currently given the label α and assigns them the label β and vice versa. The swap-move algorithm finds a local minimum such that there is no swap move, for any pair of labels α, β, that will produce a lower energy labelling. An expansion move for a label α increases the set of the nodes that are given this label. The expansion-move algorithm finds a local minimum such that there is no expansion move, for any label α, that will produce a labelling with lower energy.

3.5 Relations between directed and undirected graphical models

We have introduced two graphical frameworks for representing the probability distributions, corresponding to directed and undirected graphs, and it is instructive to discuss the relationship between these. In this section, we introduce two most common approaches: a moral graph, which converts a directed graph to an undirected graph; and a factor graph, which can represent both directed and undirected graphical models.

3.5.1 Moral graph representation

We convert the distribution specified by a factorization over a directed graph into one specified by a factorization over an undirected graph. This can be achieved if the clique potentials of the undirected graph are given by the conditional distributions of the directed graph. In order for this to be valid, we must ensure that the set of the variables that appears in each of the conditional distributions is a member of at least one clique of the undirected graph. For the nodes on the directed graph having just one parent, this is achieved simply by replacing the directed edge with an undirected edge. However, for nodes in the directed graph having more than one parent, this is not sufficient. Consider the example of a DAG in Fig. 3.3 on page 22, which is shown in Fig. 3.6 *Left* on page 29. The joint distribution takes the form $P(\boldsymbol{x}_1)P(\boldsymbol{x}_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_4)P(\boldsymbol{x}_3 \mid \boldsymbol{x}_2)P(\boldsymbol{x}_4)P(\boldsymbol{x}_5 \mid \boldsymbol{x}_4)P(\boldsymbol{x}_6 \mid \boldsymbol{x}_5)P(\boldsymbol{x}_7 \mid \boldsymbol{x}_5)$. We see that the factor $P(\boldsymbol{x}_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_4)$ involves the three variables $\underline{\boldsymbol{x}}_1$, $\underline{\boldsymbol{x}}_2$, and $\underline{\boldsymbol{x}}_4$, and so these must all belong to a single clique if this conditional distribution is to be absorbed into a clique potential. To ensure this, we add an extra edge between the pair of parents of the node 2, as shown in Fig. 3.6 *Right*.

In general, to convert a directed graph into an undirected graph, we first need to add the undirected edges between all pairs of the parents for each node in the graph and then replace all directed edges with undirected edges. This process is known as *moralization*, and the resulting undirected graph is called the *moral graph* (Cowell *et al.*, 1999; Bishop, 2006). To derive the joint probability distribution of the moral graph, we first initialize all of the clique potentials. Then, we assign each conditional probability distribution in the original directed graph to one of the clique potentials. Note that in all cases the partition function Z is 1. We see we have to discard some conditional

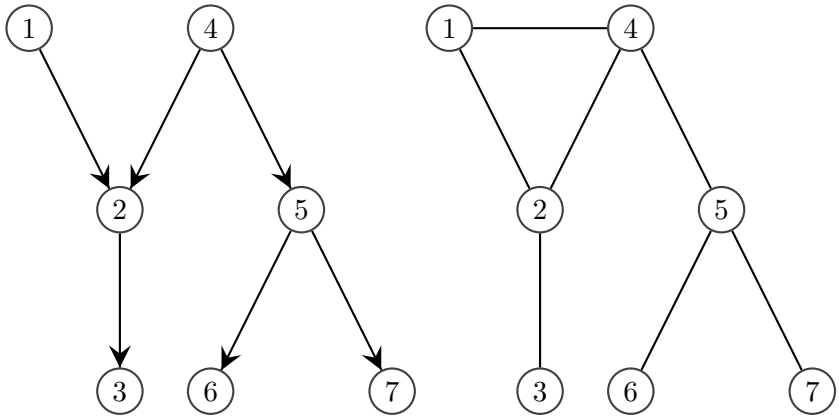


Figure 3.6: *Left*: the example of a DAG in Fig. 3.3 on page 22. *Right*: the corresponding moral graph. For nodes 3, 5, 6, 7 having just one parent, the directed edges are replaced by undirected edges. For node 2 having two parent nodes 1, 4, an extra edge has to be linked between the pair of parents, and then the directed edges are replaced by undirected edges.

independence properties from the graph from a directed graph to an undirected graph representation. The process of moralization adds the fewest extra edges and so retains the maximum number of independence properties (Cowell *et al.*, 1999).

3.5.2 Factor graph representation

As we see from previous sections, both directed and undirected graphs allow a global function of several variables to be expressed as a product of the factors over the subsets of those variables. Here we introduce a graphical construction called a factor graph (Kschischang *et al.*, 2001), which makes this decomposition explicit by introducing additional nodes for the factors themselves in addition to the nodes representing the variables.

A factor graph $\mathcal{F}$ is a bipartite graph (Bang-Jensen & Gutin, 2008) containing two types of the nodes: the variable nodes (denoted as circles), and the factor nodes (denoted as grey squares). The graph only contains the edges between the variable nodes and the factor nodes. The joint distribution P over a set of the variables can be expressed as a product of the factors

$$P(\boldsymbol{x}) = \prod_s \boldsymbol{f}_s(\boldsymbol{x}_s) \tag{3.10}$$

where $\boldsymbol{x}_s$ denotes a subset of the nodes. Each factor $\boldsymbol{f}_s$ is a function of a corresponding set of the nodes $\boldsymbol{x}_s$.

Undirected graphs, given by (3.7) on page 26, are the special cases in which the factors $\boldsymbol{f}_s(\boldsymbol{x}_s)$ are the potential functions. Directed graphs, whose factorization is

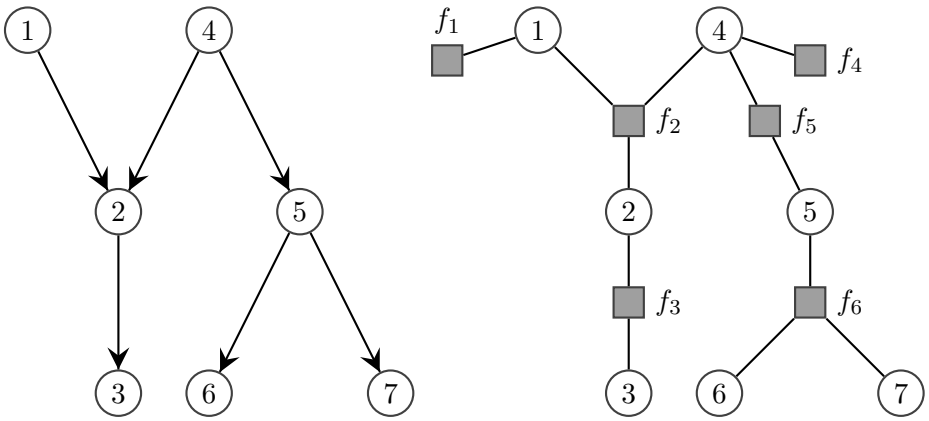


Figure 3.7: Factor graph representation of a directed graph. *Left*: a directed graph $\mathcal{D}$, same as in Fig. 3.3 on page 22, with the factorization $P(\boldsymbol{x}_1)P(\boldsymbol{x}_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_4)P(\boldsymbol{x}_3 \mid \boldsymbol{x}_2)P(\boldsymbol{x}_4)P(\boldsymbol{x}_5 \mid \boldsymbol{x}_4)P(\boldsymbol{x}_6 \mid \boldsymbol{x}_5)P(\boldsymbol{x}_7 \mid \boldsymbol{x}_5)$. *Right*: a factor graph representing the same distribution with factors $f_1(\boldsymbol{x}_1) = P(\boldsymbol{x}_1)$, $f_2(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4) = P(\boldsymbol{x}_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_4)$, $f_3(\boldsymbol{x}_2, \boldsymbol{x}_3) = P(\boldsymbol{x}_3 \mid \boldsymbol{x}_2)$, $f_4(\boldsymbol{x}_4) = P(\boldsymbol{x}_4)$, $f_5(\boldsymbol{x}_4, \boldsymbol{x}_5) = P(\boldsymbol{x}_5 \mid \boldsymbol{x}_4)$, $f_6(\boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7) = P(\boldsymbol{x}_6 \mid \boldsymbol{x}_5)P(\boldsymbol{x}_7 \mid \boldsymbol{x}_5)$.

defined by (3.1) on page 23, represent the special cases of (3.10) in which the factors are the conditional distributions.

To convert a directed graph to a factor graph, we simply create the variable nodes and the factor nodes in the factor graph, where the variable nodes are same as the nodes of the directed graph and the factor nodes correspond to the conditional distributions. Then, we add appropriate edges between appropriate variable nodes and factor nodes. The conversion of a directed graph to a factor graph is illustrated in Fig. 3.7.

It is also simple to convert an undirected graph to a factor graph. We create the variable nodes and the factor nodes in the factor graph, where the variable nodes are same as the nodes of the undirected graph and the factor nodes correspond to the maximal cliques $\boldsymbol{x}_s$. The factors $f_s(\boldsymbol{x}_s)$ are equal to the clique potentials. Note that there may be multiple factor graphs that correspond to the same undirected graph, which is illustrated in Fig. 3.8.

3.6 Summary

In this chapter, we have presented a theoretical basis needed for this thesis. We give some of the basic notations in graph theory (e. g. directed graph, undirected graph, cycle, and directed acyclic graph). Bayesian networks (BNs) is introduced briefly as one type of the directed graphical models, which is built on the directed acyclic graphs and the factorization properties. Markov random fields (MRFs) and conditional random fields (CRFs) are introduced as two types of the undirected graphical models. MRFs mainly capture the mutually dependent relationships such as the spatial neighbourhood

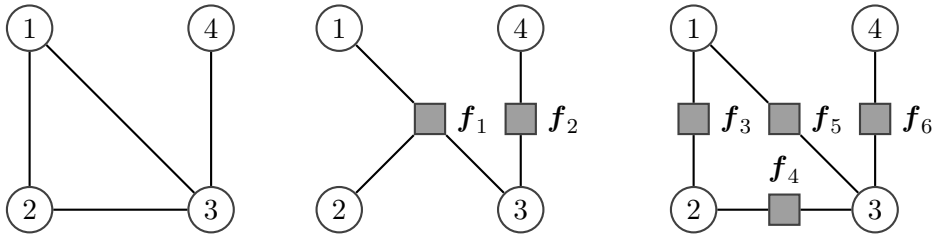


Figure 3.8: Factor graph representation of an undirected graph, illustrating the undirected graph may not yield a unique factor graph. *Left*: an undirected graph $\mathcal{H}$, same as in Fig. 3.2 on page 20. *Middle*: a factor graph with factor $\boldsymbol{f}_1(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)\boldsymbol{f}_2(\boldsymbol{x}_3, \boldsymbol{x}_4)$ representing the same distribution as the undirected graph. *Right*: a different factor graph representing the same distribution with cliques of maximum degree two, whose factors satisfy $\boldsymbol{f}_3(\boldsymbol{x}_1, \boldsymbol{x}_2)\boldsymbol{f}_4(\boldsymbol{x}_2, \boldsymbol{x}_3)\boldsymbol{f}_5(\boldsymbol{x}_1, \boldsymbol{x}_3)\boldsymbol{f}_6(\boldsymbol{x}_3, \boldsymbol{x}_4)$.

relationships. CRFs are the discriminative models that directly model the conditional distribution over the labels. At the end of this chapter, we introduce two approaches to build the relations between directed graph and undirected graph: a moral graph, which converts a directed graph to an undirected graph; a factor graph, which could represent both directed and undirected graphical models.

Chapter 4

A Generic Framework for Image Interpretation of Man-made Scenes

Between the idea
And the reality

Between the motion
And the act

Falls the Shadow

-*Thomas Stearns Eliot* (1888 - 1965)

4.1 Overview

As motivated in Section 1.1, spatial and hierarchical relationships are two valuable cues for image interpretation of man-made scenes. In this chapter we will develop a consistent graphical model representation for image interpretation that includes both information about the spatial structure and the hierarchical structure. The key idea for integrating the spatial and the hierarchical structural information into the interpretation process is to combine them with the low-level region class probabilities in a classification process by constructing the graphical model on the multi-scale image regions. We will start by constructing the graphical model. Then, the generic statistical model will be formulated as a multi-class labelling problem, where we will derive the corresponding energy function. Then, we will compare our model with the previous models and show that at certain choices of the parameters of our model, these methods fall out as the special cases. We will also derive the particular models for the energy

potentials and the conditional probability energy that are suited well for scene interpretation. We will derive the features from each region obtained from the unsupervised segmentation algorithm. Then we employ a classifier to calculate the label distribution for the local unary potential. Then we give one particular formulation for each of the pairwise potentials and the conditional probability energy. Finally, we will discuss the learning and the inference issues of this graphical model.

The complete proposed workflow for interpreting images of man-made scenes is sketched in Fig. 4.1. First, the test image is partitioned into regions by some unsupervised segmentation algorithms. Then, different features are extracted from the segmented regions. These features are passed to the learned graphical model to produce the final classification results. The graphical model is learned from the training images beforehand. The illustration in Fig. 4.1 shows that the graphical model can provide a consistent model representation including spatial and hierarchical structures, and therefore outperforms the classical local classification approach.

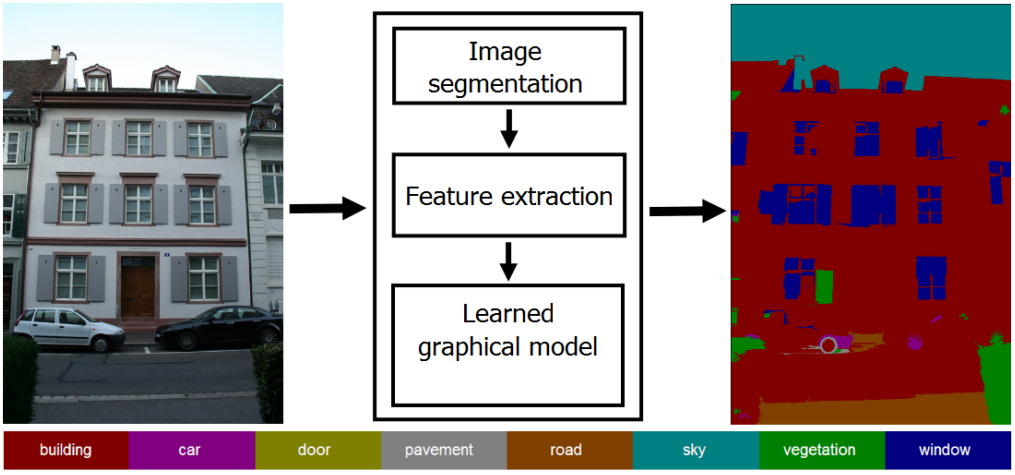


Figure 4.1: The basic dataflow for image interpretation of a test image for the graphical model framework. First, the test image is partitioned into regions by some unsupervised segmentation algorithms. Then, different features are extracted from the segmented regions. These features are passed to the learned graphical model to produce the final classification results. The graphical model is learned from the training images beforehand.

4.2 Statistical model for the interpretation problem

In the following sections, we will derive a generic model for the scene interpretation problem, which is formulated as a multi-class labelling problem. We will end up with an energy function that can be optimized approximately. Before defining the statistical model, we need to construct the graphical model first.

4.2.1 The graphical model construction and parametrization

By constructing the graphical model, we can flexibly choose either directed edges or undirected edges to model the relationships between the random variables based on the semantic meaning of these relationships.

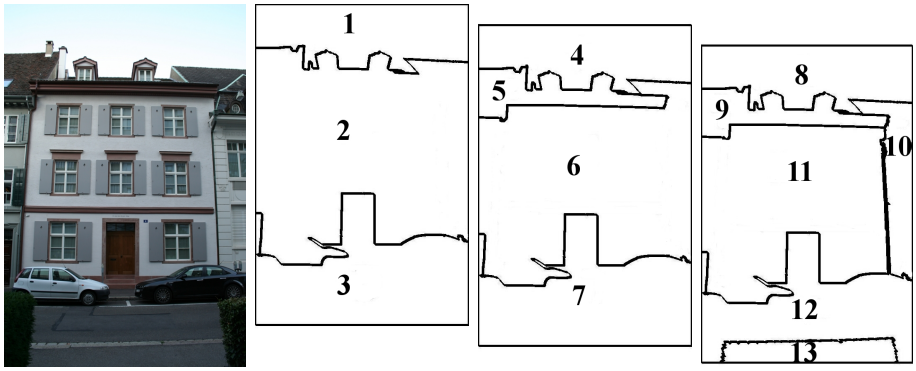
We use an example image to explain this model construction process. Given a test image, Fig. 4.2 on page 36 shows the corresponding multi-scale segmentation of the image, and the corresponding graphical model for image interpretation. Three layers are connected via a region hierarchy. The development of the regions over several scales is used to model the region hierarchy. Drauschke (2009) defined a region hierarchy with the directed edges between the regions of the successive scales. Furthermore, the relation is defined over the maximal overlap of the regions. Nodes connection and numbers correspond to the multi-scale segmentation. The blue edges between the nodes represent the neighbourhoods at one scale, and the red dashed edges represent the hierarchical relation between the regions. The pairwise interactions between the spatial neighbouring regions can be modelled by the undirected edges. The pairwise potential functions can be defined to capture the similarity between the neighbouring regions. The hierarchical relation between regions of the scene partonomy representing parent-child relations or part-of relations can be modelled by either the undirected edges or the directed edges.

The graphical model could consist of either the directed edges or the undirected edges. In general, we can parametrize the directed edges by the conditional probabilities, and the undirected edges by the potential functions. In Fig. 4.2, there are both directed edges and undirected edges. The potential functions are used to parametrize the undirected edges. The relationship between $\underline{\boldsymbol{x}}_1$ and $\underline{\boldsymbol{x}}_2$ is parametrized by the pairwise potential function $\phi(\boldsymbol{x}_1, \boldsymbol{x}_2)$. We use the local conditional probabilities to parametrize the directed edges. When the edge between node 1 and node 4 is a directed edge, the relationship between $\underline{\boldsymbol{x}}_4$ and its parent $\underline{\boldsymbol{x}}_1$ is parametrized by the conditional probability $P(\boldsymbol{x}_4 \mid \boldsymbol{x}_1)$. When the edge between node 1 and node 4 is a undirected edge, the relationship between $\underline{\boldsymbol{x}}_4$ and $\underline{\boldsymbol{x}}_1$ is parametrized by the pairwise potential function $\phi(\boldsymbol{x}_1, \boldsymbol{x}_4)$. Other edges are parametrized accordingly.

4.2.2 Representation as a multi-class labelling problem

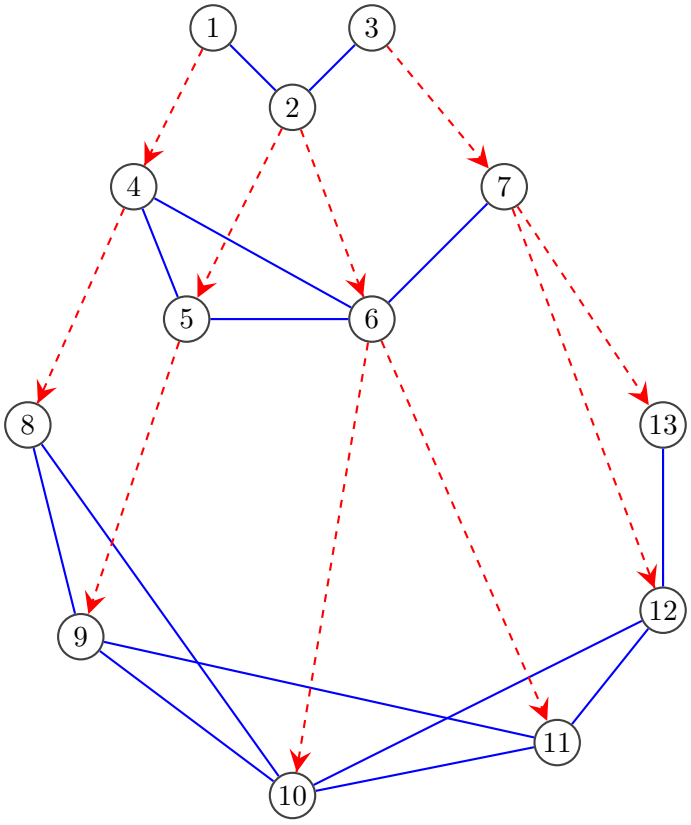
As we see from previous sections, both directed and undirected graphs allow a global function of several variables to be expressed as a product of the factors over the subsets of those variables. As in other graphical representations, the structure of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ can be used to define a factorization for a probability distribution over $\mathcal{G}$ according to the conditional independence relationships encoded in the graphical structure.

Consider a set of the random variables $\{\underline{\boldsymbol{x}}_i, i \in \mathcal{V}\}$ defined over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$. $\underline{\boldsymbol{x}} = [\underline{\boldsymbol{x}}_1; \cdots; \underline{\boldsymbol{x}}_i; \cdots; \underline{\boldsymbol{x}}_n]$. Each random variable $\underline{\boldsymbol{x}}_i$ is associated with a node $i \in \mathcal{V} = \{1, \cdots, i, \cdots, n\}$ and takes a vector value from the label set $\mathcal{L} = \{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_C\}$. Any possible assignment of the labels to the random variables is called a *labelling*, which is



(a) Example image of a man-made scene

(b) Multi-scale segmentation (from left to right: top, middle and bottom scale)



(c) The graphical model

Figure 4.2: Illustration of the graphical model architecture. (a). An example image of a man-made scene. (b). The boundary maps of the segmented image corresponding to the multi-scale segmentation of mean shift (Comaniciu & Meer, 2002) algorithm (from left to right: top, middle and bottom scale). (c). The graphical model construction, with three layers connected via a region hierarchy. Nodes in the graph, indicated by numbers, correspond to the segmented regions. The blue edges between the nodes represent the neighbourhoods at one scale (undirected edges), and the red dashed edges represent the hierarchical relation between regions (undirected or directed edges).

denoted by the vector $\boldsymbol{x}$ and takes values from the set $\mathcal{L}^n$. Therefore, we present the scene interpretation problem as a multi-class labelling problem. Given the observed data $\boldsymbol{d}$, the distribution P over a set of the variables $\underline{\boldsymbol{x}}$ can be expressed as a product of the factors [1]

$$P(\boldsymbol{x} \mid \boldsymbol{d}) = \frac{1}{Z} \prod_{i \in \mathcal{V}} \boldsymbol{f}_i(\boldsymbol{x}_i \mid \boldsymbol{d}) \prod_{\{i,j\} \in \mathcal{E}} \boldsymbol{f}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) \prod_{\langle i,k \rangle \in \mathcal{S}} \boldsymbol{f}_{ik}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{4.1}$$

where the factors $\boldsymbol{f}_i, \boldsymbol{f}_{ij}, \boldsymbol{f}_{ik}$ are the functions of the corresponding sets of the nodes, and Z is the normalization factor. The set $\mathcal{V}$ is the set of the nodes in the complete graph, and the set $\mathcal{E}$ is the set of pairs collecting the neighbouring nodes within each scale. $\mathcal{S}$ is the set of pairs collecting the parent-child relations between regions with the neighbouring scales, where $\langle i, k \rangle$ denotes nodes i and k are connected by either a undirected edge or a directed edge. Note that this model only exploits up to second-order cliques, which makes learning and inference much faster than the model involving high-order cliques.

To get a better understanding of the model, we illustrate the stochastic model of Fig. 4.2 in the form of a *factor graph*, which is previously discussed in Section 3.5.2. The factor graph representation is shown in Fig. 4.3, by omitting all the factors on each node. Each square in this factor graph corresponds to the factor which is a local function of the involved variables. For example, the square connecting nodes 1 and 2 corresponds to the factor $\boldsymbol{f}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, and the square connecting nodes 1 and 4 corresponds to the factor $\boldsymbol{f}_{14}(\boldsymbol{x}_1, \boldsymbol{x}_4)$. This graph makes obvious that the model assumes only binary cliques, without the higher order cliques among the nodes.

By simple algebra calculation, the probability distribution given in (4.1) can be written in the form

$$P(\boldsymbol{x} \mid \boldsymbol{d}) = \frac{1}{Z} \exp\left(\sum_{i \in \mathcal{V}} \log \boldsymbol{f}_i(\boldsymbol{x}_i) + \sum_{\{i,j\} \in \mathcal{E}} \log \boldsymbol{f}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{\langle i,k \rangle \in \mathcal{S}} \log \boldsymbol{f}_{ik}(\boldsymbol{x}_i, \boldsymbol{x}_k) \right) \tag{4.2}$$

where we drop the factor conditioned on the data $\boldsymbol{d}$ for simplicity. Therefore, the probability distribution for this graphical model is a *Gibbs* distribution

$$P(\boldsymbol{x} \mid \boldsymbol{d}) = \frac{1}{Z} \exp\left(-E(\boldsymbol{x} \mid \boldsymbol{d})\right) \tag{4.3}$$

The term

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = -\sum_{i \in \mathcal{V}} \log \boldsymbol{f}_i(\boldsymbol{x}_i) - \sum_{\{i,j\} \in \mathcal{E}} \log \boldsymbol{f}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{\langle i,k \rangle \in \mathcal{S}} \log \boldsymbol{f}_{ik}(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{4.4}$$

is the energy function. For the consistency with most other works (e. g. Shotton *et al.*,

[1] The formal theoretical proof is linked to a graphical model defined over a chain graph, which is a generalization of both the undirected graph and the directed graph, see Appendix A for a detail description.

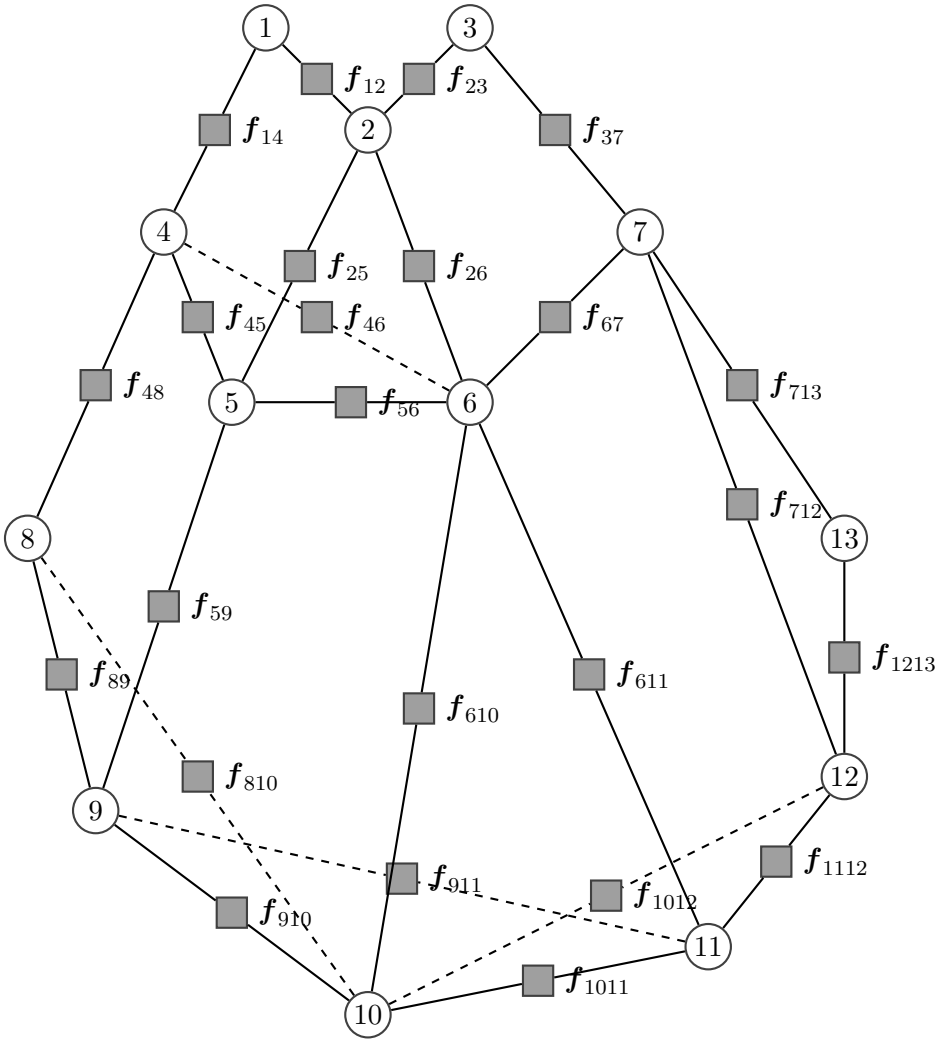


Figure 4.3: A factor graph representation of the graphical model shown in Fig. 4.2 on page 36, without depicting all the factors on each node. The dashed lines indicate the 3D structure of this graph.

2006; Kohli *et al.*, 2009; Yang & Förstner, 2011c) in the literature, in the following, the energy function in (4.4) is defined as

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) + \beta \sum_{\langle i,k \rangle \in \mathcal{S}} E_3(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{4.5}$$

where α and β are the weighting coefficients in the model. E_1 is the unary potential, which represents the relationships between the variables and the local observed data. E_2 is the pairwise potential, which represents the relationships between the variables of the neighbouring nodes within each scale. E_3 is either the hierarchical pairwise potential or the conditional probability energy, which represents the relationships between the regions of the scene partonomy with neighbouring scales. This graphical model is illustrated in Fig. 4.2 on page 36.

The most probable or maximum *a posteriori* (MAP) labelling $\boldsymbol{x}^*$ is defined as

$$\boldsymbol{x}^* = \arg \max_{\boldsymbol{x} \in \mathcal{L}^n} P(\boldsymbol{x} \mid \boldsymbol{d}) \tag{4.6}$$

and can be found by minimizing the energy function $E(\boldsymbol{x} \mid \boldsymbol{d})$.

4.3 Relation to previous models

In this section, we draw comparisons with the previous models for image interpretation (Plath *et al.*, 2009; Fulkerson *et al.*, 2009; Yang *et al.*, 2010a; Drauschke & Förstner, 2011; Yang & Förstner, 2011c) and show that at certain choices of the parameters of our framework, these methods fall out as the special cases. We will now show that our model is not only a generalization of the standard flat CRF over the image regions, but also of the hierarchical CRF and the conditional Bayesian network.

4.3.1 Equivalence to flat CRFs over regions

Let us consider the case with only one layer segmentation of the image (the bottom layer of the graphical model in Fig. 4.2 on page 36). In this case, the weight β is set to be zero, the set $\mathcal{V}^1$ is the set of nodes in the graph of the bottom layer, and the set $\mathcal{E}^1$ is the set of pairs collecting the neighbouring nodes in the bottom layer. This allows us to rewrite (4.5) as

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}^1} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}^1} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) \tag{4.7}$$

which is exactly the same as the energy function associated with the flat CRF defined over the image regions with E_1 as the unary potential and E_2 as the pairwise potential. In this case, our model becomes equivalent to the flat CRF models defined over the image regions (Gould *et al.*, 2008; Batra *et al.*, 2008; Fulkerson *et al.*, 2009; Yang & Förstner, 2011c).

4.3.2 Equivalence to hierarchical CRFs

Let us now consider the case with the multi-scale segmentation of the image. If we choose E_3 as a pairwise potential in (4.5), the energy function reads

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) + \beta \sum_{\{i,k\} \in \mathcal{S}} E_3(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{4.8}$$

which is exactly the same as the energy function associated with the hierarchical CRF defined over the multi-scale of the image regions with E_1 as the unary potential, E_2 as the pairwise potential within each scale, and E_3 as the hierarchical pairwise potential with the neighbouring scales. In this case, our model becomes equivalent to the hierarchical CRF models defined over multi-scale of image regions (He *et al.*, 2004; Yang *et al.*, 2010a; Yang & Förstner, 2011b).

If we set α to be zero, and choose E_3 as a pairwise potential in (4.5), the energy function reads

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \beta \sum_{\{i,k\} \in \mathcal{S}} E_3(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{4.9}$$

which is the same as the energy function associated with the tree-structured CRF by neglecting the direct local neighbourhood dependencies on the image regions on multiple scales. In this case, our model becomes equivalent to the tree-structured CRF models defined over multi-scale of the image regions (Reynolds & Murphy, 2007; Plath *et al.*, 2009).

4.3.3 Equivalence to conditional Bayesian networks

If we set α to be zero, and choose E_3 as the conditional probability energy in (4.5), the energy function reads

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \beta \sum_{(i,k) \in \mathcal{S}} E_3(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{4.10}$$

which is the same as the energy function associated with the tree-structured conditional Bayesian network defined over the multi-scale of the image regions. In the tree-structured conditional Bayesian network, the classification of a region is based on the unary features derived from the region and the binary features derived from the relations of the region hierarchy graph. In this case, our model becomes equivalent to the tree-structured conditional Bayesian network defined over multi-scale of the image regions (Drauschke & Förstner, 2011).

4.4 Data-driven modelling of energy potentials and conditional probability

The proposed energy function (4.5) consists of three basic elements:

1. The unary potential $E_1(\boldsymbol{x}_i \mid \boldsymbol{d})$ describes how likely it is to predict a particular class label $\boldsymbol{x}_i$, given the local observed data.

2. The local pairwise potential $E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d})$ describes the category compatibility between the neighbouring labels $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ given the data.

3. The hierarchical pairwise potential or the conditional probability energy $E_3(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d})$ describes the likelihood for a relationship between the regions of the scene partonomy with the neighbouring scales given the data.

In this section, we will derive the particular models for the energy potentials and the conditional probability energy that are suited well for scene interpretation. Note that the use of these particular models is not prescribed by our framework. They should be considered as one possible implementation of the proposed method.

We will derive the features from each region obtained from the unsupervised segmentation algorithm. Then we employ a classifier called randomized decision forest (RDF) to calculate the label distribution for the local unary potential. Then we give one particular formulation for each of the pairwise potentials and the conditional probability energy. Note that the setup of energy potentials and the conditional probability energy is identical to that used for the final experiments.

4.4.1 Features

Features contains the information needed to make the class-specific decisions while being highly invariant with respect to extraneous effects such as changing object appearance, pose, illumination and background clutter. Several well-engineered features have been experimentally found to be well fit for image classification task (Drauschke & Mayer, 2010; Yang & Förstner, 2011a). We use the following five feature sets $\boldsymbol{h} = \bigcup_{i=1}^{5} \boldsymbol{h}_i$ from each image region obtained from the unsupervised segmentation algorithms. In our experiment presented in Chapter 5, we use the mean shift segmentation (Comaniciu & Meer, 2002) and the watershed segmentation (Vincent & Soille, 1991).

Basic features $\boldsymbol{h}_1$: First feature set $\boldsymbol{h}_1$ are eleven basic features including (1) the number of the components of the region (C); (2) the number of the holes of the region (H); (3) Euler characteristic for planar figures (Lakatos, 1976) ($E = C - H$); (4) the area (A); (5) the perimeter (U); (6) the form factor ($F = U^2/(4\pi A)$); (7) the height of the bounding box; (8) the width of the bounding box; (9) the area ratio between the region and its bounding box; (10) the ratio between the center of the region and the height of the image; (11) the ratio between the center of the region and the width of the image.

Colour features $\boldsymbol{h}_2$: For representing the spectral information of the region, we use nine colour features (Barnard *et al.*, 2003) as second feature set $\boldsymbol{h}_2$: the mean and the standard deviation of R-channel, G-channel and B-channel respectively, in the RGB colour space; and the mean of H-channel, S-channel and V-channel respectively, in the HSV colour space.

Peucker features $\boldsymbol{h}_3$: Twelve Peucker features are derived from the generalization of the region's border as third feature set $\boldsymbol{h}_3$, and represent parallelity or orthogonality of the border segments. We select the four points of the boundary which are farthest away from each other. From this polygon region with four corners, we derive three central moments, and eigenvalues in direction of major and minor axis, aspect ratio of eigenvalues, orientation of the polygon region, coverage of the polygon region, and four angles of the polygon region boundary points.

Texture features $\boldsymbol{h}_4$: We use eighteen texture features derived from the Walsh transform (Petrou & Bosdogianni, 1999; Lazaridis & Petrou, 2006) as fourth feature set $\boldsymbol{h}_4$, because the features from Walsh filters are among the best texture features from the filter banks (Drauschke & Mayer, 2010). We determine the magnitude of the response of nine Walsh filters. For each of the nine filters, we determine the mean and the standard deviation for each region.

SIFT features $\boldsymbol{h}_5$: Fifth feature set $\boldsymbol{h}_5$ are mean SIFT (Scale-Invariant Feature Transform) descriptors (Lowe, 2004) of the image region. SIFT descriptors are extracted for each pixel of the region at a fixed scale and orientation, which is practically the same as the HOG descriptor (Dalal & Triggs, 2005), using the fast SIFT framework in Vedaldi & Fulkerson (2008). The extracted descriptors are then averaged into one l_1-normalized descriptor vector for each region.

These features are roughly listed in Table 4.1. The resulting 178 features are then concatenated into one feature vector.

4.4.2 Unary potential

The local unary potential E_1 independently predicts the label x_i based on the image $\boldsymbol{d}$:

$$E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) = -\log P(\boldsymbol{x}_i \mid \boldsymbol{d}) \tag{4.11}$$

The label distribution $P(\boldsymbol{x}_i \mid \boldsymbol{d})$ is usually calculated by using a classifier. Here, we employ randomized decision forest (RDF) (Breiman, 2001) as the classifier, where the derived features from the image regions for the RDF classifier are chosen from Table 4.1. Existing work has shown the power of decision forests as the classifiers (Maree *et al.*, 2005; Lepetit *et al.*, 2005; Bosch *et al.*, 2007). As illustrated in Fig. 4.4, a RDF is an ensemble classifier that consists of T decision trees (Shotton *et al.*, 2008). The feature vector $\boldsymbol{d}_i$ of image region i is classified by going down each tree. This process gives a

Table 4.1: List of the derived features from the image regions: basic features, colour features, Peucker features, texture features, SIFT features. The number indicates the feature numbers in each feature set.

$\boldsymbol{h}_1$ basic features (11) region area and perimeter, height and width of the bounding box, etc.
$\boldsymbol{h}_2$ colour features (9) mean and standard deviation of the RGB and the HSV colour spaces
$\boldsymbol{h}_3$ Peucker features (12) moments and eigenvalues of a region as orthogonality or parallelity
$\boldsymbol{h}_4$ texture features (18) texture features derived from the Walsh transform
$\boldsymbol{h}_5$ SIFT features (128) mean SIFT descriptor features

class distribution at the leaf nodes and also a path for each tree. The class distributions $P(\boldsymbol{x}_i \mid \boldsymbol{d}_i)$ is obtained by averaging the class distribution over the leaf nodes for all T trees. This classification procedure is identical to Shotton *et al.* (2008).

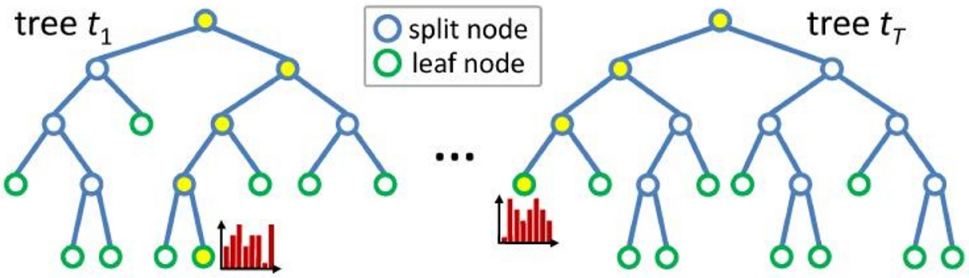


Figure 4.4: Randomized decision forest. A decision forest is an ensemble classifier that consists of T decision trees. A feature vector is classified by going down each tree. This process gives a class distribution at the leaf nodes and also a path for each tree. (Figure courtesy of Jamie Shotton (Shotton *et al.*, 2008).)

Based on the fact that the RDF classifier does not take the location information explicitly, we incorporate the location potential (similar to Shotton *et al.* (2006)) in the unary potential. The location potential $-\log Q(\boldsymbol{x}_i \mid \boldsymbol{d})$ is the negative logarithm of the function of the class labels $\boldsymbol{x}_i$ given the image coordinates $\boldsymbol{z}_i$ as the center of the region i, where

$$Q(\boldsymbol{x}_i \mid \boldsymbol{d}) = W(\boldsymbol{x}_i \mid \boldsymbol{z}_i) \tag{4.12}$$

The location potential captures the dependence of the class label on the rough location

of the region in the image. The learning of $W(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$ is described in Section 4.5.2 in detail. Therefore, the unary potential E_1 is written as

$$E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) = -\log P(\boldsymbol{x}_i \mid \boldsymbol{d}) - \log Q(\boldsymbol{x}_i \mid \boldsymbol{d}) \tag{4.13}$$

4.4.3 Pairwise potentials

The local pairwise potential E_2 describes the category compatibility between the neighboring labels $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ given the image $\boldsymbol{d}$, which takes the form (Boykov & Jolly, 2001)

$$E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) = g_{ij}(1 - \delta(\boldsymbol{x}_i = \boldsymbol{x}_j)) \tag{4.14}$$

where $\delta(\cdot)$ is the Kronecker delta. In this work, the feature function g_{ij} measures the colour difference between the neighbouring regions, as suggested by Rother *et al.* (2004),

$$g_{ij} = \frac{1 + 4\exp(-2c_{ij})}{0.5(N_i + N_j)}$$

where c_{ij} is the l_2 norm of the colour difference between the regions in the HSV colour space. N_i is the number of the regions neighbored to region i, and N_j is the number of the regions neighbored to j. The potentials E_2 are scaled by N_i and N_j to compensate for the irregularity of the graph $\mathcal{G}$. We refer the reader to Boykov & Jolly (2001); Shotton *et al.* (2006); Gould *et al.* (2008) for more details about designing the pairwise potential.

The hierarchical pairwise potential $E_{3,h}$ describes the category compatibility between the hierarchically neighbouring labels $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$ given the image $\boldsymbol{d}$, which takes the similar form as the local pairwise potential

$$E_{3,h}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) = g^{'}_{ik}(1 - \delta(\boldsymbol{x}_i = \boldsymbol{x}_k)) \tag{4.15}$$

where the feature function $g^{'}_{ik}$ relates to the hierarchical pairs of the regions (i, k), and is defined as

$$g^{'}_{ik} = (1 + 4\exp(-2c_{ik}))$$

with c_{ik} being the l_2 norm of the colour difference between the regions in the HSV colour space. The hierarchical pairwise potential acts as a link across the scale, facilitating propagation of the information in the model.

Note that here we give two simple pairwise potential formulations compared with the unary potentials. The results could be better if more sophisticated features for the pairwise potentials would be used. Furthermore, the pairwise potentials are usually represented by a weighted summation of many features functions (Shotton *et al.*, 2006), and the parameters with the size as same as feature number are learned from the training data. But this kind of parameter learning remains a difficult problem (Alahari *et al.*, 2010).

4.4.4 Conditional probability energy

The conditional probability energy $E_{3,c}$ takes the form

$$E_{3,c}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) = -\log P(\underline{\boldsymbol{x}}_i = \boldsymbol{l}_r \mid \underline{\boldsymbol{x}}_k = \boldsymbol{l}_t, \boldsymbol{d}) = -\log P(\boldsymbol{x}_r \mid \boldsymbol{x}_t) \tag{4.16}$$

where $\boldsymbol{l}_r, \boldsymbol{l}_t \in \mathcal{L}$, and $P(\boldsymbol{x}_r \mid \boldsymbol{x}_t)$ denotes the random variable $\underline{\boldsymbol{x}}_i$ is in the r-th state and its parents $\underline{\boldsymbol{x}}_k$ is in the t-th state. For the specific construction of our graphical model, the node $\underline{\boldsymbol{x}}_i$ always has one unique parent $\underline{\boldsymbol{x}}_k$, which lives in the successive scale, as illustrated in Fig. 4.2. If we have no prior information about the node labels, uniform distribution is adopted, which means there is no bias for the node label. The learning procedure of the conditional probabilities $P(\boldsymbol{x}_r \mid \boldsymbol{x}_t)$ is described in Section 4.5.3 in detail.

4.5 Learning and inference for the graphical model

In this section, we discuss the learning and the inference issues of the graphical model in (4.5). The classifier and the location potential for the unary potential, and the weighting parameters α, β, and the conditional probability energy are the model parameters that should be learned. We take the learning approach based on piecewise training (Sutton & McCallum, 2005). Piecewise training involves dividing the graphical model into pieces corresponding to the different terms in (4.5). Each of these terms is then learned independently, as if it were the only term in the model.

In (4.5), when the nodes in E_3 are connected by the directed edges, meaning E_3 is the conditional probability energy, we convert this model into a factor graph, and the inference is carried out by loopy belief propagation (Pearl, 1988; Yedidia *et al.*, 2000). When the nodes in E_3 are connected by the undirected edges, meaning E_3 is the hierarchical pairwise potential, the inference is carried out with the α-expansion (Boykov *et al.*, 2001), which is a graph cut (Boykov & Kolmogorov, 2004) based move making algorithm (Boykov *et al.*, 2001).

4.5.1 Learning the classifier

The classifier operates in the image regions defined by the unsupervised segmentation. In order to train the RDF classifier, we take the ground-truth label of each region to be the majority vote of the ground-truth pixel labels. Then a RDF is trained on the labelled data for each of the classes. According to a decision tree learning algorithm, a decision tree recursively splits left or right down the tree to a leaf node. We use the extremely randomized trees (Geurts *et al.*, 2006) as learning algorithm. Each tree is trained separately on a small random subset of the training data. The learning procedure is identical to Shotton *et al.* (2008). We refer the reader to Shotton *et al.* (2008) for more details.

4.5.2 Learning the location potential

The location potential $-\log Q(\boldsymbol{x}_i \mid \boldsymbol{d}) = -\log W(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$ takes the form of a look-up table with an entry for each class $\boldsymbol{x}_i$ and the region center location $\boldsymbol{z}_i$, where

$$W(\boldsymbol{x}_i \mid \boldsymbol{z}_i) = \left(\frac{N_{\boldsymbol{x}_i, \hat{\boldsymbol{z}}_i} + 1}{N_{\hat{\boldsymbol{z}}_i} + 1} \right)^2 \tag{4.17}$$

The index $\hat{\boldsymbol{z}}_i$ is the normalized version of the region center $\boldsymbol{z}_i$, where the normalization allows for the images of different sizes: the image is mapped onto a canonical square and $\hat{\boldsymbol{z}}_i$ indicates the pixel position within this square. $N_{\boldsymbol{x}_i, \hat{\boldsymbol{z}}_i}$ is the number of the regions of the class $\boldsymbol{x}_i$ at the normalized location in $\hat{\boldsymbol{z}}_i$, and $N_{\hat{\boldsymbol{z}}_i}$ is the total number of the regions at the location in $\hat{\boldsymbol{z}}_i$.

For example, in our experiment, we use part of the annotation images in 8-class eTRIMS dataset (Korč & Förstner, 2009) to learn the location potential, but ensure no overlap between these images and the testing images in the experimental part. Some learned location potentials are illustrated in Fig. 4.5. From Fig. 4.5, we see *sky* tends to occur at the top part of images, while *road* tends to occur at the bottom part of images, and *building* tends to occur in the middle part of images. Here, the dark blue area indicates the most likely locations of one class, while the dark red area indicates the most unlikely locations.

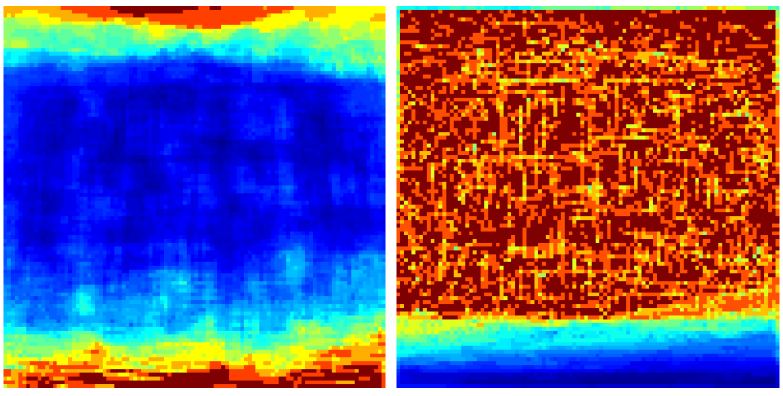
4.5.3 Learning the conditional probability energy

When the random variables involved are discrete, the conditional probability distributions in the graphical model become the conditional probability tables (CPTs) (Murphy, 1998). The conditional probability energy $E_{3,c}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) = -\log P(\boldsymbol{x}_r \mid \boldsymbol{x}_t)$ takes the form of CPTs with an entry θ_{irt} for each $\underline{\boldsymbol{x}}_i$ is in the r-th state and its parents $\underline{\boldsymbol{x}}_k$ is in the t-th state. Suppose the graphical model has s layers (cf. Fig. 4.2 on page 36). We generate $(s-1)$ CPTs of which each has $C \times C$ elements, where C is the number of the class labels. For example, in Fig. 4.2, $s = 3$; therefore, we generate two CPTs.

The parameter θ_{irt} is estimated using the maximum likelihood method. We count the co-occurrences of the parent region and the child region. We could minimize the negative logarithm of the likelihood of the parameter θ_{irt}

$$\begin{aligned} \theta^*_{irt} &= -\arg\min_{\theta_{irt}} \sum_{i,r,t} \log \theta_{irt}^{N_{irt}} \\ s.t. &\sum_r \theta_{irt} = 1 \end{aligned} \tag{4.18}$$

where N_{irt} is the number of times that $\underline{\boldsymbol{x}}_i$ appears in the r-th state and its parents $\underline{\boldsymbol{x}}_k$ in the t-th state, which is simply counted from the training samples. Minimizing (4.18)



(a) building (b) road

(c) sky (d) vegetation

Figure 4.5: Example location potentials. Part of the annotation images in 8-class eTRIMS dataset (Korč & Förstner, 2009) is used to learn the location potentials, with no overlap between these images and the testing images in the experimental part. The annotation images are mapped onto a canonical square. The size of each image is 100×100 here. *Sky* tends to occur at the top part of images, while *road* tends to occur at the bottom part of images, and *building* tends to occur in the middle part of images. Here, the dark blue area indicates the most likely locations of one class, while the dark red area indicates the most unlikely locations.

leads to the following analytical solution

$$\theta^*_{irt} = \frac{N_{irt}}{\sum_r N_{irt}} \tag{4.19}$$

This result is analogous to the standard maximum likelihood estimation for Bayesian networks (Koller & Friedman, 2009).

4.5.4 Learning the weights

Having learned the potentials and the conditional probability energy as described earlier, the problem remains of how to assign appropriate weights. In our formulation (4.5), we have two weights α and β which represent the trade-off among the confidence in the local unary potential E_1, the local pairwise potential E_2, and the hierarchical pairwise potential $E_{3,h}$ or the conditional probability energy $E_{3,c}$.

The training of model parameters in general is not an easy problem and there is a wide body of literature dealing with it, (cf. Taskar *et al.*, 2004; He *et al.*, 2006; Korč & Förstner, 2008; Alahari *et al.*, 2010). We estimate α and β by 5-fold cross validation on the training data.

4.5.5 Inference

In (4.5), when the nodes in E_3 are connected by the directed edges, the graphical model in Fig. 4.2 consists of the undirected edges and the directed edges. To perform a consistent inference, we convert this model into a factor graph (Section 3.5.2). Given the factor graph representation, we use OpenGM package provided by Andres *et al.* (2010) to perform the inference in the factor graph using loopy belief propagation (Pearl, 1988; Yedidia *et al.*, 2000).

In (4.5), when the nodes in E_3 are connected by the undirected edges, the graphical model in Fig. 4.2 only consists of the undirected edges. It has been experimentally shown (Kolmogorov & Rother, 2006; Russell *et al.*, 2010), that for most computer vision problems graph cut (Boykov & Kolmogorov, 2004) based move making algorithms (Boykov *et al.*, 2001) tend to outperform other approaches in terms of speed and quality. As the pairwise potentials of the energy function in (4.5) are composed of metrics [1] , the

[1]The potential function ϕ is called a metric (Boykov *et al.*, 2001) on the space of labels $\mathcal{L}^n$, if for any $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k \in \mathcal{L}^n$, it satisfies the following three properties

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_i) = 0$$

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_j, \boldsymbol{x}_i) \geq 0$$

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \phi(\boldsymbol{x}_i, \boldsymbol{x}_k) + \phi(\boldsymbol{x}_k, \boldsymbol{x}_j)$$

If ϕ only satisfies the first two properties, it is called a semi-metric. The α-expansion algorithm can only be used with metric term. Otherwise, the $\alpha\beta$-swap can be used with semi-metric and metric term. While the α-expansion move algorithm produces a labelling, which is within a known factor of the global minimum, the $\alpha\beta$-swap does not guarantee any closeness to the global minimum (Veksler, 1999). It is trivial to show that E_2 (4.14) and $E_{3,h}$ (4.15) are both metrics.

energy can be minimized approximately using the well known α-expansion algorithm (Boykov *et al.*, 2001). Therefore, the inference is carried out with the multi-label graph optimization library of Boykov *et al.* (2001); Kolmogorov & Zabih (2004); Boykov & Kolmogorov (2004) using α-expansion, which is explained in Section 3.4.2.

4.6 Summary

In this chapter we have presented a generalization of many previous region based methods within a principled graphical model framework. This generic graphical model is used to solve the task of scene interpretation, which is formulated as a multi-class labelling problem. The statistical model leads to an energy function that can be optimized approximately by either loopy belief propagation (Pearl, 1988; Yedidia *et al.*, 2000) or graph cut based move making algorithm (Boykov *et al.*, 2001).

Our approach enables the integration of the features and the spatial structural information and the hierarchical structural information defined over the multi-scale image segmentation in one optimization framework. We also derive three reasonable energy potentials, i. e. the local unary potential, the local pairwise potential, the hierarchical pairwise potential, and the conditional probability energy from the training data, which we will use for our particular implementation of the framework. The energy function for the statistical model for the interpretation problem is shown in (4.5) on page 39. In the experiments presented in Chapter 5, we will compare the following four different models.

Region classifier: When the weights α and β are set to be zero, and the set $\mathcal{V}_1$ is the set of nodes in the graph of the bottom layer of the graphical model, (4.5) becomes

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}_1} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) \tag{4.20}$$

which is the energy function associated with the region classifier.

Flat CRF: When the weight β is set to be zero, the set $\mathcal{V}_1$ is the set of nodes in the graph of the bottom layer, and $\mathcal{E}_1$ is the set of pairs collecting the neighbouring nodes in the bottom layer, (4.5) becomes

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}_1} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}_1} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) \tag{4.21}$$

which is the energy function associated with the flat CRF defined over the image regions.

Hierarchical CRF: If E_3 is chosen as a hierarchical pairwise potential in (4.5), the graphical model only consists undirected edges. The energy function reads

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) + \beta \sum_{\{i,k\} \in \mathcal{S}} E_{3,h}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{4.22}$$

which is the energy function associated with the hierarchical CRF defined over the multi-scale of the image regions.

Hierarchical mixed graphical model: If E_3 is chosen as the conditional probability energy in (4.5), the graphical model consists both undirected edges and directed edges. The energy function reads

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) + \beta \sum_{(i,k) \in \mathcal{S}} E_{3,c}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \quad (4.23)$$

which is the energy function associated with the hierarchical mixed graphical model.

Chapter 5

Experimental Results

A thousand miles begins with a single step.

-*Lao Tzu* (600 B.C. - 470 B.C.)

5.1 Overview

In this chapter, we will show that the framework for scene interpretation developed in Chapter 4 allows for significantly better classification results than the standard classical local classification approach on man-made scenes by incorporating spatial and hierarchical structures. We will investigate the performance of the algorithm on a public database, namely the eTRIMS dataset (Korč & Förstner, 2009), to show the relative importance of information from the spatial structure and the hierarchical structure. We will also see that the graphical model can provide a consistent model representation, and therefore appears to be the right tool for our task.

We will consider a classification result better than another one in terms of the classification accuracy. The results are evaluated by average classification accuracy across all classes. The classification accuracy for a class is given by

$$\text{classification accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}} \tag{5.1}$$

where tp, tn, fp, and fn refer to *true positives, true negatives, false positives, and false negatives*, respectively.

We rewrite the energy function for the statistical model for the interpretation problem in Chapter 4 as follows

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) + \beta \sum_{\langle i,k \rangle \in \mathcal{S}} E_3(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{5.2}$$

The set $\mathcal{V}$ is the set of the nodes in the complete graph, and the set $\mathcal{E}$ is the set of pairs collecting the neighbouring nodes within each scale. $\mathcal{S}$ is the set of pairs collecting the parent-child relations between regions with the neighbouring scales, where $\langle i, k \rangle$ denotes nodes i and k are connected by either a undirected edge or a directed edge. This stochastic model is illustrated in Fig. 4.2 on page 36.
Let us consider the case with only one layer segmentation of the image (the bottom layer of the graphical model in Fig. 4.2). When the weights α and β are set to be zero, and the set $\mathcal{V}_1$ is the set of nodes in the graph of the bottom layer, (5.2) becomes

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}_1} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) \tag{5.3}$$

which is the energy function associated with the region classifier.
When the weight β is set to be zero, the set $\mathcal{V}_1$ is the set of nodes in the graph of the bottom layer, and $\mathcal{E}_1$ is the set of pairs collecting the neighbouring nodes in the bottom layer, (5.2) becomes

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}_1} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}_1} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) \tag{5.4}$$

which is the energy function associated with the flat CRF defined over the image regions.
Let us now consider the case with the multi-scale segmentation of the image. If E_3 is chosen as a hierarchical pairwise potential in (5.2), the energy function reads

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) + \beta \sum_{\{i,k\} \in \mathcal{S}} E_{3,h}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{5.5}$$

which is the energy function associated with the hierarchical CRF defined over the multi-scale of the image regions.
If E_3 is chosen as the conditional probability energy in (5.2), the energy function reads

$$E(\boldsymbol{x} \mid \boldsymbol{d}) = \sum_{i \in \mathcal{V}} E_1(\boldsymbol{x}_i \mid \boldsymbol{d}) + \alpha \sum_{\{i,j\} \in \mathcal{E}} E_2(\boldsymbol{x}_i, \boldsymbol{x}_j \mid \boldsymbol{d}) + \beta \sum_{(i,k) \in \mathcal{S}} E_{3,c}(\boldsymbol{x}_i, \boldsymbol{x}_k \mid \boldsymbol{d}) \tag{5.6}$$

which is the energy function associated with the hierarchical mixed graphical model.
The features used for region classifier are basic features, colour features, Peucker features, texture features, and SIFT features, which is listed in Table 4.1 on page 43. The formulations of unary potential, local pairwise potential, hierarchical pairwise potential, and conditional probability energy are described in Section 4.4.

We will start by describing the setup for the following experiments. We introduce one specific image database of man-made scenes. In the interpretation workflow described in Fig. 4.1, image segmentation serves as pre-step for the system. We use two segmentation methods, namely the watershed algorithm by Vincent & Soille (1991)

and the mean shift algorithm by Comaniciu & Meer (2002), to demonstrate the role of the initial segmentation algorithms in the final classification results. Then we show the region classification results using a random forest classifier as a baseline. Incorporated with the spatial and hierarchical structures, we show the hierarchical CRF produces better classification results than both the region classifier and the flat CRF. In the end of this chapter, we will demonstrate the applicability of the graphical model for scene interpretation. We want to show that the hierarchical mixed graphical model results are comparable to the results obtained with the hierarchical CRF.

We conduct the experiments to evaluate the performance of the proposed model on eTRIMS dataset (Korč & Förstner, 2009). In all experiments, we take the ground truth label of a region to be the majority vote of the ground truth pixel labels. At the test stage, to ensure no bias in favor of our method, we compute our accuracy at the pixel level.

5.2 Experimental setup

5.2.1 Image database

We use the eTRIMS dataset (Korč & Förstner, 2009) to evaluate the image interpretation of man-made scenes in terms of building facade image region classification accuracy. The dataset is a collection of annotated images of street scenes from various European cities including: Basel, Berlin, Bonn, and Heidelberg. Several example images are shown in Fig. 1.3 on page 4. Ground truth annotation is provided on the pixel level. Each image pixel is assigned with a class label. The ground truth labelling is approximate, with foreground labels often overlapping the background objects.

There are 60 annotated images in the eTRIMS dataset. We consider all eight object classes: *building*, *car*, *door*, *pavement*, *road*, *sky*, *vegetation*, *window*. These classes are the typical objects which can appear in the images of building facades. In the experiments, we randomly divide the images into a training set with 40 images and a testing set with 20 images. Table 5.1 summarizes the number of the objects and the images for each annotated class. In total, there are 1702 annotated objects in the dataset.

The dataset is comprised of the images and the corresponding ground truth. An example image with ground truth labelling from the dataset is shown in Fig. 5.1. Ground truth is created by human interpretation of the images, it refers to the appearance of the objects in the images, not to their 3D-structure. Therefore, occluded parts of an object are not annotated as part of an object. Furthermore, the window region in a building region is not annotated as part of an building object (cf. Fig. 5.1 (b)). Ground truth labels each pixel with the ground truth class or background. The ground truth is represented as an indexed image. The pixel values $1, 2, 3, \ldots, 8$ correspond to class names in the alphabetical order (1=*building*, 2=*car*, 3=*door*, 4=*pavement*, 5=*road*, 6=*sky*, 7=*vegetation*, 8=*window*). The pixel value 0 corresponds to background. More example images with ground truth labelling from the dataset are shown in Fig. 5.2.

Table 5.1: Statistics of the 8–Class eTRIMS dataset.

Class Name	**Images**	**Objects**
Building	60	142
Car	27	67
Door	53	85
Pavement	56	76
Road	49	51
Sky	60	71
Vegetation	56	194
Window	60	1016
Total	60	1702

Note that the ground truth labelling is not pixel accurate (cf. auxiliary visualization of the object boundaries in Fig. 5.1 (c)).

(a) (b) (c)

Figure 5.1: An example image with ground truth labelling from the eTRIMS dataset. (a) Example image. (b) Ground truth showing *building, car, door, pavement, road, sky, vegetation, window* labels. The black region corresponds to background. (c) Visualization of ground truth object boundaries with polygons in pink colour.

5.2.2 Segmentation algorithms

In the experiments, our graphical model works on the region level. A region is defined by the boundary of an image partition, where each pixel only belongs to one region. Therefore, the initial unsupervised segmentation algorithms may play an important role in the final classification results. The result of image segmentation is a set of

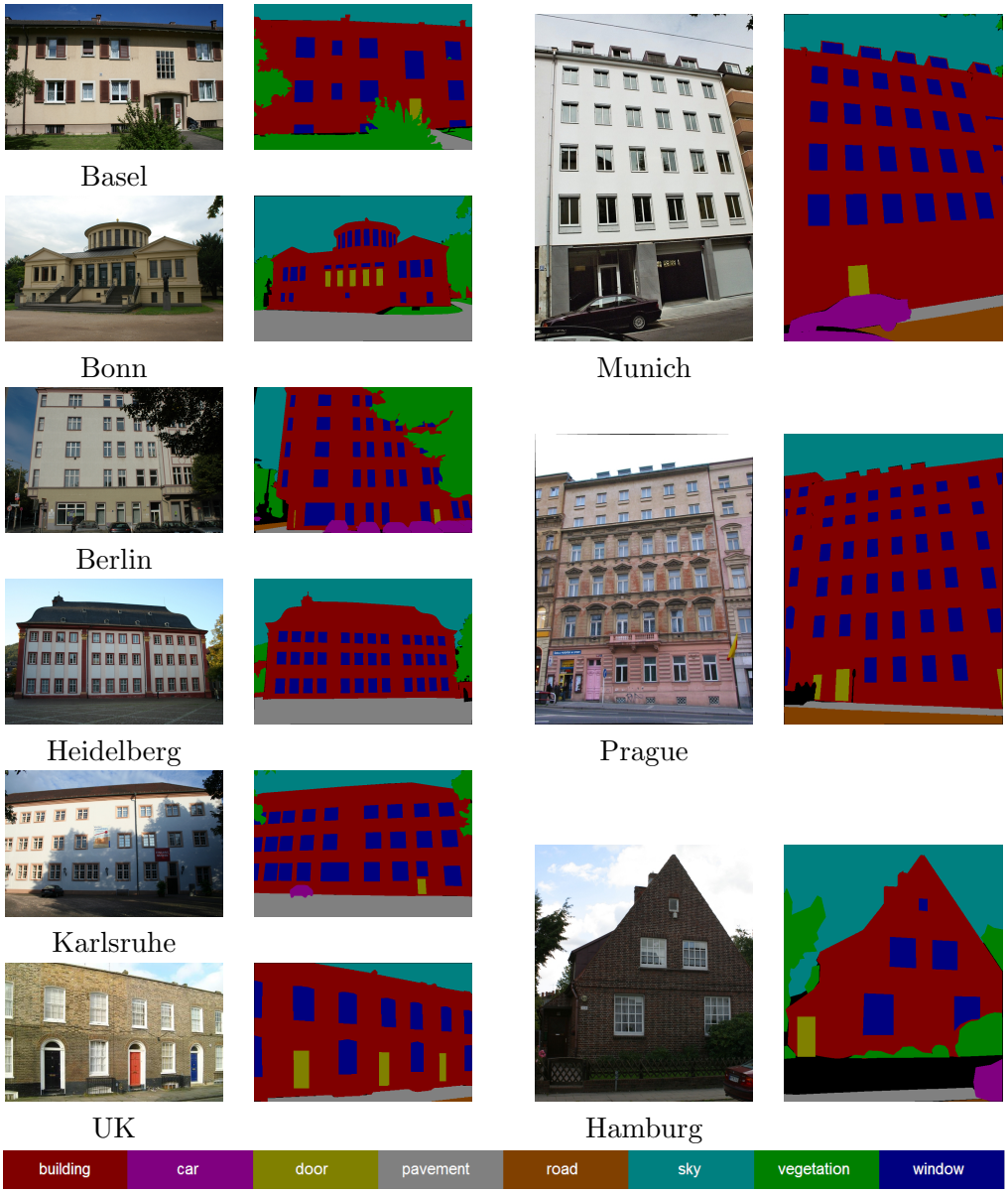


Figure 5.2: Example images from the 8–Class eTRIMS dataset (Korč & Förstner, 2009). Column 1 and 3 show the example images. Column 2 and 4 show the ground truth with *building, car, door, pavement, road, sky, vegetation, window* labels. The bottom row is the Legend. City names of origin are given below the example images.

segmented regions that cover the entire image. To test how much the influence of the segmentation algorithms would be, we employ two different segmentation methods, namely the watershed algorithm (Vincent & Soille, 1991) and the mean shift algorithm (Comaniciu & Meer, 2002), each of which has two variants, namely a baseline version and a multi-scale version.

5.2.2.1 Baseline watershed

We segment the images using the watershed method (Vincent & Soille, 1991), which turns out to give approximately 900 regions per image. As a result, we obtain an image partition, where each pixel only belongs to one region. In all 60 images, we extract around 56 000 regions. We take the ground truth label of a region to be the majority vote (above 50%) of the ground truth pixel labels. We have following statistics. Almost 34% of all the segmented regions get the class label *building*. 28% of all regions get the class label *window*. These statistics are very comprehensive, because the facade images show the facades typically contain many windows. Furthermore, 23% of the regions get the class label *vegetation*, 2% belong to *sky*, and the last 13% of the regions are spread over most of other classes. Table 5.2 summarizes the statistics for the percentage of each class label, the average size of the region of each class, and the percentage of the image covered by each class for the baseline watershed segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009).

Table 5.2: Statistics of the percentage of each class label, the average size of the region of each class, and the percentage of the image covered by each class for the baseline watershed segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

	Baseline watershed							
	b	c	d	p	r	s	v	w
class percentage	34	4	1	2	2	2	23	28
average size of region	614	268	477	684	1490	4096	209	152
class covering percentage	49	3	1	4	6	16	11	10

5.2.2.2 Baseline mean shift

We segment the images using the mean shift algorithm (Comaniciu & Meer, 2002), tuned to give approximately 480 regions per image. In all 60 images, we extract around 30 000 regions. We have following statistics. Compared to the ground truth labelling, almost 36% of all the segmented regions get the class label *building*. 26% of all regions get the class label *window*. Furthermore, 21% of the regions get the class label *vegetation*, and 2% belong to *sky*, and the last 15% of the regions are spread over most of other classes. Table 5.3 summarizes the statistics for the percentage of each class label,

the average size of the region of each class, and the percentage of the image covered by each class for the baseline mean shift segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009).

Table 5.3: Statistics of the percentage of each class label, the average size of the region of each class, and the percentage of the image covered by each class for the baseline mean shift segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

	Baseline mean shift							
	b	c	d	p	r	s	v	w
class percentage	36	5	2	2	2	2	21	26
average size of region	1014	424	569	1671	2563	6741	380	310
class covering percentage	48	3	1	4	6	16	11	11

5.2.2.3 Multi-scale watershed

We segment the images using the multi-scale watershed method (Drauschke, 2009) on the smoothed version of the original image, tuned to give approximately 1000 regions per image counting all scales. We determine the segmentation from the boundaries on the image's gradient magnitude, and then use the Gaussian scale space for obtaining the regions at several scales, which has been described by Drauschke *et al.* (2006). For each scale, we convolve each image channel with a Gaussian filter and combine the channels to compute the gradient magnitude. We determine the scale-specific neighbourhood graph on each image partition by the spatial arrangement (cf. Fig. 3.5 *Middle*). In all 60 images, we extract around 62 000 regions. We use three layers in the scale space for the experiments. The bottom layer often contains 900 or more regions, and the number decreases down to 15 in the top layer. Three layers are connected via a region hierarchy. The development of the regions over the scales is used to model the region hierarchy. Furthermore, the relation is defined over the maximal overlap of the regions (cf. Fig. 4.2). Multi-scale watershed segmentation results of one example image in eTRIMS dataset are shown in Fig. 5.3, where the region boundaries are superimposed on the smoothed versions at three different scales of the example image. Table 5.4 summarizes the statistics for the percentage of each class label, the average size of the region of each class, and the percentage of the image covered by each class for the multi-scale watershed segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009).

5.2.2.4 Multi-scale mean shift

Our approach uses the Gaussian scale-space for obtaining the regions at several scales. For each scale, we convolve each image channel with a Gaussian filter and apply the

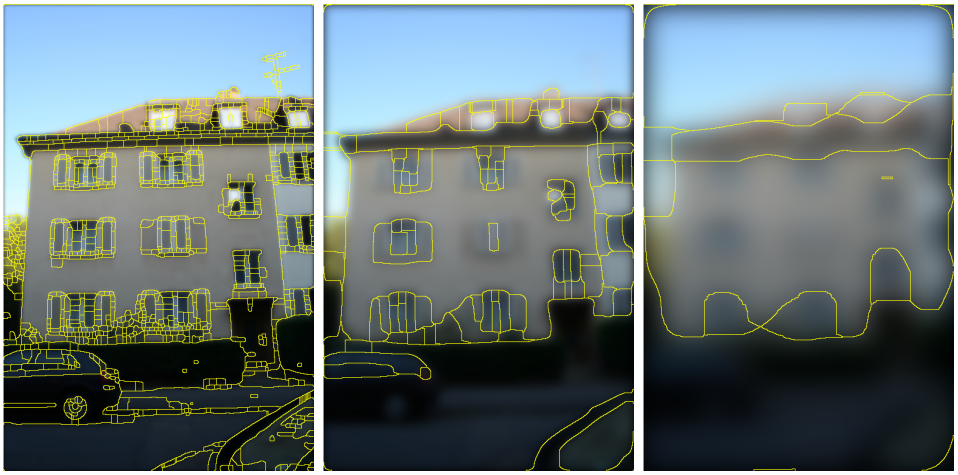


Figure 5.3: Multi-scale watershed segmentation (Drauschke, 2009) results of an example image. From *left* to *right*: the segmentation results at scale 1, 2, 3, respectively. Region boundaries, shown in *yellow*, are superimposed on the smoothed versions at different scales of the original image.

Table 5.4: Statistics of the percentage of each class label, the average size of the region of each class, and the percentage of the image covered by each class for the multi-scale watershed segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

	Multi-scale watershed							
	b	c	d	p	r	s	v	w
class percentage	34	4	1	2	2	2	22	28
average size of region	1613	449	816	1140	2432	7887	427	254
class covering percentage	48	2	1	2	4	16	8	6

mean shift algorithm (Comaniciu & Meer, 2002) to segment the smoothed image. As a result of the mean shift algorithm, we obtain a complete partitioning of the image for each scale, where every image pixel belongs to exactly one region. We determine the scale-specific neighbourhood graph on each image partition by the spatial arrangement (cf. Fig. 3.5 *Middle*). In all 60 images, we extract around 61 000 regions. We use three layers in the scale space for the experiments. The bottom layer often contains around 500 regions, and the number decreases down to 200 in the top layer. Three layers are connected via a region hierarchy. The development of the regions over the scales is used to model the region hierarchy. Furthermore, the relation is defined over the maximal overlap of the regions (cf. Fig. 4.2). Fig. 5.4 shows the region results of one example image in eTRIMS dataset from the multi-scale mean shift segmentation, where the colour of each region is assigned randomly that the neighbouring regions are likely to have different colours.

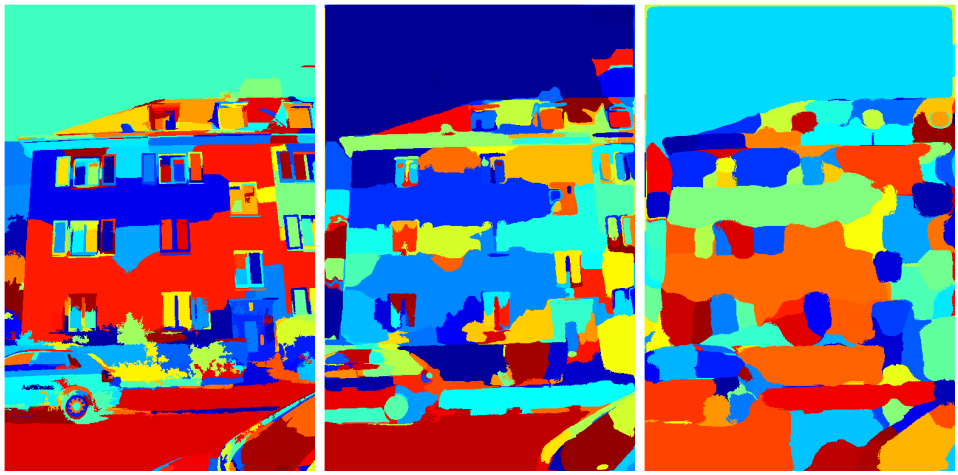


Figure 5.4: The region images of the mean shift (Comaniciu & Meer, 2002) segmentation result at scale 1, 2, 3, respectively. The colour of each region is assigned randomly that the neighbouring regions are likely to have different colours.

Table 5.5 summarizes the statistics for the percentage of each class label, the average size of the region of each class, and the percentage of the image covered by each class for the multi-scale mean shift segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009).

Table 5.5: Statistics of the percentage of each class label, the average size of the region of each class, and the percentage of the image covered by each class for the multi-scale mean shift segmentation in the 8-Class eTRIMS dataset (Korč & Förstner, 2009). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

	Multi-scale mean shift							
	b	c	d	p	r	s	v	w
class percentage	36	5	2	2	2	3	20	24
average size of region	1507	639	750	2102	3150	5239	633	473
class covering percentage	47	3	1	4	6	16	11	10

5.3 Results for the baseline region classifier

In this section, we present the experimental results for a RDF classifier as a baseline with both baseline mean shift segmentation and baseline watershed segmentation.

5.3.1 Results with baseline mean shift and the RDF classifier

We give the RDF classification results on the regions from the baseline mean shift segmentation with all the feature sets from the images in the eTRIMS dataset (Korč & Förstner, 2009). The feature sets are basic features $\boldsymbol{h}_1$, colour features $\boldsymbol{h}_2$, Peucker features $\boldsymbol{h}_3$, texture features $\boldsymbol{h}_4$, and SIFT features $\boldsymbol{h}_5$ (cf. Section 4.4.1). We run experiments five times, and obtain overall averaging classification accuracy 58.8%. The number of the decision trees is chosen as $T = 250$. Fig. 5.5 *Left* shows the classification results over all 8 classes. The classification accuracy with respect to the numbers of the decision trees T for training are shown in Fig. 5.5 *Right*. While increasing the number of the decision trees, the classification accuracy also increases. After 250 iteration, the accuracy converges. So we choose $T = 250$ for performing the experiments.

To emphasize the importance of the each feature set, we give the RDF classification results on the regions from the baseline mean shift segmentation with the each feature set. The overall classification accuracy is listed in Table 5.6, when applying the RDF classifier on each feature set. The number of the decision trees is chosen as $T = 250$. A random classifier for 8 classes, the expected classification accuracy is 12.5%.

Table 5.6: Average accuracy using a randomized decision forest (RDF) classifier with the baseline mean shift segmentation on each feature set of eTRIMS dataset (Korč & Förstner, 2009). The feature sets are basic features $\boldsymbol{h}_1$, colour features $\boldsymbol{h}_2$, Peucker features $\boldsymbol{h}_3$, texture features $\boldsymbol{h}_4$, and SIFT features $\boldsymbol{h}_5$.

feature set	$\boldsymbol{h}_1$	$\boldsymbol{h}_2$	$\boldsymbol{h}_3$	$\boldsymbol{h}_4$	$\boldsymbol{h}_5$
accuracy	43.8%	49.6%	40.9%	27.9%	54.1%

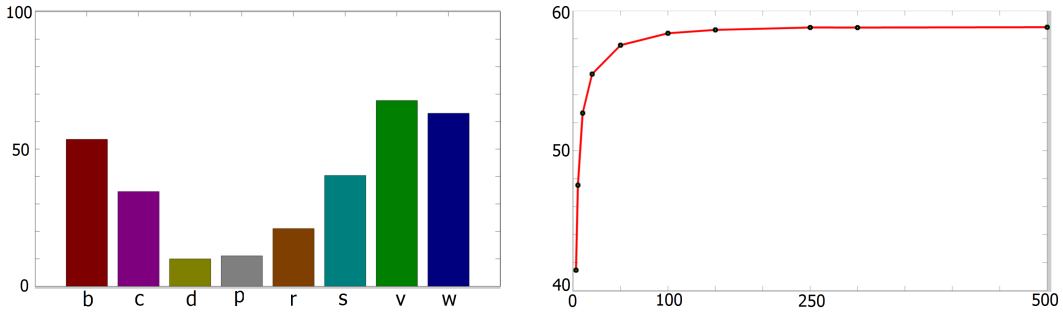


Figure 5.5: The classification accuracy of each class of the RDF classifier with baseline mean shift and the accuracy with respect to the numbers of the decision trees. *Left*: the classification accuracy of each class of the RDF classifier with baseline mean shift on the feature sets $\boldsymbol{h}$. *Right*: the RDF classification accuracy with respect to the numbers of the decision trees for training. (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Fig. 5.6 presents some result images of the RDF method. The black regions in all the result images and ground truth images correspond to background. The qualitative inspection of the results in Fig. 5.6 shows that the RDF classifier yields some reasonable results. There exists some misclassification for each class. For example, the incorrect results at windows are often due to the reflectance of vegetation and sky in the window panes. A sky region is assigned label *car* in one image (cf. the third column in Fig. 5.6). This can be resolved simply by introducing some kind of the spatial prior (Gould *et al.*, 2008), such as *sky* is above the *building*, *road* and *pavement* are below the *building*, *car* is above the *road*, and *window* is surrounded by *building*. A full confusion matrix summarizing the RDF classification results over all 8 classes is given in Table 5.7, showing the performance of this method.

Here, the features are extracted at a local scale. The classification results are achieved from bottom up on these local features by the classifier, which leads to incorrect labelling and noisy boundaries in the test images. To enforce consistency, a Markov or conditional random field (Shotton *et al.*, 2006) is often introduced for refinement, which will likely improve the performance (cf. Section 5.4).

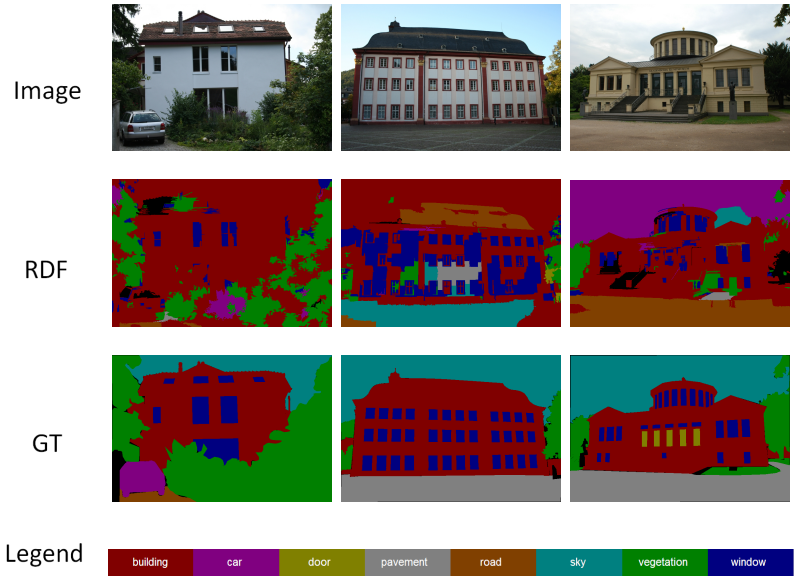


Figure 5.6: Qualitative classification results of a RDF classifier with the baseline mean shift on the testing images from the eTRIMS dataset (Korč & Förstner, 2009). (1st-row) Testing images. (2nd-row to 3rd-row) Classification results using the RDF classifier (RDF), and the ground truth (GT), respectively. (4th-row) Legend.

Table 5.7: Accuracy of RDF classifier with the baseline mean shift segmentation on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**60**	8	2	2	2	1	9	16
c	22	**40**	1	3	1	2	29	2
d	46	0	**15**	0	0	0	8	31
p	40	16	0	**12**	4	4	16	8
r	40	20	0	14	**23**	3	0	0
s	29	2	0	5	2	**48**	7	7
v	11	5	1	1	1	0	**76**	5
w	24	1	2	0	0	1	4	**68**

5.3.2 Results with baseline watershed and the RDF classifier

To test whether the classification result mainly benefits from the mean shift segmentation method, and not from the feature sets we use, we also employ another unsupervised segmentation method, namely the watershed algorithm by Vincent & Soille (1991), to segment the facade images.

The overall classification accuracy is 55.4%, with the RDF classifier on all the feature sets $\boldsymbol{h}$ and the number of the decision trees chosen as $T = 250$. The confusion matrix is given in Table 5.8.

In comparison with Table 5.7, the accuracy for each class remains similar, which shows that the type of finding image regions from the image segmentation algorithms is not critical and the low classification performance results from the lack of either good features or contextual information.

Table 5.8: Pixelwise accuracy of the image classification using the RDF classifier and the watershed segmentation on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**59**	4	1	3	5	9	11	7
c	67	**21**	0	5	2	0	3	2
d	19	0	**12**	0	0	0	62	7
p	57	3	0	**9**	30	0	0	1
r	14	1	0	58	**23**	1	3	1
s	17	0	0	6	0	**73**	2	1
v	13	4	1	2	1	13	**61**	4
w	29	1	1	1	0	6	3	**57**

5.4 Results for the hierarchical CRF

The hierarchical CRF model is defined over the multi-scale of the image regions when we choose E_3 as a pairwise potential in (4.5) on page 39, the corresponding energy function is shown in (5.5). In this section, we present the experimental results for the hierarchical CRF with both multi-scale mean shift segmentation and multi-scale watershed segmentation, and the comparison with the baseline RDF region classification results and the flat CRF classification results.

5.4.1 Results with multi-scale mean shift and the hierarchical CRF

Fig. 5.7 shows the classification results for the hierarchical CRF with the multi-scale mean shift segmentation.

Figure 5.7: One example of the classification results using the hierarchical CRF from 3-scale mean shift segmentation. From *left* to *right*: the classification result at scale 1, 2, 3, respectively.

Table 5.10 shows the confusion matrix obtained by applying the hierarchical CRF to the whole test set. Accuracy values in the table are computed as the percentage of the image pixels assigned to the correct class label, ignoring the pixels labelled as void in the ground truth. The overall classification accuracy is 69.0%. The weighting parameter settings, learned by cross validation on the training data, are $\alpha = 0.1$, $\beta = 0.65$. For comparison, the baseline RDF classifier alone gives an overall accuracy of 58.8% (cf. Section 5.3.1), and the flat CRF ($\alpha = 0.8$, $\beta = 0$) gives an overall accuracy of 65.8% (Yang & Förstner, 2011c). Therefore, the hierarchical potential increases the accuracy by 3.2%. This seemingly small numerical improvement corresponds to a large perceptual improvement (cf. Fig. 5.8).

Compared to the confusion matrix showing the flat CRF with with the baseline mean shift in Table 5.9 (Yang & Förstner, 2011c), the hierarchical CRF performs significantly better on *pavement*, *vegetation*, *road*, and *window* classes, slightly better on *car* and *sky* classes, and slightly worse on *building* and *door* classes.

Qualitative results of the hierarchical CRF with the multi-scale mean shift on the eTRIMS dataset are presented in Fig. 5.8. The qualitative inspection of the results in these images shows that the hierarchical CRF yields large improvement over the baseline RDF region classification results and the flat CRF classification results. The greatest accuracies are for classes which have low visual variability and many training examples (such as window, vegetation, building, and sky) whilst the lowest accuracies are for classes with high visual variability or few training examples (for example door, car, and pavement). We expect more training data and the use of features with better invariance properties will improve the classification accuracy. Objects such as car, door, pavement, and window are sometimes incorrectly classified as *building*, due to the dominant presence of the building in the image. Detecting windows, cars, and doors should resolve some of such ambiguities.

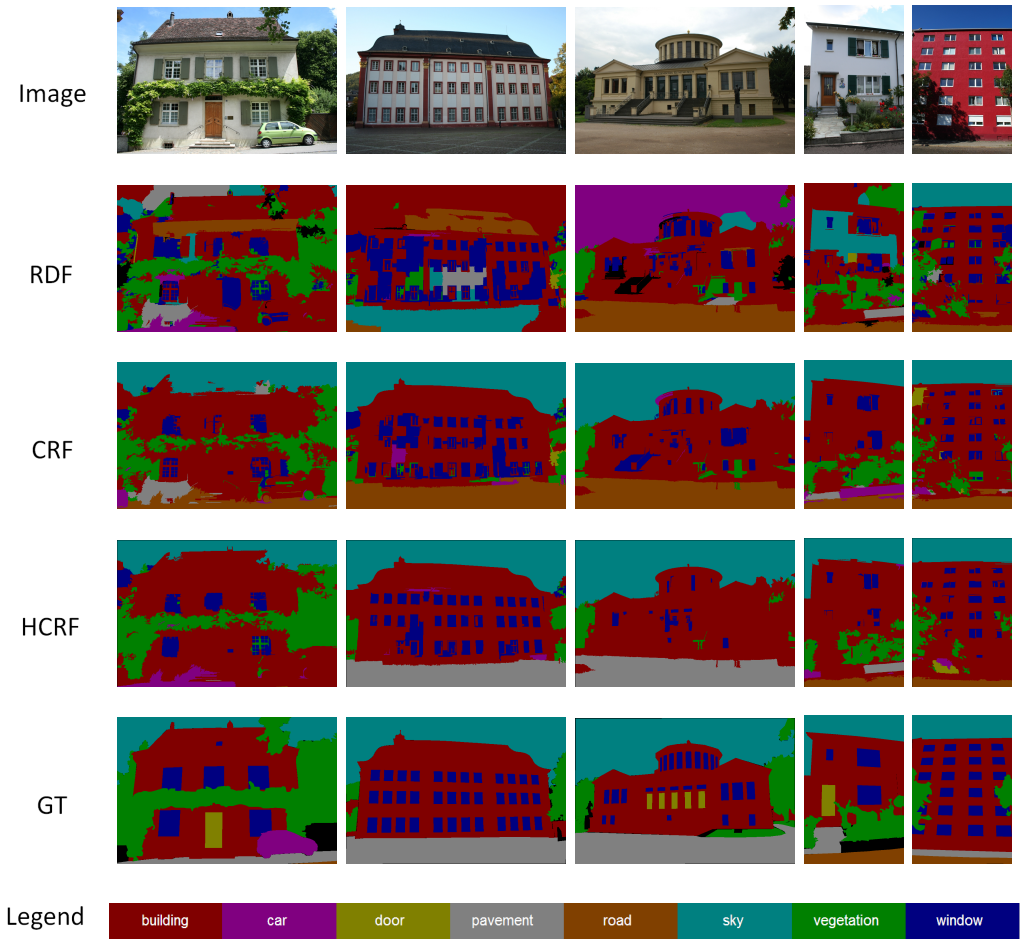


Figure 5.8: Qualitative classification results of the hierarchical CRF with the multi-scale mean shift segmentation on the testing images from the eTRIMS dataset (Korč & Förstner, 2009). The qualitative inspection of the results in these images shows that the hierarchical CRF yields large improvement over the flat CRF results and the RDF region classifier results. (1st-row) Testing images. (2nd-row to 5th-row) Classification results using the RDF region classifier (RDF), the flat CRF model (CRF) (Yang & Förstner, 2011c), the hierarchical CRF model (HCRF), and the ground truth (GT), respectively. (6th-row) Legend.

Table 5.9: Pixelwise accuracy of the image classification using the flat CRF (Yang & Förstner, 2011c) with the baseline mean shift segmentation on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**71**	2	1	1	1	2	10	12
c	12	**35**	0	12	11	0	30	0
d	42	0	**16**	1	6	0	8	27
p	11	15	0	**22**	36	0	14	2
r	4	8	0	44	**35**	0	9	0
s	13	0	0	0	0	**78**	8	1
v	18	5	2	1	1	0	**66**	7
w	19	1	1	0	0	1	3	**75**

Table 5.10: Pixelwise accuracy of the image classification using the hierarchical CRF with the multi-scale mean shift segmentation on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**67**	3	1	4	5	1	8	11
c	17	**36**	0	11	9	0	26	1
d	50	5	**14**	8	0	0	7	16
p	6	4	0	**85**	1	0	4	0
r	0	11	0	21	**53**	0	15	0
s	11	0	0	0	0	**80**	8	1
v	9	5	1	0	1	0	**78**	6
w	15	0	1	0	0	2	2	**80**

5.4.2 Results with multi-scale watershed and the hierarchical CRF

With multi-scale watershed segmentation, Table 5.12 shows the confusion matrix obtained by applying the hierarchical CRF to the whole test set. Accuracy values in the table are computed as the percentage of image pixels assigned to the correct class label, ignoring pixels labelled as void in the ground truth. The overall classification

accuracy is 65.3%. The weighting parameter settings, learned by cross validation on the training data, are $\alpha = 0.8$, $\beta = 0.1$. For comparison, the RDF classifier alone gives an overall accuracy of 55.4%, and the flat CRF ($\alpha = 1.08$, $\beta = 0$) gives an overall accuracy of 61.8% (Yang & Förstner, 2011c). Therefore, the location, local pairwise, and hierarchical potentials increase the accuracy by 7%. Compared to the confusion matrix showing the flat CRF with with the baseline watershed in Table 5.11, the hierarchical CRF gains better accuracy on *building*, *car*, *sky*, *vegetation*, and *window* classes.

Table 5.11: Pixelwise accuracy of the image classification using the flat CRF with baseline watershed on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**66**	2	1	2	3	3	12	11
c	44	**10**	2	9	23	0	7	5
d	35	0	**13**	0	1	0	36	15
p	26	8	1	**52**	5	0	4	4
r	22	10	1	15	**38**	0	13	1
s	10	0	0	0	0	**78**	10	2
v	28	11	2	2	1	0	**48**	8
w	20	1	2	0	0	0	2	**75**

Qualitative results of the hierarchical CRF on the eTRIMS dataset are presented in Fig. 5.9. The qualitative inspection of the results in these images shows that the hierarchical CRF yields large improvement over the baseline RDF region classification results and the flat CRF classification results. However, some misclassification still exists. For example, 11% of pavement is misclassified as road, and 42% of road is misclassified as pavement (cf. Table 5.12). Objects such as pavement and road can be confused with each other. This effect is partially attributable to inaccuracies in the manual ground truth labelling, where pixels are often mislabelled near object boundaries. Pavement and road have the similar appearance, therefore, no discriminative features have been found to distinguish them.

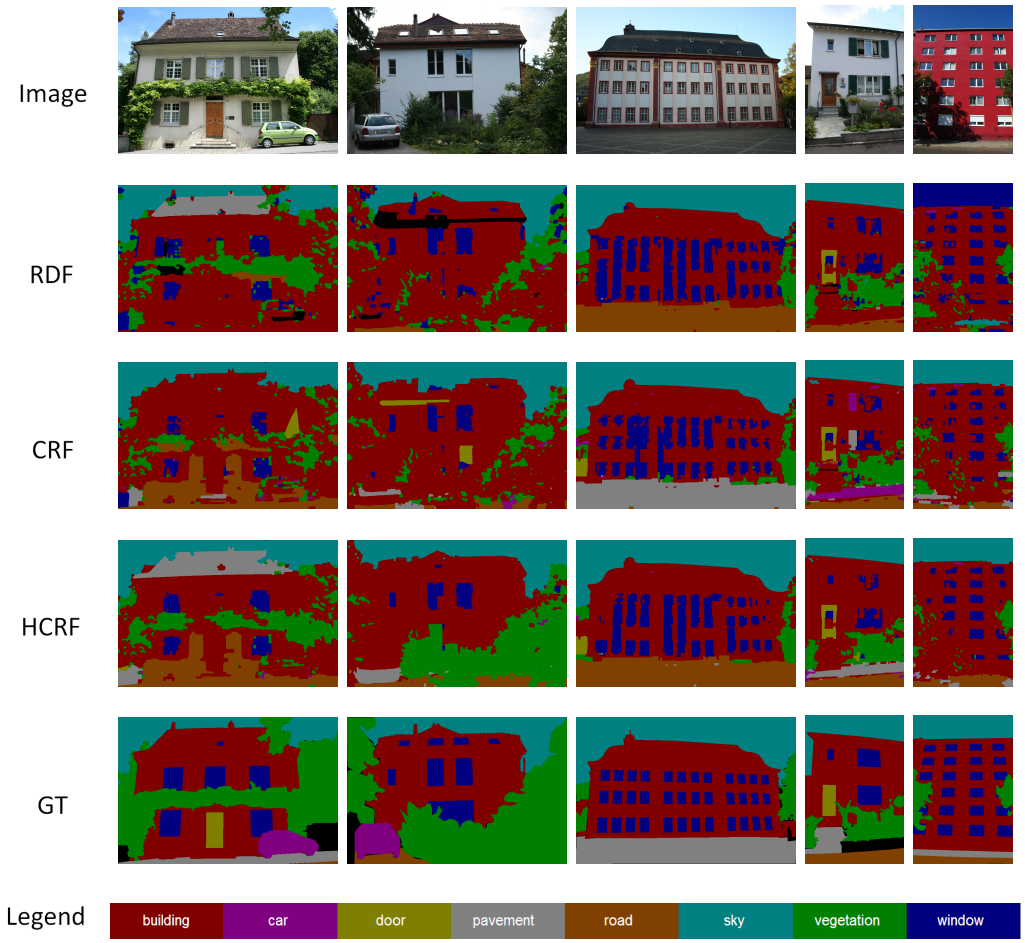


Figure 5.9: Qualitative classification results of the hierarchical CRF with the multi-scale watershed segmentation on the testing images from the eTRIMS dataset (Korč & Förstner, 2009). The qualitative inspection of the results in these images shows that the hierarchical CRF yields large improvement over the flat CRF results and the RDF region classifier results. (1st-row) Testing images. (2nd-row to 5th-row) Classification results using the RDF region classifier (RDF), the flat CRF model (CRF), the hierarchical CRF model (HCRF), and the ground truth (GT), respectively. (6th-row) Legend.

Table 5.12: Pixelwise accuracy of the image classification using the hierarchical CRF with multi-scale watershed on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**67**	4	1	3	3	3	8	11
c	48	**34**	1	7	6	0	3	1
d	26	0	**9**	0	2	0	59	4
p	49	8	0	**17**	11	0	11	4
r	8	5	0	42	**31**	0	14	0
s	9	0	0	0	0	**81**	9	1
v	11	4	1	1	1	0	**79**	3
w	20	0	1	0	0	0	1	**78**

5.5 Results for the hierarchical mixed graphical model

The hierarchical mixed graphical model is defined over the multi-scale of the image regions when we choose E_3 as the conditional probability energy in (4.5) on page 39, the corresponding energy function is shown in (5.6). In this section, we first calculate the conditional probability tables for the energy term. Then, we present the experimental results for the hierarchical mixed graphical model with both multi-scale mean shift segmentation and multi-scale watershed segmentation, and the comparison with the baseline region classifier, the flat CRF, and the hierarchical CRF classification results.

5.5.1 Conditional probability tables

Following the learning procedure presented in Section 4.5.3, we derive the conditional probability tables (CPTs).

The two tables corresponding to the three layers of the multi-scale mean shift segmentation on the training data of eTRIMS dataset (Korč & Förstner, 2009) are presented in Table 5.13 and Table 5.14. We obtain the following information regarding the probability tables. They have each $8 \times 8 = 64$ elements. All two tables have many elements equal zero or almost equal zero, which means that the relationship between two classes does not occur at all.

For the image regions resulting from the multi-scale watershed segmentation, the CPT of 1st layer and 2nd layer is given in Table 5.15. For example, if we have given a *building* region, then the probability for the target of one of its children is 0.88 for representing a *building* as well, but we find a *window* as child with a probability of 0.35.

Table 5.13: Conditional probability table (CPT) of 1st layer and 2nd layer of the multi-scale mean shift segmentation. The table shows the conditional probability for each class (rows) given its parent and is row-normalized to sum to 100%. Column labels indicate the parent class, and row labels the given class. (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

$\underline{\boldsymbol{x}}_i$ \ $\underline{\boldsymbol{x}}_k$	b	c	d	p	r	s	v	w
b	95	0	1	0	0	0	1	3
c	1	96	0	0	2	0	0	1
d	4	0	94	0	1	0	1	0
p	5	2	0	83	9	0	1	0
r	0	2	0	4	93	0	1	0
s	1	0	0	0	0	99	0	0
v	3	0	0	0	0	0	96	1
w	11	0	0	0	0	0	1	88

Table 5.14: Conditional probability table (CPT) of 2nd layer and 3rd layer of the multi-scale mean shift segmentation. The table shows the conditional probability for each class (rows) given its parent and is row-normalized to sum to 100%. Column labels indicate the parent class, and row labels the given class. (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

$\underline{\boldsymbol{x}}_i$ \ $\underline{\boldsymbol{x}}_k$	b	c	d	p	r	s	v	w
b	88	0	0	0	0	9	2	1
c	88	9	0	0	0	0	0	3
d	49	0	44	0	0	0	7	0
p	56	0	0	44	0	0	0	0
r	21	0	0	0	69	0	10	0
s	15	0	0	0	0	80	5	0
v	37	0	1	0	0	8	52	2
w	78	0	0	0	0	2	0	20

Table 5.15: Conditional probability table (CPT) of 1st layer and 2nd layer of the multi-scale watershed segmentation. The table shows the conditional probability for each class (rows) given its parent and is row-normalized to sum to 100%. Column labels indicate the parent class, and row labels the given class. (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

$\underline{\boldsymbol{x}}_i$ \ $\underline{\boldsymbol{x}}_k$	b	c	d	p	r	s	v	w
b	88	0	1	1	1	4	2	3
c	16	78	0	0	1	0	4	1
d	13	0	81	2	1	0	3	0
p	16	1	0	65	13	0	5	0
r	12	4	0	9	65	0	10	0
s	2	0	0	0	1	94	3	0
v	12	0	1	0	0	2	85	0
w	35	0	0	0	1	0	1	63

5.5.2 Results with multi-scale mean shift and the hierarchical mixed graphical model

Table 5.16 shows the confusion matrix obtained by applying the hierarchical mixed graphical model to the whole test set. The overall classification accuracy is 68.9%. The weighting parameters, learned by cross validation on the training data, are $\alpha = 0.8$, $\beta = 1$. For comparison, the RDF region classifier gives an overall accuracy of 58.8%, the flat CRF gives an overall accuracy of 65.8%, and the hierarchical CRF gives an overall accuracy of 69.0%.

Qualitative results of the hierarchical mixed graphical model with the multi-scale mean shift on the eTRIMS dataset (Korč & Förstner, 2009) are presented in Fig. 5.10. The qualitative inspection of the results in these images shows that the hierarchical mixed graphical model yields significant improvement. The hierarchical mixed graphical model yields more accurate and cleaner results than the flat CRF and the RDF region classifier, and comparable to the hierarchical CRF model.

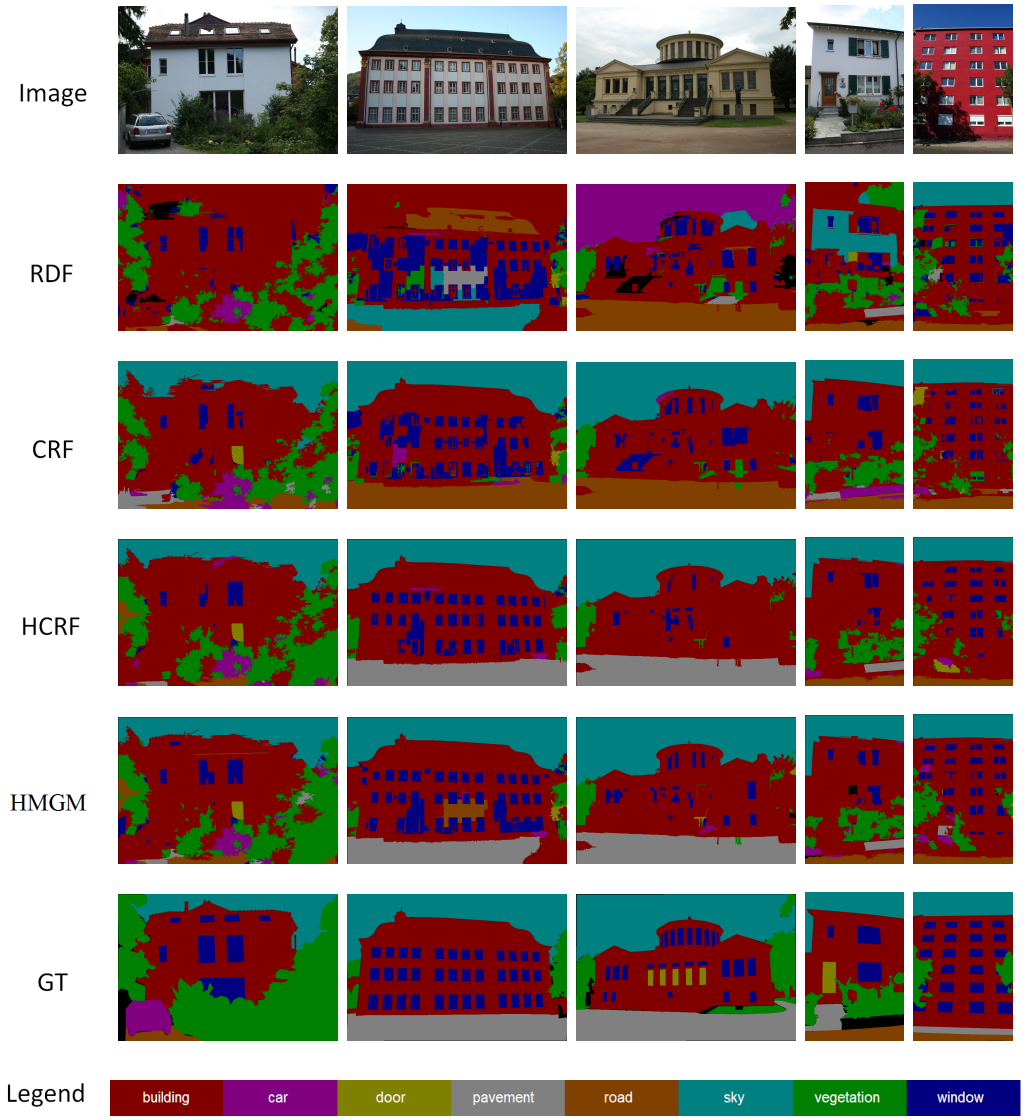


Figure 5.10: Qualitative classification results of the hierarchical mixed graphical model with the multi-scale mean shift segmentation on the testing images from the eTRIMS dataset (Korč & Förstner, 2009). The hierarchical mixed graphical model yields more accurate and cleaner results than the flat CRF and the RDF region classifier, and comparable to the hierarchical CRF model. (1st-row) Testing images. (2nd-row to 6th-row) Classification results using the RDF region classifier (RDF), the flat CRF model (CRF) (Yang & Förstner, 2011c), the hierarchical CRF model (HCRF), the hierarchical mixed graphical model (HMGM), and the ground truth (GT), respectively. (7th-row) Legend.

Table 5.16: Pixelwise accuracy of the image classification using the hierarchical mixed graphical model with the multi-scale mean shift segmentation on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**70**	3	1	3	3	1	8	11
c	37	**28**	0	8	5	0	20	2
d	66	2	**11**	2	0	0	9	10
p	8	2	0	**76**	1	1	10	2
r	4	3	0	23	**60**	0	7	3
s	12	0	0	0	0	**80**	7	1
v	10	6	0	1	2	0	**78**	3
w	18	1	2	0	0	1	3	**75**

5.5.3 Results with multi-scale watershed and the hierarchical mixed graphical model

Table 5.17 shows the confusion matrix obtained by applying the hierarchical mixed graphical model to the whole test set. Accuracy values in the table are computed as the percentage of the image pixels assigned to the correct class label, ignoring the pixels labelled as void in the ground truth. The overall classification accuracy is 68.0%. The weighting parameters, learned by cross validation on the training data, are $\alpha = 1.08$, $\beta = 1$. For comparison, the RDF region classifier gives an overall accuracy of 55.4%, the flat CRF gives an overall accuracy of 61.8%, and the hierarchical CRF gives an overall accuracy of 65.3%.

Compared to the confusion matrix showing the flat CRF with the baseline watershed in Table 5.11 on page 67, the hierarchical mixed graphical model performs significantly better on *car*, *vegetation*, and *road* classes, slightly better on *building*, *window*, and *sky* classes, and significantly worse on *door* class.

Qualitative results of the hierarchical mixed graphical model on the eTRIMS dataset are presented in Fig. 5.11. Compared to the classification results showing the flat CRF with the baseline watershed segmentation and the hierarchical CRF with the multi-scale watershed segmentation, the hierarchical mixed graphical model produces significantly better results than the results from the flat CRF, and slightly better than the results from the hierarchical CRF.

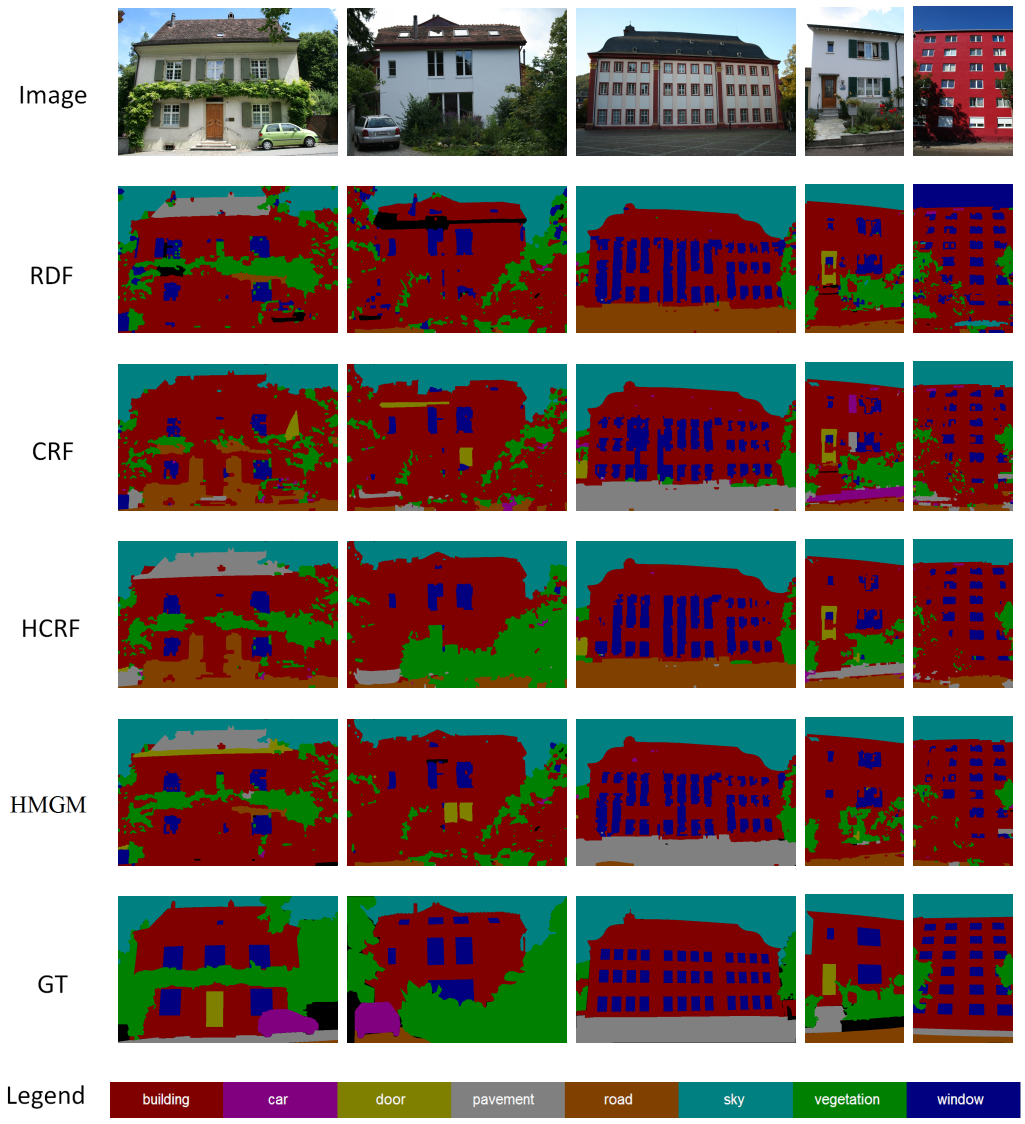


Figure 5.11: Qualitative classification results of the hierarchical mixed graphical model with the multi-scale watershed segmentation on the testing images from the eTRIMS dataset (Korč & Förstner, 2009). The qualitative inspection of the results in these images shows that the hierarchical mixed graphical model yields more accurate and cleaner results than the flat CRF and the RDF region classifier, and comparable to the hierarchical CRF model. (1st-row) Testing images. (2nd-row to 6th-row) Classification results using the RDF region classifier (RDF), the flat CRF model (CRF), the hierarchical CRF model (HCRF), the hierarchical mixed graphical model (HMGM), and the ground truth (GT), respectively. (7th-row) Legend.

Table 5.17: Pixelwise accuracy of the image classification using the hierarchical mixed graphical model with the multi-scale watershed on the eTRIMS 8-class dataset (Korč & Förstner, 2009). The confusion matrix shows the classification accuracy for each class (rows) and is row-normalized to sum to 100%. Row labels indicate the true class (Tr), and column labels the predicted class (Pr). (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

Pr \ Tr	b	c	d	p	r	s	v	w
b	**68**	3	2	3	3	1	10	10
c	26	**38**	0	5	7	0	23	1
d	35	0	**0**	2	1	0	45	17
p	31	3	1	**52**	9	0	1	3
r	12	10	1	13	**60**	0	3	1
s	8	0	0	0	0	**82**	9	1
v	8	6	0	2	2	0	**80**	2
w	21	0	1	0	0	0	1	**77**

5.6 Summary

By visual inspection of the classification results for some challenging test images, e. g. Fig. 5.10 and Fig. 5.11, we have demonstrated that our graphical model framework outperforms the method either with only spatial information (Yang & Förstner, 2011c) or without contextual information.

The overall performance of the classification methods on the eTRIMS dataset (Korč & Förstner, 2009) in terms of the pixelwise classification accuracy is listed in Table 5.18. We observe that the classification results from the mean shift segmentation are consistently better than the results from the watershed segmentation. This is probably

Table 5.18: Pixelwise accuracy comparison of four image classification methods with two segmentation algorithms on the eTRIMS 8-class dataset (Korč & Förstner, 2009). (C: classification, S: segmentation, RDF: randomized decision forest, CRF: flat conditional random field, HCRF: hierarchical conditional random field, HMGM: hierarchical mixed graphical model.)

C \ S	watershed	mean shift
RDF	55.4%	58.8%
CRF	61.8%	65.8%
HCRF	65.3%	69.0%
HMGM	68.0%	68.9%

because the mean shift preserves more consistent segmentation boundaries. By using the spatial neighbourhood information, the flat CRF (Yang & Förstner, 2011c) outperforms the RDF region classifier significantly (approx. 7% for each segmentation algorithm). Furthermore, by using additional hierarchical information, the hierarchical CRF and the hierarchical mixed graphical model outperform the flat CRF, which confirms the aforementioned visual inspection. Note that the hierarchical mixed graphical model with watershed segmentation gains accuracy of 6.2% than the flat CRF, compared to the hierarchical mixed graphical model with mean shift segmentation (3.1%). The difference in these results may be caused by the different scale-selection schemes in two segmentation algorithms. The highest scale of the watershed segmentation gives very few regions, compared to the highest scale of the mean shift segmentation.

We summarize the classification results over all eight classes on the eTRIMS dataset (Korč & Förstner, 2009) from eight confusion matrix tables in Fig. 5.12. The flat CRF

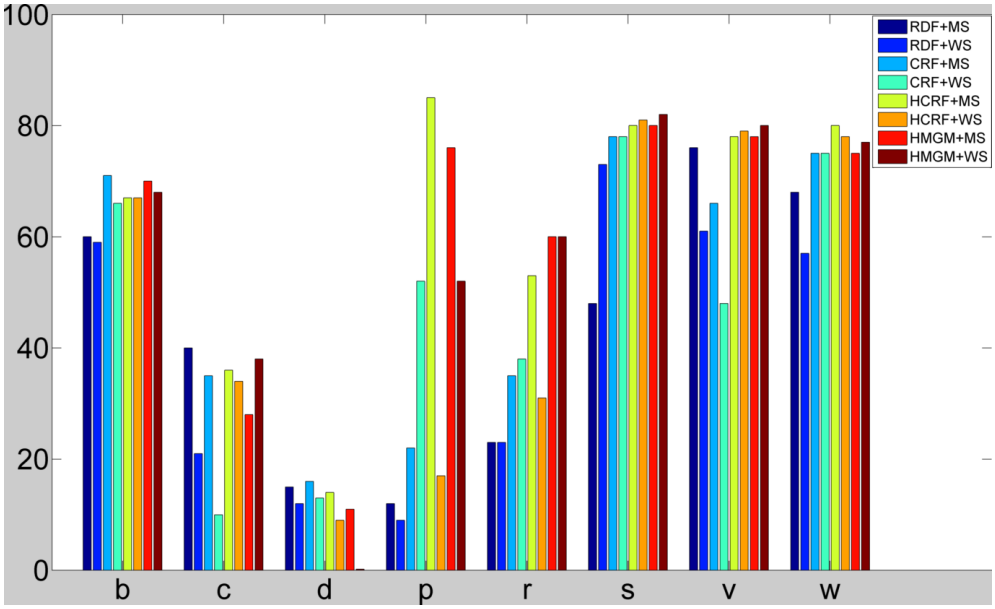


Figure 5.12: The classification results over all eight classes from all eight cases of four classification methods with two segmentation algorithms on the eTRIMS dataset (Korč & Förstner, 2009). The legend shown on the top right corner. RDF+MS: RDF region classifier with mean shift segmentation, RDF+WS: RDF region classifier with watershed segmentation, CRF+MS: flat CRF with mean shift segmentation, CRF+WS: flat CRF with watershed segmentation, HCRF+MS: hierarchical CRF with multi-scale mean shift segmentation, HCRF+WS: hierarchical CRF with multi-scale watershed segmentation, HMGM+MS: hierarchical mixed graphical model with multi-scale mean shift segmentation, HMGM+WS: hierarchical mixed graphical model with multi-scale watershed segmentation. Note that each colour represents one of the eight cases of four classification methods with two segmentation algorithms, and should not be confused with the colour in other figures. (b = *building*, c = *car*, d = *door*, p = *pavement*, r = *road*, s = *sky*, v = *vegetation*, w = *window*.)

outperforms the RDF region classifier for all eight classes except the class *car*. The hierarchical CRF and the hierarchical mixed graphical model outperforms the flat CRF for most classes. The best accuracies for each class are the flat CRF with mean shift for class *building*, the RDF classifier with mean shift for class *car*, the flat CRF with mean shift for class *door*, the hierarchical CRF with mean shift for class *pavement*, the hierarchical mixed graphical model with mean shift and watershed for class *road*, the hierarchical mixed graphical model with watershed for class *sky*, the hierarchical mixed graphical model with watershed for class *vegetation*, and the hierarchical CRF with mean shift for class *window*. The greatest accuracies are for classes which have low visual variability and many training examples, e. g. *window*, *sky*, *building*, and *vegetation*, whilst the lowest accuracies are for classes with high visual variability or few training examples, e. g. *car* and *door*.

We want to emphasize that our experiments should be seen as a demonstration of a consistent and convenient probabilistic model to incorporate the contextual information, e. g. the spatial structure and the hierarchical structure. With the current settings for the local and hierarchical pairwise potential functions, our method tends to produce rather low classification rate for the object classes with minor instances, e. g. *car* and *door*, as in all eight cases of four classification methods with two segmentation algorithms on the eTRIMS dataset (Korč & Förstner, 2009) (cf. Fig. 5.12). An investigation into more sophisticated potential functions might resolve this problem. In computer vision, the pairwise potentials are usually represented by a weighted summation of many features functions (Shotton *et al.*, 2006), and the parameters with the size as same as feature number are learned from the training data. By maximizing the conditional log-likelihood, better accuracy usually obtained. But this kind of parameter learning remains a difficult problem and also is most time-consuming part (Alahari *et al.*, 2010). While in our proposed graphical model formulation, we simply have two weighting parameters (similar to Gould *et al.* (2008); Fulkerson *et al.* (2009); Ladicky *et al.* (2009)). So this is the trade-off between efficiency and accuracy.

Compared to the higher order conditional random fields, our graphical model framework only exploits up to second-order cliques. The work on solving higher order potentials using move making algorithms has targeted the particular classes of the potential functions. Developing efficient large move making for exact and approximate minimization of general higher order energy functions is a difficult problem. Parameter learning for the higher order CRF is also a challenging problem. Furthermore, there are standard techniques for transforming arbitrary high-order factors into pairwise ones called order reduction (Ishikawa, 2009; Gallagher *et al.*, 2011). Order reduction methods operate by expressing each high order term as an expression with only the pairwise interactions by introducing auxiliary variables. Order reduction is followed by an inference procedure on the order-reduced random field. Since there are many possible ways to perform order reduction, it is difficult to ascertain a better reduction that generates easier pairwise inference problems. On the other hand, our proposed model makes learning and inference much easier.

Chapter 6

Conclusion and Future Work

The way ahead is long, I see no ending.

Yet high and low I'll search with my will unbending.

- *Qu Yuan* (340 B.C. - 278 B.C.)

In this thesis, we have addressed the problem of incorporating two different types of the contextual information, namely the spatial structure and the hierarchical structure for image interpretation of man-made scenes. Towards this, the thesis makes the following key contributions:

- We propose a statistically motivated, generic probabilistic graphical model framework for scene interpretation, which seamlessly integrates different types of the image features, and the spatial structural information and the hierarchical structural information defined over the multi-scale image segmentation. It unifies the ideas of the existing approaches, e. g. conditional random fields (CRFs) and Bayesian networks (BNs), which has a clear statistical interpretation as the MAP estimate of a multi-class labelling problem. Given the graphical model structure, we derive the probability distribution based on the factorization property implied in the model structure. The statistical model leads to an energy function that can be optimized approximately by either loopy belief propagation or graph cut based move making algorithm. The particular type of the features, the spatial structure, and the hierarchical structure however is not prescribed.

- We demonstrate the application of the proposed model on the building facade image classification task. We show that the framework for scene interpretation allows for significantly better classification results than the standard classical local classification approach on man-made scenes by incorporating spatial and hierarchical structures. We investigate the performance of the algorithms on a public dataset to show the relative importance of the information from the

spatial structure and the hierarchical structure. We present an approach for the region classification using an efficient randomized decision forest classifier as a baseline. Incorporated with the spatial structure and the hierarchical structure, we show that both the hierarchical CRF and the hierarchical mixed graphical model produce better classification results than both the baseline region classifier and the flat CRF.

In this work, we restrict our experiments on man-made scenes, however, we would like to point out that our method is general enough to be applied to other applications in photogrammetry and computer vision. As long as the spatial and hierarchical structures exist, our method can be applied. These applications includes image retrieval, image categorization, object class segmentation, object recognition, and remote sensing data classification. The original motivation for our approach was not to outperform other classification methods, but to give an integrated graphical model having both the benefits from random fields and Bayesian networks. Our method should be seen as a construction of a consistent probabilistic model to incorporate the spatial and hierarchical structures.

We want to emphasize that the choice of the crafted application-dependent features is crucial for the final success. Even more, we think that the discriminative power in the features of unary and pairwise potentials is the key to the overall performance of the graphical models. To make these graphical models applicable to the generic real-world applications, it is unavoidable to incorporate the methods for automatic feature extraction from the image and feature selection from the feature pool.

So far, our work has made some progress towards the long-term goal of scene interpretation. However, there are still plenty of work to be done. In the following, we address some possible future directions for building on our work.

First, the theory of the graphical model developed in Chapter 4 is linked to a chain graphical model defined over a chain graph, which is a generalization of both the undirected graph and the directed graph, and could be applied to other applications in photogrammetry, computer vision, and beyond these domains, such as sequence labelling, human motion recognition, gene and protein classification, rather than scene interpretation. The chain graphical model may allow integrating more complex heuristic BNs in the chain graph, rather than our intuitive graphical model for image interpretation which is too simple and specific.

Second, the occluded parts of an object are not annotated as part of an object in eTRIMS dataset (Korč & Förstner, 2009). In our models, we don't take occlusion into account. But, one important cue that we can derive from scene structure is knowing the relative location of objects. So, we are able to reason about the occlusion to a certain extend. An interesting research direction is in developing the graphical models that make better use of the geometric understanding of a man-made scene to determine what parts of an object are occluded and taking that information into account. Hoiem *et al.* (2011) believe surface information can help to recover the occlusion boundaries. Motivated by Drauschke *et al.* (2009), we believe that 3D information, either from the laser scan data or the range data derived from the multi-view images, appear to be

very useful.

Third, our methods operate on the region level resulting from certain unsupervised segmentation algorithm, which allows for fast inference. However, one disadvantage of such an approach is that the mistakes in the initial unsupervised segmentation, in which regions span multiple object classes, cannot be recovered from. For each region from the segmentation, a class label is commonly assigned to the region according to the majority vote of the ground truth pixel labels. At the starting point, ambiguity is introduced in the region ground truth labelling. One may resolve this problem by assigning a class probability vector to the region, not assigning most probable label to the region. We could result in a probability estimation model of the image segmentation regions. One could also eliminate the inconsistent regions by employing Hierarchical CRFs (Ladicky *et al.*, 2009), which allow for the integration of the region-based CRFs with a low-level pixel based CRF.

Fourth, our method presented in this thesis could be seen as a mid-level graphical model representation. An exciting direction for future work is to integrate this mid-level model with a high-level model for an incremental built-up of a context aware scene description. It will provide a smart integration of the bottom-up and the top-down reasoning and allowing to incorporate the prior knowledge. The mid-level model establishes both a spatial aggregation structure and a hierarchical partonomy. In the high-level model, one could exploit attribute grammars, which uses the attributes and the probabilities of the classified regions from the mid-level model, to control the semantic reconstruction of the scene. The result of the high-level module is a highly structured interpretation of the complete scene, given its own priors and the evidence provided by the mid-level model. The grammar model representing the semantic high-level structure again serves as a prior for the mid-level model. This bottom-up-top-down cycle is repeated until the interpretation appears stable enough. A concept for the interpretation of integrating CRFs with a stochastical attribute grammar in order to capture the structural complexity of the scene has been developed in Schmittwilken *et al.* (2009).

Fifth, the structures of the proposed graphical model is fixed. The fixed structure is in fact constructed based on expert's *a priori* knowledge about the relationships between image pixels, regions and objects. On the other hand, the problem of selecting from the exponentially large space of the possible network structures becomes of great importance. In fact, unsupervised discovery of the structured, predictive models from the sparse data is a central problem in artificial intelligence (Lee *et al.*, 2006). There are recent works that tackle this issue, which deal with either the random field model, e. g. (Lee *et al.*, 2006; Lin *et al.*, 2009; Zhu *et al.*, 2010), or the Bayesian network model, e. g. (Mansinghka *et al.*, 2006; Xie *et al.*, 2006). It would be interesting to test whether these methods are applicable to the mixed graphical model as well.

Appendix A

Chain graphical model

Chain graphical model was originally introduced in the statistic society (Lauritzen & Wermuth, 1989; Frydenberg, 1990). The basic graphical representation that underlies the chain graphical model is a chain graph, which contains both directed and undirected edges to capture different types of the relationships among the random variables.

In Section 3.5, we have introduced two approaches, a moral graph and a factor graph, to exploit the relations between directed and undirected graphical models. In this section, we introduce a chain graphical model framework, including the model parametrization and the joint probability distribution.

A.1 Chain graph and model parametrization

A chain graphical model consists of both the directed edges and the undirected edges. We can parametrize the directed edges by the conditional probabilities, and the undirected edges by the potential functions.

We give a definition of a chain graph as follows.

Definition A.1 *Chain graph. A chain graph is an acyclic graph containing both directed and undirected edges.*

We denote a chain graph with $\mathcal{K}$. Fig. A.1 shows an example of a chain graph. If we add the undirected edge $\{2, 6\}$ to $\mathcal{K}$, we have a directed path $2, 3, 6, 2$ from node 2 to itself, which breaks the acyclicity requirement, therefore, the resulting graph is not a chain graph anymore. The acyclicity requirement on a chain graph implies that the graph can be decomposed into a directed graph of the chain components $\mathcal{K}_1, \cdots, \mathcal{K}_l$, where the nodes within the chain component are connected to each other only with the undirected edges, and any edge between the nodes in two chain components can only be a directed edge. Note l is the number of chain components in $\mathcal{K}$. For example, in the chain graph of Fig. A.1, we have five chain components: $\{3, 6, 8\}, \{2, 5\}, \{1\}, \{4\}, \{7\}$. Note that when the chain graph is an undirected graph, the whole graph forms a single chain component, while when the chain graph is a directed graph, each node is its own chain component.

In Fig. A.1 (same as shown in Fig. 3.1 on page 20), there are both the directed edges and the undirected edges. We use the local conditional probabilities to parametrize the directed edges. The relationship between $\underline{\boldsymbol{x}}_2$ and its parent $\underline{\boldsymbol{x}}_1$ is parametrized by the conditional probability $P(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$. The relationship between $\underline{\boldsymbol{x}}_8$ and its parents $\underline{\boldsymbol{x}}_5$, $\underline{\boldsymbol{x}}_7$ is parametrized by the conditional probability $P(\boldsymbol{x}_8 \mid \boldsymbol{x}_5, \boldsymbol{x}_7)$. Potential functions are used to parametrize the undirected edges. The relationship between $\underline{\boldsymbol{x}}_6$ and $\underline{\boldsymbol{x}}_8$ is parametrized by the pairwise potential function $\phi(\boldsymbol{x}_6, \boldsymbol{x}_8)$. Other edges are parametrized accordingly.

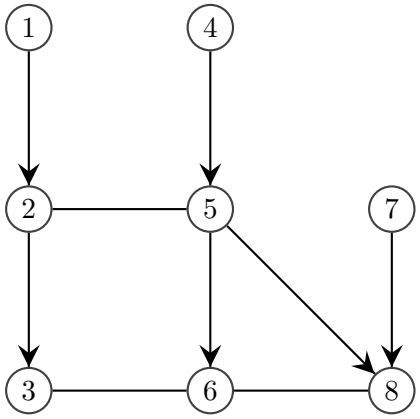


Figure A.1: A chain graph $\mathcal{K}$. There are both the directed edges and the undirected edges, but no directed cycles.

A.2 Joint probability distribution

Given a chain graphical model and the parametrization, we can derive the joint probability distribution. As we can see from previous sections, both directed and undirected graphs allow a global function of several variables to be expressed as a product of the factors over the subsets of those variables.

Similar to the moralized version of a directed graph, there exists the concept of moralization of a chain graph (Frydenberg, 1990). Let $\mathcal{K}$ be a chain graph and $\mathcal{K}_1, \cdots, \mathcal{K}_l$ be its chain components. We use $\mathrm{Pa}_{\mathcal{K}_i}$ to denote the parents of the nodes in $\mathcal{K}_i$. The moralized graph of $\mathcal{K}$ is an undirected graph. We first link any pair of the nodes using the undirected edges in $\mathrm{Pa}_{\mathcal{K}_i}$, for all $i = 1, \cdots, l$, and then convert all directed edges into undirected edges.

Consider a set of the random variables $\{\underline{\boldsymbol{x}}_i, i \in \mathcal{V}\}$ defined over a chain graph $\mathcal{K}$. $\underline{\boldsymbol{x}} = [\underline{\boldsymbol{x}}_1; \cdots; \underline{\boldsymbol{x}}_i; \cdots; \underline{\boldsymbol{x}}_n]$. Each random variable $\underline{\boldsymbol{x}}_i$ is associated with a node $i \in \mathcal{V}$. $\underline{\boldsymbol{s}}_i$ is denoted as the set of the random variables corresponding to the chain component $\mathcal{K}_i$. The set of random variables, associated with the parents of the chain component $\mathcal{K}_i$, is denoted as $\mathrm{Pa}(\underline{\boldsymbol{s}}_i)$. As in other graphical representations, the structure of a chain graph $\mathcal{K}$ can also be used to define a factorization for a probability distribution. Intuitively, the factorization for a chain graphical model represents the distribution as a product

of each of the set of the random variables $\underline{\boldsymbol{s}}_i$ given its parents $P(\boldsymbol{s}_i \mid \mathrm{Pa}(\boldsymbol{s}_i))$ (Koller & Friedman, 2009).

First, we define a set of the factors $\boldsymbol{f}_i(\boldsymbol{x}_c)$, $i = 1, \cdots, l$, $c \in \mathcal{C}$, where $\boldsymbol{x}_c = \{\boldsymbol{x}_i, i \in c\}$, such that the induced subgraph $\mathcal{H}_c$ is a complete subgraph in the moralized graph of $\mathcal{K}$. Each $\boldsymbol{f}_i(\boldsymbol{x}_c)$ corresponds to either the conditional probability or the potential function. Note l is the number of the chain components in $\mathcal{K}$, and $\mathcal{C}$ is the set of the cliques.

Then, we associate the factor $\boldsymbol{f}_i(\boldsymbol{x}_c)$ with a single chain component $\mathcal{K}_i$, where the nodes are connected to each other only with the undirected edges, $\mathcal{H}_c \subseteq \mathcal{K}_i \cup \mathrm{Pa}_{\mathcal{K}_i}$. Recalling the definition of CRFs in Section 3.4, we define $P(\boldsymbol{s}_i \mid \mathrm{Pa}(\boldsymbol{s}_i))$ as a CRF with these factors. Then, the joint probability distribution is defined as

$$\begin{aligned} P(\boldsymbol{x}) &= \prod_{i=1}^{l} P(\boldsymbol{s}_i \mid \mathrm{Pa}(\boldsymbol{s}_i)) \\ &= \prod_{i=1}^{l} \frac{1}{Z_i(\mathrm{Pa}(\boldsymbol{s}_i))} \prod_{c \in \mathcal{C}} \boldsymbol{f}_i(\boldsymbol{x}_c) \end{aligned} \tag{A.1}$$

where $Z_i(\mathrm{Pa}(\boldsymbol{s}_i)) = \sum_{\boldsymbol{s}_i} \prod_{c \in \mathcal{C}} \boldsymbol{f}_i(\boldsymbol{x}_c)$.

Eq. A.1 is called the chain rule for a chain graph. This key equation expresses the *factorization properties* of the joint distribution for a chain graphical model.

By simple algebra calculation, the joint probability distribution given in (A.1) can be written in the form

$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{l} \sum_{c \in \mathcal{C}} \log \boldsymbol{f}_i(\boldsymbol{x}_c) \right) \tag{A.2}$$

where $Z = \prod_{i=1}^{l} \frac{1}{Z_i(\mathrm{Pa}(\boldsymbol{s}_i))}$ is a normalization constant. Therefore, the joint probability distribution for a chain graphical model is a *Gibbs* distribution

$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\left(-E(\boldsymbol{x})\right) \tag{A.3}$$

The term

$$E(\boldsymbol{x}) = \sum_{i=1}^{l} \sum_{c \in \mathcal{C}} - \log \boldsymbol{f}_i(\boldsymbol{x}_c) \tag{A.4}$$

is the energy function.

A.3 Factor graph representation

In the following, we introduce a factor graph representation, which is a notion of unifying the undirected graphs, the directed graphs, and the chain graphs. The chain graphical model represents a joint probability distribution that is factorized as a prod-

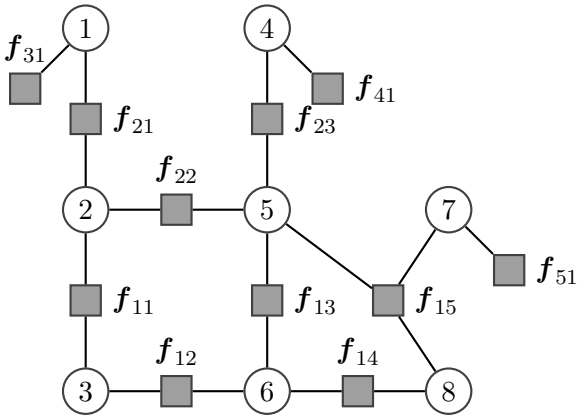


Figure A.2: A factor graph representation of a chain graph $\mathcal{K}$ in Fig. A.1 on page 84. Each square corresponds to a factor in (A.5). For example, the square connecting nodes 1 and 2 corresponds to the factor $\boldsymbol{f}_{21}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, and the square connecting nodes 8 and 5, 7 corresponds to the factor $\boldsymbol{f}_{15}(\boldsymbol{x}_5, \boldsymbol{x}_7, \boldsymbol{x}_8)$.

uct of the factors over the subsets of the variables. Therefore, we can apply rules discussed in Section 3.5.2 to convert the chain graphical model into a factor graph representation.

In Fig. A.1 on page 84, we require that the conditional distribution $P(\boldsymbol{x}_2, \boldsymbol{x}_5 \mid \boldsymbol{x}_1, \boldsymbol{x}_4)$ is defined as a normalized product of the factors $\frac{1}{Z_2(\boldsymbol{x}_1, \boldsymbol{x}_4)} \boldsymbol{f}_{21}(\boldsymbol{x}_1, \boldsymbol{x}_2) \boldsymbol{f}_{22}(\boldsymbol{x}_2, \boldsymbol{x}_5) \boldsymbol{f}_{23}(\boldsymbol{x}_4, \boldsymbol{x}_5)$, where $Z_2(\boldsymbol{x}_1, \boldsymbol{x}_4) = \sum_{\boldsymbol{x}_2, \boldsymbol{x}_5} \boldsymbol{f}_{21}(\boldsymbol{x}_1, \boldsymbol{x}_2) \boldsymbol{f}_{22}(\boldsymbol{x}_2, \boldsymbol{x}_5) \boldsymbol{f}_{23}(\boldsymbol{x}_4, \boldsymbol{x}_5)$. A similar factorization applies to $P(\boldsymbol{x}_3, \boldsymbol{x}_6, \boldsymbol{x}_8 \mid \boldsymbol{x}_2, \boldsymbol{x}_5, \boldsymbol{x}_7)$. Therefore, the joint probability distribution is given by

$$
\begin{aligned}
P(\boldsymbol{x}) =& P(\boldsymbol{x}_3, \boldsymbol{x}_6, \boldsymbol{x}_8 \mid \boldsymbol{x}_2, \boldsymbol{x}_5, \boldsymbol{x}_7) P(\boldsymbol{x}_2, \boldsymbol{x}_5 \mid \boldsymbol{x}_1, \boldsymbol{x}_4) P(\boldsymbol{x}_1) P(\boldsymbol{x}_4) P(\boldsymbol{x}_7) \\
=& \left\{ \frac{1}{Z_1(\boldsymbol{x}_2, \boldsymbol{x}_5, \boldsymbol{x}_7)} \boldsymbol{f}_{11}(\boldsymbol{x}_2, \boldsymbol{x}_3) \boldsymbol{f}_{12}(\boldsymbol{x}_3, \boldsymbol{x}_6) \boldsymbol{f}_{13}(\boldsymbol{x}_5, \boldsymbol{x}_6) \boldsymbol{f}_{14}(\boldsymbol{x}_6, \boldsymbol{x}_8) \boldsymbol{f}_{15}(\boldsymbol{x}_5, \boldsymbol{x}_7, \boldsymbol{x}_8) \right\} \\
& \left\{ \frac{1}{Z_2(\boldsymbol{x}_1, \boldsymbol{x}_4)} \boldsymbol{f}_{21}(\boldsymbol{x}_1, \boldsymbol{x}_2) \boldsymbol{f}_{22}(\boldsymbol{x}_2, \boldsymbol{x}_5) \boldsymbol{f}_{23}(\boldsymbol{x}_4, \boldsymbol{x}_5) \right\} \boldsymbol{f}_{31}(\boldsymbol{x}_1) \boldsymbol{f}_{41}(\boldsymbol{x}_4) \boldsymbol{f}_{51}(\boldsymbol{x}_7)
\end{aligned}
\tag{A.5}
$$

where $Z_1(\boldsymbol{x}_2, \boldsymbol{x}_5, \boldsymbol{x}_7) = \sum_{\boldsymbol{x}_3, \boldsymbol{x}_6, \boldsymbol{x}_8} \boldsymbol{f}_{11}(\boldsymbol{x}_2, \boldsymbol{x}_3) \boldsymbol{f}_{12}(\boldsymbol{x}_3, \boldsymbol{x}_6) \boldsymbol{f}_{13}(\boldsymbol{x}_5, \boldsymbol{x}_6) \boldsymbol{f}_{14}(\boldsymbol{x}_6, \boldsymbol{x}_8)$ $\boldsymbol{f}_{15}(\boldsymbol{x}_5, \boldsymbol{x}_7, \boldsymbol{x}_8)$, and $Z_2(\boldsymbol{x}_1, \boldsymbol{x}_4) = \sum_{\boldsymbol{x}_2, \boldsymbol{x}_5} \boldsymbol{f}_{21}(\boldsymbol{x}_1, \boldsymbol{x}_2) \boldsymbol{f}_{22}(\boldsymbol{x}_2, \boldsymbol{x}_5) \boldsymbol{f}_{23}(\boldsymbol{x}_4, \boldsymbol{x}_5)$.

Based on the joint probability distribution of (A.5), the example graph in Fig. A.1 on page 84 can be converted into a factor graph representation as shown in Fig. A.2. Each square corresponds to a factor in (A.5). For example, the square connecting the nodes 1 and 2 corresponds to the factor $\boldsymbol{f}_{21}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, and the square connecting the nodes 8 and 5, 7 corresponds to the factor $\boldsymbol{f}_{15}(\boldsymbol{x}_5, \boldsymbol{x}_7, \boldsymbol{x}_8)$. Given this factor graph, we can use principled methods, such as the max-product algorithm, to infer the optimal states of all random variables that produce the maximum joint probability (Bishop, 2006).

Bibliography

Abney, Steven. 1997. Stochastic attribute-value grammars. *Computational Linguistics*, **23**, 597–618. 11

Alahari, Karteek, Russell, Chris, & Torr, Philip. 2010. Efficient piecewise learning for conditional random fields. *Pages 895–901 of: IEEE Conference on Computer Vision and Pattern Recognition*. 18, 44, 48, 77

Andres, Björn, Kappes, Jörg H., Köthe, Ullrich, Schnörr, Christoph, & Hamprecht, Fred A. 2010. An empirical comparison of inference algorithms for graphical models with higher order factors using opengm. *Pages 353–362 of: Annual Symposium of the German Association for Pattern Recognition (DAGM)*. 48

Bang-Jensen, Jrgen, & Gutin, Gregory Z. 2008. *Digraphs: Theory, Algorithms and Applications*. 2nd edn. Springer Publishing Company, Inc. 19, 29

Barnard, K., & Forsyth, D. 2001. Learning the semantics of words and pictures. *Pages 408–415 of: International Conference on Computer Vision*, vol. 2. 13

Barnard, K., Duygulu, P., Freitas, N. D., Forsyth, D., Blei, D., & Jordan, M. 2003. Matching words and pictures. *Pages 1107–1135 of: Journal of Machine Learning Research*, vol. 3. 42

Batra, Dhruv, Sukthankar, Rahul, & Chen, Tsuhan. 2008. Learning class-specific affinities for image labelling. *Pages 1–8 of: IEEE Conference on Computer Vision and Pattern Recognition*. 39

Becker, S. 2009. Generation and application of rules for quality dependent facade reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, **64**(6), 640–653. 11

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, **B-36**(2), 192–236. 1, 12

Besag, J. 1986. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society Series B*, **48**(3), 259–302. 12, 27

Bishop, Christopher. 2006. *Pattern recognition and machine learning*. USA: Springer-Verlag New York, Inc. 23, 28, 86

BIBLIOGRAPHY

Borenstein, Eran, Sharon, Eitan, & Ullman, Shimon. 2004. Combining top-down and bottom-up segmentation. *Pages 46–53 of: CVPR Workshop on Perceptual Organization in Computer Vision.* 16, 17

Bosch, Anna, Zisserman, Andrew, & Muñoz, Xavier. 2007. Image classification using random forests and ferns. *Pages 1–8 of: IEEE International Conference on Computer Vision.* 42

Boykov, Yuri, & Jolly, Marie-Pierre. 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. *Pages 105–112 of: International Conference on Computer Vision.* 44

Boykov, Yuri, & Kolmogorov, Vladimir. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 1124–1137. 27, 45, 48, 49

Boykov, Yuri, Veksler, Olga, & Zabih, Ramin. 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 1222–1239. 27, 45, 48, 49

Breiman, Leo. 2001. Random forests. *Machine Learning*, **45**(1), 5–32. 42

Brenner, C., Haala, N., & Fritsch, D. 2001. Towards fully automated 3D city model generation. *In: Automatic Extraction of Man-Made Objects from Aerial and Space Images III.* 9

Brunn, Ansgar, & Weidner, Uwe. 1997. Extracting buildings from digital surface models. *Pages 1–8 of: IAPRS: 3D Reconstruction and Modeling of Topographic Objects.* 13

Comaniciu, Dorin, & Meer, Peter. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(5), 603–619. 36, 41, 53, 56, 59

Cowell, Robert G., Dawid, A. Philip, Lauritzen, Steffen L., & Spiegelhalter, David J. 1999. *Probabilistic networks and expert systems.* Springer-Verlag. 28, 29

Dalal, Navneet, & Triggs, Bill. 2005. Histograms of oriented gradients for human detection. *Pages 886–893 of: IEEE Conference on Computer Vision and Pattern Recognition.* 42

Delong, Andrew, Osokin, Anton, Isack, Hossam N., & Boykov, Yuri. 2010. Fast approximate energy minimization with label costs. *Pages 2173–2180 of: IEEE Conference on Computer Vision and Pattern Recognition.* 14

Dick, A. R., Torr, P. H. S., & Cipolla, R. 2004. Modelling and interpretation of architecture from several images. *International Journal of Computer Vision*, **60**, 111–134. 4, 10

Drauschke, M. 2009. An irregular pyramid for multi-scale analysis of objects and their parts. *Pages 293–303 of: IAPR-TC-15 Workshop on Graph-based Representations in Pattern Recognition.* 35, 57, 58

Drauschke, M., & Förstner, W. 2011. A bayesian approach for scene interpretation with integrated hierarchical structure. *Pages 1–10 of: Annual Symposium of the German Association for Pattern Recognition (DAGM).* 23, 39, 40

Drauschke, M., & Mayer, H. 2010. Evaluation of texture energies for classification of facade images. *Pages 257–262 of: ISPRS Technical Commission III Symposium on Photogrammetry Computer Vision and Image Analysis.* 11, 41, 42

Drauschke, M., Schuster, H.-F., & Förstner, W. 2006. Detectability of buildings in aerial images over scale space. *Pages 7–12 of: ISPRS Technical Commission III Symposium on Photogrammetry Computer Vision and Image Analysis.* IAPRS 36 (3). 57

Drauschke, M., Roscher, R., Läbe, T., & Förstner, W. 2009. Improving image segmentation using multiple view analysis. *Pages 211–216 of: Object Extraction for 3D City Models, Road Databases and Traffic Monitoring - Concepts, Algorithms and Evaluation (CMRT09).* 80

Feng, X., Williams, C. K. I., & Felderhof, S. N. 2002. Combining belief networks and neural networks for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 467–483. 6, 15, 16

Fischer, A., Kolbe, T.H., & Lang, F. 1997. Integration of 2D and 3D reasoning for building reconstruction using a generic hierarchical model. *Pages 159–180 of:* Förstner, W., & Plümer, L. (eds), *SMATI '97, Workshop on Semantic Modeling for the Acquisition of Topographic Information from Images and Maps.* 9

Fischer, A., Kolbe, T.H., & Lang, F. 1999. On the use of geometric and semantic models for component-based building reconstruction. *Pages 101–119 of:* Förstner, W., Liedtke, C.-E., & Bückner, J. (eds), *SMATI '99, Workshop on Semantic Modeling for the Acquisition of Topographic Information from Images and Maps.* 9

Frahm, J.M., Pollefeys, M., Lazebnik, S., Gallup, D., Clipp, B., Raguram, R., Wu, C., Zach, C., & Johnson, T. 2010. Fast robust large-scale mapping from video and internet photo collections. *ISPRS Journal of Photogrammetry and Remote Sensing*, **65**(6), 538–549. 10

Fröhlich, Björn, Rodner, Erik, & Denzler, Joachim. 2010. A fast approach for pixelwise labeling of facade images. *Pages 3029–3032 of: International Conference on Pattern Recognition*, vol. 7. 10, 11

Frydenberg, Morten. 1990. The chain graph Markov property. *Scandinavian Journal of Statistics*, **17**(4), 333–353. 83, 84

Fujishige, Satoru. 1990. Submodular functions and optimization. *Annals of Discrete Mathematics*, **47**. 27

Fulkerson, B., Vedaldi, A., & Soatto, S. 2009. Class segmentation and object localization with superpixel neighborhoods. *Pages 670–677 of: International Conference on Computer Vision*. 39, 77

Gallagher, Andrew, Batra, Dhruv, & Parikh, Devi. 2011. Inference for order reduction in markov random fields. *In: IEEE Conference on Computer Vision and Pattern Recognition*. 77

Geman, Stuart, & Geman, Donald. 1984. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741. 12, 26

Geurts, Pierre, Ernst, Damien, & Wehenkel, Louis. 2006. Extremely randomized trees. *Machine Learning*, **63**(1), 3–42. 45

Gülch, Eberhard, Müller, Hardo, Läbe, Thomas, & Ragia, Lemonia. 1998. On the performance of semi-automatic building extraction. *In: Proceedings of ISPRS Commission III Symposium, Columbus, Ohio*. 10

Gonfaus, J.M., Boix, X., van de Weijer, J., Bagdanov, A.D., Serrat, J., & Gonzalez, J. 2010. Harmony potentials for joint classification and segmentation. *Pages 3280–3287 of: IEEE Conference on Computer Vision and Pattern Recognition*. 13

Gould, S., Rodgers, J., Cohen, D., Elidan, G., & Koller, D. 2008. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, **80**(3), 300–316. 13, 39, 44, 61, 77

Greig, D M, Porteous, B T, & Seheult, A H. 1989. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society Series B*, **51**(2), 271–279. 27

Gruen, Armin, & Wang, Xinhua. 1999. Cybercity modeler, a tool for interactive 3-d city model generation. *Photogrammetric Week 99*, 1–11. 10

Hammersley, J. M., & Clifford, P. 1971. Markov field on finite graph and lattices. *Unpublished*. 25

Hartz, Johannes, & Neumann, Bernd. 2007. Learning a knowledge base of ontological concepts for high-level scene interpretation. *Pages 436–443 of: IEEE Conference on Machine Learning and Applications (ICMLA)*. 11

Hartz, Johannes, Hotz, Lothar, Neumann, Bernd, & Terzic, Kasim. 2009. Automatic incremental model learning for scene interpretation. *In: International Conference on Computational Intelligence (IASTED CI-2009)*. 11

He, X., Zemel, R., & Carreira-perpiñán, M. 2004. Multiscale conditional random fields for image labeling. *Pages 695–702 of: IEEE Conference on Computer Vision and Pattern Recognition.* 13, 40

He, X., Zemel, R., & Ray, D. 2006. Learning and incorporating top-down cues in image segmentation. *Pages 338–351 of: European Conference on Computer Vision.* 13, 48

Herman, Martin, & Kanade, Takeo. 1984. The 3d mosaic scene understanding system: incremental reconstruction of 3d scenes for complex images. *Pages 137–148 of: DARPA Image Understanding Workshop.* 9

Hinton, G. E., Osindero, S., & Bao, K. 2005. Learning causally linked markov random fields. *In: International Workshop Artificial Intelligence and Statistics.* 17

Hoiem, Derek, Efros, Alexei A., & Hebert, Martial. 2007. Recovering surface layout from an image. *International Journal of Computer Vision*, **75**(1), 151–172. 14

Hoiem, Derek, Efros, Alexei A., & Hebert, Martial. 2011. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, **91**(3), 328–346. 80

Ishikawa, Hiroshi. 2009. Higher-order clique reduction in binary graph cut. *Pages 2993–3000 of: IEEE Conference on Computer Vision and Pattern Recognition.* 77

Jordan, Michael I. (ed). 1998. *Learning in graphical models.* MIT Press. 19

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., & Saul, Lawrence K. 1999. An introduction to variational methods for graphical models. *Machine Learning*, **37**, 183–233. 24

Kluckner, Stefan, & Bischof, Horst. 2010. Image-based building classification and 3d modeling with super-pixels. *Pages 233–238 of: ISPRS Technical Commission III Symposium on Photogrammetry Computer Vision and Image Analysis.* 10

Kohli, P., Kumar, M.P., & Torr, P. 2007. P3 & Beyond: Solving energies with higher order cliques. *Pages 1–8 of: IEEE Conference on Computer Vision and Pattern Recognition.* 14

Kohli, P., Ladicky, L., & Torr, P. 2009. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, **82**(3), 302–324. 14, 39

Koller, D., & Friedman, N. 2009. *Probabilistic graphical models: Principles and techniques.* MIT Press. 19, 21, 23, 26, 48, 85

Kolmogorov, Vladimir. 2006. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 1568–1583. 27

Kolmogorov, Vladimir, & Rother, Carsten. 2006. Comparison of energy minimization algorithms for highly connected graphs. *Pages 1–15 of: European Conference on Computer Vision*. 48

Kolmogorov, Vladimir, & Rother, Carsten. 2007. Minimizing nonsubmodular functions with graph cuts-a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 1274–1279. 27

Kolmogorov, Vladimir, & Zabih, Ramin. 2004. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**(2), 147–159. 27, 49

Korč, Filip. 2011. *Tractable learning for a class of global discriminative models for context sensitive image interpretation*. Ph.D. thesis, University of Bonn, Bonn, Germany. 2

Korč, Filip, & Förstner, Wolfgang. 2008. Approximate parameter learning in conditional random fields: An empirical investigation. *Pages 11–20 of: Annual Symposium of the German Association for Pattern Recognition (DAGM)*. 48

Korč, Filip, & Förstner, Wolfgang. 2008. Interpreting terrestrial images of urban scenes using discriminative random fields. *Pages 291–296 of: 21st Congress of the International Society for Photogrammetry and Remote Sensing*. IAPRS 37 (B3a). 10

Korč, Filip, & Förstner, Wolfgang. 2009. eTRIMS Image Database for interpreting images of man-made scenes. *In: TR-IGG-P-2009-01, Department of Photogrammetry, University of Bonn*. 46, 47, 51, 53, 55, 56, 57, 58, 59, 60, 62, 63, 65, 66, 67, 68, 69, 71, 72, 73, 74, 75, 76, 77, 80

Koutsourakis, Panagiotis, Simon, Loic, Teboul, Olivier, Tziritas, Georgios, & Paragios, Nikos. 2009. Single view reconstruction using shape grammars for urban environments. *Pages 1795–1802 of: IEEE International Conference on Computer Vision*. 11

Kschischang, Frank R., Frey, Brendan J., & andrea Loeliger, Hans. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, **47**(2), 498–519. 29

Kumar, M. P., Torr, P. H. S., & Zisserman, A. 2005. OBJ CUT. *Pages 18–25 of: IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. 16

Kumar, Sanjiv, & Hebert, Martial. 2003a. Discriminative random fields: A discriminative framework for contextual interaction in classification. *Pages 1150–1157 of: IEEE International Conference on Computer Vision*, vol. 2. 2, 4, 6, 12, 13, 25

Kumar, Sanjiv, & Hebert, Martial. 2003b. Man-made structure detection in natural images using a causal multiscale random field. *Pages 119–126 of: IEEE Conference on Computer Vision and Pattern Recognition*. 16, 18

Ladicky, L., Russell, C., Kohli, P., & Torr, P.H.S. 2009. Associative hierarchical crfs for object class image segmentation. *Pages 739–746 of: International Conference on Computer Vision.* 14, 77, 81

Ladicky, L., Russell, C., Kohli, P., & Torr, P.H.S. 2010. Graph cut based inference with co-occurrence statistics. *Pages 239–253 of: European Conference on Computer Vision.* 14

Lafferty, J., McCallum, A., & Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Pages 282–289 of: International Conference on Machine Learning.* 2, 12, 26

Lakatos, Imre. 1976. *Proofs and refutations.* Cambridge: Cambridge University Press. 41

Lauritzen, S. L., & Spiegelhalter, D. J. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, **50**, 157–224. 24

Lauritzen, S. L., & Wermuth, N. 1989. Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31–57. 83

Lazaridis, G., & Petrou, M. 2006. Image registration using the Walsh transform. *IEEE Transactions on Image Processing*, **15**(8), 2343–2357. 42

Lee, Su-In, Ganapathi, Varun, & Koller, Daphne. 2006. Efficient structure learning of markov networks using l_1-regularization. *Pages 817–824 of:* Schölkopf, Bernhard, Platt, John C., & Hoffman, Thomas (eds), *Advances in Neural Information Processing Systems.* 81

Lepetit, Vincent, Lagger, Pascal, & Fua, Pascal. 2005. Randomized trees for real-time keypoint recognition. *Pages 775–781 of: IEEE Conference on Computer Vision and Pattern Recognition.* 42

Levin, Anat, & Weiss, Yair. 2006. Learning to combine bottom-up and top-down segmentation. *Pages 581–594 of: European Conference on Computer Vision.* LNCS, vol. 3954. 13

Liang, P., Jordan, M. I., & Klein, D. 2009. Probabilistic grammars and hierarchical dirichlet processes. *In:* O'Hagan, T., & West, M. (eds), *The Handbook of Applied Bayesian Analysis.* Oxford University Press. 11

Lin, Yuanqing, Zhu, Shenghuo, Lee, Daniel D., & Taskar, Ben. 2009. Learning sparse markov network structure via ensemble-of-trees models. *Journal of Machine Learning Research - Proceedings Track*, **5**, 360–367. 81

Liu, Fei, Xu, Dongxiang, Yuan, Chun, & Kerwin, William S. 2006. Image segmentation based on bayesian network-markov random field model and its application to in vivo plaque composition. *Pages 141–144 of: IEEE International Symposium on Biomedical Imaging: From Nano to Macro.* 17, 18

Lowe, D.G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60**(2), 91–110. 42

MacKay, David J. C. 2002. *Information theory, inference & learning algorithms.* New York, USA: Cambridge University Press. 24

Mansinghka, Vikash K., Kemp, Charles, Griffiths, Thomas L., & Tenenbaum, Joshua B. 2006. Structured priors for structure learning. *In: Uncertainty in Artificial Intelligence.* 81

Maree, Raphael, Geurts, Pierre, Piater, Justus, & Wehenkel, Louis. 2005. Random subwindows for robust image classification. *Pages 34–40 of: IEEE Conference on Computer Vision and Pattern Recognition.* 42

Mayer, Helmut. 1999. Automatic object extraction from aerial imagery: a survey focusing on buildings. *Computer Vision and Image Understanding*, **74**(2), 138–149. 9

Mayer, Helmut, & Reznik, Sergiy. 2006. MCMC linked with implicit shape models and plane sweeping for 3D building facade interpretation in image sequences. *Pages 130–135 of: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Proceedings of Photogrammetric Computer Vision.* 10

Mayer, Helmut, & Reznik, Sergiy. 2007. Building facade interpretation from uncalibrated wide-baseline image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, **61**(6), 371–380. 10

Micusik, B., & Kosecka, J. 2009. Semantic segmentation of street scenes by superpixel co-occurrence and 3d geometry. *Pages 625 – 632 of: ICCV Workshop on Video-Oriented Object and Event Classification.* 10, 14

Micusik, B., & Kosecka, J. 2010. Multi-view superpixel stereo in urban environments. *International Journal of Computer Vision*, **89**(1), 106–119. 10

Modestino, J. W., & Zhang, J. 1992. A markov random field model-based approach to image interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14**(6), 606–615. 1, 4, 13

Mortensen, Eric N., & Jia, Jin. 2006. Real-time semi-automatic segmentation using a bayesian network. *Pages 1007–1014 of: IEEE Conference on Computer Vision and Pattern Recognition.* 6, 15

Müller, Pascal, Wonka, Peter, Haegler, Simon, Ulmer, Andreas, & Van Gool, Luc. 2006. Procedural modeling of buildings. *ACM Transactions on Graphics*, **25**, 614–623. 11

Murphy, Kevin P. 1998. A brief introduction to graphical models and bayesian networks. 46

Murphy, Kevin P., Weiss, Yair, & Jordan, Michael I. 1999. Loopy belief propagation for approximate inference: An empirical study. *Pages 467–475 of: Uncertainty in Artificial Intelligence.* 24

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann. 23, 24, 45, 48, 49

Petrou, M., & Bosdogianni, P. 1999. *Image Processing: The Fundamentals.* Wiley. 42

Plath, Nils, Toussaint, Marc, & Nakajima, Shinichi. 2009. Multi-class image segmentation using conditional random fields and global classification. *Pages 817–824 of:* Bottou, Léon, & Littman, Michael (eds), *International Conference on Machine Learning.* 13, 14, 39, 40

Potts, Renfrey B. 1952. Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society*, **48**, 106–109. 26, 27

Pérez, P. 1998. Markov random fields and images. *CWI Quarterly*, **11**(4), 413–437. 24, 25

Reynolds, J., & Murphy, K. 2007. Figure-ground segmentation using a hierarchical conditional random field. *Pages 175–182 of: Canadian Conference on Computer and Robot Vision.* 13, 14, 40

Ripperda, N., & Brenner, C. 2009. Evaluation of structure recognition using labelled facade images. *Pages 532–541 of: Annual Symposium of the German Association for Pattern Recognition (DAGM).* 11

Rother, Carsten, Kolmogorov, Vladimir, & Blake, Andrew. 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, **23**, 309–314. 44

Rottensteiner, F., Trinder, J., Clode, S.P., & Kubik, K. 2007. Building detection by fusion of airborne laser scanner data and multi-spectral images: Performance evaluation and sensitivity analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, **62**(2), 135–149. 10

Russell, Bryan C., Freeman, William T., Efros, Alexei A., Sivic, Josef, & Zisserman, Andrew. 2006. Using multiple segmentations to discover objects and their extent in image collections. *Pages 1605–1614 of: IEEE Conference on Computer Vision and Pattern Recognition.* 14

Russell, Chris, Ladicky, L'ubor, Kohli, Pushmeet, & Torr, Philip. 2010. Exact and approximate inference in associative hierarchical networks using graph cuts. *Pages 501–508 of: Uncertainty in Artificial Intelligence.* 48

Sarkar, S., & Boyer, K. L. 1993. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**, 256–274. 1, 2, 15

Schmittwilken, Jörg, Yang, Michael Ying, Förstner, Wolfgang, & Plümer, Lutz. 2009. Integration of conditional random fields and attribute grammars for range data interpretation of man-made objects. *Annals of GIS*, **15**(2), 117–126. 11, 81

Schnitzspan, P., Fritz, M., & Schiele, B. 2008. Hierarchical support vector random fields: Joint training to combine local and global features. *Pages 527–540 of:* Forsyth, D., Torr, P., & Zisserman, A. (eds), *European Conference on Computer Vision.* 14, 15

Schnitzspan, P., Fritz, M., Roth, S., & Schiele, B. 2009. Discriminative structure learning of hierarchical representations for object detection. *Pages 2238–2245 of: IEEE Conference on Computer Vision and Pattern Recognition.* 13

Shotton, Jamie, JohnWinn, Rother, Carsten, & Criminisi, Antonio. 2006. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *Pages 1–15 of: European Conference on Computer Vision.* LNCS, vol. 3951. 13, 14, 17, 37, 43, 44, 61, 77

Shotton, Jamie, Johnson, Matthew, & Cipolla, Roberto. 2008. Semantic texton forests for image categorization and segmentation. *Pages 1–8 of: IEEE Conference on Computer Vision and Pattern Recognition.* 42, 43, 45

Spinello, L., Triebel, R., Vasquez, D., Arras, K.O., & Siegwart, R. 2010. Exploiting repetitive object patterns for model compression and completion. *Pages V: 296–309 of: European Conference on Computer Vision.* 12

Sutton, C., & McCallum, A. 2005. Piecewise training for undirected models. *Pages 568–575 of: Uncertainty in artificial intelligence.* 45

Taskar, B., Chatalbashev, V., & Koller, D. 2004. Learning associative markov networks. *Pages 102 – 109 of: International Conference on Machine Learning.* 48

Teboul, Olivier, Simon, Loic, Koutsourakis, Panagiotis, & Paragios, Nikos. 2010. Segmentation of building facades using procedural shape priors. *Pages 3105–3112 of: IEEE Conference on Computer Vision and Pattern Recognition.* 10, 11

Toyoda, T., & Hasegawa, O. 2008. Random field model for integration of local information and global information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(8), 1483–1489. 6, 13, 14

Tsotsos, J.K. 1988. A 'complexity level' analysis of immediate vision. *International Journal of Computer Vision*, **2**(1), 303–320. 1

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. 2010. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(9), 1582–1596. 11

Vedaldi, A., & Fulkerson, B. 2008. Vlfeat: An open and portable library of computer vision algorithms. *http://www.vlfeat.org/*. 42

Veksler, Olga. 1999. *Efficient graph-based energy minimization methods in computer vision*. Ph.D. thesis, Cornell University, Ithaca, NY, USA. 48

Vincent, Luc, & Soille, Pierre. 1991. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**(6), 583–598. 41, 52, 56, 63

Wainwright, Martin J., Jaakkola, Tommi S., & Willsky, Alan S. 2005. Map estimation via agreement on trees: message-passing and linear programming. *IEEE Transactions on Information Theory*, **51**, 3697–3717. 27

Wendel, A., Donoser, M., & Bischof, H. 2010. Unsupervised facade segmentation using repetitive patterns. *Pages 51–60 of: Annual Symposium of the German Association for Pattern Recognition (DAGM)*. 12

Wenzel, S., & Förstner, W. 2008. Semi-supervised incremental learning of hierarchical appearance models. *Pages 399–404 of: 21st Congress of the International Society for Photogrammetry and Remote Sensing*. IAPRS 37 (B3b). 12

Wu, C.C., Frahm, J.M., & Pollefeys, M. 2010. Detecting large repetitive structures with salient boundaries. *Pages II: 142–155 of: European Conference on Computer Vision*. 12

Xie, Xianchao, Geng, Zhi, & Zhao, Qiang. 2006. Decomposition of structural learning about directed acyclic graphs. *Artificial Intelligence*, **170**(4-5), 422–439. 81

Yang, L., Meer, P., & Foran, D.J. 2007. Multiple class segmentation using a unified framework over mean-shift patches. *Pages 1–8 of: IEEE Conference on Computer Vision and Pattern Recognition*. 13

Yang, Michael Ying, & Förstner, Wolfgang. 2011a. Feature evaluation for building facade images - an empirical study. *Department of Photogrammetry, Institute of Geodesy and Geoinformation, University of Bonn, TR-IGG-P-2011-02*. 41

Yang, Michael Ying, & Förstner, Wolfgang. 2011b. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. *Pages 196 – 203 of: International Conference on Computer Vision, IEEE/ISPRS Workshop on Computer Vision for Remote Sensing of the Environment*. 15, 25, 40

Yang, Michael Ying, & Förstner, Wolfgang. 2011c. Regionwise classification of building facade images. *Pages 209 – 220 of: Photogrammetric Image Analysis (PIA2011)*. LNCS 6952. Springer. 25, 39, 64, 65, 66, 67, 72, 75, 76

Yang, Michael Ying, Förstner, Wolfgang, & Drauschke, Martin. 2010a. Hierarchical conditional random field for multi-class image classification. *Pages 464–469 of: International Conference on Computer Vision Theory and Applications*. 15, 39, 40

Yang, Michael Ying, Cao, Yanpeng, Förstner, Wolfgang, & McDonald, John. 2010b. Robust wide baseline scene alignment based on 3d viewpoint normalization. *Pages 654–665 of: International Conference on Advances in Visual Computing*. Springer-Verlag. 12

Yang, Michael Ying, Cao, Yanpeng, & McDonald, John. 2011. Fusion of camera images and laser scans for wide baseline 3D scene alignment in urban environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, **66**(6, Supplement), S52 – S61. 12

Yedidia, J.S., Freeman, W.T., & Weiss, Y. 2000. Generalized belief propagation. *Pages 689–695 of: Advances in Neural Information Processing Systems*, vol. 13. 24, 27, 45, 48, 49

Zhang, Lei, & Ji, Qiang. 2010. Image segmentation with a unified graphical model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(8), 1406–1425. 15, 16, 17, 18

Zhang, Lei, & Ji, Qiang. 2011. A bayesian network model for automatic and interactive image segmentation. *IEEE Transactions on Image Processing*, **20**(9), 2582–2593. 15, 16

Zhang, Lei, Zeng, Zhi, & Ji, Qiang. 2011. Probabilistic image modeling with an extended chain graph for human activity recognition and image segmentation. *IEEE Transactions on Image Processing*, **20**(9), 2401–2413. 15, 17, 18

Zhu, Jun, Lao, Ni, & Xing, Eric P. 2010. Grafting-light: fast, incremental feature selection and structure learning of markov random fields. *Pages 303–312 of: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. 81