Conditional Random Fields for Detection of Visual Object Classes

A dissertation submitted to
TECHNISCHE UNIVERSITÄT DARMSTADT
Fachbereich Informatik

for the degree of
Doctor rerum naturalium (Dr. rer. nat.)

presented by

PAUL SCHNITZSPAN
Dipl.-Math.

born in Seeheim-Jugenheim, Germany

Prof. Stefan Roth, PhD, examiner
Prof. Dr. Bernt Schiele, co-examiner

Date of Submission: 23rd of July, 2010
Date of Defense: 3rd of September, 2010

Darmstadt, 2010
D17

ABSTRACT

High-level computer vision tasks, such as object detection in single images, are of growing importance for our every day lives. Reliable systems for object detection, in particular, may simplify our lives significantly or make them safer (e.g. in driver assistance scenarios). Graphical models lend themselves to analyze and design computer vision algorithms because of their modularity that allows to design complex models built on simpler modules. This modularity and decomposability enables a better understanding of the domain of interest that in turn enables the design of models with increased reliability.

In this dissertation we study discriminative, undirected graphical models, namely conditional random fields (CRFs), and propose extensions to standard CRFs in order to address object detection in challenging scenes. We discuss the advantages of discriminative models compared to generative variants in the presence of cluttered background, partial occlusion and viewpoint variation. While standard CRFs are restricted to fixed, local neighborhood dependencies we propose to learn arbitrary graph structures. Furthermore, we take advantage of the decomposability of graphical models and propose to interpret the random variables as object parts and develop a joint approach of part-based and monolithic object detection. This view on objects yields a better and intuitive understanding of the structure of objects, and in accordance with observations of related work we demonstrate an improved reliability of our joint system.

A secondary focus of this work is the field of search and rescue robotics. Specifically, we are concerned with victim detection in search and rescue scenarios, which requires additional demands besides reliability. In this setting we require real-time capable models, hence, we need efficient algorithms without sacrificing performance. We propose to leverage the complementarity of different sensors (visual, thermal and laser in this work) within a sensor fusion scheme for an improved victim detection performance.

ZUSAMMENFASSUNG

Diese Dissertation beschäftigt sich mit der Lokalisierung von Objekten in komplexen Szenen. Die Lokalisierung von Objekten in solchen Szenen ist von immenser Bedeutung für unser tägliches Leben, weil zuverlässige Systeme unser Leben vereinfachen oder eine höhere Sicherheit garantieren könnten (z.B. in Fahrerassistenzprogrammen). Basierend auf graphischen Modellen werden Modelle vorgeschlagen, die ein besseres Verständnis von der Struktur von Objekten liefern können. Graphische Modelle eignen sich dafür besonders wegen ihrer Faktorisierbarkeit in einfachere Module.

Diese Arbeit untersucht diskriminative, ungerichtete graphische Modelle (sogenannte Conditional Random Fields). Um die anspruchsvollen Szenen handhaben zu können, werden Erweiterungen der ursprünglichen Modelle vorgeschlagen. Diese Erweiterungen ermöglichen ein besseres Verständnis der Objekt Struktur und erzielen eine empirisch bewiesene bessere Genauigkeit. Dafür wird speziell die standardmäßige, lokal begrenzte Nachbarschaftsabhängigkeit durch beliebige Nachbarschaftsbeziehungen ersetzt. Ein effizienter Algorithmus zur Selektion der Nachbarchaften wird in das graphische Modell eingebunden. Weiterhin wird die Modularität der graphischen Modelle ausgenutzt und die einzelnen Zufallsvariablen als Objektteile interpretiert. Dadurch wird ein lokales Objektteile basiertes Modell mit einem globalen Objektmodell kombiniert, um, einhergehend mit verwandten Arbeiten, eine höhere Genauigkeit in der Lokalisierung von Objekten zu erzielen.

Ein weiterer Schwerpunkt der Arbeit ist die Entwicklung von Rettungsrobotern. Zusätzlich zu der Genauigkeit des Systems wird in diesem Szenario eine hohe Anforderung an die Laufzeit gestellt. Nur Modelle, die in Echtzeit und auf dem Roboter direkt laufen, sind hierfür adequat. In dieser Dissertation wird ein Modell vorgeschlagen, basierend auf mehreren verschiedenen Sensoren. Hier werden visuelle, Wärme- und Lasersensoren verwendet um schnelle aber trotzdem zuverlässige Modelle zu entwickeln.

ACKNOWLEDGEMENTS

First, I want to thank my advisors Bernt Schiele and Stefan Roth for their advice and guidance throughout my PhD studies. Especially during long nights before deadlines their active and productive help has been very helpful and I really enjoyed working with both of them even in tough times when nothing seemed to be working. Thank you for constantly pushing me to my limit and enhancing my horizon. For me it remains an unresolved phenomenon though, why the number of new ideas is super-exponential in the number of ideas of the advisors and not additive!?

Further, I want to thank Ursula Paeckel for her personal advice and remembering me of the things, which really matter. Without Uschi I probably would not be writing my dissertation right now. "Uschi hat den Schein ins Rollen gebracht". I am really grateful for her being the soul of the MIS group.

Special thanks go to the Postdocs accompanying me throughout my studies. Gyuri Dorko, Kristof Van Laerhoven and Diane Larlus, thank you for constantly saying "it will be fine!". Eventually, you are right.

I want to thank all of my colleagues, Mario Fritz, Edgar Seemann, Tam Huynh, Maja Stikic, Andreas Zinnen, Ulf Blanke, Christian Wojek, Mykhaylo Andriluka, Sandra Ebert, Michael Stark, Ulrich Steinhoff, Nikodem Majer, Stefan Walk, Markus Rohrbach, Zeeshan Zia, Anton Andriyenko, Christoph Vogel, Eugen Berlin and Marko Borazio. The discussions with you have been very fruitful and I enjoyed working with you. Even though we suffered during pre-deadline phases, we became stronger and wiser with every throwback and every overly negative review. Special thanks go to Mario Fritz for supervising my Diploma thesis and his advice at the beginning of my PhD thesis. Further I want to thank Christian Wojek for helpful suggestions on the HOG features and providing a considerable code base. Thanks to all my office mates for enduring my non-research comments.

Thanks go to my colleagues of GKmM and particularly to the Rescue Robot Team (Johannes Meyer, Karen Petersen, Stefan Kohlbrecher, Armin Strobel) for giving me the opportunity to work in an interdisciplinary environment. Working with you has been a nice and productive distraction from my own PhD studies.

I acknowledge financial support from the research training group GRK 1362. The unbureaucratic way of funding has been very helpful.

Further, I want to thank the students, I had the opportunity to work with: Oliver Schwahn and Konstantin Fuchs. Both of you helped a lot in the robot team.

I want to thank my family and non-research friends: my friends for distracting me from work after long days of research and my family for providing support and advice and giving me a place where I can always return to.

Last but not least and most important I want to thank my lovely girlfriend Steffi. Thank you for enduring all these tough phases of my PhD life and being there. Your love gave me the confidence to work on and finish this dissertation.

Kiwis rock!

CONTENTS

1 INTRODUCTION

Contents

Computer vision has gained increased attention in the last years, in which an increasing power of computers enabled the development of complex and powerful algorithms for automatic perception. Even though notable improvements have been made, the abilities of modern computer vision algorithms are still far behind the capabilities of human perception. Especially in high-level computer vision tasks the recently presented approaches cannot yet reach human performance. Nonetheless, high-level computer vision tasks and especially object detection, the main topic of this dissertation, are of substantial importance for our every day lives, as they may simplify them or make them safer. Consider, for example, a driver assistance scenario, in which the car warns the driver about dangerous situations or even drives completely autonomously without human interaction. In such scenarios, object detection can provide crucial information, for example, the location of pedestrians and other traffic participants. The knowledge of the locations of all traffic participants can in turn be used to decide on next steps in order to avoid accidents. Therefore, however, a generic object detector is desirable in order to detect all instances of all classes meaning that we cannot leverage prior knowledge about single classes (e.g. the configuration of pedestrians). Another example explicitly addressed in this dissertation is victim detection in search and rescue scenarios. Think of a collapsed building with possible human victims inside. In case it may be too dangerous for human rescuers to enter the building, one could send out teams of autonomous robots, searching for human victims.

Many high-level vision tasks, especially object detection, involve inference in challenging scenes, for example detection under partial occlusion, cluttered background or articulation. Consider for example Fig. 1.1 in which humans can still reliably locate partially visible object instances, even though the car in the front or the rightmost motorbike is only partially visible. Leveraging contextual information is key to the success of human perception. As an example, Fig. 1.1 shows a

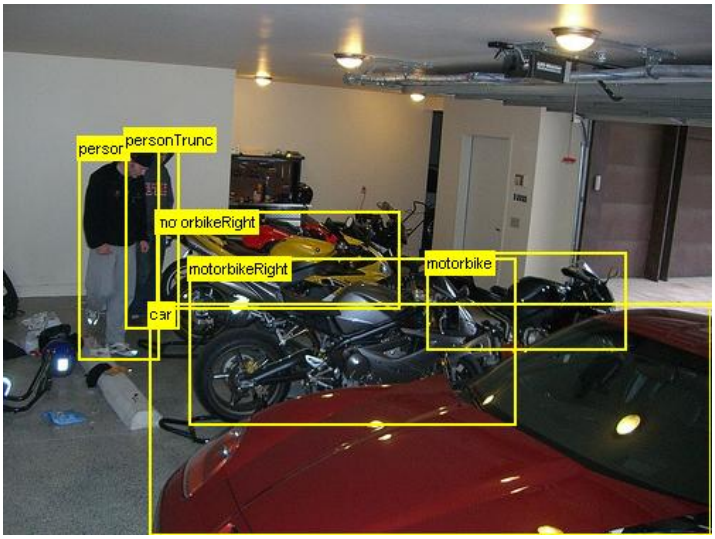


Figure 1.1: Example of a challenging scene. For humans the context of a garage-like scene helps to detect all object instances in the image. Even severely occluded and only partially visible instances can be reliably detected by humans, while computer vision approaches are still far behind the capabilities of human perception.

garage-like scene, in which one could expect object classes like cars and motorbikes. Moreover, human perception relies on contextual information on a more local level: Even though instances may only be partially visible, the visible parts yield enough evidence to infer the presence and location of objects. Here, global and local information of objects is exploited simultaneously: While the visible parts provide limited evidence, the global interpretation of the constellation of object parts and prior knowledge about the expected size of the instance yields good object detection performance. To this date, object detection models are not able to sufficiently represent such complex dependencies. This is due to lacking feature representations and noisy sensor measurements, but also due to the limitations of today's models themselves. Many models, for example, build on either local or global information, but not both simultaneously, or they do not represent dependencies among object parts, which would allow to model the structure of objects (i.e. the constellation of object parts). The main goal of this dissertation is to address both of these issues through hierarchical representations and modeling dependencies of object parts, and to propose probabilistic models that make a step toward richer feature representations and learning the constellation of object parts. Another secondary focus of this dissertation lies in advancing the field of search and rescue robotics on the sensing and planing side: This dissertation proposes a framework that builds on complementary sensor measurements such as visual and thermal data in order to overcome error-prone single sensor measurements.

In order to address the mentioned issues based on hierarchical representations and part constellations, we propose a probabilistic framework that is built on undirected graphical models. These undirected graphical models open up a natural

way to incorporate pairwise dependencies between nodes that can be interpreted as representing spatially located object parts. Let $\mathbf{y}$ describe (output) random variables that we want to infer while $\mathbf{x}$ denotes (input) random variables; in many computer vision applications $\mathbf{x}$ is interpreted as an input image while $\mathbf{y}$ is often interpreted as localized random variables that take on part labels of objects or model the presence or absence of objects. Traditionally, undirected graphical models represent the joint distribution $p(\mathbf{y}, \mathbf{x})$ of these output and input variables; these instantiations are also known as *generative* models. However, such approaches require to represent the distribution $p(\mathbf{x})$ of the input variables even though these are fixed when the model is applied, which causes difficulties if rich and overlapping feature descriptors are used. In this case, the dependencies of the features need to be modeled carefully, which could result in poor models. Consequently, *discriminative* variants have been proposed, which specifically address this issue (Lafferty *et al.*, 2001). These so-called conditional random fields model the conditioned distribution $p(\mathbf{y} \mid \mathbf{x})$ directly, which obviates representing dependencies among input variables $\mathbf{x}$ and allows for incorporating rich and overlapping feature descriptors. This flexibility in defining the feature descriptors is particularly appealing in complex high-level computer vision tasks such as object detection.

The main thesis put forward in this dissertation is that for challenging scenes discriminatively trained part-based models are more reliable than monolithic models or generatively trained variants. We study this thesis based on graphical models, which open up a natural way to model and interpret dependencies among object parts.

1.1 CHALLENGES

Object detection in real world images is a very challenging task as the characteristics of different object categories range, for example, from rigid objects such as cars to articulated objects such as cats, or they can be man-made or natural. Moreover, the appearance of objects of the same class can vary greatly across different instances. In the following we sketch some of these challenges of object detection in real world images.

Intra-class variation. The appearance of instances of the same object class often varies greatly, since the human understanding of one object class can often be very broad. In Fig. 1.2, for example, the aeroplane class varies from passenger plane to propeller-driven aircraft and the boat class contains instances of cruise ships and sail boats. This high intra-class variation imposes high demands on the expressiveness and flexibility of models that are used for object detection.

Partial occlusion. In real world images objects of interest are often partially occluded by other objects in the scene. While humans can often detect objects even in these challenging scenarios, computer vision algorithms cannot reach human

Figure 1.2: Examples of high intra-class variation. The aeroplane class shows instances of passenger plane and propeller-driven aircraft, while the boat class includes cruise ships and sail boats. This intra-class variation causes a high variation of the appearance of different object instances. The yellow boxes show the original annotations.

capabilities (see Fig. 1.3). Especially monolithic models such as (Dalal and Triggs, 2005) often suffer from incomplete visibility of objects, while part-based approaches such as (Leibe *et al.*, 2005) specifically leverage visible local information and are robust to a certain degree of partial occlusion. However, these part-based approaches are often prone to fake evidence in cluttered background, where false alarms like object-like structures may be assigned a higher confidence of the object's presence than partially occluded instances.

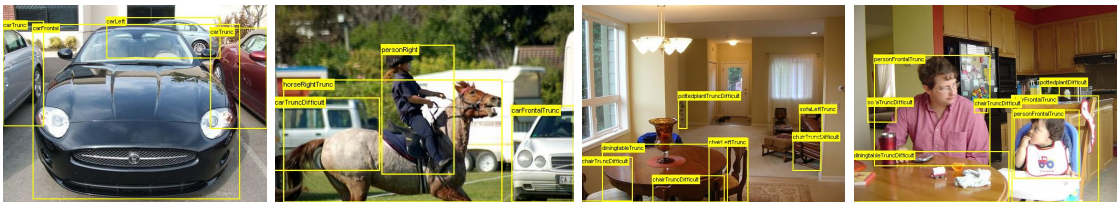


Figure 1.3: The two left looking cars in the back of the image are severely occluded and often cause difficulties to today's object detection algorithms. The chair class is specifically prone to occlusion since humans often happen to sit on chairs; in this case chair instances are often visible by only $10-30\%$. The yellow boxes show the original annotations.

Viewpoint variation. Viewpoint variation is particularly challenging for object classes for which the aspect ratio changes considerably with the viewpoint. Bicycles, for example, are relatively wide when seen from the side and narrow when seen from the front or back (see Fig. 1.4). The viewpoint furthermore influences the appearance of object instances. Canonical sideviews of cars, for example, show two wheels that are relatively salient for the presence of a car. Front or rear views, on the other hand, do not exhibit such discriminative parts. This again affects the visual appearance of objects, and with changing viewpoints, the appearance varies dramatically.

Figure 1.4: For some object classes the appearance of instances of that class changes significantly with the viewpoint. Sideviews of bicycle and cars for example show two wheels, which can discriminate objects from background while front or back views often do not show such discriminative object parts. The yellow boxes show the original annotations.

Articulation. Another challenge is added when we consider articulated objects like cats and dogs (Fig. 1.5). These object classes raise demands on the flexibility of detection models. An ideal model would adapt to physically plausible configurations of objects but not allow for impossible configurations of object parts. Ioffe and Forsyth (2001) discuss this problem for detecting pedestrians: Limb-like structure that does not correspond to people, but by chance occurs in groups resembling people. Here, we recognize a trade-off between being flexible enough to adapt to all possible articulations and being powerful enough to discriminate cluttered background from object instances.

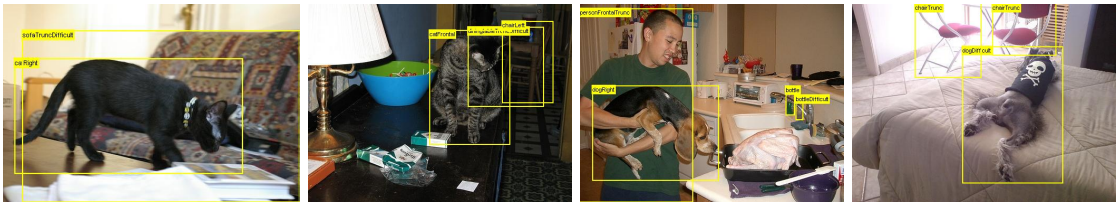


Figure 1.5: For articulated objects like cats and dogs the spatial constellation of object parts like head and legs can differ dramatically across object instances. This challenge imposes high demands on the flexibility of detection frameworks to model such variations in articulation. The yellow boxes show the original annotations.

Background clutter. Cluttered background complicates reliable object detection significantly, since object-like structures can appear accidentally in the background or make robust and reliable feature calculation difficult (clutter could distract from clear boundaries of object parts or entire objects; see Fig. 1.6). Flexible generative models often suffer from such cluttered background while discriminative monolithic models can handle these difficult scenes more reliably (this aspect will be discussed in the related work chapter in more detail). Nonetheless, the challenge of cluttered background is far from being solved and reliable models for object detection need to address this challenge explicitly.

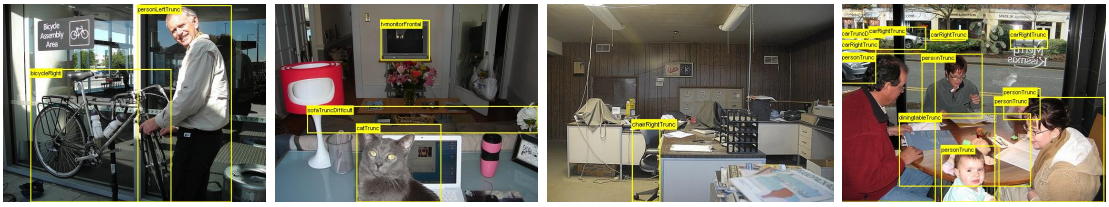


Figure 1.6: Cluttered scenes can often distract models to fake evidence in the background, in which object-like structures leads to a high confidence of an object's presence. The yellow boxes show the original annotations.

Real-time requirements. While the previously mentioned challenges are primarily addressed by models proposed in chapters 3, 4 and 5, the challenge of real-time capable models is discussed in chapter 6. For mobile robots it is crucial to detect objects of interest in real-time, since especially for search and rescue robotics the computation time is one of the most important considerations. Most of today's state-of-the-art models are not real-time capable since the primary focus is on the reliability of the system. Recently, efficient implementations and parallel techniques on the graphics processing unit have enabled speeding-up detection by two or more orders of magnitude.

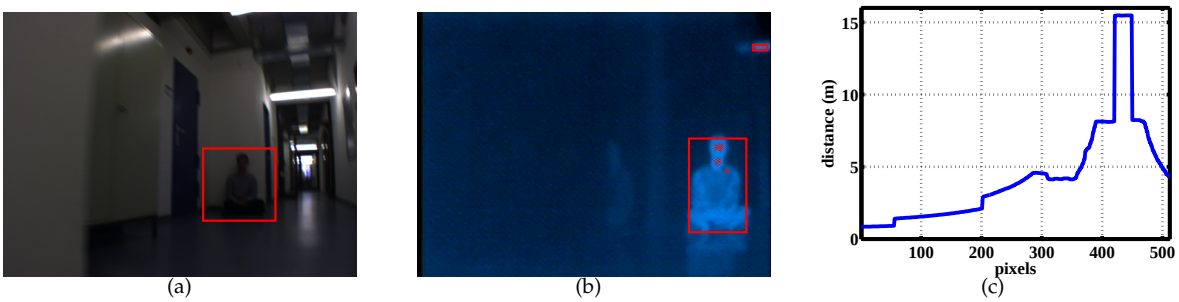
Lack of prior knowledge. Another challenging condition in robotic search and rescue scenarios is the assumption that no prior knowledge of the scene is given. Imagine a building that has collapsed and is explored by a robot: In this case we do not know in advance what the illumination conditions are and what the appearance of possible victims looks like. While in many datasets the data is captured and split afterwards into training and test images and thus the training and test images may be similar, in this scenario we are confronted with a much more challenging setting: The appearance of victims in the test images may differ dramatically from the training instances due to changes in illumination and in general no prior knowledge about the specific site and the appearance of the background is given.

1.2 DATASETS AND EVALUATION CRITERIA

This section briefly describes the datasets and evaluation criteria used throughout this dissertation. While we already introduced challenges of reliable object detection above, we now refer to the datasets that present all of the discussed challenges.

1.2.1 Datasets

In order to demonstrate the effectiveness and power of our models in chapters 3, 4, 5, we evaluate them on the so-called PASCAL VOC datasets (Everingham *et al.*, 2006, 2007). The images contained in these datasets are downloaded from flickr and comprise real world pictures taken from its users. In a sense these images are as close



(a) (b) (c)

Figure 1.7: Examples of sensor data: (a) Visual image with victim detection. (b) Thermal image with heat detections. (c) Range samples along the horizontal axis of the image.

to the real world as they could be and show most of the challenges discussed above. Among the most challenging issues are: partial occlusion, intra-class variability, viewpoint variation and background clutter. The PASCAL datasets are known to be very challenging and the best models on these datasets presented to date perform marginally above 30% average precision (this common evaluation criterion measures the area under the precision recall curve).

In order to evaluate our sensor fusion scheme of chapter 6 we recorded our own dataset since no adequate benchmarks are available to the best of our knowledge. Since we are interested in victim detection in search and rescue scenarios, we recorded an indoor setting with fairly complex instances of victims. We logged visual, thermal and laser data from the respective complementary sensors in order to evaluate the effectiveness of our sensor fusion scheme. Fig. 1.7 shows examples of visual, thermal and laser used for the sensor fusion experiments.

1.2.2 Evaluation criteria

Throughout the dissertation, we measured performance based on the PASCAL evaluation criterion (Everingham *et al.*, 2006, 2007), which is defined as follows: If the intersection area of the hypothesis and the ground truth divided by the union area of both exceeds 0.5, then the hypothesis is considered correct. The area of intersection and union is typically computed on the bounding boxes of the hypothesis and the ground truth. This evaluation criterion is also used for the PASCAL challenge and in many more settings.

1.3 CONTRIBUTIONS

The main contributions of this dissertation involve extended conditional random fields based models for object detection in challenging scenes. An overview of these contributions is given in Fig. 1.8: First we describe a part-based and hierarchical extension to standard one layer and foreground/background CRFs in chapter 3.

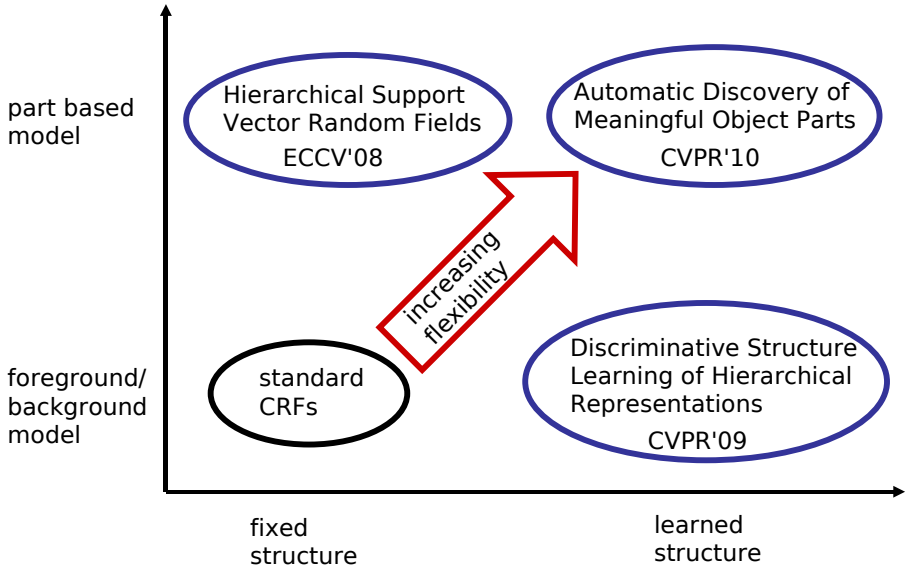
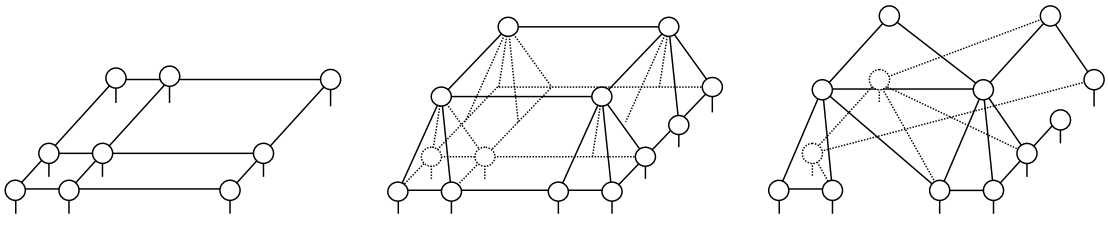


Figure 1.8: Overview over the contributions in CRFs for object detection. Starting from standard fixed structure, foreground/background CRFs we propose hierarchical and part-based representations to overcome the restriction to fixed structures by learning arbitrary pairwise graph representations.

Then we relax the restriction of fixed structure of the underlying graphical model as explained in chapter 4. Afterwards we combine both ideas (part-based + structure learning) in one single consistent framework, in which structure learning is carried forward to latent multi-class CRFs (chapter 5). An overview over the underlying graphical models is shown in Fig. 1.9: Fig. 1.9(a) depicts the fixed structure of standard one-layer CRFs, Fig. 1.9(b) shows the hierarchical structure used in chapter 3, while Fig. 1.9(c) gives an example of the general learned structure used in chapters 4 and 5. The last contribution of this work addresses sensor fusion of complementary devices for robust and reliable victim detection in search and rescue robotics (chapter 6). Another contribution in that direction has been published (Andriluka *et al.*, 2010), but is not described in this work.

Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features (Schnitzspan *et al.*, 2008). We propose a hierarchical model that combines the flexibility of part-based approaches and local representations with the expressiveness and power of monolithic feature descriptors. The model is built on a conditional random field that allows to incorporate rich feature descriptors of different scopes and enables a natural way to jointly learn and infer the entire hierarchy of the model. Therefore, our model is able to automatically learn the



(a) (b) (c)

Figure 1.9: Illustration of different graphical models: (a) Example of a standard, fixed one-layer structure. (b) example of a fixed hierarchical structure. (c) Example of a flexible, learned structure.

trade-off and interplay between local, semi-local and global feature contributions. Moreover, we rely on a joint and discriminative learning paradigm, which trains all model parameters jointly in a single consistent framework. Experimentally, we show that both the combination of local and global features as well as the joint training result in improved detection performance on challenging datasets.

Discriminative Structure Learning of Hierarchical Representations for Object Detection (Schnitzspan *et al.*, 2009). We propose a hierarchical multi-feature representation and automatically learn flexible hierarchical object models for a wide variety of object classes. To that end, we relax the restriction to fixed structures in conditional random fields by automatically selecting and modeling complex, long-range feature couplings within our model. To achieve this generality and flexibility our work combines structure learning in conditional random fields and discriminative parameter learning of classifiers using hierarchical features. We adopt an efficient gradient-based heuristic for model selection and carry it forward to discriminative, multidimensional selection of features and their couplings for improved detection performance.

Automatic Discovery of Meaningful Object Parts with Latent CRFs (Schnitzspan *et al.*, 2010). This part can be seen as a generalization of both of the previous contributions. We combine the power of discriminative models with the flexibility of part-based models and structure learning. Specifically, we propose a latent conditional random field (CRF) based on a flexible assembly of parts. By modeling part labels as hidden nodes and developing an EM algorithm for learning from class labels alone, our approach enables the automatic discovery of semantically meaningful object part representations. To increase the flexibility and expressiveness of the model, we learn the pairwise structure of the underlying graphical model at the level of object part interactions. Efficient gradient-based techniques are used to estimate the structure of the domain of interest and carried forward to the multi-label or object part case.

Object Detection in Search and Rescue Robotics (Meyer *et al.*, 2010; Andriluka *et al.*, 2010). In this part of the dissertation we focus on the challenge of onboard and real-time object detection on mobile and autonomous robots. We propose a probabilistic framework that leverages the complementary nature of different sensor types such as visual, thermal and laser devices. While sensor fusion is not new, the main focus of this part of the dissertation lies on integrating recent success in computer vision on object detection into the field of search and rescue robots. Particularly, we show that using powerful models combined with complementary information of different sensors benefits the detection performance even in the presence of erroneous sensor alignments.

While the sensor fusion scheme for ground vehicles is described in this dissertation, we want to point to an extension to aerial vehicles, which is not described in this dissertation. In (Andriluka *et al.*, 2010), we describe how inertial sensor measurements of a quad-rotored aerial vehicle can benefit detection of human victims lying on the ground.

1.4 OUTLINE

The dissertation is structured as follows:

Chapter 2 (related work) describes the notation of this dissertation and basic concepts that are important for understanding this work, and reviews related work on object detection in monocular, single images of challenging scenes and sensor fusion in robotic systems. We structure the related work along the main axes *hierarchical and part-based models*, *conditional random fields*, *structure learning in graphical models* and *sensor fusion for mobile robotics*.

Chapter 3 (hierarchical support vector random fields) describes our efforts to define a hierarchical, multi-class graphical model that learns all model parameters jointly in a single consistent framework.

The work presented in this chapter corresponds to the ECCV'08 publication (Schnitzspan *et al.*, 2008) "Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features".

Chapter 4 (discriminative structure learning) extends standard fixed structure conditional random fields to automatic selection of pairwise feature couplings. Here we show how the estimation of the underlying structure of the domain of interest improves the detection performance in challenging object detection datasets.

The work presented in this chapter corresponds to the CVPR'09 publication (Schnitzspan *et al.*, 2009) "Discriminative Structure Learning of Hierarchical Representations for Object Detection".

Chapter 5 (latent CRFs) combines the advantages of chapters 3 and 4. In this chapter

we combine the expressiveness of part-based models and the flexibility of structure learning in order to automatically learn a meaningful part-based representation in a weakly supervised framework.

The work presented in this chapter corresponds to the CVPR'10 publication (Schnitzspan *et al.*, 2010) "Automatic Discovery of Meaningful Object Parts with Latent CRFs".

Chapter 6 (sensor fusion for mobile robotics) discusses sensor fusion techniques for onboard and real-time processing on a mobile robotic platform. This chapter describes interdisciplinary and joint work conducted in the research training group "Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments". While in the other chapters the focus lies on advancing the field of object detection in challenging scenes, this chapter focuses on meeting real-time requirements of autonomous search and rescue robotics.

The work presented in this chapter corresponds to the RoboCup Symposium'10 publication (Meyer *et al.*, 2010) "A Semantic World Model for Urban Search and Rescue Based on Heterogeneous Sensors".

Chapter 7 (conclusion and future work) concludes the dissertation and discusses possible future directions that are not yet covered by this work.

RELATED WORK 2

Contents

In the context of computer vision, object detection has been approached from two opposing directions – generative and discriminative models. Both paradigms are inspired by probability theory since both of them model the probability of random variables $\mathbf{x}$ and $\mathbf{y}$; in our terminology we refer to $\mathbf{x}$ as the input variables while $\mathbf{y}$ refers to the output. We interpret $\mathbf{x}$ often as an image while $\mathbf{y}$ are spatially distributed object parts. Both, generative and discriminative models, have in common that during inference (detecting object instances in our case) they compute the probability $p(\mathbf{y} \mid \mathbf{x})$ of output variables $\mathbf{y}$ conditioned on the input $\mathbf{x}$. However, they differ in the interpretation of the learning and modeling task. A generative model represents a full probability of all (input + output) variables while a discriminative model provides a model for only the output variable conditioned on the input. While discriminative models learn this conditional distribution directly, generative variants model the joint probability of input and output variables. The joint probability is more general by allowing to generate new and artificial samples of the joint probability while the conditional probability learns the classification into positive and negative instances directly. The discussion, which paradigm is preferable, has not yet come to a conclusive end. Typically, generative models perform better if only few training samples are given while discriminative models happen to be advantageous in the presence of a large training set. In this work, we make use of discriminative models as they have been shown to be advantageous on the challenging datasets considered in this dissertation (Dalal and Triggs, 2005; Felzenszwalb *et al.*, 2008). Another important aspect of this dissertation is the study of part-based models. Recently, exploiting part-based models has been shown to considerably support monolithic object detection (Felzenszwalb *et al.*, 2008; Desai *et al.*, 2009). Combining

both worlds in one model yields the discriminative power of monolithic models but also a higher flexibility due to deformable constellations of object parts. We study the applicability of graphical models for learning such representations. In general, graphical models provide a natural way to analyze and design machine learning and computer vision models. One major advantage of graphical models lies in their decomposability that allows to compose complex models of simpler modules. The probabilistic nature of graphical models ensures consistency of the model and enables the linkage of the model to the data. The graph theoretical part defines the way the different modules interact with each other. In our setting we leverage the graph structure to represent dependencies among spatially located object parts. As we show in chapter 5 we take advantage of the modularity of graphical models by marginalization over the vertices that is beneficial for an increased reliability.

Recently, conditional random fields (Lafferty *et al.*, 2001) have been proposed as discriminative graphical models and leveraged for object detection (Quattoni *et al.*, 2004; Kapoor and Winn, 2006; Winn and Shotton, 2006; Hoiem *et al.*, 2007). These conditional random field approaches show favorable properties by opening a natural way to directly model neighborhood and longer-range dependencies within and between object instances and between object and background. While most models are restricted to a fixed pairwise connected graph structure, recently discriminatively learning arbitrary structures have been shown to be feasible (Schmidt *et al.*, 2008). In our work we are not only interested in the pure object detection problem, but also want to automatically learn a meaningful object part representation of instances of a given class since this has proven to increase the reliability of object detection frameworks (Felzenszwalb *et al.*, 2008). As a desirable side product we can infer a decomposition of objects into parts, which in turn can be used for further interpretations of an instance, for example the viewpoint or the pose of objects. Therefore, we discuss part-based models in the related work section, too, again looking at the discrepancy between generative and discriminative models. In the remainder of the chapter, we review the basic concepts this thesis is built on and describe related work that we roughly structure along four main axes namely *hierarchical, part-based and multi-feature models*, *conditional random fields*, *structure learning in graphical models* and *sensor fusion for mobile robotics*.

2.1 BACKGROUND AND NOTATION

In the following we describe basic concepts and briefly summarize the notation used in this thesis. We try to keep the same notation throughout this work as especially in chapters 3, 4 and 5 we build our model on conditional random fields (CRFs), making a unified notation desirable and valuable. In this section we describe the basic aspects, which are most important to understand the conditional random field models described in this thesis. An extensive discussion is clearly beyond the scope of this thesis and we refer to other more comprehensive work for further readings: (Wainwright and Jordan, 2003; Yedidia *et al.*, 2003; Bishop, 2006). We start with

describing graphical models, conditional random fields and support vector machines and end with summarizing the notation.

2.1.1 Graphical models

Graphical models have the appealing property that they help in understanding and formalizing probability distributions over many variables. Consequently, graphical models find widespread use in many fields, among them computer vision. An interesting aspect about graphical models is that the associated graphical structure provides insight into the factorization of the probability of the random variables. The graphical model is denoted by $\mathcal{G} = (V, E)$, where V indexes the set of random variables $\mathbf{x} \cup \mathbf{y}$, where $\mathbf{x}$ is typically referred to as the input (e.g. an image) while $\mathbf{y}$ denotes the output variables. E denotes the set of edges, where each edge (i, j) connects two nodes $i, j \in V$. Graphs can either be directed or undirected as described in the following.

Directed graphical models. Directed graphical models are also known as Bayesian networks. In these models, the edges are directed and the graph needs to be acyclic in order to guarantee a factorization, which obeys the causality of the random variables. A directed edge defines a parent-child hierarchy, in which the parents $\pi(i)$ of node i have an edge pointing to node i. The joint probability distribution of random variables $\mathbf{x}, \mathbf{y}$ is defined over these parent-child dependencies

$$p(\mathbf{y}, \mathbf{x}) = \prod_{i \in V} p(v_i | \pi(i)) \ , \tag{2.1}$$

where $\pi(i)$ denotes the set of random variables belonging to the parents of node i. v_i refers to a random variable out of $\mathbf{x} \cup \mathbf{y}$. An appealing property of directed graphical models is the inherent causal structure modeled with the parent-child dependencies. However, the restriction to acyclic graphs prohibits widespread use in complex computer vision problems.

Undirected graphical models. Undirected graphical models generalize the directed instantiations in that they allow for arbitrary structures including cyclic structures. In this case, the factorization of the model is described over maximal cliques that form fully connected subsets of nodes.

$$p(\mathbf{y}, \mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_{c \in C} \psi_c(\mathbf{y}_c, \mathbf{x}_c) \ , \tag{2.2}$$

where ψ_c defines a potential function of the random variable associated to clique $c \in C$. C refers to the set of all maximal cliques. The partition function (a normalization factor) $\mathcal{Z}$ ensures integration of $p(\mathbf{y}, \mathbf{x})$ (or summation) to 1 and is calculated as

$$\mathcal{Z} = \int_{\mathbf{y} \cup \mathbf{x}} \prod_{c \in C} \psi_c(\mathbf{y}_c, \mathbf{x}_c) d(\mathbf{y} \cup \mathbf{x}) \ . \tag{2.3}$$

Modeling the joint probability of $\mathbf{x}$ and $\mathbf{y}$ with an undirected graphical model is also referred to as Markov random fields if the random variables obey the Markov property. This Markov property requires a random variable to be independent of all other variables given its neighbors.

The general undirected formalism involves a major drawback. Exact computation of the partition function $\mathcal{Z}$ is computational expensive since the space, which needs to be integrated over, is exponentially large making brute force computation intractable. As long as the graph obeys a tree structure exact inference is feasible with smart inference techniques. However, for arbitrary and highly connected graphs, on which we often rely in our work, exact inference is typically intractable, drawing the need for approximate inference techniques.

Inference. Inference provides information about (unobserved) variables in the model. Typically, we want to compute the marginal distribution $p(y_i)$ of single nodes (or sets of nodes). Moreover in undirected graphical models, inference yields (an approximation of) the partition function $\mathcal{Z}$. While tree-structured graphs can be inferred efficiently (Pearl, 1988), in general graphs (especially cyclic graphs) exact inference is intractable. In this case, we need to resort to approximate inference techniques. In our case we use the loopy variant of belief propagation that was shown to perform often surprisingly well even for complex cyclic graphs (Murphy *et al.*, 1999).

2.1.2 Conditional random fields

Conditional random fields (CRFs) are discriminative models (in contrast to generative Markov random fields) building on undirected graphical models. CRFs represent the conditional probability $p(\mathbf{y} \mid \mathbf{x})$ of discrete random variables $\mathbf{y}$ given an input $\mathbf{x}$; typically $\mathbf{x}$ is an image in our terminology. The conditioned modeling of the random variables given the input shows convenient properties especially for complex high-level tasks like object detection. Since the distribution of random variables is conditioned on the input, we do not need to explicitly represent dependencies among the input variables. This aspect is particularly interesting in the case of hierarchical and overlapping feature descriptors as considered in this thesis.

An undirected graphical model is a conditional random field if all output variables obey the Markov property, meaning a random variable is conditionally independent of all variables given its neighbors. Generally, CRFs can be written as

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \frac{1}{\mathcal{Z}}(\mathbf{x}; \theta) \prod_{c \in C} \psi_c(\mathbf{y}_c, \mathbf{x}; \theta) \ , \tag{2.4}$$

where θ denotes the model parameters to be learned. Now the random variables $\mathbf{y}$ are conditioned on the input $\mathbf{x}$ and the clique potential function depends on the entire input and not only subsets of the input. Instead of using the clique notation we often build on a pairwise structure of the graph. In this case we explicitly differentiate between types of potentials (unary and pairwise). The conditional

probability is then defined as

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \frac{1}{\mathcal{Z}(\theta, \mathbf{x})} \prod_{i \in V} \psi_i(y_i, \mathbf{x}; \theta) \prod_{(i,j) \in E} \phi_{ij}(y_i, y_j, \mathbf{x}; \theta) \ , \tag{2.5}$$

where ψ refers to the unary potentials, whereas ϕ refers to the pairwise potentials.

The main difference to Markov random fields (MRFs) lies in the direct definition of the conditioned probability. In contrast, MRFs model the joint distribution $p(\mathbf{y}, \mathbf{x}; \theta)$, which requires careful definition of the feature functions in order to guarantee tractable learning of the model. This is due to the need to model $p(\mathbf{x})$ in MRFs, in which dependencies among the features has to be covered by the graphical model rather than only among output variables $\mathbf{y}$ as is the case for CRFs. Sutton and McCallum (2007) discuss the main advantage of CRFs over MRFs as follows: CRFs allow for rich and overlapping features. This can be explained by the fact that we do not need to model $p(\mathbf{x})$ and thus CRFs make independence assumptions among $\mathbf{y}$ but not among $\mathbf{x}$.

Originally, CRFs have been paired with linear feature functions in the potential functions:

$$\psi_i(y_i, \mathbf{x}; \theta) = \exp\left(\theta_{y_i}^T f_i(\mathbf{x})\right) \ , \tag{2.6}$$

with a feature function f_i that defines the features of node i. The pairwise potentials are typically defined similarly:

$$\phi_{ij}(y_i, y_j, \mathbf{x}; \theta) = \exp\left(\theta_{y_i y_j}^T g_{ij}(\mathbf{x})\right) \ , \tag{2.7}$$

with a feature function g_{ij} that maps the input to the joint feature of nodes i and j.

The learning task is presented as maximizing either the conditional log-likelihood or the log-posterior over a set of M images $X = (\mathbf{x}^1, \ldots, \mathbf{x}^M)$:

$$\mathcal{L}(\theta) = \log p(Y|X; \theta) = \sum_{m \in M} \log p(\mathbf{y}^m | \mathbf{x}^m; \theta) \ , \tag{2.8}$$

where $Y = (\mathbf{y}^1, \ldots, \mathbf{y}^M)$ refers to all random variables. The log-posterior adds an additional prior $\log p(\theta)$ imposed on the model parameters. In the case of linear potentials, the optimization problem is convex and optimization techniques such as gradient descent are guaranteed to reach the global optimum.

Hidden variables. The CRF model above assumes that all random variables $\mathbf{y}$ are observed. However, in some cases it is useful to define hidden random variables for which the label assignment is not known in advance. These hidden variables are marginalized out in order to compute the probability of observed variables given the input. Let $\mathbf{z}$ denote latent random variables:

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \mathbf{x}; \theta) \ . \tag{2.9}$$

In this case, the potential functions can be defined to additionally depend on the latent variables $\mathbf{z}$. Unfortunately the optimization problem with latent variables in

the graphical model is no longer convex, which may result in local optima. In this case good initializations of the hidden variables can be helpful to avoid these local optima.

2.1.3 Support vector machines

Since we use support vector machines (SVMs) in our model, we want to introduce the notion of such models as well (Boser *et al.*, 1992; Vapnik, 1998). SVMs are discriminative models that aim to optimally separate positive examples from the negative ones. The foundation of their success on many challenging computer vision tasks lies in the maximum margin paradigm that aims to fit a separating hyperplane to the training data with maximum distance to opposing classes. The dual optimization problem is defined as

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} t_i t_j \alpha_i \alpha_j K(\mathbf{x}^i, \mathbf{x}^j) \\ &\text{s.t. } 0 \leq \alpha_i \leq C, \ \forall i = (1, \ldots, l) \qquad , \\ &\sum_{i=1}^{l} t_i \alpha_i = 0 \end{aligned} \tag{2.10}$$

where $\boldsymbol{\alpha}$ refers to the support vector coefficients. $t_i \in \{-1, 1\}$ denotes the label of training sample $\mathbf{x}^i$. $K(\cdot, \cdot)$ describes a Mercer kernel - a scalar product in a potentially high dimensional vector space. l limits the number of training samples while C refers to a penalty factor. Classifying a new test sample $\mathbf{x}$ can be computed as:

$$F(\mathbf{x}) = \sum_{i \in S} \alpha_i K(\mathbf{x}^i, \mathbf{x}) + \alpha_0 \ , \tag{2.11}$$

where S denotes the set of support vectors - those training examples $\mathbf{x}^i$, for which $\alpha_i \neq 0$. α_0 denotes an offset. $F(\mathbf{x})$ is the margin of test sample $\mathbf{x}$ - the signed distance to the separating hyperplane. The estimated class can be computed via sign $(F(\mathbf{x}))$.

2.1.4 Summary of notation

In summary, the terminology of the most important variables, parameters and functions is defined as follows.

- m is an index to the (training) images while M denotes the number of images.
- $V = (1, \ldots, N)$ refers to the nodes of our graphical model.
- $E \subset V \times V$ denotes the set of edges connecting two nodes.
- Images or bounding boxes are indicated by $X = (\mathbf{x}^1, \ldots, \mathbf{x}^M)$.

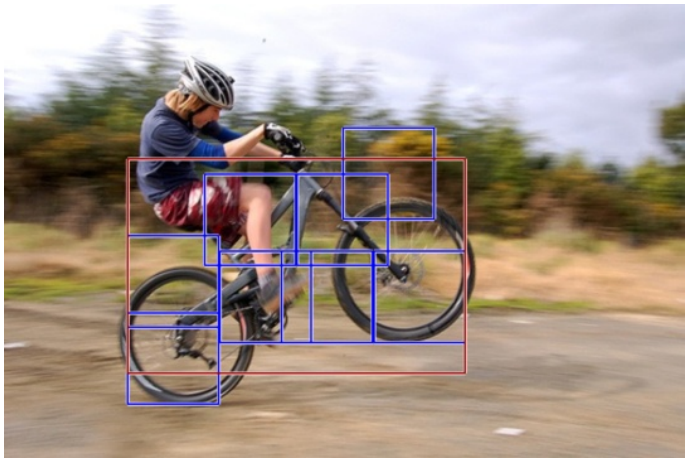
- Observed random variables are denoted by $Y = (\mathbf{y}^1, \ldots, \mathbf{y}^M)$ with $\mathbf{y}^m = (y_1^m, \ldots, y_N^m)$.
- Hidden random variables are referred to as $Z = (\mathbf{z}^1, \ldots, \mathbf{z}^M)$ with $\mathbf{z}^m = (z_1^m, \ldots, z_N^m)$.
- Both y_i^m and z_i^m can take on the labels $\{0, \ldots, P\}$
- The labels $\{0, \ldots, P\}$ can be interpreted as object parts where 0 stands for background. In the simpler foreground/background case, the random variables are drawn out of $\{-1, 1\}$ or $\{0, 1\}$.
- $\mathcal{Z}$ refers to the partition function - a normalization factor.
- θ denotes the set of model parameters that are optimized during the training phase.
- ψ and ϕ refer to the unary and pairwise potentials respectively.
- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are used for support vector coefficients while S refers to a set of support vectors in the terminology of support vector machines.

2.2 HIERARCHICAL, PART-BASED AND MULTI-FEATURE MODELS

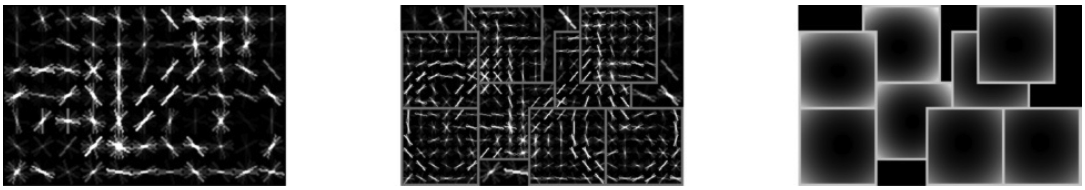
In the following we discuss related work on hierarchical, part-based and multi-feature models that have been primarily used for object detection and object recognition.

Part-based models. The foundation of part-based models goes back to the seminal work of Fischler and Elschlager (1973). This work initiated an extensive and still ongoing discussion about part-based models that are considered powerful due to their expressiveness and their intuitive interpretation.

Generative part-based models show favorable properties like factorizability into different components that can often be interpreted as object parts, and a certain robustness to partial occlusion (Fergus *et al.*, 2003; Leibe *et al.*, 2004; Felzenszwalb and Huttenlocher, 2005). This robustness stems from the respective local representation of objects and the interaction between this local representation either directly or implicitly over an anchor point. These local interactions of parts usually enable another favorable property, namely robustness to articulation and local deformations of object instances (Felzenszwalb and Huttenlocher, 2005; Amit and Trouve, 2007). Consequently, richer local representations have been developed yielding hierarchical part representations, which increase the semantic interpretability as well as the robustness to viewpoint changes (Bouchard and Triggs, 2005; Epshtein and Ullman, 2007). Moreover, interpreting local representations as object parts has been recently leveraged for knowledge transfer across object classes, allowing to share object parts such as for example wheels of cars, motorbikes and bicycles (Stark *et al.*, 2009; Sudderth *et al.*, 2008). Sharing of parts means that we want to learn a general class



(a)



(b)

Figure 2.1: (a) Example of a part-based model taken from (Felzenszwalb *et al.*, 2008). (b) Representation of global features (left), representation of located object parts (middle), and (right) canonical spatial locations of parts relative to the object instance.

of parts that can occur across different object classes, for example a wheel can either occur on cars or bicycles. The different constellations of object parts then describe the different object classes.

Unfortunately, all these generative part-based models comprehend certain drawbacks. Factorizing object instances into local representations often results in weakly structured and simplifying models (Leibe *et al.*, 2004) that could lead to hallucinated and superfluous object parts that correspond to fake evidence. In the case of fully connected models (Fergus *et al.*, 2003) training and inferring the model is computationally expensive leading to restrictions in the number of considered object parts. These observations have led to extensions of generative models that impose a global verification stage on top of the locally inferred hypothesis (Leibe *et al.*, 2005).

While generative models are said to perform better than discriminative variants if only few training data are available, discriminatively trained models typically enable higher levels of performance with increasing number of training instances (Dollár *et al.*, 2008; Kumar *et al.*, 2009). Consequently, Maji and Malik (2009) extended the generative model of Leibe *et al.* (2004) to a discriminatively trained max margin model that directly optimizes the classification problem instead of modeling the (potentially complex) distribution of positive samples.

This discriminative learning paradigm is often used in monolithic models, too. Contradictory to part-based models, these monolithic models encode objects in a

single typically high dimensional feature descriptor. Paired with discriminative classifiers such as SVMs, these models enable high levels of performance as shown by one of the most prominent representative of such models (Dalal and Triggs, 2005). However, monolithic models often require large amounts of training data and typically degrade in the presence of partial occlusion. In an effort to combine the advantages of both worlds (part-based as well as monolithic models) monolithic models have been enriched with the notion of parts and both have been presented to a discriminative classifier (Desai *et al.*, 2009; Felzenszwalb *et al.*, 2008). Fig. 2.1 shows an example detection and the feature representations inferred with (Felzenszwalb *et al.*, 2008). The additional notion of object parts yields a much higher flexibility than pure monolithic models and, at the same time, accounts for a better alignment of object instances.

Hierarchical and multi-feature models. Hierarchical models have been designed to strike a balance between local and global information of objects or images. Similar to part-based models the lower layers of the hierarchy are typically responsible for encoding a local view on objects or images while the higher layers encode a holistic view. Such hierarchical models lack the flexibility of purely part-based methods as they can be seen as monolithic models enhanced with local information. However, the two most prominent representatives of such models encode the hierarchical representation within a spatial pyramid kernel that is used in a discriminative classifier (Grauman and Darrell, 2007; Lazebnik *et al.*, 2006). The kernel definition within SVMs is replaced with a more general and powerful kernel:

$$K(f(\mathbf{x}^1), f(\mathbf{x}^2)) = \sum_{l \in L} \frac{1}{2^{L-l}} K_l(f(\mathbf{x}^1), f(\mathbf{x}^2)) \ , \tag{2.12}$$

where L refers to the number of layers and each K_l denotes a layer-specific kernel function. Typically, the kernels encode local information (bottom layer) to holistic views on the image or object (top layer). Even though the local features are encoded in a rather rigid spatial layout, the models showed promising performance on a variety of tasks.

An interesting extension to the hierarchical models mentioned so far is given in the work of Harzallah *et al.* (2009). While approaches based on spatial pyramid were mainly used for image classification the work of Harzallah *et al.* (2009) combines the latter task with object localization. The global information about the presence or absence of a specific object class in an image can benefit the local search for specific instances of that class and vice versa. Another model which uses spatial pyramid representations for object detection is discussed in (Lampert *et al.*, 2009). In this work the hierarchical representation of objects is enhanced by additional more fine grained layers, and even though this is in general computational expensive, the model remains applicable as it is combined with an efficient subwindow search.

In the work of Varma and Ray (2007) a multi-feature framework has been proposed building on the idea of multiple kernel learning (MKL) (Bach *et al.*, 2004). The appealing idea behind this approach lies in the automatic combination and impor-

tance weighting of different complementary feature descriptors. MKL generalizes the SVM model to consider weighted sums of different kernels, which can be seen as a generalization of the pyramid kernel above. Let d_k denote the weight of kernel k:

$$K(f(\mathbf{x}^1), f(\mathbf{x}^2)) = \sum_k d_k K_k(f^k(\mathbf{x}^1), f^k(\mathbf{x}^2)) \ , \tag{2.13}$$

with $\sum_k d_k = 1$. If all kernels K_k are Mercer, than the weighted sum is a Mercer kernel as well. Here, every kernel could refer to a kernel computed on one of several complementary feature descriptors. The use of such complementary features is considered beneficial since the complementary nature is exploited directly and drawing wrong conclusions due to lacking feature descriptors could be avoided. Consequently, an intensive but rather confusing discussion about the power of MKL in computer vision has been opened. Recently, Gehler and Nowozin (2009a) shed some light into the discussion damping the excitement and expectation on MKL, since it was shown that MKL is only advantageous in the presence of many uninformative features. However, the features used in computer vision are often engineered for the specific task making uninformative features unlikely. Subsequently, generalized variants of MKL have been proposed putting MKL back to success (Varma and Babu, 2009; Gehler and Nowozin, 2009b).

The work of Bosch *et al.* (2007) represents a first effort to combine hierarchical representations and multi-feature approaches. This idea was further traced in (Vedaldi *et al.*, 2009) showing promising levels of performances on a variety of computer vision tasks. Even though both approaches build on the original MKL formalism, the automatic combination of several complementary and rich feature descriptors happened to be responsible for enabling performance improvements.

Inspiration for our work. In this dissertation, we combine the power of hierarchical representations with the rich notation of part-based models. In contrast to sparsely represented models of related work, we aim at inferring a dense representation of objects and the object parts and explicitly consider partially occluded instances in our model. Inspired by recent success of discriminative models, we take a discriminative learning approach within a graphical model representation. As an outlook for future work, we show preliminary experiments with multi-feature models, which consistently improves the performance of object detection. Due to the generality of our work, more complementary feature descriptors can be easily adapted and integrated.

2.3 CONDITIONAL RANDOM FIELDS

As described above CRFs allow for rich feature descriptors that may overlap significantly. This property is especially appealing for incorporating hierarchical representations as is done in this dissertation. Further, due to the modularity and intuitive interpretability of the model components, CRFs can be modified easily in different aspects. Several related works have focused on different aspects and in the following

we cluster and describe related work on CRFs in the field of computer vision along three measures namely *powerful unary classifiers*, *contextual models* and *latent models*. As a brief reminder, the conditional probability of random variables $\mathbf{y}$ given input $\mathbf{x}$ based on a pairwise connectivity can be written as:

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \frac{1}{\mathcal{Z}(\theta, \mathbf{x})} \prod_{i \in V} \psi_i(y_i, \mathbf{x}; \theta) \prod_{(i,j) \in E} \phi_{ij}(y_i, y_j, \mathbf{x}; \theta) \ , \tag{2.14}$$

with model parameters θ, and unary and pairwise potentials ψ and ϕ.

Powerful unary classifiers. Standard CRFs usually build on linear classifiers for modeling the distribution of the random variables. However, in complex tasks like scene segmentation and object detection incorporating powerful classifiers has been proven beneficial in recent work. While the following methods agree on the benefit of incorporating powerful classifiers, the underlying learning paradigms vary greatly among the different approaches.

In (Lee *et al.*, 2005, 2006a) the linear classifiers are replaced with discriminative support vector machines (SVMs) for modeling the unary potential functions. In the notation of CRFs the unary potentials are defined as

$$\psi_i(y_i, \mathbf{x}; \boldsymbol{\alpha}) = \frac{\exp\left(F(\boldsymbol{\alpha}(y_i), f_i(\mathbf{x}))\right)}{\sum_c \exp\left(F(\boldsymbol{\alpha}(c), f_i(\mathbf{x}))\right)} \ , \tag{2.15}$$

with (S refers to the set of support vectors)

$$F(\boldsymbol{\alpha}(c), f_i(\mathbf{x})) = \sum_{j \in S} \alpha_j(c) K(f_j(\mathbf{x}), f_i(\mathbf{x})) + \alpha_0(c) \ . \tag{2.16}$$

In (Lee *et al.*, 2005) and (Lee *et al.*, 2006a) the authors train the SVM classifiers separately from the imposed pairwise feature couplings, resulting in an efficient training procedure and showing improved performance over standard linear models, even though the model alternates between the maximum margin formalism and the likelihood maximization paradigm of CRFs.

Another framework building on SVMs that comes particularly close to our joint learning paradigm discussed in chapter 3 is discussed in (Taskar *et al.*, 2003). In this work the authors specifically address the issue of different learning paradigms by transforming the primal random field formulation into its equivalent dual representation and training all model parameters simultaneously in a maximum margin learning scheme. This joint learning paradigm was proven beneficial showing improved performance compared to using different learning schemes.

Tsochantaridis *et al.* (2005) think the other way around and extend standard support vector machines to consider structured output variables. In contrast to adding powerful unary classifiers to a graphical model, this can be seen as adding a dependency structure to powerful classifiers like SVMs. The authors cast the entire learning problem into a consistent max margin support vector machine formulation.

In (Torralba *et al.*, 2004) and (Shotton *et al.*, 2009) boosting was integrated as a discriminative and powerful classifier replacing the linear potentials. Both methods relax the restriction to local neighborhood dependencies of standard CRFs by explicitly using longer-range dependencies of either neighboring objects of interest or neighboring image regions.

All of these CRF-based models replace the standard linear classifiers with more powerful and potentially non-linear classifiers. Especially the combinations with large margin classifiers have been shown to be appealing because they all lead to improved performance on the respective tasks compared to the standard linear formulation.

Contextual models. The foundation of using CRFs in computer vision was laid by Kumar and Hebert (2003) and more recently Kumar *et al.* (2006). These works have proposed to leverage contextual information of neighboring image regions yielding an improved detection and segmentation performance compared to non-contextual models. The authors propose a pairwise potential that specifically leverages discontinuities in the image:

$$\phi_{ij}(y_i, y_j, \mathbf{x}; \theta) = \exp\left(y_i y_j + \sigma\left(y_i y_j \theta^T g_{ij}(\mathbf{x})\right)\right) \quad , \tag{2.17}$$

where σ is a sigmoid mapping to the interval $[-1, 1]$ and g refers to the pairwise feature function. The first term $y_i y_j$ is a smoothness term favoring the same labels in neighboring nodes. Ideally, the data-dependent term (second term) will act as a discontinuity adaptive model that will moderate smoothing when the data from two sites is 'different'. The focus of this early CRF model lies on exploiting the discriminative nature of CRFs that has been shown to yield superior performance compared to generative Markov random fields.

As already briefly mentioned Torralba *et al.* (2004) and Shotton *et al.* (2009) integrate richer pairwise dependencies by directly exploiting contextual information in images. Torralba *et al.* (2004) directly relate objects or regions of interest to each other being able to discard unlikely constellations of objects. Shotton *et al.* (2009) trace a similar approach induced on a more local level, namely the class labels of neighboring image regions are set in context inferring only plausible combinations. The approaches of He *et al.* (2004) and He *et al.* (2006) are similar in spirit in that they leverage contextual information of the entire image in order to improve the reasoning about present object classes. While in (He *et al.*, 2004) an explicit global potential function is defined to model the spatial occurrence of object classes, the authors leverage super-pixel segmentations in (He *et al.*, 2006) and model likely cooccurrences at the longer-range super-pixel level. The work of He and Zemel (2008) extends the model of He *et al.* (2006) in that a latent topic model is introduced additionally to the longer-range super-pixel segmentation. This topic model exploits global image information as cooccurring image topics are modeled to restrict the cooccurrence of labels at the pixel level.

Larlus and Jurie (2008) combines a local Markov random field with a global object presentation. In this work inferred contours in the image are aligned with object

hypothesis in order to benefit from both, local as well as longer-range dependencies.

Kohli *et al.* (2009) describe a higher order CRF that combines the power of Shotton *et al.* (2009) with the strong notion of context at super-pixel level of He *et al.* (2006). The standard CRF model is augmented with an additional potential function, which works on super-pixels:

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \frac{1}{\mathcal{Z}(\theta, \mathbf{x})} \prod_{i \in V} \psi_i(y_i, \mathbf{x}; \theta) \prod_{(i,j) \in E} \phi_{ij}(y_i, y_j, \mathbf{x}; \theta) \prod_{c \in C} \lambda_c(c, \mathbf{x}; \theta) \ , \tag{2.18}$$

where C denotes the set of super-pixels and λ denotes a super-pixel potential function that favors the same labels for all nodes in one super-pixel. The authors explicitly mention the need for higher order potentials, which work on longer-range dependencies in order to incorporate a strong interpretation of context within images. In (Ladicky *et al.*, 2009) the model was further extended to a hierarchical model, which respects multiple super-pixel segmentations at the same time.

Another approach that leverages contextual information at a more local level is given in the work of Levin and Weiss (2006). The focus of this work lies on combining local and global information of objects in one consistent framework, which highly motivated our hierarchical feature representation described in chapters 3, 4 and 5. In this spirit the contextual information is integrated at the object scale and the holistic object information supports the description and arrangement of local features.

Similar in spirit to (Levin and Weiss, 2006) the work of Winn and Shotton (2006) goes beyond the notion of context at the object instance level. The authors describe a principled framework to guarantee so-called layout consistency of objects that ensures that all object instances are structured similarly. The layout consistency is exploited in the pairwise potentials, where all possible state transitions are distinguished:

$$-\log \psi_{ij}(y_i, y_j, \mathbf{x}; \theta) = \begin{cases} \theta_{\text{bg}}, & \text{background-background} \\ 0, & \text{consistent foreground} \\ \theta_{\text{fb}} g_{ij}(\mathbf{x}), & \text{foreground-background} \\ \theta_{\text{co}} g_{ij}(\mathbf{x}), & \text{class occlusion} \\ \theta_{\text{io}} g_{ij}(\mathbf{x}), & \text{instance occlusion} \\ \theta_{\text{ic}} g_{ij}(\mathbf{x}), & \text{inconsistent foreground} \end{cases} . \tag{2.19}$$

This model is particularly interesting because the challenging case of partially occluded object instances is treated directly by defining different valid transition states of direct neighbors (e.g. self occlusion, object boundary, class occlusion). Hoiem *et al.* (2007) extends the notion of layout consistency to the three-dimensional case. This generalization comes closer to the real world since under varying viewpoints the neighborhood dependency of consistent object parts obeys rather a 3D than a projected 2D propagation scheme. In the general case even complex occlusion scenarios and differing viewpoints can be handled reliably.

Latent models. The work of Quattoni *et al.* (2004) and Quattoni *et al.* (2007) comes particularly close to our framework presented in chapter 5 in terms of the model representation. Quattoni *et al.* (2004) extends CRFs to hidden-state CRFs by introducing latent nodes to the model:

$$p(\mathbf{y} \mid \mathbf{x}; \theta) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \mathbf{x}; \theta) = \frac{1}{\mathcal{Z}(\mathbf{x}, \theta)} \prod_{i \in V} \psi_i(y_i, z_i, \mathbf{x}; \theta) \prod_{(i,j) \in E} \phi_{ij}(y_i, y_j, z_i, z_j, \mathbf{x}; \theta) \ , \tag{2.20}$$

where $\mathbf{z}$ refers to the hidden random variables that are marginalized out. Especially for object detection these latent nodes are promising, since in most object detection datasets only bounding box labels at the object scale are given. Introducing hidden nodes opens up the opportunity to include a notion of object parts that is known to be advantageous in object detection. Since assignments of local features to object parts is not known from bounding box labels, the weakly supervised model of Quattoni *et al.* (2004) is shown to adapt to spatially well located object parts.

Kapoor and Winn (2006) extend the notion of hidden nodes to so-called located hidden random fields. While in (Quattoni *et al.*, 2004) the authors have inferred mostly local neighborhood dependencies, the work of Kapoor and Winn (2006) successfully introduces longer-range dependencies to the graphical model. The authors superimpose a global node at the object scale and define all local nodes to depend on this global state. Thereby, the resulting model is capable to build stronger relationships between the object parts and infer more meaningful object parts.

Inspiration for our work. Inspired by the expressive power of replacing linear potentials with powerful classifiers, we integrate SVMs as unary potentials within our CRF-based model. Since the discriminative nature of CRFs enables potentially high levels of performances, we decided to use the discriminative variant of graphical models instead of generative random fields. Since in most object detection challenges only bounding box labels are provided, we discuss a latent model in chapter 5, which is able to learn meaningful object part representations.

2.4 STRUCTURE LEARNING IN GRAPHICAL MODELS

Structure learning in graphical models is promising for object detection because it allows to learn the structure of the domain of interest and a deeper understanding of the structure or objects. Graph structure learning is motivated by feature selection and both are related in that in the problem statements the most meaningful or significant feature or feature coupling is to be selected. While feature selection has a long history in computer vision, several authors succeeded in structure learning in general graphs only recently. It turned out that many successful structure learning approaches have adapted fundamental ideas from feature selection methods such as the used selection criteria that in most cases can be adopted directly. However, in this review of related work we focus on recent advances in graphical structure learning rather than feature selection since an extensive study of related feature

selection models is beyond the scope of this dissertation.

The problem statement of structure learning in graphical models is particularly interesting because the optimal solution would estimate the implicit dependencies of nodes in the graphs and therefore the structure of the domain of interest, for example, short- and long-range dependencies of object parts. Generally, optimal structure learning is NP hard making approximations or relaxations necessary to cope with large scale problems. Given the graphical model $\mathcal{G} = (V, E)$ the task involves the selection of optimal pairwise dependencies depicted by E. Typically, this learning task is interpreted as selecting the optimal pairwise couplings out of the complete candidate set $V \times V$.

In the following we cluster the related work describing those approximations or relaxations along the axes *Bayesian networks*, in which graphical structure learning was initially described, and *random fields*. Generally speaking, the Bayesian network models estimate the structure of the graphical model decoupled from parameter learning while the random field variants are advantageous in that they enable simultaneous parameter training in a single consistent framework. Moreover, Bayesian networks account for directed acyclic graphs while random fields comprise general graphs. The random field discussion involves MRFs as well as CRFs. Since CRFs are discriminative classifiers they are often paired with discriminative structure learning while in MRFs the generative nature is often exploited for structure learning as well.

A notable exception that cannot be intuitively associated to graphical models is the work of Tran and Forsyth (2007). In this work the configuration of pedestrians is included in the histograms of oriented gradients (Dalal and Triggs, 2005) and significant performance gains are reported by doing so. The configuration of pedestrians is learned directly within the model which can be seen as a variant of structure learning that is not based on a graphical model.

Bayesian networks. Heckerman *et al.* (1995) describes a generative framework, in which the user provides prior knowledge about the structure of the directed acyclic graph. This prior structure is then modified by greedily adding and removing edges based on the Bayesian information criterion (BIC) of the provided statistical data. While this model is not guaranteed to yield the optimal solution, Chickering (2003) proved the prerequisites under which the optimal solution can be found with greedy equivalence search. In the limit of large sample size this method converges to the optimal solution if the optimal solution can be modeled with a directed acyclic graph. The runtime of both of these approaches is super-exponential in the size of variables - the nodes or vertices of the graphical model. Koivisto and Sood (2004) specifically address the runtime issue and introduce an efficient and exact algorithm to yield the optimal solution for moderately many variables. In the case of a huge amount of variables the proposed algorithm remains feasible if a suitable prior restriction is imposed in the structure. The idea of efficient algorithms is further traced in (Guo and Schuurmans, 2006), in which a convex relaxation of the structure learning problem is defined. Since the relaxed problem is convex it is guaranteed to yield the optimal (relaxed) solution. Experimentally, the authors show that the relaxed

solution outperforms greedy heuristic search.

All of the mentioned structure learning models need to restrict the number of edges directing into one and the same node in order to bound the complexity of the inferred graph. Tsamardinos *et al.* (2006) overcome this restriction by exploiting the local parent-child search within the global structure learning framework. By first inferring the parents and children of all nodes and then greedily orienting the edges with gradient descent, the authors are able to efficiently estimate and approximate the underlying structure of the domain of interest. Schmidt *et al.* (2007) further extend the latter model by imposing an L1 sparsity prior on the learned structure and inferring the Markov blanket as candidates for selecting pairwise connections. This combination of local search and global structure learning showed promising results even for many variables.

An interesting extension to learning the structure of Bayesian networks is presented in (Greiner *et al.*, 2005), in which structure learning of Bayesian networks is paired with discriminative parameter learning. This approach was proven advantageous in most cases compared to generative parameter learning. Another extended approach that inverts the idea of Greiner *et al.* (2005) is discussed in (Grossman and Domingos, 2004). Here discriminative structure learning is paired with generative parameter learning.

A more comprehensive study of the extended approaches is given in (Pernkopf and Bilmes, 2005). Here, many combinations of generative/discriminative parameter learning with generative/discriminative structure learning are evaluated. The authors conclude that in general the discriminative methods perform better than the generative ones.

Random fields. Recently, the Markov blanket scheme for candidate selection of pairwise couplings was adapted to Markov random fields (Wainwright *et al.*, 2006; Meinshausen and Buhlmann, 2006). In (Meinshausen and Buhlmann, 2006) a Gaussian model is assumed that learns the dependency network

$$p(y_i|y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_N) \ .$$

Even though the model was shown to yield consistent topologies, it cannot be used for classification since it is not a joint density estimator $p(\mathbf{y})$. Wainwright *et al.* (2006) theoretically infer a lower bound on the training samples in terms of number of nodes and maximum neighborhood size in order to guarantee convergence. Additionally an L1 regularization scheme is discussed that leads to sparse solutions even for high dimensional problems.

Opposing to the local Markov blanket search, Lee *et al.* (2006b) discuss a global random field variant, in which the joint density $p(\mathbf{y})$ is estimated. This variant enables the estimation of the graphical model structure and simultaneous training of the associated parameters in one consistent model. This work is particularly interesting as it discusses and compares different efficient feature selection heuristics and adapts those for pairwise feature selection in random fields. Specifically, the gain-based heuristic of Pietra *et al.* (1997) was shown to perform marginally better

than the gradient-based heuristic of Perkins *et al.* (2003), while the latter heuristic is more efficient to compute. Moreover, Lee *et al.* (2006b) includes an L1 sparsity prior to structure learning, which results in desirably sparse (tractable) solutions.

Höfling and Tibshirani (2009) generalize the pseudo-likelihood formalism to a fast and exact method for parameter as well as structure learning in Markov random fields. The authors show improved runtime compared to (Lee *et al.*, 2006b) while not sacrificing performance.

Johnson *et al.* (2007) propose a maximum entropy relaxation to the structure learning problem in order to yield a convex optimization problem. The model is tractable to learn thin graphs including tolerance specifications that are suitable to address the trade-off between data-fidelity and sparsity. Another relaxed problem discussion can be found in (Banerjee *et al.*, 2008). The authors define a convex problem that is guaranteed to reach the global optimum and sparse solutions since an L1 regularizer is included like in other related work.

Schmidt and Murphy (2010) discuss an interesting generalization of structure learning in Markov random fields that relaxes the restriction to pairwise couplings. The authors describe an efficient hierarchical embedding scheme that prunes all higher order potentials, which contain an inactive lower order potential. With this hierarchical embedding scheme the model is guaranteed to converge and the generalization to higher order potentials is shown to perform significantly better than the restricted pairwise models.

Using CRFs for structure learning is not as well explored as the MRF variants. Parise and Welling (2006) evaluates both a MRF on synthetic data and a linear chain CRF on real data, where both models are paired with generative structure learning. Interestingly, the MRF shows promising results for small datasets but deteriorates with growing number of training samples while the CRF's performance improves with larger datasets.

The most recent work on structure learning in CRFs and the only related model to date, which combines the discriminative nature of CRFs with discriminative structure learning is the work of Schmidt *et al.* (2008). This model was the first to allow for the general case of multiple states per random variables as an extension to binary random variables considered in other related work on structure learning. This generalization requires block regularizations since each pairwise coupling can be associated with several parameter sets (one for each state/class pair). The authors conclude with a comprehensive comparison of generative and discriminative structure learning: discriminative structure learning was shown to outperform the generative variant.

Inspiration for our work. We adapt the discriminative random field notation of Schmidt *et al.* (2008), since in most object detection datasets no valuable prior knowledge about (or even the ground truth of) the structure of the domain of interest is given. Therefore, the random field variant is favorable since we specifically want to allow for cycles if they help to discriminate objects from background. Since we are facing a complex learning problem, we aim at adapting an efficient gradient-based

pairwise coupling selection scheme (Lee *et al.*, 2006b) and modify it for discriminative learning (see chapter 4). In chapter 5 we discuss an extension to structure learning of hidden part-based models.

2.5 SENSOR FUSION

Our work on sensor fusion is mainly focused on the deployment of mobile robots. Thus, the review of related work not only touches upon computer vision but goes into more detail in the robotic fields. Since we primarily explore the use of visual, thermal and laser sensors and the combination of recent advances in computer vision with sensor fusion, we restrict the review of related work to those models.

Hall and Llinas (1997) give an introduction to sensor fusion distinguishing *decision level fusion*, *feature level fusion* and *data level fusion*, see Fig. 2.2. Data level fusion associates incoming data of different sensors directly to each other, followed by feature extraction and then detection/recognition. Hariharan *et al.* (2006), for example, show favorable properties of the complementary nature of the sensors: While for a visual sensor some information of a scene is occluded, a thermal device may be sensible for this particular cue and vice versa. Feature level fusion associates the different data sources after decoupled feature extraction. Klein *et al.* (2009), for example, propose to enrich the visual cues with additional laser-range data and cast all features forward to AdaBoost. Decision level fusion associates different sensor data after detection in each sensor process as discussed, for example, in (Zivkovic and Kröse, 2007; Kleiner and Kümmerle, 2007; Spinello *et al.*, 2008). In all these proposed frameworks a complete detection process is inferred for each sensor device and afterwards all detections are combined to one single output. Intuitively we recognize that the complexity of the data association problem increases from decision level over feature level to data level schemes. In data level fusion we need synchronized sensors and a pixel-by-pixel correspondence, while decision level frameworks are inherently robust to small offsets in synchronization and minor shifts in correspondence of pixels.

An interesting modification of decision level fusion for people detection is discussed in (Doherty and Rudol, 2007). The authors propose to scan images with a thermal camera and extract regions of interest wherever regions with human temperature are detected. These regions of interest are then analyzed (i.e. verified) with a visual model based on a boosted part-based model.

Inspiration for our work. In our work we are interested in combining recent advances in computer vision with the progress on exploring the complementarity of different sensors. In our setting we are facing a noisy sensor synchronization implying to use decision level fusion. We found that this scheme already leads to improved performance and we leave the comparison to other techniques for future work. We evaluate our model in a search and rescue application, in which a mobile robot explores an area, searching for human victims.

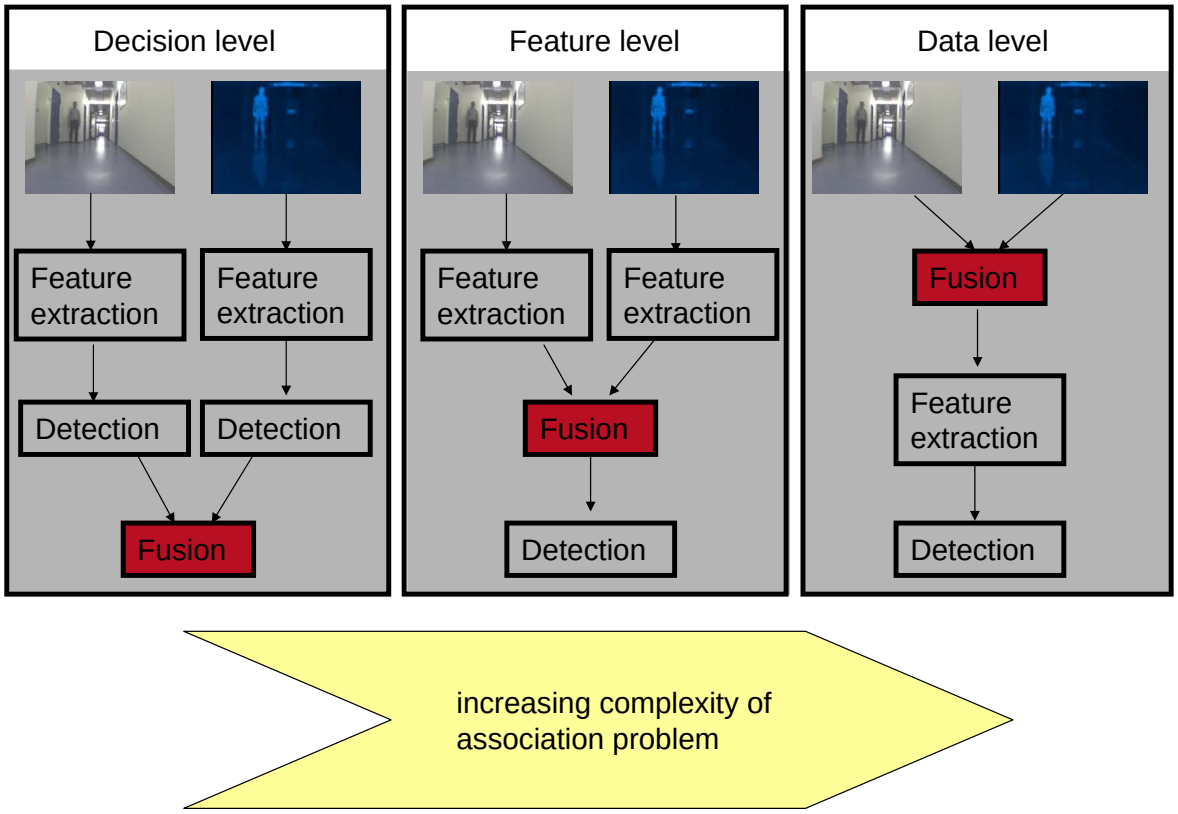


Figure 2.2: Illustration of different sensor fusion schemes. The terminology is adapted from (Hall and Llinas, 1997). Intuitively, decision level fusion is inherently more robust to offsets and errors in data association than feature or data level fusion.

Inspired by (Doherty and Rudol, 2007), we search for human body temperature, while not adapting the sequential region of interest verification scheme, but rather stick to the original decision level fusion, in which the decisions or detections of human evidence are inferred independently from each other. The data fusion then takes place by associating visual and thermal detections to each other. We further extend this approach by verifying the estimated distance to victims with the measured distance from a laser scanner. While the thermal evidence is inferred with a simple generative heat blob detector, we adapt the monolithic, discriminative model of Dalal and Triggs (2005) for visual evidence (we integrate the efficient and real-time capable implementation of Wojek *et al.* (2008)).

HIERARCHICAL SUPPORT VECTOR RANDOM FIELDS 3

Contents

In this chapter we extend standard foreground-background, one-layer conditional random fields (CRFs) (Lafferty *et al.*, 2001) to a multi-layer and multi-class model. Our goal is to incorporate richer object representations in a hierarchical model and propose a modified learning paradigm that enables joint training of all model parameters in a single consistent framework. The underlying graphical model structure that is discussed here realizes a step towards incorporating longer-range dependencies in CRFs, while standard CRFs typically consider only local neighborhood relations. The focus in this chapter is on evaluating the usefulness of hierarchical, part-based models for object detection. While in this chapter the graphical model structure is fixed, we discuss in the following chapter, how flexible structures can be learned and incorporated automatically. This chapter describes the work published in (Schnitzspan *et al.*, 2008).

3.1 INTRODUCTION

While impressive results have been reported for the detection of objects in challenging real world scenes, the underlying models vary greatly even between the most successful approaches. Methods using a global feature descriptor (e.g. (Dalal and

Triggs, 2005)) paired with discriminative classifiers such as SVMs enable high levels of performance, but require large amounts of training data and typically degrade in the presence of partial occlusions. Local feature-based approaches (e.g. (Felzenszwalb and Huttenlocher, 2005; Fergus *et al.*, 2003; Leibe *et al.*, 2005)) are more robust in the presence of partial occlusions but often produce a significant number of false positives. This chapter discusses an extension to standard CRFs called hierarchical support vector random field that allows 1) to combine the power of global feature-based approaches with the flexibility of local feature-based methods in one consistent multi-layer framework and 2) to automatically learn the trade-off and the optimal interplay between local, semi-local and global feature contributions. In order to achieve this, we leverage the ability of CRFs (Lafferty *et al.*, 2001) to model neighborhood dependencies not only between local image features, but also between object subparts and parts using a multi-layer CRF. On the top-layer we incorporate a global object detector while on the layers below we employ smaller apertures in terms of object-parts and local features or subparts. The layers are connected via intra-layer potentials to benefit from simultaneous bottom-up and top-down propagation schemes. This allows to set up a joint and hierarchical model of local and global discriminative methods that augments CRFs to a multi-layer model with powerful unary classifiers.

The contributions of this chapter are the following: First, we extend classical one-layer CRFs to multi-layer CRFs while maintaining computational tractability. Second, this work shows how to integrate local, semi-local and global information in a powerful model. Third, we extend CRFs to a consistent framework, which allows to jointly train the parameters of non-linear classifiers and the CRF parameters. Fourth, we experimentally show the contributions of the various components of the model on challenging datasets.

The chapter is structured as follows. In section 3.2 we introduce our multi-layer model, the respective potential functions and the parameters to be optimized. In section 3.3 we explain how we apply the model to object detection and verification. Finally, in section 3.4 we evaluate various aspects of our work on two different datasets.

3.2 HIERARCHICAL SUPPORT VECTOR RANDOM FIELDS (HSVRF)

While global detectors have been shown to achieve impressive results in object detection for unoccluded object instances, part-based approaches tend to be more successful in dealing with partial occlusion. Since adjacent regions in images are not independent from each other, CRFs model these dependencies directly by introducing pairwise clique potentials. However, standard CRFs work on a very local level and long-range dependencies are not addressed explicitly in simple one-layer models. In this chapter we discuss how SVM learning and multiple connected layers can be incorporated in one consistent framework in order to overcome restrictions to local neighborhoods and combine both, local as well as long-range dependencies. In the following we will describe how we set up the multi-layer model step by step

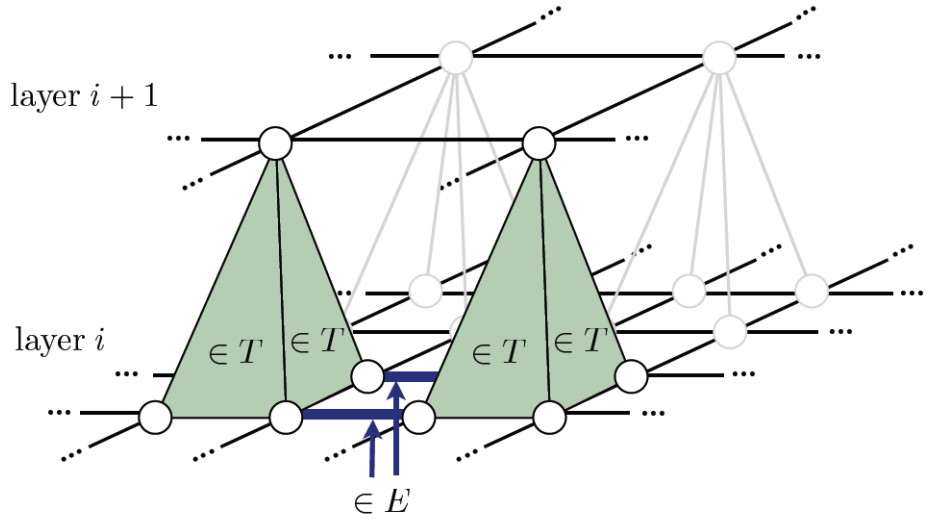


Figure 3.1: Illustration of the model architecture. Two layers are connected via the ternary cliques T. The alternation between pairwise cliques E and ternary cliques T is key to the computationally feasibility while a high degree of of interconnectedness is introduced.

starting from the one-layer case.

3.2.1 One-layer CRF model

We denote the set of all images with $X = (\mathbf{x}^1, \ldots, \mathbf{x}^M)$. Each image $\mathbf{x}^m$ is overlaid with a grid of nodes where each node is linked to the evidence in the image via unary ψ and pairwise ϕ potential functions. We denote the set of grid nodes by $i \in V$ in which each node i is associated to a certain region of the image $\mathbf{x}^m$. $(i,j) \in E$ refers to the pairwise cliques connecting two adjacent nodes i and j. Each node $i \in V$ will be assigned a label from $y_i \in \{0, \ldots, P\}$ which indicates the parts of an object $\{1, \ldots, P\}$ or background $\{0\}$. We denote the set of all labels by $\mathcal{Y}$. Therefore, the factorization of the conditional probability distribution can be written as

$$p(\mathbf{y} \,|\mathbf{x}^m; \theta) = \frac{1}{\mathcal{Z}(\mathbf{x}^m, \theta)} \prod_{i \in V} \psi_i(y_i, \mathbf{x}^m; \theta) \prod_{(i,j) \in E} \phi_{ij}(y_i, y_j, \mathbf{x}^m; \theta) \ . \tag{3.1}$$

Here, $\mathcal{Z}(\cdot, \cdot)$ refers to a normalization factor called the partition function. θ denotes the model parameters as explained below. Here and in the remainder of the chapter we drop superscript m for notational simplicity wherever applicable.

3.2.2 Multi-layer CRF model

As motivated before, one-layer CRFs act at a very local level and represent a single view on the data typically represented with unary and pairwise potentials. In order to overcome those local restrictions, we introduce multiple layers $l \in \{1, \ldots, n_L\}$ with associated unary potentials ψ^l and pairwise potentials ϕ^l, to enhance the model by evidence aggregation on a local ($l = 1$) to a global level ($l = n_L$). Different numbers of parts are deployed to different layers $\{0, \ldots, P^l\}$. We propose a connectivity between the layers as displayed in Figure 3.1, which provides a high degree of interconnectedness and yet results in a computationally tractable model, which is highly desirable for both inference and training. The key to this is the alternation between pairwise cliques $(i, j) \in E^l$ and ternary cliques $(i, j, k) \in T^l$ that omit the introduction of higher (higher than third) order cliques. The conditional distribution for this multi-layer model resolves into:

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}; \theta) \quad = \quad & \tfrac{1}{\mathcal{Z}(\mathbf{x},\theta)} \prod_{l=1}^{L} \left[\prod_{i \in V^l} \psi_i^l(y_i, \mathbf{x}; \theta) \prod_{(i,j) \in E^l} \phi_{ij}^l(y_i, y_j, \mathbf{x}; \theta) \right] && \text{(intra-layer)} \\ & \prod_{l=1}^{L-1} \prod_{(i,j,k) \in T^l} \lambda_{ijk}^l(y_i, y_j, y_k, \mathbf{x}; \theta) && \text{(inter-layer)} \end{aligned} \tag{3.2}$$

where additional to the one-layer notation $\lambda_{ijk}^l(\cdot, \cdot, \cdot, \cdot; \cdot)$ denotes the ternary clique potentials that connect layer l to layer $l + 1$ using third-order cliques. T^l describes the set of all ternary cliques between layer l and layer $l + 1$ (see Figure 3.1 for illustration).

This model combines different views on the data by layer-specific potentials and the hierarchical structure accounts for longer-range dependencies.

3.2.3 Potentials

As described in Eq. 3.1 and 3.2 the conditional probabilities factor into unary potentials ψ^l, pairwise potentials ϕ^l and additional ternary potentials λ^l required for the multi-layer model. Due to the flexibility of CRFs, the layer-specific feature functions $f^l(\mathbf{x})$, $g^l(\mathbf{x})$ and $h^l(\mathbf{x})$ for the unary, pairwise and ternary potentials respectively can be chosen freely. Those deployed in the experiments are detailed in section 3.3.1.

Unary potentials. The discriminative power in the unary potentials is key to the overall performance of the CRF. In some cases, a CRF using less powerful classifiers such as the commonly used logistic regression can even be outperformed by an SVM employing no connectivity at all (Lee *et al.*, 2005).

Therefore, we build our unary potentials on SVMs to leverage previous results on robust large margin classification. We adapt the one-against-all strategy which results in training one SVM for each class. $f^l(\cdot)$ refers to the feature function for the node features and ρ_c^l denotes the offset. Then, the potential of node y_i being of class

c is defined as

$$\psi_i^l(y_i = c, \mathbf{x}; \boldsymbol{\beta}_c^l, S_c^l, \rho_c^l) = \exp\left(\sum_{j \in S_c^l} \left(\boldsymbol{\beta}_c^l\right)_j K(f_j^l(\mathbf{x}), f_i^l(\mathbf{x})) + \rho_c^l\right) \quad , \tag{3.3}$$

where S_c^l indexes the set of support vectors for class c and layer l and $\left(\beta_c^l\right)$ refers to model parameters to be optimized. We use RBF kernels to define the kernel function $K(f_j^l(\mathbf{x}), f_i^l(\mathbf{x})) = \exp\left(-\gamma \left\| f_j^l(\mathbf{x}) - f_i^l(\mathbf{x}) \right\|^2\right)$ with bandwidth parameter γ. Note, that this approach employs multi-class one-against-all SVMs.

Pairwise and ternary potentials. We define the pairwise and ternary potentials using a linear classification model, which is a popular choice in the CRF literature. For the pairwise potentials we set

$$\phi_{ij}^l(y_i = c_1, y_j = c_2, \mathbf{x}; \mathbf{e}^l) = \exp\left(\left(\mathbf{e}_{c_1 c_2}^l\right)^T g_{ij}^l(\mathbf{x})\right) \quad , \tag{3.4}$$

where i and j are two adjacent nodes and c_1 and c_2 refer to any label from $\{0, \ldots, P^l\}$. $g^l(\cdot)$ denotes the feature function for the pairwise potentials of layer l. $e_{c_1 c_2}^l$ refers to the parameters to be trained. The ternary potentials are defined as

$$\lambda_{ijk}^l(y_i = c_1, y_j = c_2, y_k = c_3, \mathbf{x}; \mathbf{t}^l) = \exp\left(\left(\mathbf{t}_{c_1 c_2 c_3}^l\right)^T h_{ijk}^l(\mathbf{x})\right) \quad , \tag{3.5}$$

where i, j, k belong to one three-wise connected clique (i, j, k) and $c_1, c_2 \in \{0, \ldots, P^l\}$ and $c_3 \in \{0, \ldots, P^{l+1}\}$. $h^l(\cdot)$ denotes the feature function for the ternary potentials at layer l. $\mathbf{t}_{c_1 c_2 c_3}^l$ refers to the parameters to be optimized.

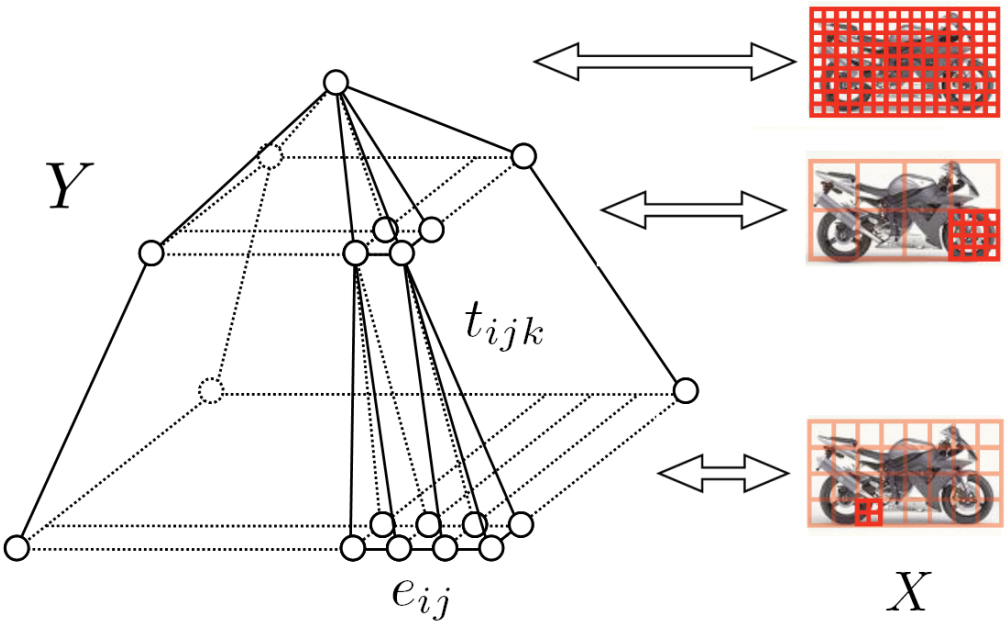
3.2.4 Parameter learning and inference

We jointly optimize all model parameters in contrast to related CRF literature (Hoiem *et al.*, 2007; Winn and Shotton, 2006; Lee *et al.*, 2005; Shotton *et al.*, 2006). . Given M training images $\mathbf{x}^m, m = \{1, \ldots, M\}$ we optimize the conditional log-likelihood

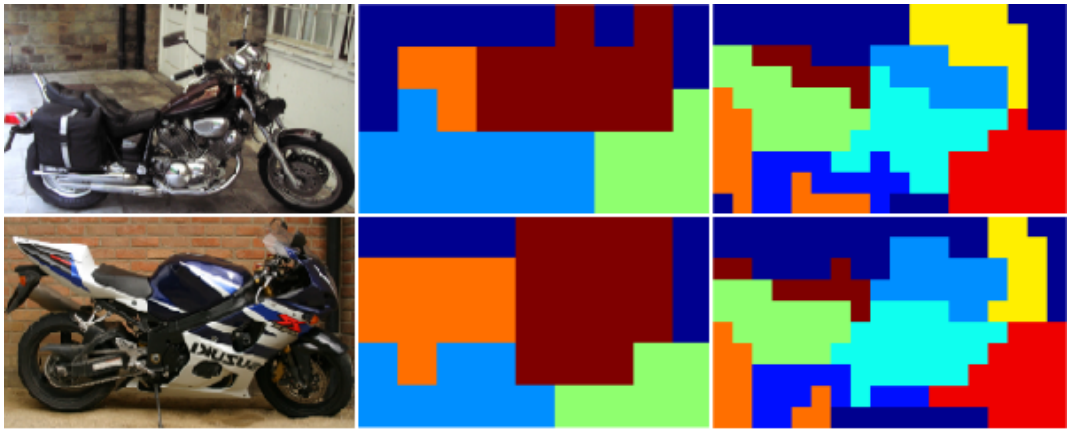
$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{e}, \mathbf{t}) = \sum_{m=1}^{M} \log p\left(\mathbf{y}^m | \mathbf{x}^m; \boldsymbol{\beta}, \mathbf{e}, \mathbf{t}\right) \tag{3.6}$$

via gradient descent for pairwise and ternary clique potentials. The unary potentials are trained with Newton optimization.

This joint training is facilitated by the primal SVM training proposed by Chapelle (2007) showing competitive results compared to common quadratic programming in the dual formalism. We make use of that idea and incorporate primal SVM training in the CRF framework.



(a)



(b)

Figure 3.2: (a) Three-layer instantiation of our model. The evidence aggregation is sketched: Starting from local information like fragments of a wheel over whole wheels to entire objects at the top layer. (b) Example of the part assignment of training data (left: training image; middle: part assignment of middle layer; right: part assignment of bottom layer). Colors encode assignments of parts; dark blue indicates background

Primal SVM training. As described in (Chapelle, 2007) the constraints of the original primal optimization problem can be integrated with a loss function in the objective function, yielding an unconstrained optimization problem. As long as this loss function is differentiable with respect to the model parameters, the optimization can be solved by Newton optimization. Originally, the non-differentiable hinge

loss is used for SVM training in the dual, but Chapelle (2007) showed competitive results using the differentiable quadratic loss or the Huber loss (a differentiable approximation of the hinge loss). The primal optimization problem for kernel SVMs is denoted by:

$$\min_{\boldsymbol{\beta}_c^l} \mathcal{Q} = \min_{\boldsymbol{\beta}_c^l} \left(\sum_{i,j \in S_c^l} \left(\boldsymbol{\beta}_c^l\right)_i \left(\boldsymbol{\beta}_c^l\right)_j K(f_i(\mathbf{x}), f_j(\mathbf{x})) + C \sum_{i=1}^{N} L\left(y_i, F_c^l\left(f_i^l(\mathbf{x})\right)\right) \right) \quad , \tag{3.7}$$

where L denotes a suitable loss function and C the penalty term. The target function $F_c^l(\cdot)$ is of the form (representer theorem (Kimeldorf and Wahba, 1970)):

$$F_c^l\left(f_i^l(\mathbf{x})\right) = \sum_{j \in S_c^l} \left(\boldsymbol{\beta}_c^l\right)_j K(f_j^l(\mathbf{x}), f_i^l(\mathbf{x})) + \rho_c^l \quad , \tag{3.8}$$

where $f_i^l(\mathbf{x})$ denotes a feature vector to be classified. $K(\cdot,\cdot)$ denotes the kernel function and $\boldsymbol{\beta}_c^l$ refers to the parameters to be optimized (note that these are not Lagrange multipliers). We consider the differentiable quadratic loss:

$$L(y_i, F_c^l(f_i(\mathbf{x}))) = \left(\max\left\{0; 1 - (\delta(y_i, c))\left(F_c^l\left(f_i^l(\mathbf{x})\right)\right)\right\}\right)^2 \quad , \tag{3.9}$$

where $\delta(y_i, c) \in \{-1, 1\}$ refers to whether y_i belongs to class c $(= 1)$ or not $(= -1)$.

Chapelle (2007) proposed to optimize the parameters $\boldsymbol{\beta}_c^l$ with Newton optimization:

$$\boldsymbol{\beta}_c^l \leftarrow \boldsymbol{\beta}_c^l - \eta \left(H^l\right)^{-1} \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\beta}_c^l} \quad , \tag{3.10}$$

where η denotes the learning rate and the Hessian H^l equals $2\left(\frac{1}{C}K + KI^0K\right)$ with kernel matrix K. I^0 is a diagonal matrix, where the entries are 1 for $\left|\left(\boldsymbol{\beta}_c^l\right)_i\right| > 0$ and 0 otherwise. The number of non-zero entries equals the number of support vectors. In order to update the offset ρ_c^l the Hessian can be augmented by an additional row and column and the offset term can be concatenated with the parameters $\boldsymbol{\beta}_c^l$ (see (Chapelle, 2007) for details).

Joint training of hSVRF. In order to account for joint training of the hSVRF parameters, we adapt the loss function $L(\cdot,\cdot)$ to consider unary SVM classifications as well as joint CRF classifications , which respects the entire multi-layer model. In that sense, object evidence, local neighborhood dependencies as well as longer-range dependencies are taken into account to optimize the unary parameters. We achieve this by adapting the loss function to consider the belief of node y_i belonging to class c inferred with loopy belief propagation (Pearl, 1988).

$$L(y_i, b_c(y_i), F_c^l(f_i(\mathbf{x}))) = \left[\max\left\{0; (1 - \delta(y_i, c)b_i(c))\left(1 - \delta(y_i, c)F_c^l\left(f_i^l(\mathbf{x})\right)\right)\right\}\right]^2 \quad , \tag{3.11}$$

where the belief $b_i(c)$ of node i belonging to class c is scaled to the range $[-1, 1]$. Whenever the CRF votes for the wrong class ($(1 - \delta(y_i, c)b_i(c)) > 1$) the original primal SVM loss function is amplified for calculating the Newton step. Otherwise ($(1 - \delta(y_i, c)b_i(c)) < 1$) the impact of the original primal loss function on the Newton step is reduced. Note, the Hessian is not affected by our changes in the loss function and $\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\beta}_c^l}$ can be computed similar to (Chapelle, 2007).

The parameters of the pairwise and ternary clique potentials can be optimized via gradient descent. Similar to (Kumar *et al.*, 2006), the gradient with respect to the pairwise parameters $\{\mathbf{e}_{c_1 c_2}^l\}$ of layer l can be expressed as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{e}_{c_1 c_2}^l} = \sum_{(i,j) \in E} \left(\delta(y_i, c_1)\delta(y_j, c_2) - b_{ij}(c_1, c_2) \right) g_{ij}^l(\mathbf{x}) \ , \tag{3.12}$$

where $\delta(\cdot, \cdot)$ refers to the Kronecker-delta and $b_{ij}(c_1, c_2)$ denotes the pairwise belief of two adjacent nodes belonging to class c_1 and c_2.

Analogously, the gradient with respect to the ternary clique parameters $\mathbf{t}_{c_1 c_2 c_3}^l$ of layer l can be written as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{t}_{c_1 c_2 c_3}^l} = \sum_{(i,j,k) \in T} \left(\delta(y_i, c_1)\delta(y_j, c_2)\delta(y_k, c_3) - b_{ijk}(c_1, c_2, c_3) \right) h_{ijk}^l(\mathbf{x}) \ , \tag{3.13}$$

where $b_{ijk}(c_1, c_2, c_3)$ denotes the ternary beliefs of three connected nodes.

This concept for updating the parameters of our model alternates between the maximum margin formalism of SVM training and the likelihood maximization of CRFs. Although imposing potential restrictions to the learning framework we stick to this alternating scheme and show performance gains with our joint training model. Future directions might involve an adaption of the max margin formalism of random fields of Taskar *et al.* (2003) in order to incorporate one unique optimization scheme.

We use quadratic programming (the common SVM training) decoupled from the CRF to initialize the parameters $\boldsymbol{\beta}^l$. The dual support vector coefficients $\boldsymbol{\alpha}_c^l$ and parameters $\boldsymbol{\beta}_c^l$ are connected via $\left(\boldsymbol{\beta}_c^l\right)_j = \delta(y_j, c) \left(\boldsymbol{\alpha}_c^l\right)_j$ as described in (Chapelle, 2007). Given the starting solution for $\boldsymbol{\beta}^l$ we start the joint optimization by Newton optimization for unary classifiers and gradient descent for pairwise and ternary clique parameters.

Inference. Given the parameters $\boldsymbol{\beta}^l$, $\mathbf{e}^l$ and $\mathbf{t}^l$ we seek to infer probabilities of the nodes belonging to the different classes. Loopy belief propagation (LBP) (Pearl, 1988) infers beliefs that one node y belongs to class c while respecting the pairwise and three-wise dependencies of adjacent nodes.

3.3 APPLICATION TO COMPUTER VISION TASKS

To support our claims about the benefits of the local to global CRF model and the presented joint optimization, we evaluate the approach on two challenging computer

vision tasks: object detection and hypothesis verification. But first we describe in detail how the method is adapted to the specific settings and show how to obtain part annotations for the training phase. We consider a 3-layer instantiation of the presented model as visualized in Figure 3.2(a) and detailed below.

3.3.1 Feature functions

Until now, we have not defined the feature functions $f^l(\cdot)$, $g^l(\cdot)$, $h^l(\cdot)$ that are specific to each layer in the CRF we propose. They link the potentials to the actual image evidence and account for local neighborhood and long-range dependencies. We build on the concept of computing histograms of oriented gradients that has been shown to be very successful on a local level, describing interest points (Lowe, 2004), as well as on a global level (Dalal and Triggs, 2005), describing full objects in a holistic manner. However, due to the generality of our work, any suitable feature function can be deployed to our model.

Unary potential feature functions. We calculate histograms of oriented gradients for a grid of non-overlapping 8×8 pixel regions and concatenate 4 neighboring histograms of gradients to one block descriptor as described in (Dalal and Triggs, 2005). This results in a 36-dimensional feature for each node that we define to be the unary feature function on the first level $f^1(\cdot)$. For the higher levels $f^2(\cdot), \ldots, f^L(\cdot)$ we successively double the number of considered blocks in horizontal and vertical directions until on the highest level, we encode the full object as in (Dalal and Triggs, 2005). As illustrated in Figure 3.2(a), the motivation behind this scheme is to aggregate evidence for an object class from different spatial localities ranging from fragments (e.g. fragment of a wheel), parts (e.g. whole wheel) to a holistic view on the object (e.g. whole motorbike).

Pairwise potential feature functions. Intuitively, pairwise potentials are responsible for modeling local dependencies by supporting or inhibiting label propagation to the neighboring nodes. In computer vision, simple pixel-based gradient-based measures are often used to inhibit propagation across potential object borders (Shotton *et al.*, 2006). Our approach goes beyond that by taking into account the change in the gradient orientation histograms between the neighboring nodes.

$$g_{ij}^l(X) = \Big(\left| f_i^l(\mathbf{x}) - f_j^l(\mathbf{x}) \right| \quad , \quad 1 \Big)^T \quad . \tag{3.14}$$

Here, we extended each difference by an offset for being capable of eliminating small isolated regions.

Ternary potential feature functions. Similar to the pairwise potentials, ternary potentials encode local dependencies, too. But furthermore, they act as a link between layers, facilitating propagation of information across locality and position in our model. Due to the computational tractability of the hierarchy we can propagate

object evidence across layers and thereby manage efficient bottom-up and top-down reasoning during inference.

To allow the ternary potential to assess the compatibility of a particular labeling of a three-wise connected clique, we define the ternary potential feature function to be the stacked pairwise difference of the feature vector associated to the 3 relevant nodes. Since higher level nodes involve more HOG blocks and are higher dimensional than lower level ones, we calculate the average over connected blocks (denoted by operation $avg(\cdot)$) in order to fit the dimension of lower level nodes.

$$h^l_{ijk}(x) = \begin{pmatrix} | & f^l_i(\mathbf{x}) & - & f^l_j(\mathbf{x}) & | \\ | & f^l_i(\mathbf{x}) & - & avg(f^{l+1}_k(\mathbf{x})) & | \\ | & f^l_j(\mathbf{x}) & - & avg(f^{l+1}_k(\mathbf{x})) & | \\ & & 1 & & \end{pmatrix} , \tag{3.15}$$

where nodes i and j are on layer l and node k is on layer $l+1$.

3.3.2 Part assignment

For optimizing the conditional log-likelihood during training, ground truth part labels are required for each training instance in order to be able to train the multi-class potentials of our model. While the labeling for the top (object) layer is given by a bounding box annotation or segmentation of the objects, the part annotation on the lower layers is not obvious. Inspired by (Bouchard and Triggs, 2005) we obtain part labels in a data-driven way by applying k-means clustering across images to infer part annotations. Instead of mere spatial clustering, we append to the image coordinates the features described in 3.3.1. In this fashion the importance is on the cluster appearance and the 2 coordinate dimensions act as regularization for the clustering to maintain a rough spatial layout. Despite the simple data-driven approach, we obtain a sensible partitioning of our training instances that exposes appearance-based though well localized assignment of parts as exemplified in Fig. 3.2(b).

3.3.3 Object detection and verification

In this paragraph we show how we infer object locations of one object class. As described in section 3.2.4, LBP yields a label assignment across layers taking into account beliefs that nodes are associated with parts (bottom and middle layer), object (top layer) or background (all layers). Given a test image we could initialize our model at every pixel location for being able to infer all possible object hypotheses. However, to reduce computational effort we first deploy the bottom and middle layer of our model. This step produces a part map of the whole image while respecting the dependencies of the bottom and middle layer. From the training set we know possible part constellations and we search for those constellations in the part map of test images to infer hypotheses of object locations. This approach resembles the

Method	EER
Multi-layer	97.5
One-layer part-based joint training	96.0
One-layer part-based without joint training	93.0
Global object detector (Dalal and Triggs, 2005)	87.0
(Mutch and Lowe, 2006)	99.9
(Leibe *et al.*, 2005)	97.5
(Hoiem *et al.*, 2007)	approx. 93.5
(Winn and Shotton, 2006)	approx. 92.9

Table 3.1: Results of the detection task on the UIUC car dataset

ISM voting of Leibe *et al.* (2005) despite that the evidence of parts of our model are inferred simultaneously and therefore the parts interact with each other. Thus, the evidence of one subpart conditions the constellation of direct neighbors via links in the bottom layer and via the middle layer it also affects the evidence of further image regions. This step generates initial hypotheses that still need to be validated by the complete model. Since the part-based approach of the lower layers showed to yield a high recall, this approach makes sense as we first search for possible locations and then infer the complete model for the hypothesized bounding boxes. The approach of coupling generative models with a discriminative verification stage has been shown to be fruitful (Fritz *et al.*, 2005). In this spirit, we address a hypothesis verification task by only inferring our model at hypothesized bounding boxes. LBP simultaneously infers beliefs of all nodes of our model. Since at the top layer we only deploy one node, we can directly use the belief of that node belonging to the object class as a score. For the layers underneath we compute object probabilities similar to the ISM part voting scheme (Leibe *et al.*, 2005) as described above and multiply them to the global belief. Thereby we distinguish between left and right facing objects and consider the maximum of the deduced scores.

3.4 EXPERIMENTS

In all experiments we used SVM^{light} (Joachims, 1999) for initial SVM training. Training the model took approximately 12 hours while we were able to infer 15 hypotheses per second.

Object detection. For the detection task we evaluated our model on the UIUC single scale car dataset. We trained the whole model on 250 bounding boxes containing cars and 200 negative crops. This experiment contains performance measurements of i) only the global object detector, ii) the one-layer model of our approach while training the SVM and CRF parameters separately, iii) the one-layer model while training the parameters jointly and iv) the complete multi-layer model. For the part labeling during training we deployed k-means clustering with 8 means

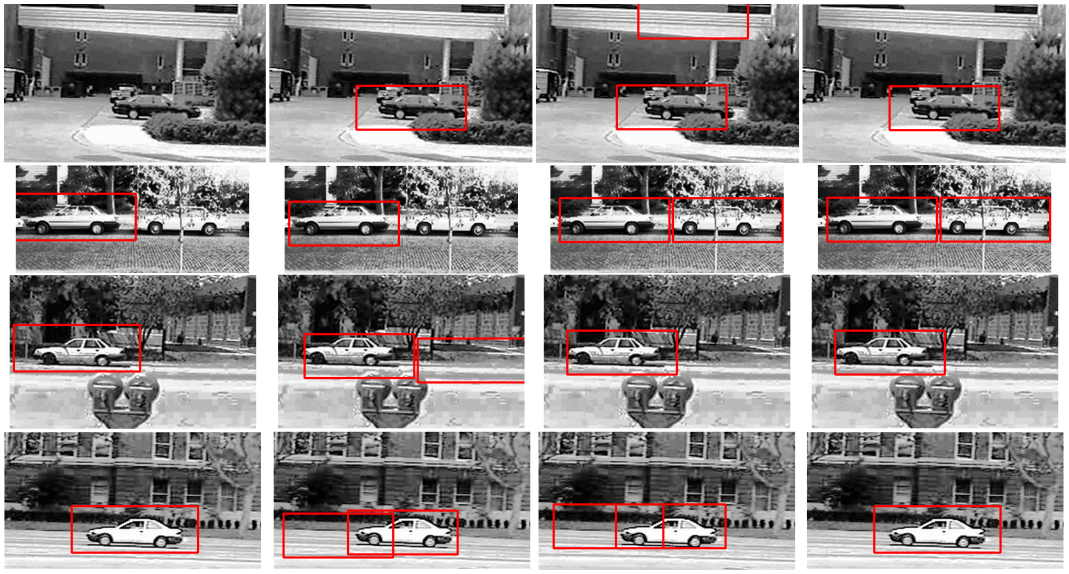


Figure 3.3: Examples on the UIUC dataset. The columns show results at EER of HOG detector, one-layer separate training, one-layer joint training and multi-layer model.

for the bottom layer and 4 means for the middle layer. For the one-layer model we used 8 means in the clustering step. The detection performance was evaluated on the 170 UIUC test images. Figure 3.4 compares the different aspects described in the previous sections. Both the joint training and the multi-layer approach consistently improved the performance. Especially note the large performance gap between the complete model (97.5% in equal error rate) and the HOG detector (87.0% in EER). Figure 3.4 shows some example images where the HOG detector can not detect all cars due to partial occlusion and the one-layer models infer false positives, while the multi-layer model detects all cars correctly.

These results expose the benefits of joint training and integration of local to global information. Our model successfully learns the trade-off between global vs. local object detection and improves the performance of both ideas by combining powerful global descriptors and flexible local feature approaches.

Further note the performance improvement between training the SVM independently from the other CRF parameters (93.0%) and training them jointly (96.0%) for the one-layer model. This evaluation highlights the advantage of training all model parameters jointly as proposed in section 3.2. In Tab. 3.1 we compare our model to the state-of-the-art in object detection on this dataset. As it can be seen, we achieve competitive results compared to other well performing models. Only Mutch and Lowe (2006) outperform our model while we obtained the same performance as Leibe *et al.* (2005). Further, we outperform the CRF-based approaches of Hoiem *et al.* (2007) and Winn and Shotton (2006).

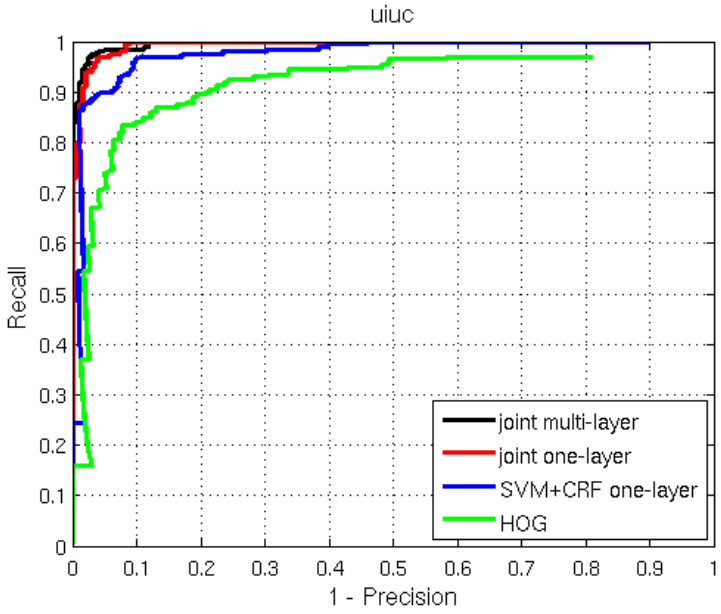


Figure 3.4: UIUC detection performance of the different aspects.

Verification of HOG detector hypothesis. For the hypothesis verification task we evaluated our model on the PASCAL 2006 motorbikes dataset (Everingham *et al.*, 2006) containing challenging multiscale, partially occluded and multiview instances. Since we want to explore the combination of an initial detector with our model acting as a verification stage, we trained a HOG detector on the provided training set and generated initial hypotheses on the test set. We set the parameters to allow for high recall at the drawback of more false positives. We also trained our joint multi-layer model on the training set and calculated the score of our approach on the hypotheses of the HOG detector (see Figure 3.4). Thereby, our multi-layer model achieved 43.7% in average precision (the common performance measure in (Everingham *et al.*, 2006)) improving the state-of-the-art by 4.7%. Note in particular that we outperformed the global HOG detector that reported an average precision of 39%, which emphasizes the benefit of combining global and local features. The next best performance for the motorbikes is 37.1% achieved by the approach of Chum and Zisserman (2007) which we outperform by 6.6%. Furthermore, our model shows a high performance improvement (more than 10% in average precision) compared to the remaining approaches. Particularly, the high precision for high scores of bounding boxes is promising; with no false positives 16% of all motorbikes are extracted while none of the other state-of-the-art approaches obtained such high recall at perfect precision.

Fig. 3.6(a) shows precision-recall-curves from which the contributions of different aspects of our model to the overall performance gain can be deduced. Consistent with the results obtained on the UIUC database, the jointly trained multi-layer model improves the performance to 43.7% while the non-jointly trained model with fixed SVM coefficients obtained 42% in average precision. After the publication of the work, Felzenszwalb *et al.* (2008) reported a performance of 58.2% on the motorbike

Figure 3.5: Example images for detecting sideviews of motorbikes: (Left) one-layer part-based model; (middle) HOG sideviews; (right) joint multi-layer model.

class of the 2006 PASCAL challenge.

Verification of generative ISM object detector hypothesis. For further testing the different aspects of our model, we decided to test our discriminative model on hypotheses obtained by the ISM model. Since the latter model was shown to yield promising results for the subset of left or right facing instances, we trained our model on sideviews of motorbikes, but evaluated the aspects on the complete multiview dataset. Overall the ISM model extracted 4238 hypotheses and achieved an average precision of 15.3%. We trained our model on 100 rightfacing and respective mirrored left views and 200 randomly cropped background images. As it can be seen in Fig. 3.4 our multi-layer model could improve the performance compared to other settings of our approach.

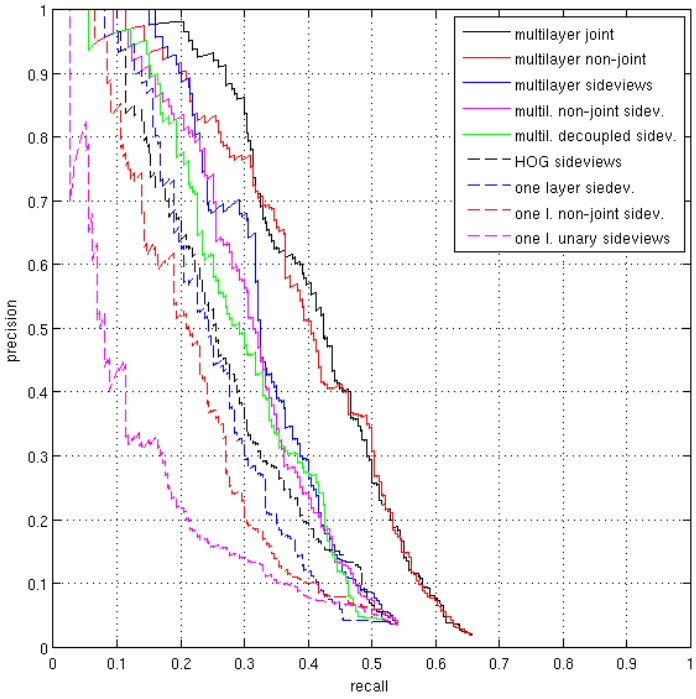
Concerning the average precision performance measure the jointly trained multi-

layer model (36.0%) significantly improved the results of the non-jointly trained model with fixed SVM parameters (33.5%), completely decoupled layers (32.3%), the HOG detector trained on sideviews (30%) and the one-layer settings of our model: jointly trained SVM and CRF parameters (27.7%), fixed SVM parameters (24.2%) and the unary classifier (SVM) (19.0%).

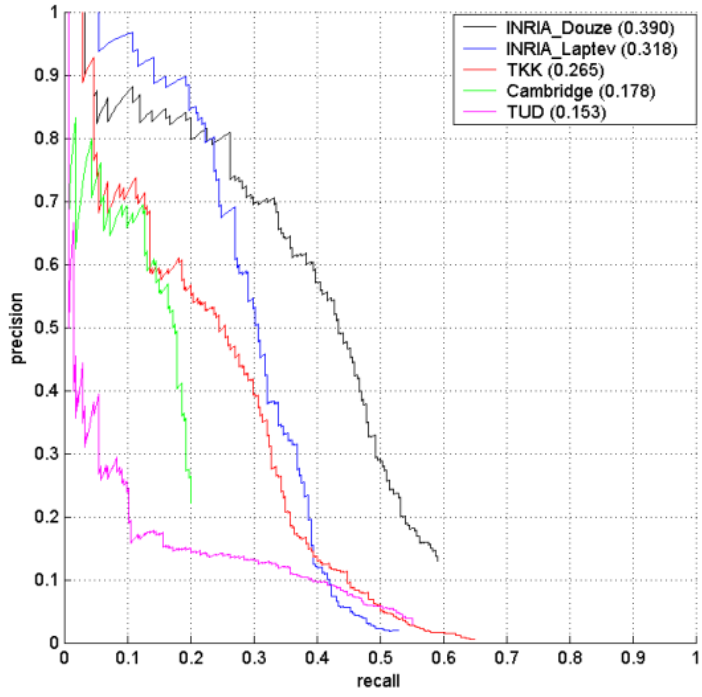
Fig. 3.4 shows some example detections for training on sideviews. Partially occluded objects can not be detected by the global detector, while the part-based approach and our multi-layer model infer them correctly. Furthermore, false detections of the part-based approach can be removed by the global detector for correct detections of our multi-layer model. However, the rear view of the motorbike (third row) can not be detected correctly due to the focus on sideviews. The measured improvements for joint training and the multi-layer approach are consistent with respect to both tested databases.

Method	AP	AP trained on sidev.
Multi-layer	43.7	36.0
Multi-layer not jointly trained	42.0	33.5
Decoupled multiple layers	-	32.2
One-layer model	-	27.7
One-layer not jointly trained	-	24.2
One-layer unary classifier	-	19.0
(Dalal and Triggs, 2005)	39.0	30.0
(Chum and Zisserman, 2007)	37.1	-
(Laptev, 2006)	31.8	-
(Viitaniemi and Laaksonen, 2006)	26.5	-
(Shotton *et al.*, 2006)	17.8	-
(Leibe *et al.*, 2005)	15.3	-
(Felzenszwalb *et al.*, 2008)	58.2	-

Table 3.2: Results for the motorbike PASCAL06 challenge (AP = average precision).



(a)



(b)

Figure 3.6: (a) PASCAL06 detection performance of our model. (b) State-of-the-art approaches on the PASCAL06 motorbikes

3.5 CONCLUSION

This chapter presents a novel multi-layer CRF which combines the power of global object detectors and flexible local feature approaches. Our model successfully learns the trade-off between local and global feature contributions for improved performance. Furthermore, we show how SVM classifiers can be incorporated into this multi-layer CRF framework and how training can be performed jointly. Experiments show that performance improves consistently. Finally, we show state-of-the-art performance on the challenging PASCAL06 motorbike detection task. Our model is kept general to allow for integration of more layers and deployment of different tractable hierarchies. Moreover, different (or additional complementary) features can be considered as well, from which we expect further performance gains.

4 DISCRIMINATIVE STRUCTURE LEARNING

Contents

In the previous chapter we extended standard conditional random fields to a multi-layer and multi-class model in order to learn richer representations of objects. Even though we incorporated longer-range dependencies within the proposed hierarchical structure, the approach remains restricted in the usage of neighborhood dependencies. In this chapter we overcome the restriction of building on a fixed structure by introducing structure learning of short- as well as long-range feature couplings. This structure learning framework is able to learn the structure of the domain of interest and experimentally shows an increased flexibility in detecting object instances compared to other approaches. The first instantiation of structure learning in graphical models, as discussed here, goes back to a foreground-background framework, neglecting the notion of object parts, whereas the combination of both, structure learning and object parts representations, is discussed in the following chapter. This chapter describes the work published in (Schnitzspan *et al.*, 2009).

4.1 INTRODUCTION

A variety of flexible models have been proposed to detect objects in challenging real world scenes. Motivated by some of the most successful techniques, we propose a hierarchical multi-feature representation and automatically learn flexible hierarchical object models for a wide variety of object classes. To that end, we do not only rely on automatic selection of relevant individual features, but go beyond previous work by automatically selecting and modeling complex, long-range feature couplings within

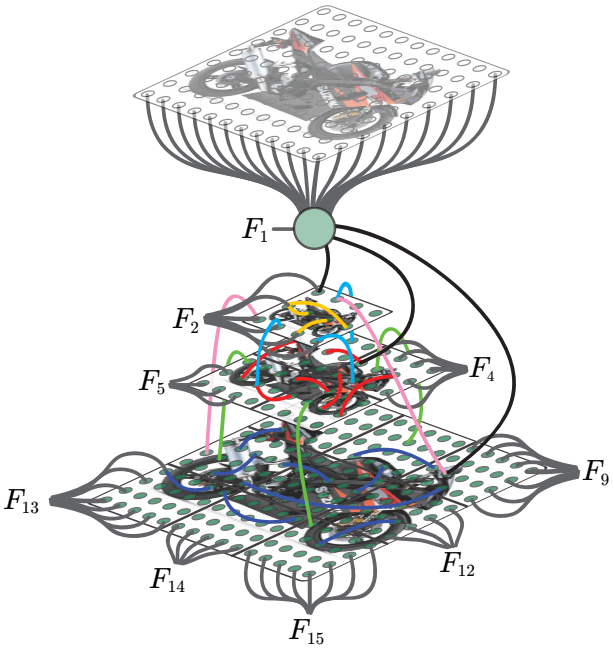


Figure 4.1: Schematic overview of our hierarchical model (Best viewed in color). The nodes of our graphical model are indicated as green dots; learned feature couplings are represented as colored lines. F refers to the discriminative unary classifiers.

this model. To achieve this generality and flexibility our work combines structure learning in conditional random fields and discriminative parameter learning of classifiers using hierarchical features. We adopt an efficient gradient-based heuristic for model selection and carry it forward to discriminative, multidimensional selection of features and their couplings for improved detection performance.

Hierarchical and multi-feature representations have shown to be a powerful basis for achieving impressive results in object detection and recognition across a variety of different datasets (Grauman and Darrell, 2007; Lazebnik *et al.*, 2006; Varma and Ray, 2007; Zhang *et al.*, 2007). The use of multiple features requires appropriate determination of the relative importance (i.e. weighting) of the features. Beyond doing this manually, a number of recent approaches have attempted to learn these weights automatically using variants of multiple kernel learning (Varma and Ray, 2007; Vedaldi *et al.*, 2009).

These learning mechanisms, however, only allow to identify and weigh the most discriminant features, but do not allow to identify and model the interplay between features that may prove to be important to representing objects well. In fact, one may posit that for many object classes the coupling between different features might be the key to discriminate object classes from others. A number of proposed conditional random field approaches allow modeling local as well as simple hierarchical couplings of features (Hoiem *et al.*, 2007; Kapoor and Winn, 2006; Lee *et al.*, 2006a; Schnitzspan *et al.*, 2008; Winn and Shotton, 2006). In particular, these approaches associate a label with each localized feature and model label

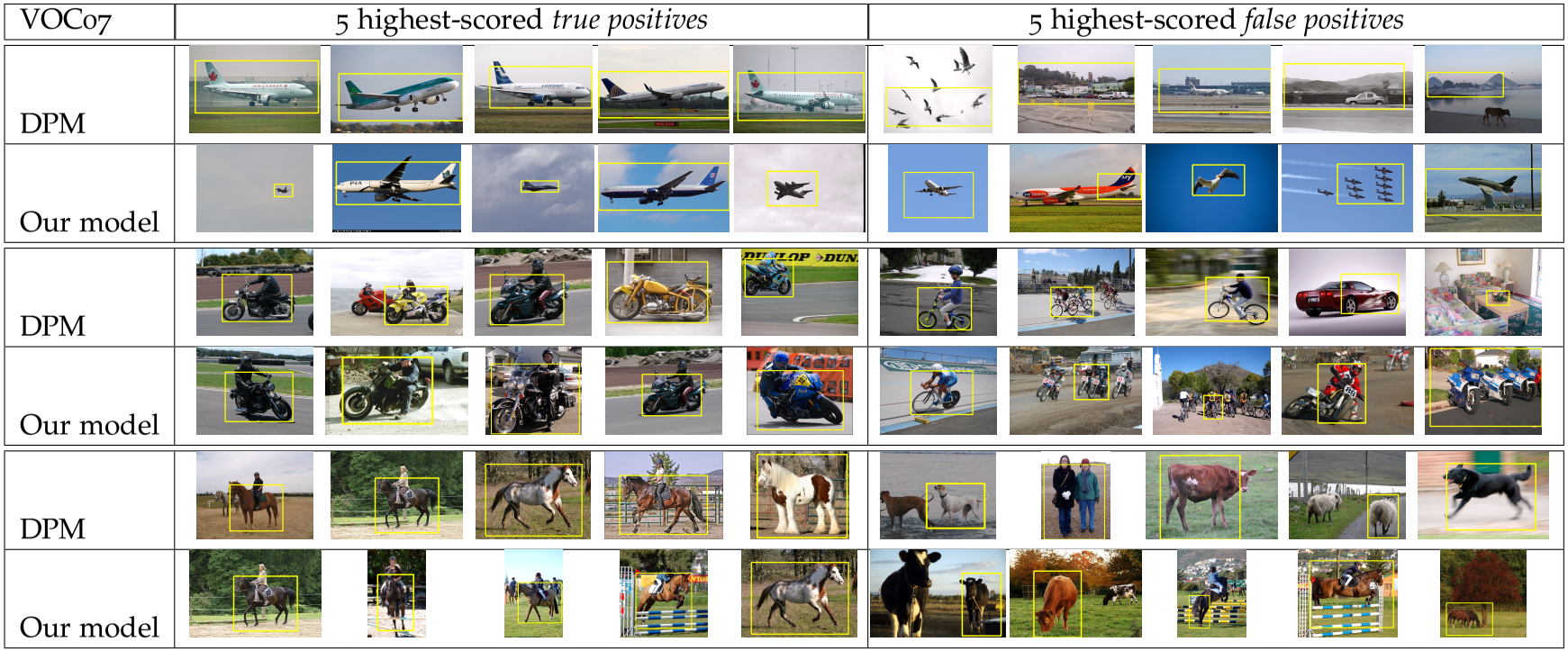


Figure 4.2: Highest-scored true and false positives of DPM (Felzenszwalb *et al.*, 2008) and our model for the PASCAL VOC 2007 challenge (aeroplane, motorbike, horse). Our framework is more flexible in modeling viewpoint, appearance and articulation changes.

dependencies by leveraging the interplay of the corresponding features. These approaches are limited, however, in that they model only simple, short-range and mid-range dependency structures in the label space, meaning they assume a fixed graphical model structure; the corresponding features are typically neighboring in space or scale.

Based on the hierarchical feature representation introduced in the preceding chapter, we address these limitations by discussing an approach that allows to learn short-range, mid-range as well as long-range dependencies, where the structure of these dependencies is identified and learned in a fully automatic manner. Unlike previous work, our approach does not require any notion of locality of the coupled features, but instead allows to find and model relevant (i.e. discriminant) couplings among arbitrary pairs of features. To enable learning of the interplay of features we cast the problem as one of structure learning in graphical models (Lee *et al.*, 2006b; Parise and Welling, 2006; Schmidt *et al.*, 2008). Specifically, we use a conditional random field to predict local labels from the image features and employ discriminative structure learning to identify dependencies whose modeling improves the discriminative power of the model. The resulting graphical model reflects the interplay between different features and therefore provides additional insights in the domain of interest in terms of feature selection and feature coupling.

Based on the hierarchical feature representation discussed in the previous chapter, we also choose a hierarchy of HOG descriptors (histograms of oriented gradients (Dalal and Triggs, 2005; Felzenszwalb *et al.*, 2008)), but additionally integrate a hierarchical bag of visual words (BoW) representation (Fergus *et al.*, 2003; Grauman and Darrell, 2007; Lazebnik *et al.*, 2006) for capturing the appearance of the object and its parts (see Fig. 4.1 for a schematic overview). Our model extends previous work by learning the contribution of the different feature types and simultaneously including relevant long-range as well as short-range couplings between arbitrary pairs of image features. As such our framework is able to model the dependency between the prediction from a HOG feature at a certain level in the hierarchy and a BoW feature at another level (cf. Fig. 4.1), if that improves the discriminative power of the model.

We apply our approach to the problem of object detection and show that it consistently outperforms SVM classifiers, which may be seen as the de facto standard in discriminant object model learning. On the PASCAL VOC 2007 detection challenge, the proposed approach outperforms one of the leading SVM-based techniques (Felzenszwalb *et al.*, 2008) on all 20 object categories. Moreover, we report the most accurate results in the literature on 16 of the 20 classes.

As the experimental results below show, our model profits from the use of powerful hierarchical and multi-feature representations. However, it is important to note that the proposed approach is very general and can be used for any local or global feature representation, not just the features used here.

Fig. 4.2 shows the first true positives (TP) and false positives (FP) of our model as well as of DPM (Felzenszwalb *et al.*, 2008), one of the leading methods on the dataset. DPM typically assigns high scores to canonical sideviews, while our work

seems to show more flexibility in modeling variations in viewpoint, appearance and articulation. Instead of being wholly misclassified, many FPs are due to misaligned bounding boxes.

4.2 CRF MODEL

In our approach we rely on conditional random fields (CRFs), which has several motivations, among them that structure learning in graphical models is a well-established field. Our approach represents each object class as a CRF with a pairwise graph structure, which models the posterior probability $p(\mathbf{y}|\mathbf{x})$ of labels $\mathbf{y}$ given an image $\mathbf{x}$. Each node $i \in V$ of the underlying graph represents a binary label $y_i \in \{1, -1\}$ encoding the presence or absence of an object of a specific class. The set of all possible edges $\Omega = V \times V$ connecting the nodes is partitioned into the *active set* $\mathcal{A} \subset \Omega$ and the *inactive set* $\mathcal{I} \subset \Omega$ (with $\mathcal{A} \cup \mathcal{I} = \Omega$ and $\mathcal{A} \cap \mathcal{I} = \emptyset$). The active set $\mathcal{A}$ defines the edge structure of our CRF model. Later we will see how to learn $\mathcal{A}$ automatically from training data; for now we assume that $\mathcal{A}$ is already given. The posterior distribution is then defined as

$$p(\mathbf{y}|\mathbf{x};\theta,\mathcal{A}) = \frac{1}{\mathcal{Z}(\mathbf{x},\theta,\mathcal{A})} \prod_{i \in V} \psi_i(y_i,\mathbf{x};\theta) \cdot \prod_{(i,j)\in\mathcal{A}} \phi_{ij}(y_i,y_j,\mathbf{x};\theta) \ , \tag{4.1}$$

where ψ_i are the unary potentials, ϕ_{ij} are the pairwise or edge potentials, θ are the parameters of the model, and $\mathcal{Z}(\mathbf{x},\theta,\mathcal{A})$ is the partition function (a normalization factor). The set of parameters $\theta = \{\boldsymbol{\alpha}, \mathbf{w}, \mathbf{e}\}$ includes parameters of the unary potentials $\boldsymbol{\alpha}$ and $\mathbf{w}$, as well as the parameters $\mathbf{e}$ of the edge potentials.

Unary potentials. The unary potentials in the CRF allow for local and global evidence aggregation; each potential ψ_i models the evidence from considering a specific image feature $f_i(\mathbf{x})$. Our representation relies on several levels of features in a hierarchy, where the feature functions at the lowest level extract local representations and the feature functions at higher levels aggregate a larger area until a global view of the object is obtained at the top level (cf. Fig. 4.1 for the hierarchical view on objects). The features will be explained in more detail in Section 4.2.1.

We define the unary potential for a node i using the softmax function (cf. (Kumar *et al.*, 2006))

$$\psi_i(y_i,\mathbf{x};\theta) = \frac{\exp\left(y_i \cdot \mathbf{w}_i^T \mathbf{F}\left(\boldsymbol{\alpha}_i, f_i(\mathbf{x})\right)\right)}{\sum_{c\in\{-1,1\}} \exp\left(c \cdot \mathbf{w}_i^T \mathbf{F}\left(\boldsymbol{\alpha}_i, f_i(\mathbf{x})\right)\right)} \ , \tag{4.2}$$

based on a weighted combination of the output of a bank of N different classifiers $\mathbf{F}\left(\boldsymbol{\alpha}_i, f_i(\mathbf{x})\right) = \left(F\left(\boldsymbol{\alpha}_{i,1}, f_i(\mathbf{x})\right), \ldots, F\left(\boldsymbol{\alpha}_{i,N}, f_i(\mathbf{x})\right)\right)^T$. Each classifier is assumed to yield a continuous-valued score. $\boldsymbol{\alpha}_i$ are the parameters of the classifier, and $\mathbf{w}_i$ are the weights. In Fig. 4.1 the classifiers are denoted with F. Interestingly, such a formulation can be seen as a probabilistic analog to multiple kernel learning (Varma and Ray, 2007) as it allows for a weighted combination of different classifiers.

Edge potentials. The edge potentials ϕ_{ij} model the interaction of two labels y_i and y_j based on the interaction of two features $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$. These pairwise potentials are crucial for our model, as they allow us to capture the interplay of features and therefore to define the structure of objects. To that end, we realize the pairwise potentials with a linear classification of concatenated unary features that is passed through a softmax non-linearity:

$$\phi_{ij}(y_i, y_j, \mathbf{x}; \theta) = \frac{\exp\left((f_i(\mathbf{x}), f_j(\mathbf{x}))^T \mathbf{e}_{ij}^{y_i y_j}\right)}{\sum_{c,d\in\{-1,1\}} \exp\left((f_i(\mathbf{x}), f_j(\mathbf{x}))^T \mathbf{e}_{ij}^{cd}\right)} \ . \tag{4.3}$$

We use a specific classification vector $\mathbf{e}_{ij}^{cd}$ for each possible edge and each combination of labels that allows to model spatial dependencies and relations of different feature types. It is important here to emphasize that these potentials may also involve long-range dependencies of distant nodes and are not restricted to modeling only local neighborhood structures as in many recent approaches (Hoiem *et al.*, 2007; Kumar *et al.*, 2006; Lee *et al.*, 2006a; Levin and Weiss, 2006; Schnitzspan *et al.*, 2008; Shotton *et al.*, 2006; Winn and Shotton, 2006).

Both, the unary and pairwise potentials, contribute to our discriminative framework in the sense that the unary potentials classify nodes in the hierarchy independently while the pairwise potentials encode dependencies and thus the spatial configuration and underlying structure of objects. What sets this work apart from previous approaches is that we are able to learn the graph structure $\mathcal{A}$ automatically, which gives us a sound and efficient way of modeling complex, long-range dependencies. This allows us to determine the structure of the underlying domain and simultaneously consider a powerful hierarchical view on objects.

4.2.1 Hierarchical features

Before showing how to learn the parameters and structure of the model, we will first introduce the features and classifiers that the CRF model is based on.

We use a hierarchical representation of objects, which provides a powerful descriptor and yet is flexible enough to capture appearance, articulation and viewpoint changes. It is furthermore based on a dense representation of multiple descriptors in order to aggregate different cues on objects. We include both hierarchical HOG (hHOG) (Dalal and Triggs, 2005) and hierarchical bag of words (hBoW) (Lazebnik *et al.*, 2006) features to account for local and global representations of objects. In the following, we assume that each local classifier $F(\boldsymbol{\alpha}_n, f(\mathbf{x}))$ is actually the concatenation of a HOG and BoW classifier $F(f(\mathbf{x})) = \left(F^{\mathcal{H}}\left(\boldsymbol{\alpha}_n, f^{\mathcal{H}}(\mathbf{x})\right), F^{\mathcal{B}}\left(\boldsymbol{\alpha}_n, f^{\mathcal{B}}(\mathbf{x})\right)\right)$, which will be described in turn[1].

[1]Here and in the remainder of the chapter we drop the subscript i for parameters $\boldsymbol{\alpha}$ and feature functions f for notational simplicity.

Hierarchical HOG descriptors. For computing the hierarchical HOG features, we compute a dense grid of non-overlapping cells of oriented gradients (Dalal and Triggs, 2005) over the image. As in (Maji *et al.*, 2008), we extract multiple layers of those cell grids with increasing cell size at higher levels. Four neighboring cells are concatenated and normalized to one block, resulting in a dense grid of blocks (neighboring blocks overlap by 50%). We concatenate several blocks to form our local descriptors (details in Section 4.4). The global descriptor captures a holistic view on the object, since we concatenate all blocks of the bottom layer into a single global feature. The various descriptors are associated with the nodes of our model, indicated as green dots in Fig. 4.1.

Based on the local and global HOG descriptors, we train discriminative classifiers in order to represent local deformations as well as global statistics of objects. Therefore, we divide the grid of nodes in rectangular subregions (3×3 at the bottom layer) and train one SVM per subregion. Within the hierarchy, we reduce the number of subregions at higher levels: 2×2 at the second level and 1×1 at all other levels. Each classifier is a kernel-based SVM (cf. Fig. 4.1):

$$F^{\mathcal{H}}\left(\boldsymbol{\alpha}_n^{\mathcal{H}}, f^{\mathcal{H}}(\mathbf{x})\right) = \sum_{\mathbf{s} \in S_n^{\mathcal{H}}} \alpha_{n,\mathbf{s}}^{\mathcal{H}} K\left(\mathbf{s}, f^{\mathcal{H}}(\mathbf{x})\right) + \alpha_{n,0}^{\mathcal{H}} \ , \tag{4.4}$$

where $S_n^{\mathcal{H}}$ refers to the set of support vectors, K is an appropriate Mercer kernel, $\alpha_{n,\mathbf{s}}^{\mathcal{H}}$ denotes the support vector coefficients, and $\alpha_{n,0}^{\mathcal{H}}$ an offset. We employ linear kernels, though any Mercer kernel can be used.

In Fig. 4.3(b) we show the hierarchical HOG (hHOG) features of the shown object weighted with the parameters of our model and in Fig. 4.3(c) weighted with the weights of a linear SVM trained on the concatenation of all features. Note, with our model the real structure and shape is better represented, since our framework is able to learn feature couplings for capturing spatial dependencies of features.

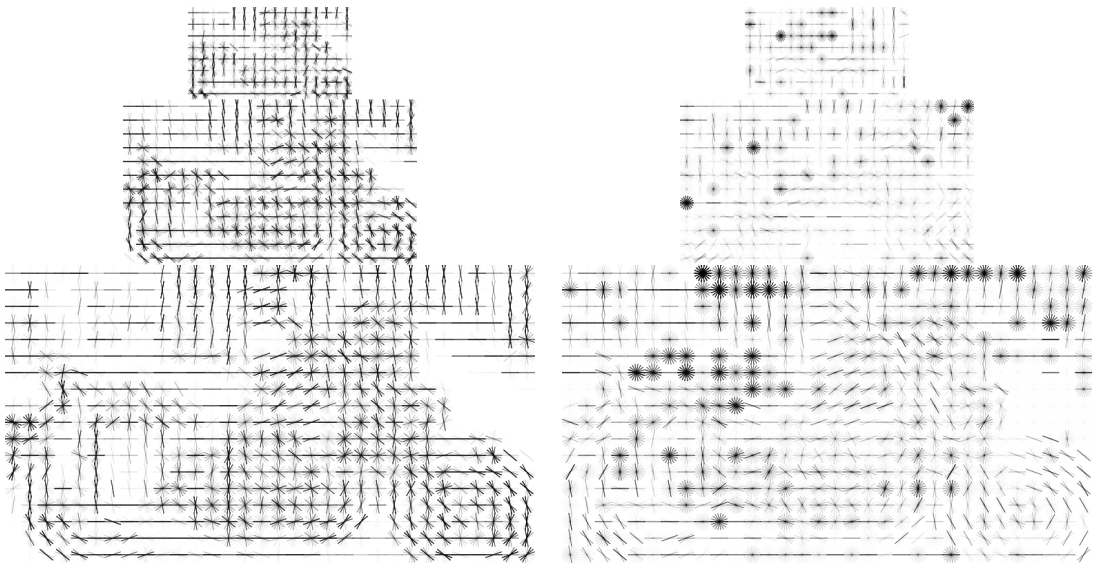
Hierarchical BoW descriptors. For integrating a hierarchical bag of words (hBoW) approach (Lazebnik *et al.*, 2006) in our model we calculate SIFT descriptors (Lowe, 2004) with radii $(5, 10, 15)$ and spacing of 10 pixels. Those descriptors are vector quantized with k-means clustering over the positive training instances. We calculate one global BoW descriptor over the entire image and subsequently divide the image in regions according to the number of nodes of every level of our hierarchical model. In every subregion, we build a histogram of word occurrences and use it as the feature $f^{\mathcal{B}}(\mathbf{x})$. The hBoW features are also classified using a kernel-based SVM:

$$F^{\mathcal{B}}\left(\boldsymbol{\alpha}_n^{\mathcal{B}}, f^{\mathcal{B}}(\mathbf{x})\right) = \sum_{\mathbf{s} \in S_n^{\mathcal{B}}} \alpha_{n,\mathbf{s}}^{\mathcal{B}} K\left(\mathbf{s}, f^{\mathcal{B}}(\mathbf{x})\right) + \alpha_{n,0}^{\mathcal{B}} \ , \tag{4.5}$$

where $S_n^{\mathcal{B}}$ refers to the set of support vectors, K is again a Mercer kernel, $\alpha_{n,\mathbf{a}}^{\mathcal{B}}$ denote the support vector coefficients, and $\alpha_{n,0}^{\mathcal{B}}$ is the offset.

Bootstrapping hard negatives. We bootstrap hard negative examples from the negative images and train the SVMs again with the additional negative images.

(a)



(b) (c)

Figure 4.3: (a) Object instance. (b) Hierarchical HOG features of the instance weighted with parameters of our model. (c) Hierarchical HOG features of the instance weighted with linear SVM weights.

4.3 MODEL LEARNING

Given training data consisting of a set of images X and the corresponding set of node labels Y, our goal is to estimate the model parameters $\theta = \{\boldsymbol{\alpha}, \mathbf{w}, \mathbf{e}\}$ and to identify a suitable graph structure represented by the active set $\mathcal{A}$.

4.3.1 Parameter learning

For now assuming a fixed graph structure $\mathcal{A}$, our goal is to train the parameters of the CRF model in a discriminative fashion. To that end, we consider the log-posterior

of the parameters

$$\mathcal{L}(\theta) = \log p(Y|X;\theta,\mathcal{A}) + \log p(\theta) \ , \tag{4.6}$$

which we aim to maximize. Here, $p(\theta) = p(\mathbf{w}) \cdot p(\mathbf{e})$ denotes a prior over the model parameters that regularizes parameter estimation to avoid overfitting. The SVM classifiers including the parameters $\boldsymbol{\alpha}$ are trained ahead of time decoupled from the rest of the model using standard quadratic programming, as for example in (Hoiem *et al.*, 2007; Lee *et al.*, 2006a; Shotton *et al.*, 2006; Winn and Shotton, 2006). This step attempts to optimally separate each object region from the background independently from other nodes in the hierarchy. Note that it would be also possible to train $\boldsymbol{\alpha}$ during CRF training based on the primal form of the SVM as described in the previous chapter 3, but we stick to the easier decoupled training procedure, since the focus of this chapter is on structure learning.

As usual in CRFs (Lafferty *et al.*, 2001), it is not possible to find a closed form estimate for the parameters. Hence we rely on gradient ascent (see e.g. (Levin and Weiss, 2006)) on the log-posterior to determine $\mathbf{w}$ and $\mathbf{e}$. Moreover, at each iteration we only consider a subset of the training data to improve efficiency, which yields a stochastic gradient ascent procedure.

Unary potentials. Assuming a Gaussian prior for the unary parameters ($P(\mathbf{w}) \sim \mathcal{N}(0,1)$), we derive the gradient of the log-posterior with respect to $\mathbf{w}_i$ as

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = \Big[\sum_{\mathbf{x}\in X} & E_{Y|\mathbf{x}} \left[\mathbf{F}(\boldsymbol{\alpha}_i, f_i(\mathbf{x})) \cdot y_i \cdot \psi_i(y_i,\mathbf{x})\right] - \\ & E_{p(\mathbf{y}|\mathbf{x})} \left[\mathbf{F}(\boldsymbol{\alpha}_i, f_i(\mathbf{x})) \cdot y_i \cdot \psi_i(y_i,\mathbf{x})\right] \Big] - \mathbf{w}_i \ ,\end{aligned} \tag{4.7}$$

where $E_{Y|\mathbf{x}}[\cdot]$ denotes the empirical expectation and $E_{p(\mathbf{y}|\mathbf{x})}[\cdot]$ denotes the expectation value under the posterior probability of our model. While the empirical expectation can be easily computed by plugging in the training label corresponding to $\mathbf{x}$, the expectation over the model distribution requires computing the marginal distribution $p(y_i|\mathbf{x})$. For a loopy graph as used here, this marginal cannot be computed in closed form. Consequently, we approximate it using loopy sum-product belief propagation (LBP) (Yedidia *et al.*, 2003), as is widely done in the literature (e.g. (Levin and Weiss, 2006)).

Note that learning the unary parameters corresponds to a simple form of structure learning that determines the relative importance of the features, much like multiple kernel learning does in SVMs. Intuitively, the weight of a node should be small, if that node is classified incorrectly for most of the training instances. Otherwise, the weight should be high, if a node helps to discriminate foreground from background training instances.

Pairwise potentials. For the pairwise potentials, we proceed in a similar fashion. We put a Laplace prior $p(\mathbf{e}) \propto \exp(-||\mathbf{e}||)$ on the weights corresponding to a

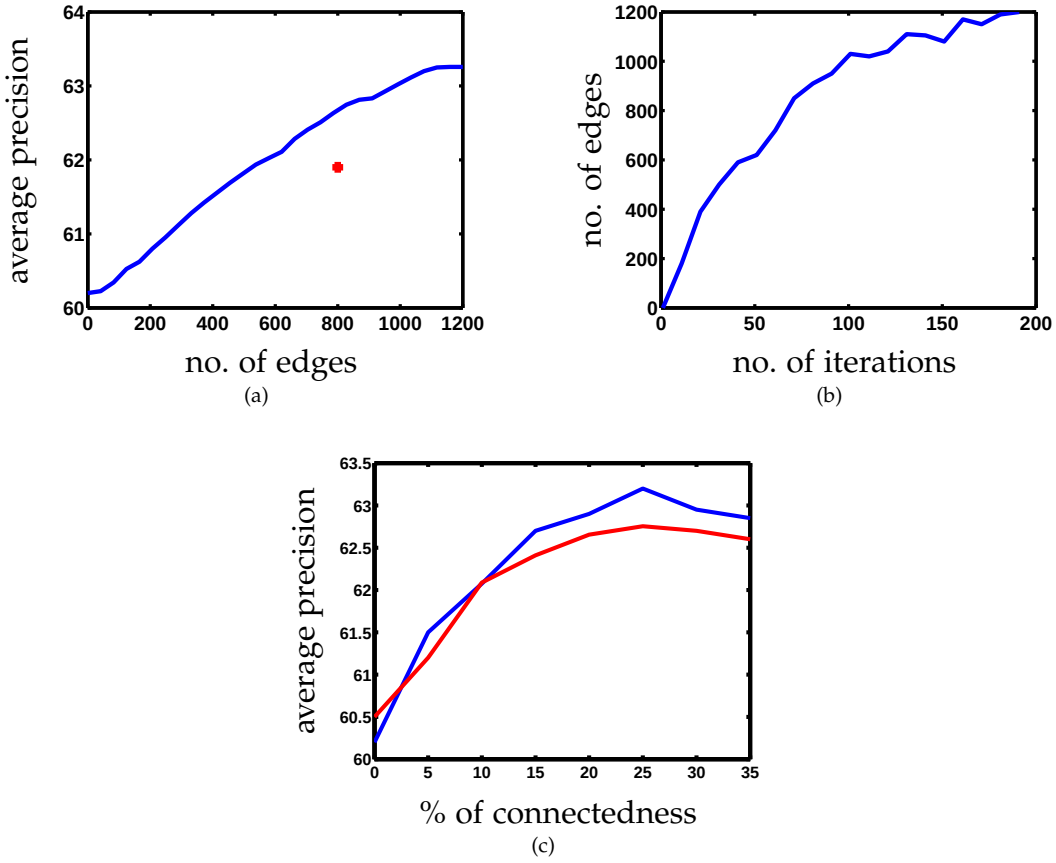


Figure 4.4: (a) Comparison of the average precision for learned (blue) and fixed (red) structure. The learned structure is plotted vs. number of edges. The fixed structure accounts for 800 edges. (b) Number of edges vs. number of iterations. (c) AP vs. different percentage of connectedness for binary labels (blue) and multi-labels (red)

L1-regularization (see below), and derive the gradient of the log-posterior as:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{e}_{ij}^{y_i y_j}} = \Big[\sum_{\mathbf{x} \in X} E_{Y|\mathbf{x}} \Big[(f_i(\mathbf{x}), f_j(\mathbf{x}))^T \phi_{ij}(y_i, y_j, \mathbf{x}) \Big] - E_{p(\mathbf{y}|\mathbf{x})} \Big[(f_i(\mathbf{x}), f_j(\mathbf{x}))^T \phi_{ij}(y_i, y_j, \mathbf{x}) \Big] \Big] - \mathrm{sgn}(\mathbf{e}_{ij}^{y_i y_j}) \ . \tag{4.8}$$

To compute the expectation over the model distribution, we require the marginals $p(y_i, y_j|\mathbf{x})$, which we again approximate using the beliefs from LBP.

The L1-regularization term not only avoids overfitting, but more importantly favors sparse solutions, where the majority of edges are inactive because of small weights (Lee *et al.*, 2006b). Care needs to be taken near 0 since the L1-regularizer is non-differentiable there. We avoid numerical problems by approximating the L1-norm by $\sqrt{\|\mathbf{e}\|^2 + \epsilon}$.

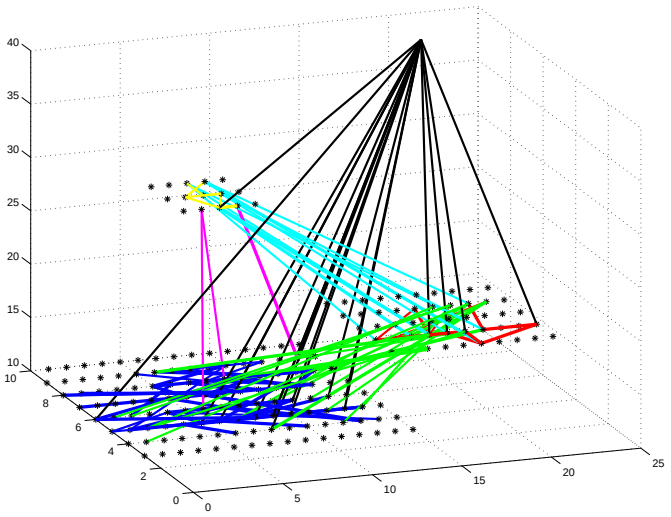


Figure 4.5: 2.5d visualization of the learned structure of our model.

4.3.2 Structure learning

The key contribution of this chapter compared to other CRF models is that we not only learn the parameters, but also the appropriate graph structure. In particular, our goal is to find a sparse set of edges that best describes the relevant dependencies and feature interactions for a particular class of objects (we learn one active set $\mathcal{A}$ per class). Similar to (Lee *et al.*, 2006b), we do this in an iterative fashion, where at each iteration we add meaningful pairwise features to the active set $\mathcal{A}$ from the large pool of candidate edges (the inactive set $\mathcal{I}$) and simultaneously remove features from the model that have become irrelevant. Since any change of the graph structure may render the current set of parameters θ inappropriate, we interleave each update of the graph structure with parameter learning as described above (100 iterations of gradient ascent). The procedure starts with a fully disconnected graph ($\mathcal{A} = \emptyset$ and $\mathcal{I} = \Omega$) and iteratively adds and removes edges.

Adding pairwise couplings. Since optimal feature selection is NP-hard, we use a gradient-based heuristic for estimating which feature most likely improves the model. We adapt the heuristic of Perkins *et al.* (2003), where at each step the feature with the largest likelihood-gradient is added to the active set. However, this method is only defined for generative models; here we carry this heuristic forward to discriminative structure learning with high-dimensional features. While such gradient-based heuristics are suboptimal, Lee *et al.* (2006b) showed that information gain-based heuristics provide only slight improvements compared to gradient-based heuristics, but the latter are more efficient to compute.

The intuition behind this is that edge (i, j) with the largest log-likelihood gradient

$\partial \log p(\mathbf{y} = \mathbf{1}|\mathbf{x}, \theta)/\partial \mathbf{e}_{ij}$ has the largest impact on changes of the target function (foreground likelihood) (Perkins *et al.*, 2003). In a generative setting, this would help explaining the foreground object, because we can expect the highest increase in likelihood by adding that edge, and thus the largest improvement of the model. Here we take a discriminative approach instead, and not only look at the importance of explaining the object, but rather link the likelihood assuming object and the likelihood assuming background to each other. To that end, we consider the log-likelihood ratio $\left(\log \frac{p(\mathbf{y}|\mathbf{x},\theta)}{p(-\mathbf{y}|\mathbf{x},\theta)}\right)$, and find the edge from the inactive set that maximizes the log-likelihood ratio:

$$(i^*, j^*) = \arg\max_{(i,j)\in\mathcal{I}} \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{e}_{ij}^{11}} - \frac{\partial \mathcal{L}}{\partial \mathbf{e}_{ij}^{-1-1}} \right\| \quad . \tag{4.9}$$

This criterion approximately finds the edge whose feature combination provides the largest improvement in discriminative power. The edge is subsequently added to the model ($\mathcal{A} \leftarrow \mathcal{A} \cup \{(i^*, j^*)\}$ and $\mathcal{I} \leftarrow \mathcal{I} \setminus \{(i^*, j^*)\}$).

So far, we argued for selecting edges according to (4.9), which requires computing the parameter gradient from (4.8). However, this is difficult to do as long as the edge is not added to the graph, but simply adding each potential candidate edge to the graph for computing (4.8) is infeasible. The underlying issue is that we need the pairwise marginals $p(y_i, y_j|\mathbf{x})$ to compute (4.8). We can, however, approximate this pairwise marginal using LBP as described in (Wainwright *et al.*, 2002):

$$\tilde{b}_{ij}(y_i, y_j) \propto \psi_i(y_i, \mathbf{x}) \cdot \psi_j(y_j, \mathbf{x}) \cdot \phi_{ij}(y_i, y_j, \mathbf{x}) \cdot \prod_{k\in\Gamma_i\setminus j} M_{ki}(y_i) \prod_{k\in\Gamma_j\setminus i} M_{kj}(y_j) \quad . \tag{4.10}$$

Here Γ_i refers to the neighborhood of i that is all nodes in $\mathcal{A}$ that are connected to i, and $M_{ki}(y_i)$ denotes the message that is passed from node k to node i.

Removing feature couplings. In order to avoid the model from becoming overly complex, which would make it inefficient and prone to overfitting, we follow two different strategies. The first is to use L1-regularization for the edge parameters, which encourages sparsity as discussed above. The other is to remove edges after each iteration of the structure learning procedure that are not crucial to the discriminative power. Whenever the weight of an active edge $(i, j) \in \mathcal{A}$ drops below a threshold ($\|\mathbf{e}_{ij}^{c_i c_j}\| \leq \tau_1$) and the weight gradient is below a threshold as well ($\|\frac{\partial \mathcal{L}}{\partial \mathbf{e}_{ij}^{c_i c_j}}\| \leq \tau_2$), we remove it from the active set ($\mathcal{A} \leftarrow \mathcal{A} \setminus \{(i, j)\}$ and $\mathcal{I} \leftarrow \mathcal{I} \cup \{(i, j)\}$). In this case the edge has no major influence on the log-likelihood ratio and since the gradient is small, one would expect the weights not to change significantly with more iterations of parameter learning. Thus, the edge and the coupling of the features can be removed without deteriorating the discriminative power considerably.

4.4 EXPERIMENTS

We report experiments on the challenging PASCAL VOC 2007 dataset (Everingham *et al.*, 2007) to support our claims about the benefits of our structure learning approach. For all experiments we report the average precision (AP), the common evaluation criterion of the PASCAL challenge. Due to computational reasons we prefiltered object hypotheses $\tilde{X}$ with the model of Felzenszwalb *et al.* (2008) and rescored them with our framework. Note, we do not leverage misclassifications of Felzenszwalb *et al.* (2008), but train our model on the provided training and validation bounding boxes and randomly cropped negative bounding boxes. Given our learned model (i.e. active edges and parameters $\boldsymbol{\alpha}, \mathbf{w}, \mathbf{e}$), for every $\tilde{\mathbf{x}} \in \tilde{X}$ we compute the log-likelihood that the object of interest is present, $\log p(\mathbf{y} = \mathbf{1}|\tilde{\mathbf{x}})$, in the hypothesized bounding box $\tilde{\mathbf{x}}$ and use this as the score. Note that we do not perform inference during testing, which is due to the fact that we are interested in an efficient way of obtaining a detection score. Inference during testing would additionally allow us to obtain a segmentation.

In all experiments we used SVM^{light} (Joachims, 1999) with linear kernels to train the parameters $\boldsymbol{\alpha}$. We did not add only a single edge per iteration but estimated and added the 20 best edges. In terms of pairwise parameters $\mathbf{e}$ we only optimize $\mathbf{e}^{+1,+1}$ and $\mathbf{e}^{-1,-1}$, and set $\mathbf{e}^{+1,-1} = \mathbf{e}^{-1,+1} = 0$, since we aim to classify whether bounding boxes contain the object or not. Thus, the case of changing signs is not represented in the training set and unlikely to appear during testing. Training the model takes approx. 10h while calculating the score for one bounding box takes approx. 0.3s.

Feature descriptors. In our experiments the global HOG descriptor is the same as in (Dalal and Triggs, 2005), though we use different sizes of local HOG descriptors. They are specific to each object class and depend on the aspect ratio and the average size of the bounding box. We used sizes between 4×2 or 2×4 blocks and 9×5 or 5×9 blocks of local gradient histograms. Thus, each local descriptor covers an area between 40×24 and 80×48 pixels (or 24×40 and 48×80). The local descriptors are sampled densely over the bounding box and may overlap up to $\frac{2}{3}$. For the experiments using the hierarchical representation we deployed 3 levels of local descriptors and one global descriptor (see Fig. 4.3.1).

PASCAL VOC 2006 motorbikes. A preliminary experiment on the motorbikes class of the PASCAL VOC 2006 challenge serves to shed light on the different aspects of our model. This dataset contains challenging multiscale, partially occluded and multiview instances. We trained our model on the provided training and validation set. In Tab. 4.1 and Fig. 4.6 results are summarized and detailed below. In Fig. 4.3.1 the most relevant feature couplings are shown. As it can be seen, our model includes short-range as well as long-range dependencies within but also between layers.

Our complete model (hHOG + hBoW features) yields a performance of 64.0% (histogram intersection kernel) and 63.2% (linear kernel), outperforming the baseline (Felzenszwalb *et al.*, 2008) (58.2%) by more than 5% AP. This emphasizes the benefit

VOC 2006 motorbikes	lin. SVM/ RBF-SVM	Unary poten.	lin. SVM on unary	Our model structure lin. / HI
BoW	36.1 / 38.0	20.3	23.7	42.7 / **45.1**
hBoW	49.0 / 50.1	45.0	47.1	52.4 / **53.5**
HOG	49.1 / 50.3	47.3	48.5	51.0 / **53.3**
hHOG	60.1 / 61.2	59.1	60.0	62.8 / **63.4**
hHOG + hBoW	61.0 / 61.7	60.2	61.4	63.2 / **64.0**
train on (Felzenszwalb *et al.*, 2008)	-	-	-	**64.2** / -
sliding window	-	-	-	60.1 / -
DPM (Felzenszwalb *et al.*, 2008)	58.2	-	-	-

Table 4.1: Summary of the results of different aspects of our model on the PASCAL VOC 2006 motorbikes. HI denotes the use of histogram intersection kernels.

of learning the structure of objects, since in (Felzenszwalb *et al.*, 2008) a fixed structure is assumed. When we train on the output of Felzenszwalb *et al.* (2008) the performance of our model increases to 64.2% with linear kernels. When not using (Felzenszwalb *et al.*, 2008) as a pre-filter, but sliding window instead we achieve 60.1% still outperforming (Felzenszwalb *et al.*, 2008).

In Fig. 4.4(a) we compare our structure learning method vs. an instantiation with local, fixed pairwise couplings (as described in the previous chapter 3), which amount to 800 pairwise edges. The model with fixed structure showed a performance of 61.9%, while our structure learning scheme achieved the same performance with fewer edges. When we look at the performance of structure learning with 800 automatically discovered edges, our framework achieved 62.7% AP.

For further investigating the stability of our model we experimented with different initializations of the active set $\mathcal{A}$ (empty and the structure described in the previous chapter 3), with different thresholds for removing edges and with different numbers of edges to be added to $\mathcal{A}$ in each iteration. For all these experiments our model learned similar structures and achieved similar performance. In Fig. 4.4(c) (blue line) we plot the performance vs. different degrees of connectedness (resulting from different thresholds). As it can be seen the performance does not differ dramatically for different thresholds when a certain level of connectedness is reached.

Furthermore, we compared our work against several baseline methods: SVM classification (linear and RBF kernels) on the concatenation of all features (column one of Tab. 4.1), unary classification alone (column two), and SVM classification on the output of our unary potentials (column three). SVM-based classification of the concatenation of our features showed 61.0% AP for linear kernels and 61.7% for RBF kernels, which we outperform by 3.0% and 2.3% AP respectively. Concerning unary classification alone, we calculated only the unary potentials (i.e. no active edges) and added them up to one classification score yielding 60.2% AP. Compared to the latter, our complete model showed an improvement of 3.8% AP. In a different setting we train a support vector machine on the output of our unary potentials, yielding a comparable performance (61.4%) as pure SVM classification. Note that our model outperforms all other corresponding learning methods on the challenging dataset,

VOC 2007	aero	bicyc	bird	boat	bottle	bus	car	cat	chair	cow
hHOG+hBoW	**31.7**	**56.3**	1.7	**15.1**	**27.6**	**41.3**	**48.0**	15.2	9.5	**18.3**
hHOG	30.0	56.1	1.5	15.0	27.2	41.1	47.5	14.5	9.5	18.1
DPM	28.1	55.4	1.4	14.5	25.4	38.9	46.6	14.3	9.4	16.0
Best VOC07	26.2	40.9	**9.8**	9.4	21.4	39.3	43.2	**24.0**	**12.8**	14.0

	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	average
hHOG+hBoW	**26.1**	11.3	**48.5**	**38.9**	**35.8**	**14.8**	**17.7**	**18.8**	**34.1**	**39.8**	**27.5**
hHOG	25.2	10.8	47.3	37.4	35.5	13.7	16.3	18.6	32.4	37.6	26.8
DPM	22.8	10.6	44.1	37.0	35.2	13.6	16.1	18.5	31.8	36.9	25.9
Best VOC07	9.8	**16.2**	33.5	37.5	22.1	12.0	17.5	14.7	33.4	28.9	23.3

Table 4.2: Results of our algorithm on the PASCAL VOC 2007 challenge.

which supports our claims about the flexibility and advantage of structure learning.

For further insights into our work, we evaluated our model when only using BoW features with one layer and with the hierarchy (hBoW), using only HOG features with one layer and with the hierarchy (hHOG). As can be seen in Tab. 4.1 our structure learning scheme consistently outperforms the other corresponding baseline models across all evaluated features.

Preliminary experiments with a multi-label setting as in the previous chapter 3 showed slightly worse performance than our binary label setting. This loss of performance might stem from the need to define the part descriptors need carefully, since the part descriptors are not given in the annotation of objects. Therefore, the learning algorithm has to be adapted to handle hidden nodes, which is described in the following capter 5. In Fig. 4.4(c) (red line) the performance vs. different degrees of connectedness is plotted.

PASCAL VOC 2007. In order to further support our claims about the advantages of our structure learning scheme, we evaluated our model on all 20 classes of the PASCAL VOC 2007 challenge. We compare our complete model using hierarchical HOG and hierarchical BoW features against using only hierarchical HOG features. Furthermore, we show the performance of the baseline of Felzenszwalb *et al.* (2008) and the best performance of the original challenge (Everingham *et al.*, 2007). Note, we used (Felzenszwalb *et al.*, 2008) as the baseline, since it is one of the leading methods on the PASCAL dataset and is most similar to our approach in terms of used features and basic representations. All results are summarized in Tab. 4.2.

On average across classes, our model achieved a performance of 27.5% outperforming the baseline of Felzenszwalb *et al.* (2008) (25.9%) by 1.6% AP. Furthermore, our structure learning model consistently improves the detection performance of the baseline across all categories between 0.1% AP for chairs and 4.4% AP for horses. As can be seen in Fig. 4.2 our work is more flexible in terms of modeling different viewpoints, appearances, and articulated instances. Furthermore, the highest

scored false positives of our model mainly account for misaligned bounding boxes containing the object of interest or sensible false alarms like bicycles recognized as motorbikes and cows recognized as horses. Thus, we conclude that our model helps in understanding the domain of interest and successfully discriminates object instances from background.

When comparing against the original VOC 2007 challenge, we achieved the best results for 16 of 20 classes. On average, we improved the best performance of the challenge (23.3%) by 4.2% AP. Note, in that measure we do not compare against one single model, but against the performance of the best model for every object class.

Furthermore, we tested our complete model (hBoW and hHOG features) in comparison to only using hHOG features. On average, using hBoW and hHOG improves the performance of using only hHOG (26.8%) by 0.7% AP. Again, the complete model consistently shows equal (chairs) or better performance (all other classes) up to an improvement of 2.2% AP. Thus, including different features helps our framework to model complex object classes and increases the detection performance. This observation is consistent with the work of Vedaldi *et al.* (2009) who included more complementary features yielding an increased performance. We expect additional performance gains by incorporating additional orthogonal features.

4.5 CONCLUSIONS

This chapter presented a novel discriminative structure learning framework applied to hierarchical representations for object detection. Our model is defined as a structure learning extension to standard CRF models that allows to preserve the discriminative notion and increase the expressiveness of the model for object detection. The model is capable of capturing inherent structure of the domain of interest, as it flexibly learns local as well as long-range feature couplings. Paired with discriminative hBoW- and hHOG-based classification, our scheme lends itself to modeling the spatial layout of objects, which is crucial for detection in challenging real world scenes. As the experiments show, our model can represent a higher variation in viewpoint, appearance and articulation than the currently leading method on the PASCAL VOC challenge.

Future directions might involve exploration of global context information and other complementary information. Furthermore, the joint learning paradigm as discussed in the previous chapter 3 could be adapted to train all model parameters in a single consistent framework.

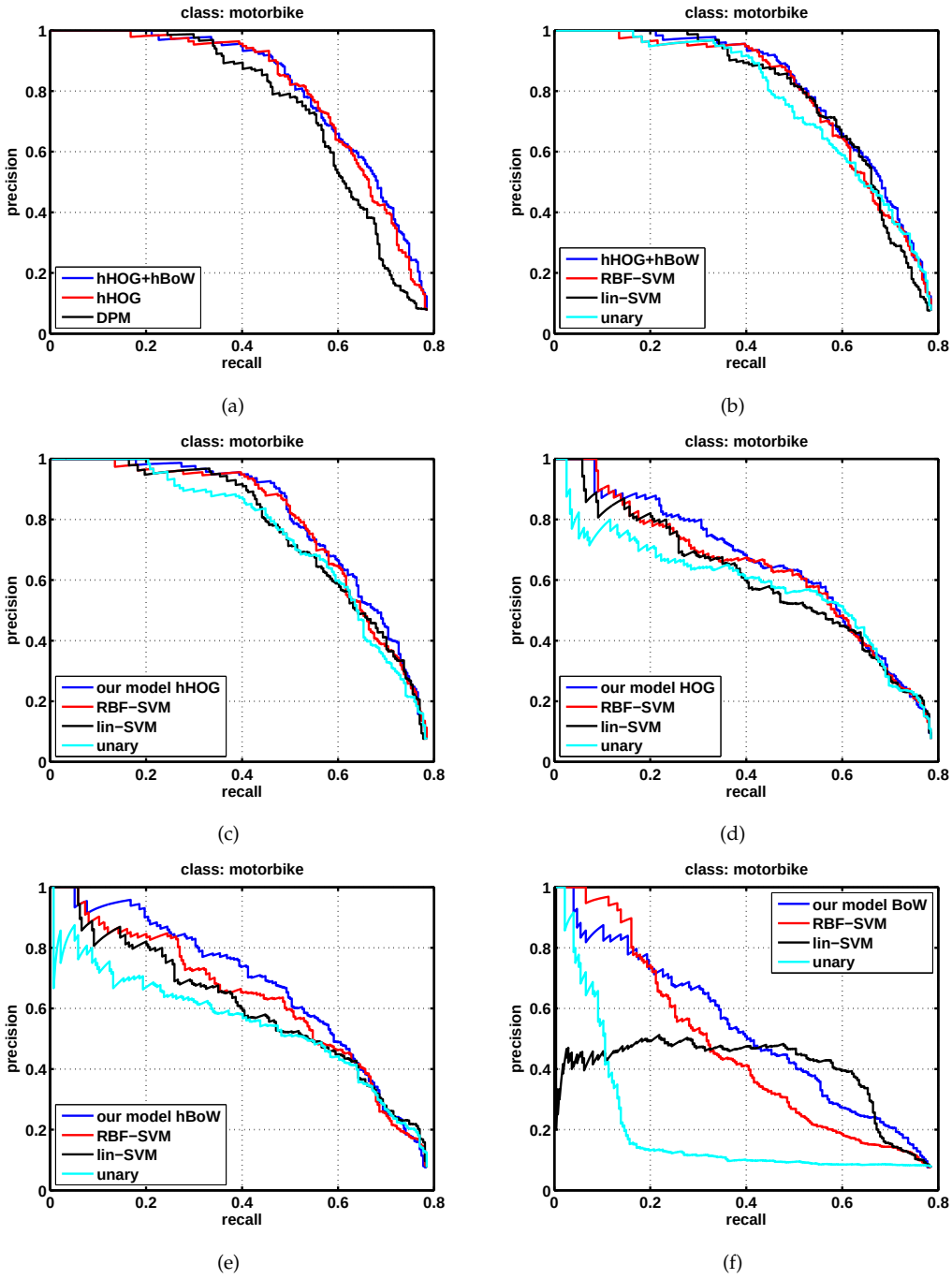


Figure 4.6: Precision-recall curves for the PASCAL VOC 2006 motorbikes dataset . (a) Our model using shape and appearance features (hHOG+hBoW) vs. our model using shape features (hHOG) vs. the model of Felzenszwalb *et al.* (2008) (DPM). (b) Our model (hHOG+hBoW) vs. RBF-kernel SVM classification on the concatenation of all of our unary features (RBF-SVM) vs. linear kernel SVM classification on the concatenation of all of our unary features (lin-SVM) vs. additively combined classifier scores from the unary potential of our model. (c) as in b) when only using hierarchical shape features. (d) as in b) but only using non-hierarchical shape features. (e) as in b) but only using hierarchical appearance features. (f) as in b) but only using non-hierarchical appearance features.

5 LATENT CRFS

Contents

This chapter can be seen as an extension to (or a generalization of) chapters 3 and 4. While the preceding chapters are restricted to either fixed structures or foreground-background instantiations, this chapter combines both ideas for incorporating richer notations (object parts) and flexible models (structure learning) in one consistent framework. We aim at learning the object parts in a weakly supervised fashion since in most object detection datasets only bounding box labels at the object level are provided. Experimentally we show that our model is able to learn meaningful parts, their spatial extent and the topological structure. This chapter describes the work published in (Schnitzspan *et al.*, 2010).

5.1 INTRODUCTION

Object recognition is challenging due to high intra-class variability caused, for example, by articulation, viewpoint changes, and partial occlusion. Successful methods need to strike a balance between being flexible enough to model such variation and discriminative enough to detect objects in cluttered, real world scenes. Motivated by these challenges we propose a latent conditional random field (CRF) based on a flexible assembly of parts.

The goal of this chapter is to introduce a model for object classes that brings together the competitive power of discriminative learning with the flexibility and expressiveness of part-based models. An important consideration is that we do

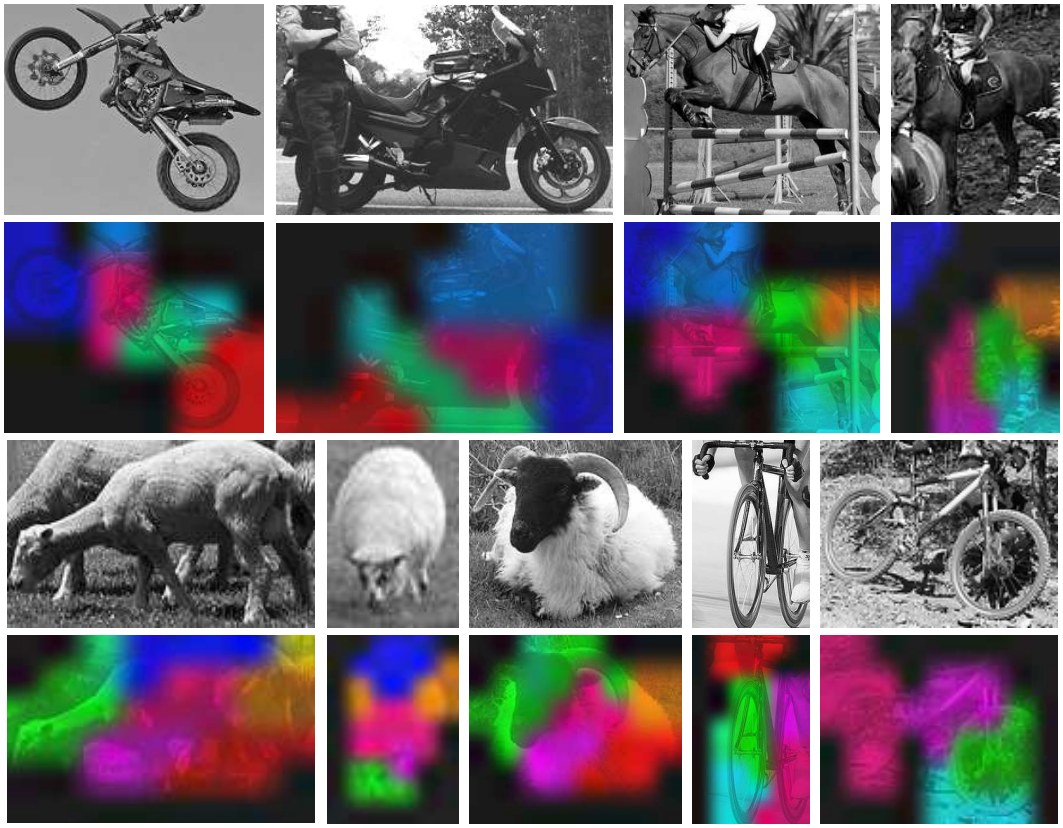


Figure 5.1: Parts of motorbikes, horses, bikes and sheep automatically discovered by our approach. Note how different viewpoints, articulation, and partial occlusions can be handled.

not want to provide supervision at the part level, but instead train the model in a weakly supervised fashion from class labels alone (cf. (Crandall and Huttenlocher, 2006; Felzenszwalb *et al.*, 2008)). In order to enable the automatic discovery of semantically meaningful part representations of objects, we model part labels as hidden nodes in a graphical model. We rely on two major components: A multi-label conditional random field (CRF) that aggregates image evidence and predicts object part occurrences, and a probabilistic classifier that predicts object or background occurrence from the spatial part configuration (see Fig. 5.2). In order to avoid having to provide part labels, we take a Bayesian approach and marginalize out the latent part configuration. Thus, our detector is a mixture of part-driven classifiers, which can take advantage of the uncertainty of bottom-up part discovery. To increase the flexibility and expressiveness of the model, we learn the pairwise structure of the underlying graphical model at the level of object part interactions. Efficient gradient-based techniques are used to estimate the structure of the domain of interest and carried forward to the multi-label or object part case.

Fig. 5.1 shows examples of how our approach automatically discovers semantically meaningful parts for PASCAL VOC 2007 motorbikes, horses, bikes and sheep. Note how the part interpretation stays consistent across different viewpoints and

other intra-class variations, and how articulation and partial occlusions are detected and handled. To train both the part CRF as well as the part-driven object classifier, we develop an expectation maximization (EM) algorithm that only requires bounding box labels as given in many object detection datasets. Our experimental results demonstrate that our model not only enables learning of semantically meaningful parts, but also obtains competitive results on the difficult PASCAL VOC 2007 dataset.

5.2 LATENT CRF MODEL

In our object detection scenario we are given a set of M images (or more precisely bounding boxes) $X = (\mathbf{x}^1, \ldots, \mathbf{x}^M)$, which contains objects from a particular class and background images. We are also given observed variables $Y = (y^1, \ldots, y^M)$, $y^m \in \{-1, 1\}$, which specify whether the corresponding image contains the object of interest ($y^m = 1$) or not ($y^m = -1$). We additionally introduce latent variables $Z = (\mathbf{z}^1, \ldots, \mathbf{z}^M)$, which refer to inferred part labelings for each image. Every part labeling $\mathbf{z}^m$ consists of N variables $\mathbf{z}^m = (z_1^m, \ldots, z_N^m)$, which denote localized part labels and are represented by the output nodes of the part CRF. Each $z_i^m \in \{0, \ldots, P\}$ takes on one of the possible part labels. P refers to the (maximum) number of object parts; part 0 represents the background. Without yet specifying the CRF in detail, we denote its nodes as $V = (1, \ldots, N)$ and let E refer to the edges of the graph. For simplicity of notation, we drop the superscript m indicating the training instance wherever applicable.

Since in object detection we are interested in the probability of presence or absence of objects, we model the posterior directly by marginalizing out the latent variables $\mathbf{z}$:

$$p(y|\,\mathbf{x}; \theta, E) = \sum_{\mathbf{z}} \underbrace{p(y|\mathbf{z}; \boldsymbol{\gamma})}_{\text{part-driven classifier}} \underbrace{p(\mathbf{z}|\mathbf{x}; \boldsymbol{\alpha}, \mathbf{e}, E)}_{\text{part CRF}} , \tag{5.1}$$

where the set of parameters is given by $\theta = \{\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{e}\}$. Here we assume that $p(y|\mathbf{z}; \boldsymbol{\gamma})$ is conditionally independent of $\mathbf{x}$ given $\mathbf{z}$, which implies that the object classifier only relies on the inferred part configuration rather than on the image itself. The part CRF models the distribution of object parts, and by marginalizing over $\mathbf{z}$ we obtain a mixture of part-driven object classifiers (see Fig. 5.2). The marginalization has the advantage that rather than committing to a possibly wrong configuration early on and drawing the wrong conclusion in consequence, we can consider all possible part configurations and take advantage of the inherent uncertainty of bottom-up part prediction.

5.2.1 Part CRF

We build our part model on CRFs since they provide a direct way of using multiple labels to represent spatially distributed parts in an image, and allow to model the uncertainty of the part configuration. The nodes in our graphical model are distributed over the image plane in an arbitrary layout, and are linked to features

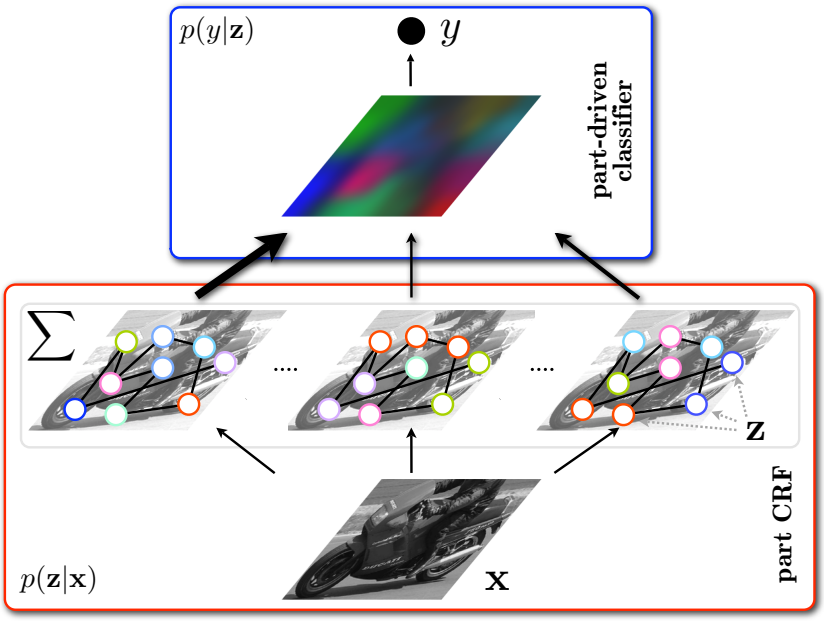


Figure 5.2: Model architecture consisting of a part CRF for bottom-up part detection, and part-driven object classifier. The part variables are marginalized out, taking advantage of their uncertainty.

computed from certain image regions. We associate a part label with each node, no matter where the corresponding feature is located or how small or large the feature size relative to the bounding box is. Several nodes can be associated with the same part label, thus allowing our model to flexibly adapt the spatial extent of object parts. During learning, the spatial extent of the object parts, their hierarchical feature representation, as well as the graph structure (= object topology) are determined automatically. In that sense our model is more powerful and far more flexible than related work (Felzenszwalb *et al.*, 2008; Kapoor and Winn, 2006; Kumar *et al.*, 2009). Moreover, it can be seen as a combination or generalization of the two preceding chapters and (Schnitzspan *et al.*, 2008, 2009).

Through the connection of nodes to image regions, the assigned node labels can be interpreted as a semantic representation of localized object parts. To deal with this flexible domain, we automatically learn the pairwise structure of feature couplings to allow arbitrary part interactions within the spatial extent of objects. The posterior distribution $p(\mathbf{z}|\mathbf{x})$ of part labels $\mathbf{z}$ given an image $\mathbf{x}$ is modeled as a CRF with unary and pairwise potentials:

$$p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) = \frac{1}{\mathcal{Z}(\boldsymbol{\alpha},\mathbf{e},\mathbf{x},E)} \prod_{i\in V} \psi_i(z_i,\mathbf{x};\boldsymbol{\alpha}) \cdot \prod_{(i,j)\in E} \phi_{ij}(z_i,z_j,\mathbf{x};\mathbf{e}) \quad . \tag{5.2}$$

The unaries ψ_i aggregate part evidence from a single image feature, while the pairwise potentials ψ_{ij} allow taking advantage of pairwise feature couplings. $\mathcal{Z}(\boldsymbol{\alpha},\mathbf{e},\mathbf{x},E)$ is the partition function, which ensures normalization. For now we assume that the

graph structure is given, meaning the set of edges $E \subseteq V \times V$ is fixed, and describe in Sec. 5.3 how the graph structure is learned as well.

Unary potentials. The unary potentials in our model are responsible for modeling part occurrences in an image based on single features $f_i(\mathbf{x})$. To supply robust discriminative part classifiers we train one support vector machine (SVM) $F(\cdot,\cdot)$ per object part and feature type. For now we assume that a part assignment is given for SVM training; in Sec. 5.3 we describe in detail how the part assignment is estimated from the class labels. Note that the part classifier is shared between all possible locations of a localized feature, which allows the object parts to occur anywhere in the bounding box and enables the model to capture articulations, viewpoint changes, and partial occlusions. In contrast to global object descriptors, this allows the representation to more easily adapt to positional variations and enables more specific appearance models. Based on the SVM classifier, we define the unary potential of node i for part label z_i using a softmax as

$$\psi_i(z_i, \mathbf{x}; \boldsymbol{\alpha}) = \frac{\exp\left(F(\boldsymbol{\alpha}(z_i), f_i(\mathbf{x}))\right)}{\sum_{c=0}^{P} \exp\left(F(\boldsymbol{\alpha}(c), f_i(\mathbf{x}))\right)} \ , \tag{5.3}$$

where the $f_i(\mathbf{x})$ refer to the features as described in Sec. 5.4, and $\boldsymbol{\alpha}(z_i)$ denotes the support vector coefficients of part label (class) z_i. The background "part" is modeled as

$$F(\boldsymbol{\alpha}(0), f_i(\mathbf{x})) = const \ . \tag{5.4}$$

This constant controls the uncertainty in part estimation of the unary SVM predictions.

Pairwise potentials. The pairwise potentials capture the structure of the domain of interest by modeling the cooccurrence of parts at connected nodes (i, j). Based on the interaction of the corresponding image features $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$, we capture the interplay of parts by computing softmax classifiers on the concatenated features. On the one hand, these linear classifiers on concatenated features are able to learn the appearance of object parts, and on the other hand, they can also model the cooccurrence of two object parts:

$$\phi_{ij}(z_i, z_j, \mathbf{x}; \mathbf{e}) = \frac{\exp\left((f_i(\mathbf{x}), f_j(\mathbf{x}))^T \mathbf{e}_{ij}^{z_i z_j}\right)}{\sum_{c_1,c_2=0}^{P} \exp\left((f_i(\mathbf{x}), f_j(\mathbf{x}))^T \mathbf{e}_{ij}^{c_1 c_2}\right)} \ . \tag{5.5}$$

Here, the parameters $\mathbf{e}_{ij}^{z_i z_j}$ are specific to each pairwise coupling and each combination of part labels. By allowing connections between arbitrary pairs of nodes, we obtain the flexibility to represent spatial relations not only of object parts in local neighborhoods, but also between distant locations within the spatial extent of objects. This topology is more flexible than simple star-shaped part models (Felzenszwalb *et al.*, 2008).

5.2.2 Part-driven object classifier

Given a spatial distribution of object parts the part-driven object classifier estimates object or background occurrence. We set up this object classifier in a non-parametric way that allows to model the contribution of each localized part label toward the object or background hypothesis. This holistic interpretation of part occurrences has the advantage that it allows for ambiguities in part localization as well as in part annotation. Such ambiguities are inevitable particularly in our latent variable setting, in which parts are inferred automatically. The object classifier can be written as

$$p(y|\mathbf{z};\boldsymbol{\gamma}) = \begin{cases} \sum_{i\in V} \gamma_i(z_i), & y = 1 \\ 1 - \sum_{i\in V} \gamma_i(z_i), & y = -1 \end{cases} , \tag{5.6}$$

with $\sum_{i\in V} \sum_{c=0}^{P} \gamma_i(c) = 1$ to ensure normalization. Note that by defining the classifier through weighted sums of part occurrences, it remains robust to an occasional absence of parts. This is in contrast to hidden CRFs that instead assume a factorized model (Kapoor and Winn, 2006; Quattoni *et al.*, 2007). During training, as will be explained in Sec. 5.3, the parameters $\boldsymbol{\gamma}$ are learned from the inferred part occurrences in the training set.

5.2.3 Detecting object instances

In order to evaluate whether an object instance is present in the bounding box $\mathbf{x}$ or not, we need to compute $p(y = 1|\mathbf{x};\theta,E) = \sum_{\mathbf{z}} p(y = 1|\,\mathbf{z};\boldsymbol{\gamma})p(\mathbf{z}\,|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E)$, which involves marginalization over all part configurations. The posterior object probability simplifies to

$$p(y = 1|\mathbf{x};\theta,E) = \sum_{i\in V} \sum_{z_i=0}^{P} \gamma_i(z_i) p(z_i|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \ . \tag{5.7}$$

Proof.

$$\begin{aligned} p(y = 1|\mathbf{x};\theta,E) &= \sum_{\mathbf{z}} p(y = 1|\,\mathbf{z};\boldsymbol{\gamma}) p(\mathbf{z}\,|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) && (5.8)\\ &= \sum_{\mathbf{z}} \left(\left(\sum_{i\in V} \gamma_i(z_i) \right) p(\mathbf{z}\,|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \right) && (5.9)\\ &= \sum_{i\in V} \sum_{z_i=0}^{P} \left(\gamma_i(z_i) \left(\sum_{\mathbf{z}\setminus z_i} p(\mathbf{z}\,|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \right) \right) && (5.10)\\ &= \sum_{i\in V} \sum_{z_i=0}^{P} \gamma_i(z_i) p(z_i|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \ . && (5.11) \end{aligned}$$

□

Since exact computation of the marginals $p(z_i|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E)$ is intractable, we approximate them using the beliefs $b_i(z_i)$ from sum-product belief propagation. For efficiency reasons, we pre-filter candidate bounding boxes with a HOG detector (Dalal and Triggs, 2005) and compute the score of the pre-filtered windows with our full model. For the sake of completeness we also report the performance without pre-filtering.

5.3 LEARNING THE MODEL

To train our latent variable model, we rely on the well-known expectation maximization (EM) algorithm. This allows our approach to discover semantically interpretable part annotations from object class labels alone. The combination of the flexibility of our model and the power of EM to infer and adapt to soft assignments of hidden nodes, and therefore part labels, yields a theoretically sound, yet practically scalable approach. Our objective is to maximize the expected complete log-likelihood with respect to $\theta = \{\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{e}\}$:

$$Q(\theta, \theta^{\text{old}}) = \sum_{Z} \Big[p(Z|Y, X; \theta^{\text{old}}, E^{\text{old}}) \cdot \log \big(p(Y|Z;\boldsymbol{\gamma}) p(Z|X;\boldsymbol{\alpha},\mathbf{e},E) \big) \Big] \ , \qquad (5.12)$$

where θ^{old} refers to the parameters from the last M-step.

Initialization. It is well known that EM requires proper initialization to work well. We infer an initial part labeling using k-means clustering over the positive training instances, which yields a hard assignment of nodes to object parts (cf. (Schnitzspan *et al.*, 2008) and previous chapter). We use these hard part assignments from the positive instances and randomly sampled negative instances to initially train the part classifiers (i.e. SVMs), which yields our initialization for $\boldsymbol{\alpha}$. Note here that this procedure only provides an initialization; later the part classifiers are re-trained as required by the part representation. The parameters of the part-driven classifier $\boldsymbol{\gamma}$ are initialized by counting part occurrences in the inferred hard assignment and normalizing. The edge parameters $\mathbf{e}$ require no initialization, as at the beginning no edges are present in the graph (see below).

E-Step. In the E-step we compute expected (i.e. soft) assignments of part labels to nodes for the training set of observed class labels Y and images X:

$$p(Z|Y, X; \theta^{\text{old}}, E^{\text{old}}) = \prod_{m=1}^{M} \frac{p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}^{\text{old}}) p(\mathbf{z}^m|\mathbf{x}^m;\boldsymbol{\alpha}^{\text{old}},\mathbf{e}^{\text{old}},E^{\text{old}})}{p(y^m|\mathbf{x}^m;\theta^{\text{old}},E^{\text{old}})} \ . \qquad (5.13)$$

Here, $p(y^m|\mathbf{x}^m;\theta^{\text{old}},E^{\text{old}})$ is computed approximately using belief propagation as in Eq. (5.7). The E-step thus yields probabilities of nodes belonging to certain object parts, which flexibly adapts toward a meaningful representation of parts as learning proceeds.

Generalized M-Step. After computing the soft part assignments, we maximize the expected complete log-likelihood $Q(\theta, \theta^{\text{old}})$ with respect to θ. We use gradient ascent, since there is no closed form solution for the parameters γ. In the following let b_i^{old} denote the part beliefs from the previous iteration. Per M-step we use one gradient update. The gradient can be approximated as

$$\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial \gamma_i(c)} \approx \sum_{m=1, y^m=1}^{M} \frac{b_i^{\text{old}}(c)}{p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} - \sum_{m=1, y^m=-1}^{M} \frac{b_i^{\text{old}}(c)}{p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} \ , \tag{5.14}$$

Proof. In the proof we make use of the approximation $\frac{p(y|\,\mathbf{z};\gamma^{\text{old}})}{p(y|\,\mathbf{z};\gamma)} \approx 1$. Also recall that $p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\alpha}, \mathbf{e}, E)$ does not depend on γ.

$$\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial \gamma_i(c)} = \sum_{Z} \left(\prod_{j=1}^{M} p(\mathbf{z}^j | y^j, \mathbf{x}^j; \theta^{\text{old}}, E^{\text{old}}) \right) \left(\sum_{m=1}^{M} \frac{1}{p(y^m|\mathbf{z}^m;\gamma)} \frac{\partial p(y^m|\mathbf{z}^m;\gamma)}{\partial \gamma_i(c)} \right) \tag{5.15}$$

$$= \sum_{m=1}^{M} \sum_{\mathbf{z}^m} \frac{p(\mathbf{z}^m|y^m, \mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})}{p(y^m|\mathbf{z}^m;\gamma)} \frac{\partial p(y^m|\mathbf{z}^m;\gamma)}{\partial \gamma_i(c)} \underbrace{\left(\sum_{Z \setminus \mathbf{z}^m} \prod_{\substack{j=1 \\ j \neq m}}^{M} p(\mathbf{z}^j | y^j, \mathbf{x}^j; \theta^{\text{old}}, E^{\text{old}}) \right)}_{=1} \tag{5.16}$$

$$= \sum_{m=1}^{M} \sum_{\mathbf{z}^m} \frac{p(y^m|\,\mathbf{z}^m;\gamma^{\text{old}}) p(\mathbf{z}^m|\mathbf{x}^m; \boldsymbol{\alpha}^{\text{old}}, \mathbf{e}^{\text{old}}, E^{\text{old}})}{p(y^m|\mathbf{z}^m;\gamma) p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} \frac{\partial p(y^m|\mathbf{z}^m;\gamma)}{\partial \gamma_i(c)} \tag{5.17}$$

$$\approx \sum_{m=1}^{M} \sum_{\mathbf{z}^m} \frac{p(\mathbf{z}^m|\mathbf{x}^m; \boldsymbol{\alpha}^{\text{old}}, \mathbf{e}^{\text{old}}, E^{\text{old}})}{p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} \frac{\partial p(y^m|\mathbf{z}^m;\gamma)}{\partial \gamma_i(c)} \ . \tag{5.18}$$

Here we observe that $\frac{\partial p(y^m|\mathbf{z}^m;\gamma)}{\partial \gamma_i(c)}$ does not depend on z_j^m for $j \neq i$ since

$$\frac{\partial}{\partial \gamma_i(c)} p(y^m|\mathbf{z}^m;\gamma) = \begin{cases} 1, & \text{if } y^m = 1 \text{ and } z_i^m = c \\ -1, & \text{if } y^m = -1 \text{ and } z_i^m = c \\ 0, & \text{otherwise} \ . \end{cases} \tag{5.19}$$

In the following we factor out $\frac{\partial}{\partial \gamma_i(c)} p(y^m|\mathbf{z}^m;\gamma)$ and approximate $p(z_i^m|\mathbf{x}^m; \boldsymbol{\alpha}^{\text{old}}, \mathbf{e}^{\text{old}}, E^{\text{old}}) \approx b_i^{\text{old}}(z_i)$:

$$\frac{\partial Q(\theta, \theta^{\text{old}})}{\partial \gamma_i(c)} \approx \sum_{m=1}^{M} \sum_{z_i^m=0}^{P} \left(\left(\frac{\partial}{\partial \gamma_i(c)} p(y^m|\mathbf{z}^m;\gamma) \right) \sum_{\mathbf{z}^m \setminus z_i^m} \frac{p(\mathbf{z}^m|\mathbf{x}^m; \boldsymbol{\alpha}^{\text{old}}, \mathbf{e}^{\text{old}}, E^{\text{old}})}{p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} \right) \tag{5.20}$$

$$= \sum_{m=1}^{M} \sum_{z_i^m=0}^{P} \left(\left(\frac{\partial}{\partial \gamma_i(c)} p(y^m|\mathbf{z}^m;\gamma) \right) \frac{p(z_i^m|\mathbf{x}^m; \boldsymbol{\alpha}^{\text{old}}, \mathbf{e}^{\text{old}}, E^{\text{old}})}{p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} \right) \tag{5.21}$$

$$\approx \sum_{m=1, y^m=1}^{M} \frac{b_i^{\text{old}}(c)}{p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} - \sum_{m=1, y^m=-1}^{M} \frac{b_i^{\text{old}}(c)}{p(y^m|\mathbf{x}^m; \theta^{\text{old}}, E^{\text{old}})} \ . \tag{5.22}$$

□

We optimize the parameters $\boldsymbol{\alpha}$ by re-training the SVM part classifier. To exploit the uncertainty of the part labeling expressed by the soft assignments, we weigh each part occurrence in the training set with its soft assignment. Here we alternate between the maximum margin objective of SVM training and likelihood maximization. As also mentioned in the preceding chapter, this potential restriction could be avoided by adapting the maximum margin Markov network formalism of Taskar *et al.* (2003). However, the focus of this chapter is on learning the hidden part labels and the structure between them.

As is the case even for fully observed training of CRFs, there is no closed form solution for the edge parameters $\mathbf{e}$. We thus use gradient ascent that locally maximizes $Q(\theta, \theta^{\text{old}})$. The gradient with respect to edge parameters for parts c_1 and c_2 is given as

$$\frac{\partial Q(\theta,\theta^{\text{old}})}{\partial \mathbf{e}_{ij}^{c_1 c_2}} = \sum_{m=1}^{M} \left[\sum_{\mathbf{z}^m} \left(\frac{\partial \log(p(\mathbf{z}^m|\mathbf{x}^m,\boldsymbol{\alpha},\mathbf{e},E))}{\partial \mathbf{e}_{ij}^{c_1 c_2}} p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}^{\text{old}}) \right.\right.$$
$$\left.\left. p(\mathbf{z}^m|\mathbf{x}^m,\boldsymbol{\alpha}^{\text{old}},\mathbf{e}^{\text{old}},E^{\text{old}}) / p(y^m|\mathbf{x}^m,\theta^{\text{old}},E^{\text{old}}) \right) \right] . \tag{5.23}$$

Proof. In order to derive this equation we simplify the notation of $Q(\cdot,\cdot)$.

$$Q(\theta,\theta^{\text{old}}) = \sum_{m=1}^{M} \sum_{Z} \Big[\Big(\log p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}) + \log p(\mathbf{z}^m|\mathbf{x}^m;\boldsymbol{\alpha},\mathbf{e},E) \Big) \cdot p(Z|Y,X;\theta^{\text{old}},E^{\text{old}}) \Big] \tag{5.24}$$

$$= \sum_{m=1}^{M} \sum_{\mathbf{z}^m} \Big[\Big(\log p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}) + \log p(\mathbf{z}^m|\mathbf{x}^m;\boldsymbol{\alpha},\mathbf{e},E) \Big) \cdot \sum_{Z\setminus\mathbf{z}^m} p(Z|Y,X;\theta^{\text{old}},E^{\text{old}}) \Big] \tag{5.25}$$

$$= \sum_{m=1}^{M} \sum_{\mathbf{z}^m} \Big[\Big(\log p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}) + \log p(\mathbf{z}^m|\mathbf{x}^m;\boldsymbol{\alpha},\mathbf{e},E) \Big) \cdot p(\mathbf{z}^m|y^m,\mathbf{x}^m;\theta^{\text{old}},E^{\text{old}}) \Big] . \tag{5.26}$$

Following this we can easily derive the gradient with respect to the edge parameters $\mathbf{e}$. Recall that $p(y|\mathbf{z};\boldsymbol{\gamma})$ does not depend on the parameters $\mathbf{e}$.

$$\frac{\partial Q(\theta,\theta^{\text{old}})}{\partial \mathbf{e}_{ij}^{c_1 c_2}} = \sum_{m=1}^{M} \sum_{\mathbf{z}^m} \left(\left(\frac{\partial \log(p(\mathbf{z}^m|\mathbf{x}^m;\boldsymbol{\alpha},\mathbf{e},E))}{\partial \mathbf{e}_{ij}^{c_1 c_2}} \right) \frac{p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}^{\text{old}}) p(\mathbf{z}^m|\mathbf{x}^m;\boldsymbol{\alpha}^{\text{old}},\mathbf{e}^{\text{old}},E^{\text{old}})}{p(y^m|\mathbf{x}^m;\theta^{\text{old}},E^{\text{old}})} \right) . \tag{5.27}$$

□

Computing the gradient of $Q(\theta,\theta^{\text{old}})$ thus reduces to calculating the conditional log-likelihood gradient, which is also required for training standard CRFs. The

gradient of the conditional log-likelihood $\mathcal{C}(\mathbf{e}) = \log p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\alpha}, \mathbf{e}, E)$ with respect to the edge parameters is given as:

$$\begin{aligned}\frac{\partial \mathcal{C}(\mathbf{e})}{\partial \mathbf{e}_{ij}^{c_1 c_2}} =& E_{\mathbf{z}_{\{z_i=c_1, z_j=c_2\}}|\mathbf{x}}\left[(f_i(\mathbf{x}), f_j(\mathbf{x}))^T \phi_{ij}(c_1, c_2, \mathbf{x})\right] - \\ & E_{p(\mathbf{z}_{\{z_i=c_1, z_j=c_2\}}|\mathbf{x})}\left[(f_i(\mathbf{x}), f_j(\mathbf{x}))^T \phi_{ij}(c_1, c_2, \mathbf{x})\right] \ ,\end{aligned} \tag{5.28}$$

where $E_{\mathbf{z}|\mathbf{x}}$ denotes the empirical expectation and $E_{p(\mathbf{z}|\mathbf{x})}$ refers to the expectation under the posterior distribution of our part CRF. A similar notation is used in (Schnitzspan *et al.*, 2009) and the previous chapter and more details on this gradient can be found there.

In order to cope with differing major orientations of the instances (left-right), we initialize our model with the original orientations of the dataset. In each iteration we evaluate our model on the original and a mirrored image and choose the one with the highest score for the next iteration.

5.3.1 Structure learning

In order to learn the spatial relationship between object parts, we use discriminative structure learning to find the edges in the CRF that maximize the discriminative power of the overall model. In particular, we develop a multi-label extension of L1-regularized gradient-based discriminative structure learning (Schmidt *et al.*, 2008; Schnitzspan *et al.*, 2009). In our scenario we are interested in improving the discriminative power of our approach in terms of the object's presence or absence, and therefore consider the log-posterior ratio

$$\mathcal{R}(\mathbf{e}_{ij}) = \max_{(c_1,c_2)\in\{0..P\}^2} \left\| \sum_{m=1, y^m=1}^{M} \frac{\partial \log p(y^m=1|\mathbf{x}^m;\theta,E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}} - \sum_{m=1, y^m=-1}^{M} \frac{\partial \log p(y^m=-1|\mathbf{x}^m;\theta,E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}} \right\| \ . \tag{5.29}$$

The edges that maximize this ratio likely improve our model, and therefore should be added to it. At the same time, edges that have small ratio and small absolute edge weights can be removed from the current active edge set, because they have only a small impact on the objective. In the preceding chapter, we discussed the binary case, in which the maximum has to be calculated only over all candidate edges. Here we assume the more complex case, in which we additionally have to consider part labellings. Therefore the structure learning objective $\mathcal{R}(\mathbf{e}_{ij})$ considers the gradient over all possible edges and possible part constellations and takes the maximum. The gradient can be written as

$$\frac{\partial \log p(y|\,\mathbf{x};\theta,E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}} = \frac{1}{p(y|\mathbf{x};\theta,E)} \left(\sum_{\mathbf{z}} p(y|\mathbf{z};\boldsymbol{\gamma}) p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \frac{\partial \log p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}} \right) \ . \tag{5.30}$$

Proof. By exploiting that

$$\frac{\partial}{\partial \mathbf{e}_{ij}^{c_1 c_2}} p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) = p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \cdot \frac{\partial}{\partial \mathbf{e}_{ij}^{c_1 c_2}} \log p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \ , \tag{5.31}$$

we derive the ratio $\mathcal{R}(\cdot)$ for a fixed m and drop this index for notational simplicity:

$$\frac{\partial \log p(y|\mathbf{x};\theta,E)}{\partial \mathbf{e}_{ij}^{c_1c_2}} = \frac{1}{p(y|\mathbf{x};\theta,E)} \frac{\partial}{\partial \mathbf{e}_{ij}^{c_1c_2}} \left(\sum_{\mathbf{z}} \underbrace{p(y|\mathbf{z};\gamma)}_{\text{independent of } \mathbf{e}_{ij}^{c_1c_2}} p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \right) \tag{5.32}$$

$$= \sum_{\mathbf{z}} \left(\frac{p(y|\mathbf{z};\gamma)}{p(y|\mathbf{x};\theta,E)} \left(\frac{\partial}{\partial \mathbf{e}_{ij}^{c_1c_2}} p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) \right) \right) \tag{5.33}$$

$$= \sum_{\mathbf{z}} \frac{p(y|\mathbf{z};\gamma)p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E)}{p(y|\mathbf{x};\theta,E)} \frac{\partial \log p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E)}{\partial \mathbf{e}_{ij}^{c_1c_2}} \ . \tag{5.34}$$

We can now compute $\mathcal{R}(\mathbf{e}_{ij})$ using the conditional log-likelihood gradient of a standard CRF. □

The gradient on the right hand side is computed as in Eq. (5.28). We proceed by finding the best edge to add:

$$\mathbf{e}_{i^*j^*} = \underset{(i,j)\in V\times V\setminus E}{\arg\max} \ \mathcal{R}(\mathbf{e}_{ij}) \ . \tag{5.35}$$

We start the learning process with no pairwise couplings of nodes and iteratively add the ten best edges (highest ratio of gradient norm) to the model at the end of each M-step. At the same time, we remove edges with absolute weight below a threshold τ_1 that also have an absolute gradient norm below the threshold τ_2. In combination with L1-regularization, this scheme leads to sparsely connected graphs, and at convergence has a connectedness of approximately 20%. Experiments with different τ_1 and τ_2 showed that the determined structure is robust to changes in τ_1 and τ_2 even though these parameters control the connectedness of our model – higher thresholds yield lower connectedness.

5.4 IMAGE FEATURES

The aggregation of features and their linkage to image regions defines the basic spatial layout of the latent nodes of our model. We build a dense representation of objects that includes both histograms of oriented gradients (HOG) (Dalal and Triggs, 2005) and bag of words (BoW) (Lazebnik *et al.*, 2006) descriptors. In our experiments these specific feature descriptors emerged to be suitable to capture local deformations and viewpoint changes. Note, however, that our model allows for an arbitrary layout of nodes, which means that we can rely on any feature aggregation scheme, both local and global. This allows for future integration of orthogonal features, which appears promising due to the success of combining several features (Vedaldi *et al.*, 2009). In our experiments we fixed the extent of our features (the size of the linked image region) and leave it for future work to automatically select optimal feature scopes.

HOG descriptors. The HOG descriptors are computed by calculating a dense grid of non-overlapping cells of oriented gradients (Dalal and Triggs, 2005) - each cell being 8×8 pixels in size. A dense block grid with 50% overlap between blocks is built by concatenating and normalizing four neighboring cells. Similar to (Felzenszwalb *et al.*, 2008), we rely on local views of objects by concatenating several neighboring blocks to form one feature descriptor. In our experiments we concatenate 5×5 neighboring blocks into one local descriptor. In addition, we compute a global descriptor that comprises all blocks of the grid, thus aggregating evidence from the entire object.

BoW descriptors. The bag of words (BoW) descriptors (Lazebnik *et al.*, 2006) are formed by densely calculating SIFT features (Lowe, 2004) with radii $(5, 10, 15)$ and a spacing of 10 pixels. We vector-quantize these features with k-means clustering over the positive training instances. We divide the image into overlapping regions, which each forms a feature descriptor. In our experiments we use regions of 50×50 pixels with an overlap of 50%. A local BoW descriptor is then formed by measuring the word occurrences in one specific region. A global BoW descriptor is calculated by measuring the word occurrences in the entire bounding box.

Part classifiers. For each local as well as global feature and feature type $\mathcal{T} \in \{\mathcal{H}, \mathcal{B}\}$ ($\mathcal{H}$ denotes HOG, and $\mathcal{B}$ BOW) we train one SVM

$$F^{\mathcal{T}}(\boldsymbol{\alpha}^{\mathcal{T}}(c), f_i^{\mathcal{T}}(\mathbf{x})) = \sum_{\mathbf{s} \in S^{\mathcal{T}}(c)} \alpha_s^{\mathcal{T}}(c) K\left(\mathbf{s}, f_i^{\mathcal{T}}(\mathbf{x})\right) + \alpha_0^{\mathcal{T}}(c) \ , \tag{5.36}$$

where $S^{\mathcal{T}}(\cdot)$ refers to the set of support vectors and $K(\cdot, \cdot)$ denotes an appropriate Mercer kernel. In our experiments we make use of the histogram intersection kernel (Maji *et al.*, 2008). Each part classifier is then defined as a sum of HOG and BoW classifiers $F(\boldsymbol{\alpha}(c), f_i(\mathbf{x})) = F^{\mathcal{H}}(\boldsymbol{\alpha}^{\mathcal{H}}(c), f_i^{\mathcal{H}}(\mathbf{x})) + F^{\mathcal{B}}(\boldsymbol{\alpha}^{\mathcal{B}}(c), f_i^{\mathcal{B}}(\mathbf{x}))$.

5.5 EXPERIMENTS

We evaluated our model on the PASCAL VOC 2007 dataset with the common average precision (AP) metric (Everingham *et al.*, 2007). This dataset includes images from 20 object classes and is challenging due to partial occlusion, articulation, and viewpoint changes. Training our model takes approximately 8 hours while computing detection scores for an entire image takes approximately 15 sec on a 2 GHz AMD Opteron machine, when using a global HOG pre-filter. Without the pre-filter we apply belief propagation for all locations and scales, which increases the computation time to approximately 450 sec per image. We use SVM^{light} for training the SVMs (Joachims, 1999).

Qualitative observations. Fig. 5.3 shows mean images over the positive training instances of motorbikes (top) and horses (bottom). The left column shows the

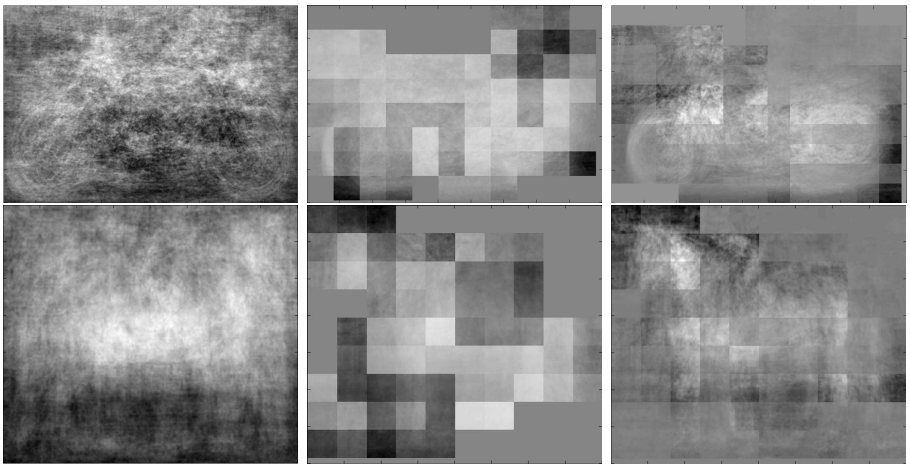


Figure 5.3: Mean image averaged over *(left)* all instances, *(middle)* part occurrences of *k*-means clustering, *(right)* part occurrences learned with EM. *(top)* VOC 2007 motorbikes, *(bottom)* horses.

VOC 2007	fixed structure ("no sl")	structure learning ("sl")
global	30.2	-
k-means	32.1	33.2
maximization	33.4	35.0
marginalization	34.2	36.3

Table 5.1: Comparison of different model instantiations on a subset of PASCAL VOC 2007 motorbikes (in average precision).

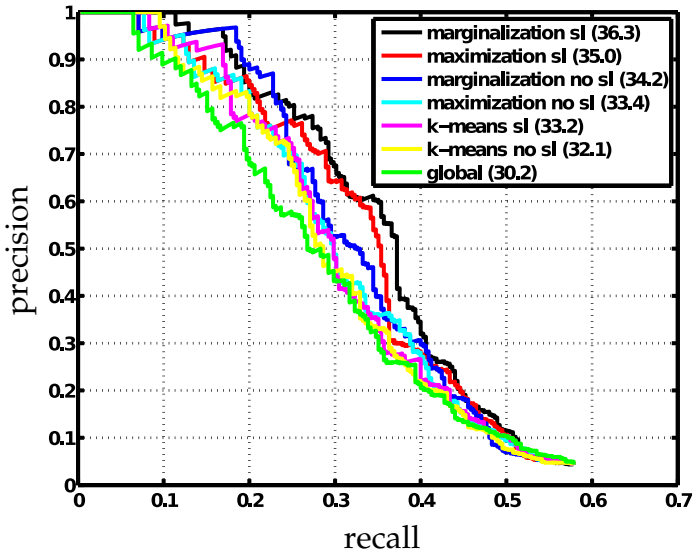


Figure 5.4: Evaluation of different instantiations of our model on a subset of PASCAL VOC 2007 motorbikes.

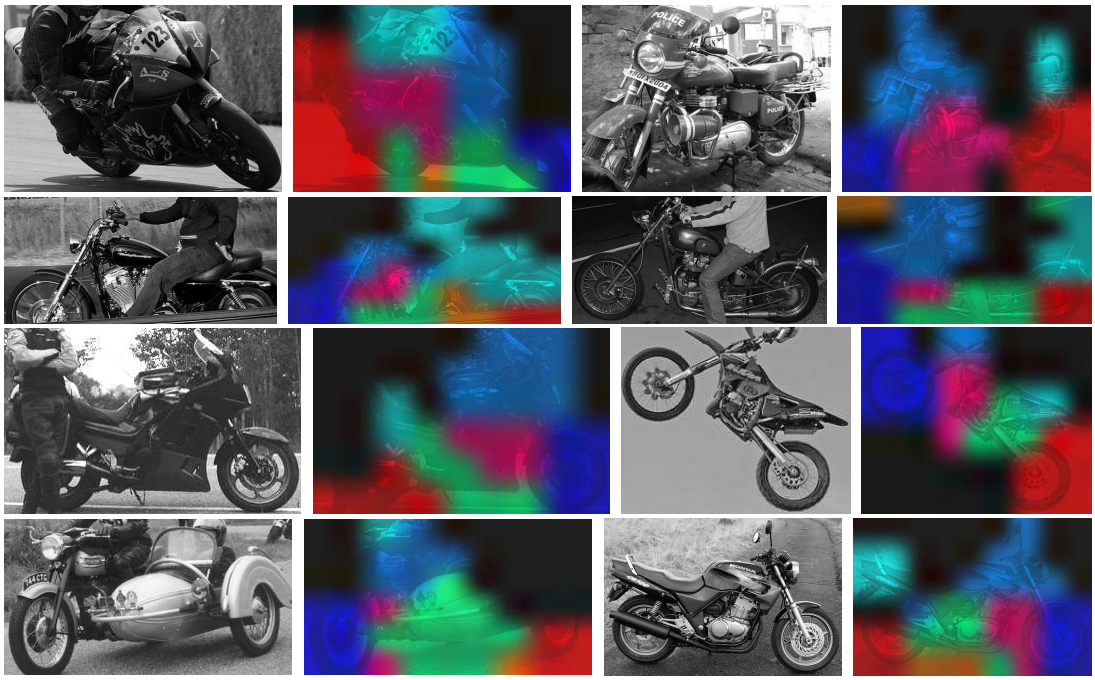


Figure 5.5: Motorbike segmentation examples (see text for details).

global mean image over all bounding boxes, where it is challenging to see any real object class structure. The middle column shows mean part occurrences of a fixed k-means part assignment (cf. (Schnitzspan *et al.*, 2008)) weighted by their probability and shifted to their canonical location. Even though one can recognize a trend towards the discovery of object parts, the object structure is still rather weak. The right column shows part occurrences of our latent CRF model, where the parts are weighted by their probability from the part CRF and shifted to their canonical location (the spatial mean of the classifier weights $\gamma_i(c)$). It becomes clearly visible that our model automatically discovers object parts, such as wheels of motorbikes or the head of horses, allowing for a much better alignment of instances and parts.

Figs. 5.1, 5.5, and 5.6 show object segmentations, where the color-coded part labelings are automatically inferred by our model. The color saturation encodes the probability of each part. As can be seen in Fig. 5.5 (motorbikes) our model is capable of handling viewpoint variation (row one) as well as partial occlusion (row two left and row three left). These segmentations illustrate one major benefit of our framework: Our model implicitly handles partial occlusions by considering all possible configurations simultaneously and weighing them according to their probability. This avoids relying on the most probable and possibly misleading part labeling. For articulated object classes like horses (Fig. 5.6) we can observe the same. Our model captures articulation (row one left and row three left), viewpoint variation (row four left) as well as partial occlusion (row two right, row four left). Note how our model adapts to a meaningful representation of parts even for articulated object classes like horses.

Quantitative evaluation on VOC 2007 motorbikes. In Tab. 5.1 and Fig. 5.4 we compare different components and settings of our model on a subset (left and right facing) of the images of the motorbike class of the PASCAL VOC 2007 challenge. We show the performance of (i) using only global part descriptors, (ii) using a fixed k-means assignment of parts, (iii) using the most probable part (MAP) per node instead of marginalization, and (iv) using marginalization. All part-based settings are evaluated with a fixed and a learned graph structure. The fixed structure accounts for local neighborhood dependencies that connect each node to its four neighbors in a regular grid, as in standard CRFs.

As can be seen, our full model outperforms the global template model by 6.1% AP, which emphasizes the importance of enriching global models with a semantically meaningful notion of parts. Moreover, treating part labels as hidden nodes is clearly advantageous to fixing them based on k-means clustering (AP increase of 3.1%). This holds true for the case of fixed graph structure as well (gain of 2.1% AP), which shows that the higher expressiveness of latent models results in superior performance. This quantitative evaluation is consistent with our qualitative observations, where parts inferred by our model showed a much better alignment of instances than the k-means instantiation.

In order to show the benefit of marginalizing out all possible part configurations, we compare the marginalization scheme against considering only the maximum part assignment (MAP) for each node. Marginalization shows a gain of 1.3% AP over maximization, which emphasizes the benefit of considering all possible part configurations instead of relying only on possibly misleading maximal responses. This observation agrees with the work of Kapoor and Winn (2006), which assumed the hypothesis of max-product, that the posterior mass is concentrated at the mode, being inaccurate due to the uncertainty in the latent part variables Note that structure learning always led to better results than a fixed graph structure, which demonstrates the increased flexibility of the learned structure.

Quantitative evaluation on all VOC 2007 classes. In order to further evaluate the contribution of our work, we show results of different instantiations of our model (using only global parts, using our full model but only HOG features with and without pre-filter, and our full model with HOG and BoW descriptors). Tab. 5.5 compares those with state-of-the-art approaches. As can be seen our model achieves competitive performance (28.7% AP on average).

Using only HOG features allows a fair comparison to (Felzenszwalb *et al.*, 2008), who use similar features. On average over all classes, our flexible part-based approach shows an improvement over (Felzenszwalb *et al.*, 2008) of 1.0% AP. We achieve better results on 16 of 20 classes, which emphasizes the benefits of the flexible object topology and marginalizing over all part constellations. Note, that our model without applying the pre-filter (27.1% AP on average across classes) is on par or slightly better than inferring our model on pre-filtered hypotheses (26.9% AP on average across classes).

We achieve an improvement of 1.5% AP and better results on 17 of 20 classes

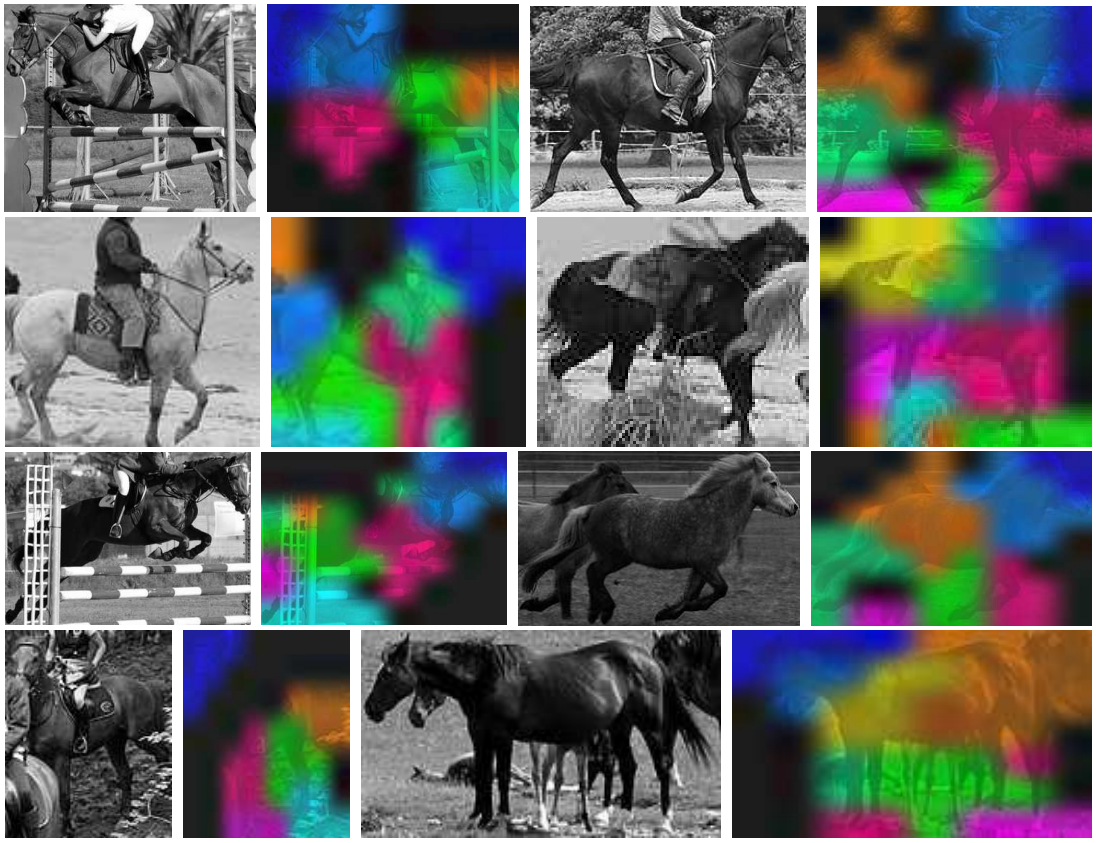


Figure 5.6: Horse segmentation examples (see text for details).

compared to (Desai *et al.*, 2009), who considered detections of all object classes within an image by simultaneously inferring a notion of multi-class layout and context. Such additional information is orthogonal to our model and is likely to improve the performance further.

Compared to our binary structure learning approach as discussed in the previous chapter and in (Schnitzspan *et al.*, 2009), we could improve the performance by 1.2% AP, which shows the advantage of integrating part labels in a structure learning framework.

We achieve better performance than (Vedaldi *et al.*, 2009) on 5 object categories, even though the latter approach gives better performance on average. It is likely that a large part of this increased performance is due to integrating more complementary feature descriptors, which could also be done in our model as sketched in Sec. 5.4. Since our model remains general and allows for integration of more features, we expect a substantial performance gain by doing so.

Compared to the original VOC 2007 challenge we achieve a performance gain of 5.4% AP and show better results on 17 of 20 object classes. Here we compare against the best method per class and not against a single model.

VOC 2007	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
Our model	31.9	57.0	9.1	15.2	26.0	42.7	49.3	14.5	15.2	18.5
Our model (HOG only)	29.1	56.4	4.6	13.0	25.2	40.7	47.3	13.5	10.1	18.8
Our model (HOG only) no pre-filter	28.2	54.8	8.9	10.8	27.2	42.5	45.9	12.2	8.0	20.1
Our model (global)	20.1	45.5	1.8	9.0	19.3	34.6	36.6	11.5	7.3	13.0
DPM (VOC07) (Felzenszwalb *et al.*, 2008)	20.6	36.9	9.3	9.4	21.4	23.2	34.6	9.8	12.8	14.0
DPM (Felzenszwalb *et al.*, 2008)	28.1	55.4	1.4	14.5	25.4	38.9	46.6	14.3	9.4	16.0
Multi-class layout (Desai *et al.*, 2009)	28.8	56.2	3.2	14.2	29.4	38.7	48.7	12.4	16.0	17.7
Bin. struct learning (Schnitzspan *et al.*, 2009)	31.7	56.3	1.7	15.1	27.6	41.3	48.0	15.2	9.5	18.3
MKL multi-feature (Vedaldi *et al.*, 2009)	37.6	47.8	15.3	15.3	21.9	50.7	50.6	30.0	17.3	33.0
Best VOC07 (Everingham *et al.*, 2007)	26.2	40.9	9.8	9.4	21.4	39.3	43.2	24.0	12.8	14.0

	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	average
Our model	24.2	11.8	49.1	41.9	35.7	14.5	18.9	23.3	34.3	41.3	28.7
Our model (HOG only)	23.1	10.9	48.0	38.4	34.7	14.3	17.1	21.0	32.7	38.8	26.9
Our model (HOG only) no pre-filter	22.5	12.5	49.2	40.1	32.9	15.5	18.0	22.9	31.5	37.5	27.1
Our model (global)	10.4	5.8	35.0	32.0	20.1	12.1	13.5	11.3	24.1	29.7	19.6
DPM (VOC07) (Felzenszwalb *et al.*, 2008)	0.2	2.3	18.2	27.6	21.3	12.0	14.3	12.7	13.4	28.9	17.1
DPM (Felzenszwalb *et al.*, 2008)	22.8	10.6	44.1	37.0	35.2	13.6	16.1	18.5	31.8	36.9	25.9
Multi-class layout (Desai *et al.*, 2009)	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	27.2
Bin. struct learning (Schnitzspan *et al.*, 2009)	26.1	11.3	48.5	38.9	35.8	14.8	17.7	18.8	34.1	39.8	27.5
MKL multi-feature (Vedaldi *et al.*, 2009)	22.5	21.5	51.2	45.5	23.3	12.4	23.9	28.5	45.3	48.5	32.1
Best VOC07 (Everingham *et al.*, 2007)	9.8	16.2	33.5	37.5	22.1	12.0	17.5	14.7	33.4	28.9	23.3

Table 5.2: Results on the PASCAL VOC 2007 object detection challenge.

5.6 CONCLUSIONS

This chapter presented a novel discriminative framework that successfully combines powerful discriminative learning techniques with the flexibility and expressiveness of part-based models and discriminative pairwise structure learning. We relied on weakly supervised training by treating part labels as hidden nodes, and letting our approach automatically discover semantically meaningful part representations. Our model lends itself to modeling the spatial layout of objects even in the presence of heavy articulation and viewpoint variation, and provides an implicit occlusion reasoning. Quantitatively our scheme achieves competitive performance on the difficult PASCAL VOC 2007 challenge, and qualitatively yields object segmentations with meaningful part labelings that reoccur across object instances.

SENSOR FUSION FOR MOBILE ROBOTS 6

Contents

While in the previous chapters we described object detection with graphical models, we want to focus on onboard and real-time object detection for mobile robotic platforms in this chapter. This chapter mainly addresses interdisciplinary work within the research training group 1362 funded by the Deutsche Forschungsgemeinschaft (DFG). The goal of this chapter is to combine several complementary sensors for an increased reliability while at the same time respecting real-time requirements of search and rescue robotics. This chapter describes the work published in (Meyer *et al.*, 2010).

6.1 INTRODUCTION

Modeling the world in complex environments is a crucial aspect on the way toward reliable, intelligent, and autonomous search and rescue robots. It is desirable not only to infer a geometrically interpretable map, but also to integrate semantic attributes to enable high-level scene interpretation as motivated in related work (Asada and Shirai, 1989; Burgard and Hebert, 2008; Kumar *et al.*, 2004). In urban search and rescue (USAR), reliable robots have to provide a semantically meaningful interpretation of objects within a scene (e.g. victims in collapsed buildings) (Tadokoro *et al.*, 2000). In unconstrained environments (as is the case in USAR scenarios) relying only on one type of sensor is often insufficient, while fusing complementary information (i.e. information from different types of sensors) enables semantic interpretability of scenes and superior reliability.

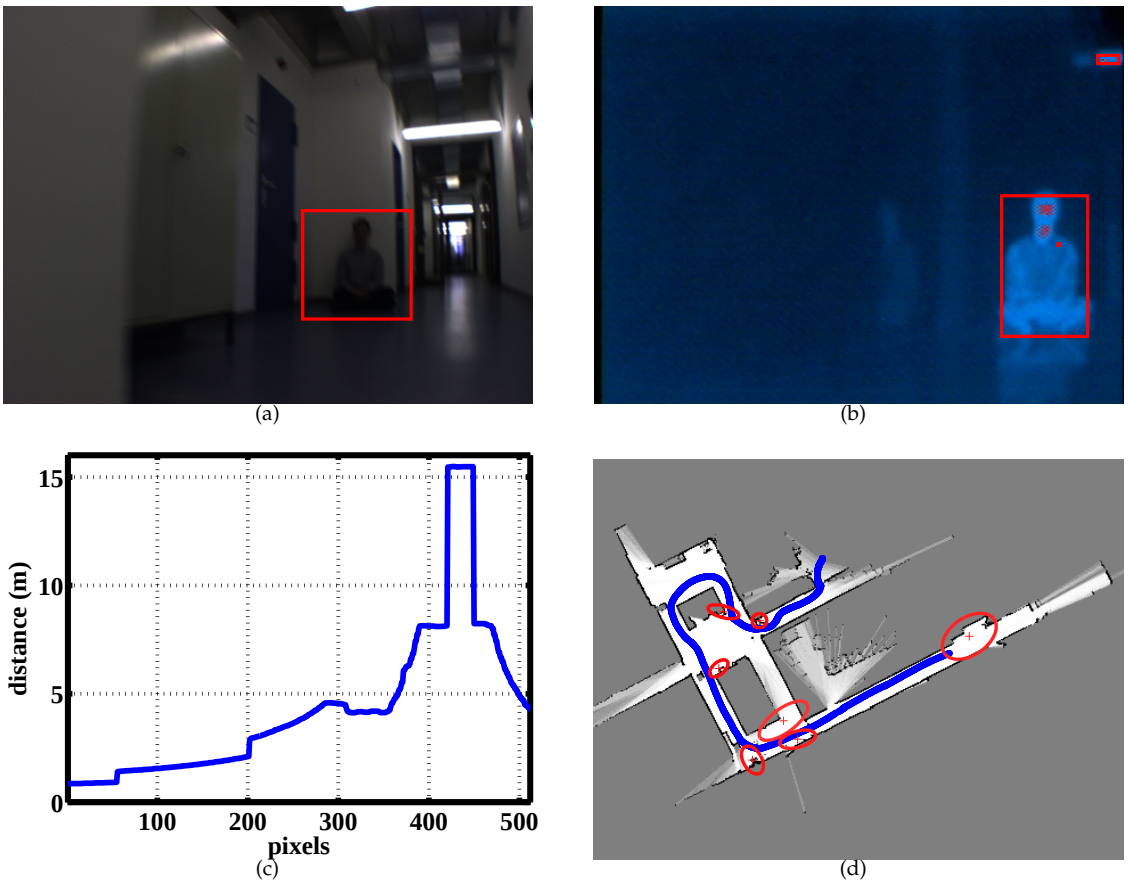


Figure 6.1: Examples of sensor and world model data: (a) Visual image with victim detection. (b) Thermal image with heat detections. (c) Range samples along the horizontal axis of the image. (d) Snapshot of the semantic world model with estimated victim locations denoted by the red covariance ellipses.

The main goal of this chapter is to propose a mobile robot system for autonomous detection of victims in USAR scenarios. The system is capable of autonomous navigation and map learning, and localizes victims and objects of interest in a 3D world coordinate system. Our setting approaches actual search and rescue operations in realism and complexity: Real human victims have to be localized in unstructured environments, even in the presence of background clutter and multiple thermal sources such as office equipment and heating. By adding objects of interest to the metric map of the environment, we augment the robot's environment with semantic information, which then can be utilized for decision making by human operators. Our system is able to achieve high performance even for cluttered and complex datasets (see Fig. 6.1). All information is processed onboard and in real-time, as is crucial for realistic rescue deployments.

In our system visual information is supplied to a generic object detector that allows detecting structured objects and assigns them a semantic meaning (e.g. upper bodies of victims or hazardous material signs). To avoid relying on a single source of

information, we consider the information from all sensors simultaneously, and derive a generic model that is able to leverage complementary information. As motivated in (Kleiner and Kümmerle, 2007), merging different sources of information helps achieving higher levels of performance in victim detection.

Sec. 6.2 describes our system, while in Sec. 6.3 the sensors and detection algorithms for victim and object detection are introduced. Afterwards two approaches for sensor fusion are presented in Sec. 6.4. Experimental validation is presented in Sec. 6.5.

6.2 SYSTEM OVERVIEW

Since the locations of victims need to be specified in world coordinates, the robot pose and a metric map have to be estimated using a combination of inertial sensing and simultaneous localization and mapping (SLAM). These estimates are continuously updated over time and used for integrating victim hypotheses obtained from different sensor types. Simultaneously with the pose and map estimation at each update step, our system generates a set of object hypotheses using a visual object detector and thermal-camera-based detector, which are used as a basis for sensor fusion.

We explore two complementary approaches for sensor fusion. In the first, which we denote as explicit sensor fusion, we integrate information from different sensors directly in the sensor space using known transformations between different sensor modalities (i.e. the mapping between thermal and visual images). This allows to use known dependencies of sensor signals to either amplify or attenuate the confidence in the measurements.

In the second approach, which we denote as implicit sensor fusion, the global belief is updated independently for each observation. The advantage of using complementary sensors is realized through accumulation of positive evidence in the world model. Integrating hypotheses into the model requires an association step for matching an hypothesis to an already known object. The case in which previously unknown objects have been found must be considered separately. Once association is established, the matching is taken for granted and the corresponding victim location estimate and evidence is updated by using an extended Kalman filter (EKF). Integration of observations into the global belief state can therefore be considered to be a method for temporal sensor fusion. Additionally, confidence in a hypothesis is influenced by negative observations, where the absence of expected detections or contradictory measurements reduce the confidence value.

When applying implicit sensor fusion the observations from different sensors are integrated independently into the world map, while explicit sensor fusion is an optional step that precedes the integration step, and is primarily used to increase the reliability of observations.

6.2.1 World model

World models generally account for a mathematical description of the environment, with different aspects being considered important depending on the application. In our USAR scenario the model is formed by a representation of building geometry and additional semantic information, like the location of people and objects of interest. By applying additional high-level knowledge to the model, it can be easily enriched with more detailed information in future work, for example classification of places, a graph of passable paths through a building, or estimates of hazardousness of specific locations. Based on this high-level description of the environment, the robot is able to plan reasonable future actions and – when integrating human operators – is able to deliver valuable information to rescue teams, for example to guide them to detected victims.

The robot state vector $\mathbf{y}_k$ contains the estimated 6DOF robot pose as well as translational and angular velocities in the global coordinate system and is updated at discrete timesteps $t = t_k$. The location of objects, including the victims, is referred to as $\mathbf{x}_k^j$ with j being an index variable over the estimates. The objects are modeled as points, ignoring their spatial extent. In this chapter we assume the world to be static apart from the movement of the robot itself. The number of objects is not known in advance. Besides the location information we introduce the probability π_k^j that object j is detected correctly as a measure of confidence, which is incrementally updated with each new sensor reading and typically increases when more detections of the same object occur.

The process of world modeling requires inference in state space from measurements given in sensor-space. Since sensors are error-prone, a probabilistic model description is used here. We choose a Gaussian representation for the continuous state variables, with estimated means $\hat{\mathbf{y}}_k$ and $\hat{\mathbf{x}}_k^j$, and variances C_k and P_k^j, respectively.

6.2.2 Simultaneous localization and mapping (SLAM)

State estimation of the vehicle and a map is performed by two components. A 2D pose and map estimate is provided by a module using incremental maximum likelihood alignment of laser scans with the estimated map. The map is represented by a discrete grid and updated using the log-odd probabilities of occupancy (Schiele and Crowley, 1994; Thrun *et al.*, 2005).

Estimation of the robot state $\mathbf{y}_k$ is performed by an extended Kalman filter (EKF) integrating observations from all available sensors. Attitude estimation is provided by a built-in IMU and compass, while position estimation is provided by wheel encoders and the 2D pose estimation updates from the SLAM module. For the USAR scenarios described in this work, our approach is sufficiently accurate as to not require multiple map hypotheses (e.g. using a Rao-Blackwellized particle filter), or explicit loop closure.

6.3 VICTIM AND OBJECT DETECTION

In order to enrich the map with semantic information, we perform onboard detection of objects of interest, which in our case correspond to people and dangerous materials marked with hazmat signs. In this chapter we focus on the detection of upper bodies of people, since this allows to detect both standing people as well as possibly injured people sitting on the ground (see Fig. 6.1), and leave more complex cases for future work. Due to background clutter, partial occlusions and complex articulations, visual people detection is a difficult problem even in this somewhat restricted setting. In particular, state-of-the-art computer vision methods are still severely challenged by this task (Ferrari *et al.*, 2009).

Object detection. In order to find initial hypotheses of people and hazmat signs in camera images, we use the popular sliding window approach. In this approach every image is exhaustively scanned over a range of positions and scales; for each position and scale a discriminative SVM classifier is used to make binary decisions about the presence or absence of an object. While seemingly expensive, the sliding window approach is especially suitable for parallel implementation, since each object location can be examined independently of the rest of the image.

In order to describe the contents of the image at each particular location, we leverage recent results in computer vision and rely on a histogram of oriented gradients (HOG) descriptor (Dalal and Triggs, 2005). In our system we scan the image with steps of 8 pixels and relative scale factors of 1.05. We use a GPU implementation developed in our group, which allows to achieve real-time performance without sacrificing recognition performance (Wojek *et al.*, 2008).

The confidence $s_k^{\text{vis}} \in [0, 1]$ of a hypothesis is calculated via a sigmoidal mapping:

$$s_k^{\text{vis}} = \frac{1}{1 + \exp(a \cdot f_k + b)} \ , \tag{6.1}$$

where f_k is the SVM score of the hypothesis, and a and b are parameters that are estimated by cross-validation (Platt, 1999).

Object classification. A HOG descriptor is especially well suited for capturing the characteristic shape of an object. However, it has shortcomings when it is necessary to distinguish between objects with similar shape, such as different hazmat signs, all of which have a rhombus shape and differ mainly in color, internal patterns and text. In order to identify hazmat signs we augment the HOG descriptors with color histograms. For each hazmat sign hypothesis of the HOG-based detector, we compute a color histogram in LAB color space and use this to perform the final classification of hazmat signs by applying a k-nearest-neighbor approach in combination with the χ^2-distance. As our experiments demonstrate, the combination of HOG and color histograms yields good performance for hazmat sign classification (Sec. 2).

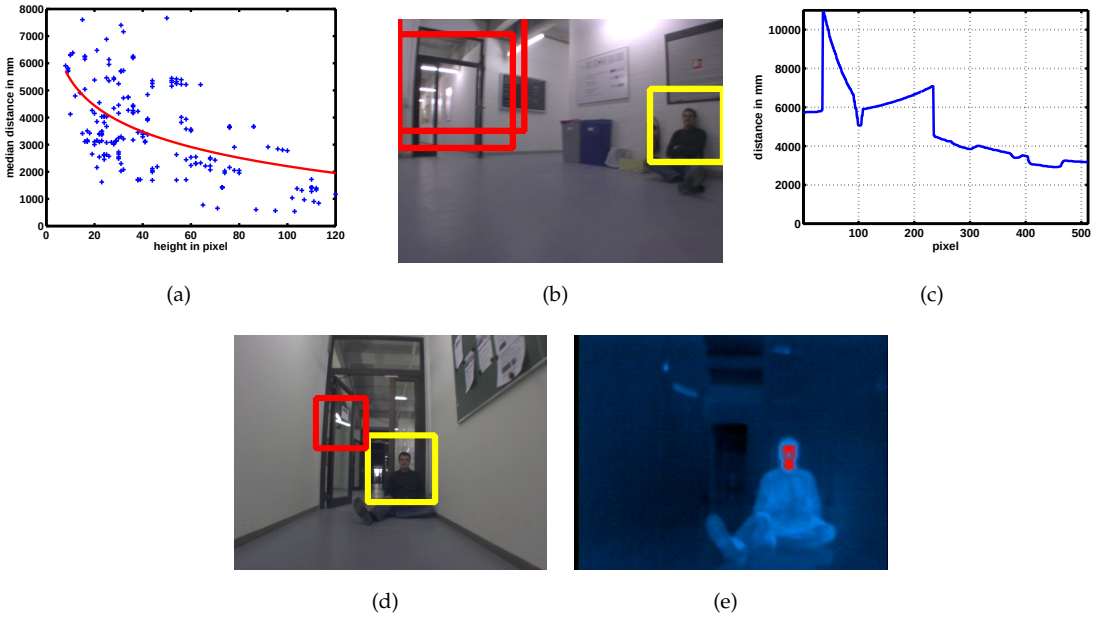


Figure 6.2: (a) Correspondence between annotation height and distance. (b) Example detections of frame 561. (c) Scanline of frame 561. (d) Example detections of frame 287. (e) Thermal image of frame 287.

Thermal victim detection. In addition to victim hypotheses from visible light camera images, our system also creates a set of hypotheses based on images from a thermal camera. These thermal hypotheses are generated with a simple procedure that searches the images for large enough groups of connected pixels with temperature values within the human body temperature range. Each such group of pixels is used to generate a hypothesis. Although hypotheses generated by the thermal camera alone are significantly less reliable compared to hypotheses produced by the visual object detector, we found them to be effective in reducing the number of false positives.

We define a simple model for the confidence $s_k^{\text{therm}} \in [0,1]$ by counting the number of pixels within the person's bounding box that have a temperature close to human bodies. This model is robust to small offsets in corresponding locations in visual and thermal images, which arise due to imprecise synchronization between these modalities.

6.4 SENSOR FUSION

The reliability of the entire victim detection framework can be increased by fusing victim hypotheses from different sensors and across time steps. Intuitively, the confidence of a detected victim should be increased if it is observed in several update steps or by different sensors and on the other hand stay below a certain

threshold when it is only spotted once. We employ an extended Kalman filter (EKF) in order to update the locations $\mathbf{x}_k^j$ of victims in our world model and integrate several hypotheses across update steps. In parallel, the confidence π_k^j that the victim is present at the respective location is updated in a separate filter with the respective measurement confidence as described below.

For simplification of notation we assume without loss of generality that at most one hypothesis is observed in every update step k. We define the measurement $\mathbf{z}_k$ to consist of the distance d_k, the bearing angle α_k, and the relative vertical angle β_k between the hypotheses and the robot:

$$\mathbf{z}_k = \begin{bmatrix} d_k, & \alpha_k, & \beta_k \end{bmatrix}^T = h(\mathbf{x}_k^j, \mathbf{y}_k) + \mathbf{v}_k \ , \tag{6.2}$$

where $h(\cdot,\cdot)$ refers to a non-linear measurement function that projects the victim's position into the world model. $\mathbf{x}_k^j$ and $\mathbf{y}_k$ denote the victim's position estimate and robot state vector respectively. The random vector $\mathbf{v}_k$ is unbiased and uncorrelated Gaussian measurement noise with hand-tuned variance R. The measurement function also depends on the robot's state $\mathbf{y}_k$, which in turn is estimated with an EKF independently.

Data association. In order to find an optimal matching between measurements and existing estimates of victim locations we use the following probability of measurement $\mathbf{z}_k$ given the index j and the position estimate $\hat{\mathbf{x}}_{k-1}^j$ with variance P_{k-1}^j:

$$p(\mathbf{z}_k|j) \propto \mathcal{N}(\mathbf{z}_k; h(\hat{\mathbf{x}}_{k-1}^j, \hat{\mathbf{y}}_k), R + H_k^j P_{k-1}^j (H_k^j)^T) \ , \tag{6.3}$$

with a first order approximation H_k^j of the measurement function $h(\hat{\mathbf{x}}_{k-1}^j, \hat{\mathbf{y}}_k)$ at the current estimated means.
Whenever this probability $p(\mathbf{z}_k|j)$ is above a previously defined threshold, we associate the new measurement to the best matching estimate with index $j_k^* = \arg\max_j p(\mathbf{z}_k|j)$. Otherwise, a new estimate is added to the world model as of a previously unobserved victim.

Kalman filter updates. We assume the victims to be static in our setting and therefore no explicit prediction step is needed. The measurement update equations of the Kalman filter are defined as:

$$K_k^j = P_{k-1}^j (H_k^j)^T \left(H_k^j P_{k-1}^j (H_k^j)^T + R \right)^{-1} \tag{6.4}$$

$$\hat{\mathbf{x}}_k^j = \hat{\mathbf{x}}_{k-1}^j + \lambda_k(\mathbf{z}_k, s_k) \cdot K_k^j \left(\mathbf{z}_k - h(\hat{\mathbf{x}}_{k-1}^j, \hat{\mathbf{y}}_k) \right) \tag{6.5}$$

$$P_k^j = \left(I - \lambda_k(\mathbf{z}_k, s_k) \cdot K_k^j H_k^j \right) P_{k-1}^j \ , \tag{6.6}$$

where I denotes the identity matrix and s_k refers to the initial score of hypothesis k as will be explained below. The measurement update uses the confidence $\lambda_k(\cdot,\cdot) \in [0,1]$ of an observation as an additional factor to the gain matrix K_k^j to honor the

observation quality and discard uncertain updates. The measurement confidence is of the form

$$\lambda_k(\mathbf{z}_k, s_k) = s_k \cdot \phi(\mathbf{z}_k, d^{\text{laser}}) \cdot \psi(\mathbf{z}_k) \ , \tag{6.7}$$

where d^{laser} is the distance to the next obstacle measured with the laser scanner. $\phi(\cdot,\cdot)$ imposes a prior on the estimated distance to $\mathbf{z}_k$ and measured distance d^{laser} to the next obstacle and is proportional to a Gaussian with manually defined variance according to the uncertainty in the sensor measurements. $\psi(\cdot)$ refers to a Gaussian height prior with mean 80cm (height of upper bodies) and manually defined variance. By employing $\phi(\cdot,\cdot)$, we ensure that the size of an hypothesis approximately matches the size that we expect, and avoid false positives with inappropriate estimated and measured distances (see Fig. 6.2(b)). $\psi(\cdot)$ guarantees that all objects appear at the expected height from the robot, while unlikely pitch angles are discarded.

Label confidence update. In the case that a new measurement is associated to a given estimate, we update the estimate's label confidence according to the disjunctive combination of two binary random events, so that confidence is increased with every new measurement:

$$\pi_k^j = \pi_{k-1}^j + \lambda_k \cdot (1 - \pi_{k-1}^j) \ . \tag{6.8}$$

If no measurement in time step k is available we decrease the label confidence of all victim estimates within the field of view by employing "negative evidence". Negative evidence is information that arises from the fact that the confidence of an estimate can decrease if it is not confirmed by sensor observations. Applying negative evidence to our algorithm helps to decrease the number of false alarms, as many false positives do not reoccur in consecutive time steps.

The negative update is applied to all objects j that should be visible in the image according to the current estimated map and positions, but have no detection event associated for the current time step. Their label confidence is reduced according to

$$\pi_k^j = \frac{p_{\text{miss}} \cdot \pi_{k-1}^j}{p_{\text{miss}} \cdot \pi_{k-1}^j + (1 - \pi_{k-1}^j)} \ . \tag{6.9}$$

The probability p_{miss} of missed detections is approximated as the inverse probability of the detector's recall on the trained dataset.

Implicit vs. explicit integration We evaluate two different fusion schemes: implicit and explicit fusion. These two approaches differ in the way the complementary information of sensors is integrated. In our model this boils down to the treatment of the initial score s_k of an hypothesis in Eq. (6.7).

In the implicit fusion scheme we consider each hypothesis from both the visual light and thermal sensor as a new measurement that is either associated to a given estimate or enters the world model as a new estimate. In this setting the sensor fusion is implicitly handled with the Kalman filter, since the measurements of both sensors can be used for data association and updating the confidence. Here we

directly use the visual or thermal score as initial score s_k:

$$s_k = \begin{cases} s_k^{\text{vis}}, & \text{if visual light hypothesis} \\ s_k^{\text{therm}}, & \text{if thermal hypothesis} \end{cases} . \tag{6.10}$$

In the explicit integration scheme we compute the overall detection confidence as a weighted sum from the individual scores of complementary sensors, yielding a single observation model where bearing and distance information is taken from the visible light bounding box only:

$$s_k = \gamma_1 \cdot s_k^{\text{vis}} + \gamma_2 \cdot s_k^{\text{therm}} + \gamma_3 \cdot s_k^{\text{laser}} , \tag{6.11}$$

where $\sum_i \gamma_i = 1$ and the coefficients γ_i are trained with cross-validation. This additive formulation makes the model robust to sensor failures (e.g. due to partial occlusions) and takes relative importance of different sensors into account. The first two components of the mixture correspond to the probability of correct detection given the score of the SVM classifier and the output probability of the thermal victim detector for the same bounding box as defined in Sec. 6.3.

While more detailed integration of different sensor modalities is possible and we plan to explore it in the future, we opt for this re-estimation approach since it allows to decouple training of the visual object detector from the rest of the system, does not require exact synchronization between different sensor streams, and allows to use simple algorithms for integration of thermal and laser sensors.

In order to model s_k^{laser}, we fit a log-linear model to a set of jointly observed bounding boxes and laser-range measurements as shown in Fig. 6.2(a). s_k^{laser} is set to a Gaussian computed at the difference between the predicted distance from the log-linear model and the median distance measured with the laser-range finder. The variance is set by hand.

6.5 EXPERIMENTS

We evaluate the performance of our system on the tasks of people and hazmat sign detection. In particular we quantify performance gains due to fusion of multiple sensor modalities and evaluate both detection in single frames and performance of the full system. For evaluation we use the dataset which consists of daylight images, thermal images, and laser-range scanner and odometry measurements collected while the robot was driving along the closed path of approximately 120 meters within the office building. For the sake of single-frame evaluation, we have annotated all people appearing in the daylight images, which are larger then 40 pixels in height. The resulting dataset contains 1480 daylight images with 300 annotated victims corresponding to 10 distinct subjects. Due to difficult illumination conditions, motion blur and large variability in viewpoints, visual people detection in such data is very challenging. At the same time detection of people in thermal images is complicated by the presence of multiple background heat sources, such as

heating and illumination equipment, computers and other office devices. In order to evaluate the robustness of our system to partial occlusions, we have collected an additional dataset of 28 images with 115 annotated people, 69 of which are partially occluded. We denote these datasets as "Hector Data 1" and "Hector Data 2" respectively [2]. In order to demonstrate the generality of our method we do not adapt the visual people detector to our scenario, although this would likely lead to improved performance, and train our visual detector on the INRIA pedestrian dataset (Dalal and Triggs, 2005), where we have re-annotated the upper bodies of people. For the detection experiments we report the average precision (AP), which measures the area under the precision-recall curve. This is a common comparison measure, for which a perfect detector would achieve 100% AP. In the following we first present the results of single-frame detection of people and hazmat signs, and then evaluate performance of the full system. For the single frame detection the confidence of an object hypothesis is computed according to the Eq. 6.11, while for the full system the confidence is based on the measurements over multiple frames.

Single frame people detection. Fig. 6.3(a) and Fig. 6.3(b) show the results of single-frame people detection of our system on the "Hector Data 1" and "Hector Data 2" datasets in form of recall/precision curves.

On the "Hector Data 1" dataset, the detector based on visual information achieves 36.3% AP, integration of visual and laser-range measurements results in 42.9% AP, and integration of visual and thermal measurements results in 43.0% AP. Integration of all three sensors leads to the best performance of 45.0% AP. The missing detections on this dataset mainly correspond to either very small or blurry instances.

Similar trends can be observed on the "Hector Data 2", where images contain less motion blur, but a significant number of people is partially occluded. On this dataset we obtain 36.5% AP using the visual detector alone, which improves to 39.1% AP by integration visual and thermal detectors. When evaluating only on the partially occluded people we obtain 27.2% AP with the visual detector, and 31.1% AP with the combination of visual and thermal detectors. These results show that, despite some drop in performance, our system is still producing meaningful detection results even in the case when people are partially occluded. The integration of thermal sensor measurements results in consistent improvement of performance of around 4% AP.

Hazmat sign detection and classification. Fig. 6.3(e) shows a precision-recall curve quantifying single-frame hazmat sign detection performance. On this type of objects we obtain 60.1% AP. Due to smaller intra-class variability the results for hazmat signs are somewhat better than results for people detection. The missing detections are often due to motion blur and hazmat signs at small scales.

We further investigate the performance of our system on the hazmat sign classification task, in which the goal is to distinguish between one of the nine hazmat sign classes depicted in Fig. 6.3(f). For that purpose we take the detection windows at maximum recall and assign them to one of the given classes or background.

[2] Both datasets are available at http://www.gkmm.tu-darmstadt.de/rescue

For the classification we follow the procedure based on color histograms, described in Sec. 6.3. We evaluate two approaches to histogram computation, one in which a color histogram is calculated on the entire detection window, and another in which the detection window is subdivided into four subregions and a separate histogram is computed for each of them. In the latter approach the final descriptor is formed by concatenating histograms of each subregion. We obtain the recognition rate of 37.5% using histograms based on the entire window, and 58.3% using subregion-based histograms. The improvement is mainly due to a better discrimination between hazmat classes with globally similar color distribution, for example white/red hazmat signs "Combustible" and "Flamable Solid" shown on Fig. 6.3(f). Region-based histograms provide better representation of the image in such difficult cases, since they are also capable of capturing the spatial distribution of colors within the detection window.

Full system performance. Finally, we evaluate the capability of our full system to correctly detect and localize people in the environment map. The predicted location and detection confidence of each person hypothesis is infered by temporal integration of sensor measurements according to the filtering procedure described in Sec. 6.4. In contrast to single frame evaluation, the detection performance is reported for the whole series of measurements contained in the dataset. The victim is considered to be localized correctly if its predicted location on the map is within 1 meter radius of the ground truth annotation, obtained my manual labeling. Multiple hits on the same ground truth annotation are only counted once, where each subsequent hit is considered a false positive.

As can be seen in Figs. 6.3(d) and 6.3(c) merging complementary information of heterogeneous devices (vis+laser+therm) outperforms all other settings by a large margin. It achieves 78.7%AP (explicit sensor fusion) and 87.2% AP (implicit sensor fusion) outperforming vis+laser by 2.9% AP and 18% AP respectively. When not using the laser, our framework suffers from placing the victims too far from ground truth annotations. vis+therm achieves 29.6% AP for explicit fusion and 49.7% AP for implicit fusion. The baseline of using only visual information achieves 29.2% AP. The implicit integration scheme achieves a higher precision for vis+thermal+laser and vis+thermal than explicit integration while the latter fusing scheme yields higher levels of recall. Note that in contrast to single-frame evaluation, where recall levels are below 60%, the complete system has recall of 90% for implicit and 100% for explicit sensor fusion schemes. This is an important result for search and rescue applications, in which the ultimate goal is to find all of the victims.

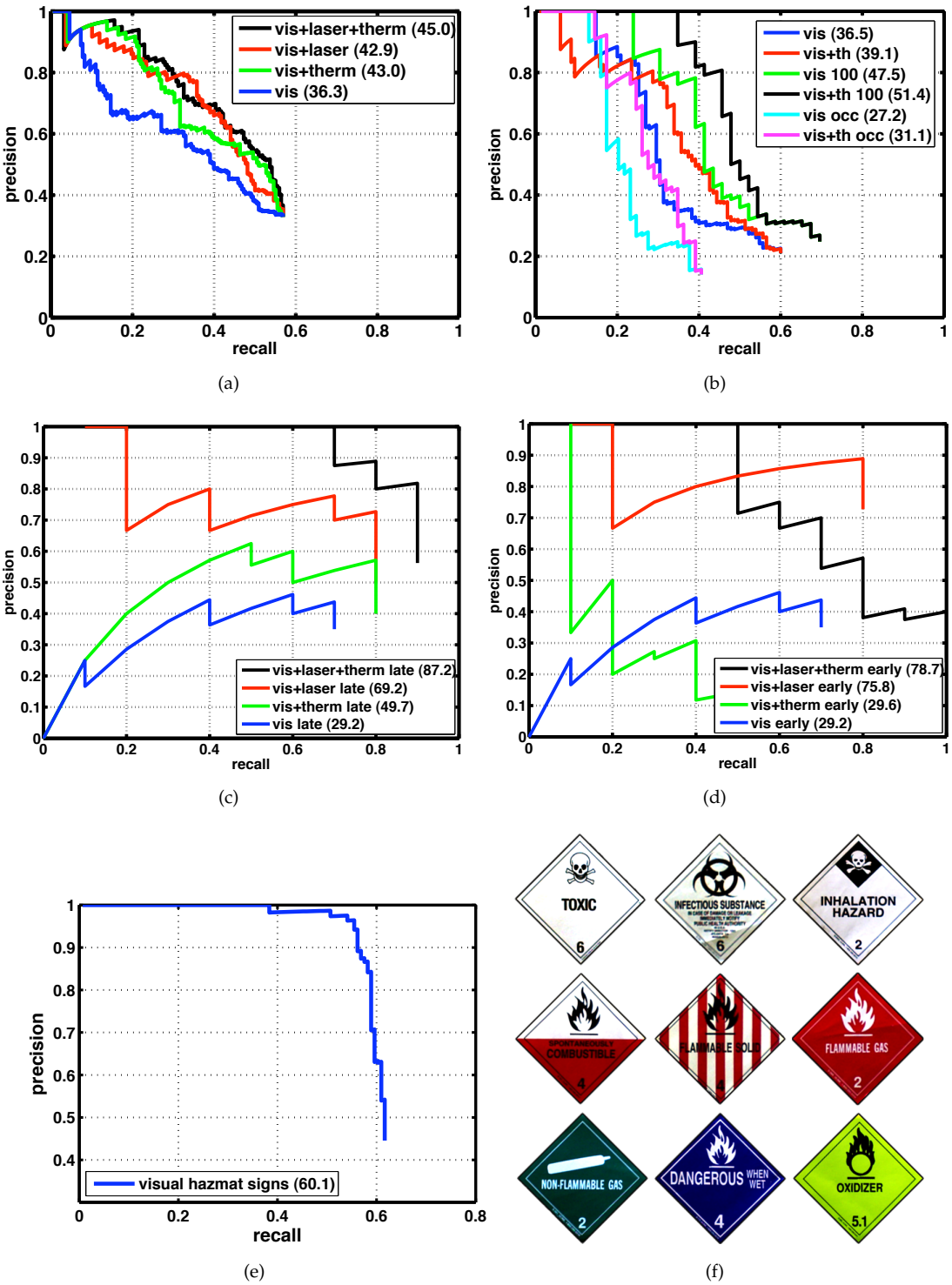


Figure 6.3: (a, b) Single-frame people detection performance for different combination of sensors on "Hector Data 1" and "Hector Data 2" datasets. (c) Single-frame hazmat sign detection performance on "Hector Data 1" dataset. (d, e) People detection performance of the full system on the "Hector Data 1" dataset for explicit and implicit sensor fusion schemes. (f) Collection of different hazmat signs.

6.6 CONCLUSION

This chapter addresses sensor fusion of heterogeneous sensors with a generic semantic world model. Our framework is able to leverage complementary information for increased reliability in complex USAR scenarios. Geometric maps are enriched with semantic interpretation of scenes by detecting victims and possibly hazardous areas. The importance of sensor fusion and the expressiveness of our model are experimentally evaluated on a complex real world dataset. In future work we will address distributed sensor fusion by using multiple robots.

7 CONCLUSION AND OUTLOOK

Contents

Graphical models are playing an increasingly important role in the design and analysis of machine learning and computer vision algorithms. Key to the success are two basic ideas that are merged coherently in the underlying theory: On one hand, the probabilistic nature of graphical models allows to treat the uncertainty in many computer vision tasks and let us interface the model with the provided data. On the other hand, the graph theoretical part allows to naturally model the structure of the domain of interest. The accompanying modularity of graphical models is particularly interesting for computer vision because a complex model can be decomposed into simpler parts that allow a deeper understanding of the domain of interest.

Therefore, this dissertation has investigated and leveraged graphical models for object detection in challenging scenes. Such scenes require computer vision approaches to be powerful, yet flexible enough to reliably detect object instances under, for example, partial occlusion, cluttered background and articulation. We have addressed these challenges by developping and studying graphical models along our main thesis that discriminatively trained, part-based models outperform monolithic, generative approaches. To that end, we have taken advantage of the modularity and intuitive decomposability of graphical models and have interpreted the random variables to represent object parts or entire objects.

7.1 DISCUSSION OF CONTRIBUTIONS

As we have discussed, many generative part-based models for object detection have been proposed in the literature, but recently, discriminative variants have gained increased attention as they have been shown to outperform generative models on a variety of challenging datasets, especially if a considerable number of training instances is given. The likely advantage of discriminative approaches lies in their direct modeling of the classification task, which is easier than to learn the joint probability of input and output variables, which is modeled by generative models. Another advantage of discriminative CRFs over generative MRFs is that rich and possibly overlapping feature representations can be included in the model

without needing to worry about of the independence assumptions among the feature descriptors.

This aspect is particularly interesting when we consider hierarchical feature representations that may overlap severely. We have argued that these hierarchical representations are advantageous to both, purely local and purely monolithic models, as they allow to learn the trade-off between flexible local and powerful global object representations. However, for high-level computer vision tasks like object detection graphical models have frequently fallen short of simpler (non-linear) classifiers, which is due to the inferior linear potential function used in standard CRFs. We have addressed this issue by incorporating large margin classifiers into the model and have demonstrated the effectiveness of such classifiers within our graphical model in accordance with observations of related work. We have extended related work by developing a new learning method, which simultaneously learns all model parameters, and we demonstrated the effectiveness of such joint learning paradigms.

Object detection has been approached from two opposing directions: part-based approaches and monolithic models have been proposed to handle the omnipresent challenges in detection scenarios. For example, part-based approaches are able to handle articulation but tend to degrade in the presence of cluttered background while monolithic models are more robust in cluttered scenes but often fail to represent articulation. While both interpretations of object representations have long been studied separately from each other, recently, joint models unifying the advantages of both worlds have become more and more important. Thereby, recent models have revealed superior modeling power and increased performances especially in datasets, for which flexible yet powerful models are required. We have studied this aspect by taking advantage of the modular nature of graphical models: By interpreting the random variables as object parts and developing multi-layer models, we have demonstrated the effectiveness of graphical models to simultaneously represent part-based and global views on objects. Our model lends itself to design complex object models, yet is simple enough to allow interpreting it intuitively.

As we have seen, graphical models are particularly interesting since they enable an explicit definition of pairwise dependencies among object parts and between parts and entire objects. However, standard conditional random fields, our choice of discriminative graphical models, are restricted to local neighborhood dependencies that are not suitable to represent complex dependencies among object parts. We have addressed this issue by first exploiting longer-range dependencies in a hierarchical setting even though this still accounts only for fixed dependencies. Since we do not know in advance, which pairwise dependencies best reflect the structure of the domain of interest and we would like to avoid the tedious manual definition of the graph structure, we have overcome the restriction to a fixed graphical structure in a second step by adapting an efficient structure learning approach to our case. Experimentally, we have demonstrated the effectiveness of learning arbitrary structures compared to restricted fixed structures.

Many recent part-based models have focused on one particular object class and therefore the parts have been defined manually. In our work, we have studied the

applicability to a wide range of object classes, and hence the goal has been to avoid tedious hand-labelling of thousands of object instances. We have developed a weakly supervised model, which not only allows to learn discriminative part representations but also to discover meaningful object parts. To that end, we have adopted a latent interpretation of random variables and have learned the object models from bounding box labels alone. By pairing the hidden model with discriminative learning techniques we have shown that we are able to successfully learn and represent the spatial layout of objects even in the presence of articulation and partial occlusion.

While single feature approaches are error-prone and often cannot represent the versatility of object appearance, multi-feature approaches have been recently proposed to overcome such restrictions. Going into the same direction, we have developed a principled and modular framework that allows to easily integrate different feature descriptors; we have shown preliminary results with two complementary feature descriptors. In accordance with related work on multi-feature approaches we have reported consistent performance improvements by leveraging the complementarity of different feature types. A similar idea has been elaborated for search and rescue robotics but at a different level: We have not integrated different feature descriptors, but studied the use of complementary sensor information such as visual, thermal and laser data. To that end, we have developed a principled framework to integrate different sensor data with the focus on increased reliability of search and rescue robotics.

7.2 OUTLOOK

In this dissertation, we have shown how standard CRFs can be enhanced to achieve an increased modeling power. However, the presented approaches have some limitations that may foster future research on graphical models for object detection. One limitation of our models is that we infer each object instance separately and consider image context only in a local surrounding of an object. In order to address this issue one could infer entire images instead of only bounding boxes of objects. In this respect, the presence and location of one object could be used to explain the presence and location of another object. For example, people tend to sit on horses or bicycles, and chairs and dining tables often appear next to each other. Enhancing our approach to model entire images would allow to model such object-object interactions. Moreover, one could infer global image statistics (e.g. indoor vs. outdoor scenarios and urban vs. natural environments), which in turn favor the presence or absence of certain object classes.

Another issue is the two-dimensional treatment of objects. When we infer a bounding box of an object we assume that the object appears at the same scale relative to the bounding box. However, certain viewpoints of objects (e.g. 45 degree views of cars) rather follow a three-dimensional propagation scheme and object parts appear differently under these viewpoints. This issue is not yet addressed in our models; especially when considering entire images and object-object relations we are in need of a three-dimensional interpretation. In this case the object-object relations

are no longer two-dimensional making a three-dimensional understanding of the entire scene necessary.

We have combined two different learning paradigms in our model, namely the max margin objective of SVMs and maximum likelihood in case of CRFs. This might be a more technical limitation since intuitively one consistent and unique learning objective is desirable. In order to address this issue one could adopt the max margin formalism of random fields (Taskar *et al.*, 2003) to our model.

A promising idea that we have not considered yet is sharing features and parts across categories. Many object categories have certain parts in common, for example, quadrupeds stand on four legs, and cars, busses, bicycles and motorbikes have wheels. These shared parts have similar appearances but the object classes differ largely by the composition of object parts or the presence or absence of some additional parts that are unique to each class. This observation motivates to jointly learn and share common object parts across classes and distinguish the different classes through different constellations of these parts.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

Y. Amit and A. Trouve (2007). POP: Patchwork of Parts Models for Object Recognition, *International Journal of Computer Vision*, vol. 75(2), pp. 267–282. 19

M. Andriluka, P. Schnitzspan, J. Meyer, S. Kohlbrecher, K. Petersen, O. von Stryk, S. Roth, and B. Schiele (2010). Vision Based Victim Detection from Unmanned Aerial Vehicles, in *International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan (submitted) 2010*. 8, 10

M. Asada and Y. Shirai (1989). Building a World Model for a Mobile Robot Using Dynamic Semantic Constraints, in *International Joint Conference on Artificial Intelligence 1989*. 87

F. Bach, G. Lanckriet, and M. Jordan (2004). Multiple kernel learning, conic duality, and the SMO algorithm, in *Proceedings of the 21st International Conference on Machine Learning 2004*. 21

O. Banerjee, L. El Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, *Journal of Machine Learning Research*, vol. 9, pp. 485–516. 29

C. Bishop (2006). *Pattern Recognition and Machine Learning*, Springer. 14

A. Bosch, A. Zisserman, and X. Munoz (2007). Representing Shape with a Spatial Pyramid Kernel, in *International Conference on Image and Video Retrieval 2007*. 22

B. Boser, I. Guyon, and V. Vapnik (1992). A Training Algorithm for Optimal Margin Classifiers, in *Fifth Annual Workshop on Computational Learning Theory 1992*. 18

G. Bouchard and B. Triggs (2005). Hierarchical Part-Based Visual Object Categorization, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005*. 19, 42

W. Burgard and M. Hebert (2008). World Modeling, in *Springer Handbook of Robotics 2008*, pp. 853–869. 87

O. Chapelle (2007). Training a Support Vector Machine in the Primal, *Neural Computation*, vol. 19, pp. 1155–1178. 37, 38, 39, 40

D. Chickering (2003). Optimal structure identification with greedy search, *Journal of Machine Learning Research*, vol. 3, p. 554. 27

O. Chum and A. Zisserman (2007). An Exemplar Model for Learning Object Classes, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2007*. 45, 47

D. Crandall and D. Huttenlocher (2006). Weakly supervised learning of part-based spatial models for visual object recognition, in *Proceedings of the Ninth European Conference on Computer Vision 2006*. 70

N. Dalal and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005*. 4, 13, 21, 27, 31, 33, 41, 43, 47, 54, 56, 57, 63, 75, 79, 80, 91, 96

C. Desai, D. Ramanan, and C. Fowlkes (2009). Discriminative models for multi-class object layout, in *Proceedings of the Twelfth IEEE International Conference on Computer Vision 2009*. 13, 21, 84, 85

P. Doherty and P. Rudol (2007). A UAV Search and Rescue Scenario with Human Body Detection and Geolocalization, in *20th Australian Joint Conference on Artificial Intelligence 2007*. 30, 31

P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu (2008). Multiple Component Learning for Object Detection, in *Proceedings of the Tenth European Conference on Computer Vision 2008*. 20

B. Epshtein and S. Ullman (2007). Semantic Hierarchies for Recognizing Objects and Parts, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2007*. 19

M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman (2007). *The PASCAL VOC Challenge 2007*. 6, 7, 63, 65, 80, 85

M. Everingham, A. Zisserman, C. Williams, and L. Van Gool (2006). *The PASCAL Visual Object Classes Challenge 2006 Results*. 6, 7, 45

P. Felzenszwalb and D. Huttenlocher (2005). Pictorial Structures for Object Recognition, *International Journal of Computer Vision*, vol. 61(1), pp. 55–79. 19, 34

P. Felzenszwalb, D. McAllester, and D. Ramanan (2008). A Discriminatively Trained, Multiscale, Deformable Part Model, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2008*. 13, 14, 20, 21, 45, 47, 53, 54, 63, 64, 65, 67, 70, 72, 73, 80, 83, 85, 106, 107

R. Fergus, A. Zisserman, and P. Perona (2003). Object Class Recognition by Unsupervised Scale Invariant Learning, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2003*. 19, 20, 34, 54

V. Ferrari, M. Marin, and A. Zisserman (2009). Progressive Search Space Reduction for Human Pose Estimation, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2009*. 91

M. Fischler and R. Elschlager (1973). The Representation and Matching of Pictorial Structures, *IEEE Transactions on Computers*, vol. C-22(1), pp. 67–92. 19

M. Fritz, B. Leibe, B. Caputo, and B. Schiele (2005). Integrating Representative and Discriminant Models for Object Category Detection, in *Proceedings of the Tenth IEEE International Conference on Computer Vision 2005*. 43

P. V. Gehler and S. Nowozin (2009a). Let the Kernel Figure it Out: Principled Learning of Pre-processing for Kernel Classifiers, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2009*. 22

P. V. Gehler and S. Nowozin (2009b). On Feature Combination for Multiclass Object Classification, in *Proceedings of the Twelfth IEEE International Conference on Computer Vision 2009*. 22

K. Grauman and T. Darrell (2007). The Pyramid Match Kernel: Efficient Learning with Sets of Features, *Journal of Machine Learning Research*, vol. 8, pp. 725–760. 21, 52, 54

R. Greiner, X. Su, B. Shen, and W. Zhou (2005). Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers, *Machine Learning*, vol. 59(3), pp. 297–322. 28

D. Grossman and P. Domingos (2004). Learning Bayesian network classifiers by maximizing conditional likelihood, in *Proceedings of the 21st International Conference on Machine Learning 2004*. 28

Y. Guo and D. Schuurmans (2006). Convex structure learning for Bayesian networks: Polynomial feature selection and approximate ordering, in *Conference on Uncertainty in Artificial Intelligence (UAI) 2006*. 27

D. Hall and J. Llinas (1997). An introduction to multisensor data fusion, *Proceedings of the IEEE*, vol. 85(1), pp. 6–23. 30, 31, 106

H. Hariharan, A. Gribok, M. Abidi, and A. Koschan (2006). Image fusion and enhancement via empirical mode decomposition, *Journal of Pattern Recognition Research*, vol. 1(1), pp. 16–32. 30

H. Harzallah, F. Jurie, and C. Schmid (2009). Combining efficient object localization and image classification, in *Proceedings of the Twelfth IEEE International Conference on Computer Vision 2009*. 21

X. He, R. Zemel, , and D. Ray (2006). Learning and incorporating top-down cues in image segmentation, in *Proceedings of the Ninth European Conference on Computer Vision 2006*. 24, 25

X. He, R. Zemel, and M. Carreira-Perpinan (2004). Multiscale Conditional Random Fields for Image Labeling, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2004*. 24

X. He and R. S. Zemel (2008). Latent Topic Random Fields: Learning using a taxonomy of labels, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2008*. 24

D. Heckerman, D. Geiger, and D. Chickering (1995). Learning Bayesian networks: The combination of knowledge and statistical data, *Machine learning*, vol. 20(3), pp. 197–243. 27

H. Höfling and R. Tibshirani (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods, *Journal of Machine Learning Research*, vol. 10, pp. 883–906. 29

D. Hoiem, C. Rother, and J. Winn (2007). 3D Layout CRF for Multi-View Object Class Recognition and Segmentation, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2007*. 14, 25, 37, 43, 44, 52, 56, 59

S. Ioffe and D. Forsyth (2001). Probabilistic methods for finding people, *International Journal of Computer Vision*, vol. 43(1), pp. 45–68. 5

T. Joachims (1999). *Advances in kernel methods: support vector learning*, chapter Making Large-Scale SVM Learning Practical, pp. 169 – 184, MIT Press Cambridge, MA, USA. 43, 63, 80

J. Johnson, V. Chandrasekaran, and A. Willsky (2007). Learning Markov Structure by Maximum Entropy Relaxation, in *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics 2007*. 29

A. Kapoor and J. Winn (2006). Located Hidden Random Fields: Learning Discriminative Parts for Object Detection, in *Proceedings of the Ninth European Conference on Computer Vision 2006*. 14, 26, 52, 72, 74, 83

G. S. Kimeldorf and G. Wahba (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *Annals of Mathematical Statistics*, vol. 41, pp. 495–502. 39

D. Klein, D. Schulz, and S. Frintrop (2009). Boosting with a Joint Feature Pool from Different Sensors, in *7th International Conference on Computer Vision Systems 2009*. 30

A. Kleiner and R. Kümmerle (2007). Genetic MRF model optimization for real-time victim detection in Search and Rescue, in *IEEE/RSJ International Conference on Intelligent Robots and Systems 2007*. 30, 89

P. Kohli, L. Ladicky, and P. Torr (2009). Robust Higher Order Potentials for Enforcing Label Consistency, *International Journal of Computer Vision*, vol. 82(3), pp. 302–324. 25

M. Koivisto and K. Sood (2004). Exact Bayesian structure discovery in Bayesian networks, *The Journal of Machine Learning Research*, vol. 5, p. 573. 27

M. P. Kumar, A. Zisserman, and P. H. Torr (2009). Efficient Discriminative Learning of Parts-based Models, in *Proceedings of the Twelfth IEEE International Conference on Computer Vision 2009*. 20, 72

S. Kumar, J. August, and M. Hebert (2006). Discriminative Random Fields, *International Journal of Computer Vision*, vol. 68(2), pp. 179–201. 24, 40, 55, 56

S. Kumar, J. Guivant, and H. Durrant-Whyte (2004). Informative Representations of Unstructured Environments, in *IEEE International Conference on Robotics and Automation 2004*. 87

S. Kumar and M. Hebert (2003). Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification, in *Proceedings of the Ninth IEEE International Conference on Computer Vision 2003*. 24

L. Ladicky, C. Russell, P. Kohli, and P. Torr (2009). Associative Hierarchical CRFs for Object Class Image Segmentation, in *Proceedings of the Twelfth IEEE International Conference on Computer Vision 2009*. 25

J. Lafferty, A. McCallum, and F. Pereira (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in *Proceedings of the Eighteenth International Conference on Machine Learning 2001*. 3, 14, 33, 34, 59

C. Lampert, M. Blaschko, and T. Hofmann (2009). Efficient subwindow search: A branch and bound framework for object localization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(12), pp. 2129–2142. 21

I. Laptev (2006). Improvements of object detection using boosted histograms, in *Proceedings of the British Machine Vision Conference 2006*. 47

D. Larlus and F. Jurie (2008). Combining Appearance Models and Markov Random Fields for Category Level Object Segmentation, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2008*. 24

S. Lazebnik, C. Schmid, and J. Ponce (2006). Beyond Bag of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2006*. 21, 52, 54, 56, 57, 79, 80

C. H. Lee, R. Greiner, and M. Schmidt (2005). Support Vector Random Fields for Spatial Classification, in *European Conference on Principles and Practice of Knowledge Discovery in Databases 2005*. 23, 36, 37

C. H. Lee, R. Greiner, and O. Zaianen (2006a). Efficient Spatial Classification using Decoupled Conditional Random Fields, in *European Conference on Principles and Practice of Knowledge Discovery in Databases 2006*. 23, 52, 56, 59

S.-I. Lee, V. Ganapathi, and D. Koller (2006b). Efficient Structure Learning of Markov Networks using L1-Regularization, in *Advances in Neural Information Processing Systems 2006*. 28, 29, 30, 54, 60, 61

B. Leibe, A. Leonardis, and B. Schiele (2004). Combining Categorization and Segmentation with an Implicit Shape Model, in *Proceedings of the Eighth European Conference on Computer Vision 2004*. 19, 20

B. Leibe, E. Seemann, and B. Schiele (2005). Pedestrian Detection in Crowded Scenes, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005*. 4, 20, 34, 43, 44, 47

A. Levin and Y. Weiss (2006). Learning to Combine Bottom-Up and Top-Down Segmentation, in *Proceedings of the Ninth European Conference on Computer Vision 2006*. 25, 56, 59

D. Lowe (2004). Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, vol. 60(2), pp. 91–110. 41, 57, 80

S. Maji, A. C. Berg, and J. Malik (2008). Classification Using Intersection Kernel Support Vector Machines is efficient, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2008*. 57, 80

S. Maji and J. Malik (2009). Object detection using a max-margin hough transform, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2009*. 20

N. Meinshausen and P. Buhlmann (2006). High Dimensional Graphs and Variable Selection with the Lasso, *Annals of Statistics*, vol. 34(2), pp. 1436–1462. 28

J. Meyer, P. Schnitzspan, S. Kohlbrecher, K. Petersen, O. Schwahn, M. Andriluka, U. Klingauf, S. Roth, B. Schiele, and O. von Stryk (2010). A Semantic World Model for Urban Search and Rescue Based on Heterogeneous Sensors, in *RoboCup Symposium, Singapore 2010*. 10, 11, 87

K. Murphy, Y. Weiss, and M. Jordan (1999). Loopy belief propagation for approximate inference: An empirical study, in *Proceedings of Uncertainty in AI 1999*. 16

J. Mutch and D. Lowe (2006). Multiclass Object Recognition with Sparse, Localized Features, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2006*. 43, 44

S. Parise and M. Welling (2006). Structure Learning in Markov Random Fields, in *Advances in Neural Information Processing Systems 2006*. 29, 54

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann. 16, 39, 40

S. Perkins, K. Lacker, J. Theiler, I. Guyon, and A. Elisseeff (2003). Grafting: Fast, incremental feature selection by gradient descent in function space, *Journal of Machine Learning Research*, vol. 3, pp. 1333–1356. 29, 61, 62

F. Pernkopf and J. Bilmes (2005). Discriminative versus Generative Parameter and Structure Learning of Bayesian Network Classifiers, in *Proceedings of the 22nd International Conference on Machine Learning 2005*. 28

S. D. Pietra, V. D. Pietra, and J. Lafferty (1997). Inducing Features of Random Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(4), pp. 380–393. 28

J. C. Platt (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, in *Advances in Large Margin Classifiers 1999*. 91

A. Quattoni, M. Collins, and T. Darrell (2004). Conditional Random Fields for Object Recognition, in *Advances in Neural Information Processing Systems 2004*. 14, 26

A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell (2007). Hidden Conditional Random Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29(10), pp. 1848–1852. 26, 74

B. Schiele and J. Crowley (1994). A Comparison of Position Estimation Techniques Using Occupancy Grids, in *IEEE International Conference on Robotics and Automation 1994*. 90

M. Schmidt and K. Murphy (2010). Convex Structure Learning in Log-Linear Models: Beyond Pairwise Potentials, in *Proceedings of the International Workshop on Artificial Intelligence and Statistics 2010*. 29

M. Schmidt, K. Murphy, G. Fung, and R. Rosales (2008). Structure Learning in Random Fields for Heart Motion Abnormality Detection, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2008*. 14, 29, 54, 78

M. Schmidt, A. Niculescu-Mizil, and K. Murphy (2007). Learning graphical model structure using l1-regularization paths, in *Proceedings of the National Conference on Artificial Intelligence 2007*. 28

P. Schnitzspan, M. Fritz, S. Roth, and B. Schiele (2009). Discriminative Structure Learning of Hierarchical Representations for Object Detection, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2009*. 9, 10, 51, 72, 78, 84, 85

P. Schnitzspan, M. Fritz, and B. Schiele (2008). Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features, in *Proceedings of the Tenth European Conference on Computer Vision 2008*. 8, 10, 33, 52, 56, 72, 75, 82

P. Schnitzspan, S. Roth, and B. Schiele (2010). Automatic Discovery of Meaningful Object Parts with Latent CRFs, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2010*. 9, 11, 69

J. Shotton, J. Winn, C. Rother, and A. Criminisi (2006). TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation, in *Proceedings of the Ninth European Conference on Computer Vision 2006*. 37, 41, 47, 56, 59

J. Shotton, J. Winn, C. Rother, and A. Criminisi (2009). TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context., *International Journal of Computer Vision*, vol. 81, pp. 2–32. 24, 25

L. Spinello, R. Triebel, and R. Siegwart (2008). Multimodal people detection and tracking in crowded scenes, in *Proceedings of the 23rd International Conference on Machine Learning 2008*. 30

M. Stark, M. Goesele, and B. Schiele (2009). A Shape-Based Object Class Model for Knowledge Transfer, in *Proceedings of the Twelfth IEEE International Conference on Computer Vision 2009*. 19

E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky (2008). Describing Visual Scenes Using Transformed Objects and Parts, *International Journal of Computer Vision*, vol. 77(1–3), pp. 291–330. 19

C. Sutton and A. McCallum (2007). An Introduction to Conditional Random Fields for Relational Learning, in L. Getoor and B. Taskar (eds.), *Introduction to Statistical Relational Learning 2007*, MIT Press. 17

S. Tadokoro, H. Kitano, T. Takahashi, I. Noda, H. Matsubara, A. Shinjoh, T. Koto, I. Takeuchi, H. Takahashi, F. Matsuno, *et al.* (2000). The robocup-rescue project: A robotic approach to the disaster mitigation problem, in *IEEE International Conference on Robotics and Automation 2000*. 87

B. Taskar, C. Guestrin, and D. Koller (2003). Max margin Markov networks, in *Advances in Neural Information Processing Systems 2003*. 23, 40, 77, 104

S. Thrun, W. Burgard, and D. Fox (2005). *Probabilistic Robotics*, MIT Press, Cambridge. 90

A. Torralba, K. Murphy, and W. Freeman (2004). Contextual models for object detection using boosted random fields., in *Advances in Neural Information Processing Systems 2004*. 24

D. Tran and D. Forsyth (2007). Configuration Estimates Improve Pedestrian Finding, in *Advances in Neural Information Processing Systems 2007*. 27

I. Tsamardinos, L. Brown, and C. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm, *Machine learning*, vol. 65(1), pp. 31–78. 28

I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, and Y. Singer (2005). Large margin methods for structured and interdependent output variables, *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484. 23

V. Vapnik (1998). *Statistical Learning Theory*, Wiley-Interscience. 18

M. Varma and B. R. Babu (2009). More generality in efficient multiple kernel learning. 22

M. Varma and D. Ray (2007). Learning The Discriminative Power-Invariance Trade-Off, in *Proceedings of the Eleventh IEEE International Conference on Computer Vision 2007*. 21, 52, 55

A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman (2009). Multiple Kernels for Object Detection, in *Proceedings of the Twelfth IEEE International Conference on Computer Vision 2009*. 22, 52, 66, 79, 84, 85

V. Viitaniemi and J. Laaksonen (2006). Techniques for Still Image Scene Classification and Object Detection, in *Proceedings of the 16th International Conference on Artificial Neural Networks 2006*. 47

M. Wainwright and M. Jordan (2003). Graphical models, exponential families and variational inference, Technical report, Department of Statistics, University of California, Berkeley. 14

M. Wainwright, P. Ravikumar, and J. Lafferty (2006). Inferring Graphical Model Structure using L1-regularized pseudo-likelihood, in *Advances in Neural Information Processing Systems 2006*. 28

M. J. Wainwright, T. Jaakkola, and A. S. Willsky (2002). Tree-based reparameterization for approximate estimation on graphs with cycles, in *Advances in Neural Information Processing Systems 2002*. 62

J. Winn and J. Shotton (2006). The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2006*. 14, 25, 37, 43, 44, 52, 56, 59

C. Wojek, G. Dorkó, A. Schulz, and B. Schiele (2008). Sliding-Windows for Rapid Object Class Localization: A Parallel Technique, in *Pattern Recognition, Proceedings of the 30th DAGM-Symposium 2008*. 31, 91

J. Yedidia, W. Freeman, and Y. Weiss (2003). Understanding Belief Propagation and Its Generalizations, in *Exploring Artificial Intelligence in the New Millennium 2003*, Morgan Kaufmann Publishers. 14, 59

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid (2007). Local features and kernels for classification of texture and object categories: A comprehensive study, *International Journal of Computer Vision*, vol. 73(2), pp. 213–238. 52

Z. Zivkovic and B. Kröse (2007). Part based people detection using 2D range data and images, in *IEEE International Conference on Robotics and Automation 2007*. 30

VITA

Born in		Seeheim-Jugenheim, Germany
Education	2007–2010	TU Darmstadt, Germany PhD student in computer science
	2001–2006	TU Darmstadt, Germany Diplom in mathematics and computer science
Positions	2007–2010	TU Darmstadt, Germany Research assistent
	2007–2010	TU Darmstadt, Germany Scholarship, research training group
	2009	TU Darmstadt, Germany Teaching assistent, Maschinelles Lernen 2
	2006	Prague, Czech Republic Intern, Praszka Energetica
	2005	Frankfurt, Germany Intern, Lufthansa
	2004	Böblingen, Germany Intern, IBM

PUBLICATIONS

[IROS'10] Mykhaylo Andriluka, Paul Schnitzspan, Johannes Meyer, Stefan Kohlbrecher, Karen Petersen, Oskar von Stryk, Stefan Roth and Bernt Schiele: Vision Based Victim Detection from Unmanned Aerial Vehicles, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, 2010

[RCS'10] Johannes Meyer, Paul Schnitzspan, Stefan Kohlbrecher, Karen Petersen, Oliver Schwahn, Mykhaylo Andriluka, Uwe Klingauf, Stefan Roth, Bernt Schiele and Oskar von Stryk: A Semantic World Model for Urban Search and Rescue Based on Heterogeneous Sensors, RoboCup Symposium (RCS), Singapore, 2010

[CVPR'10] Paul Schnitzspan, Stefan Roth and Bernt Schiele: Automatic Discovery of Meaningful Object Parts with Latent CRFs, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, 2010

[CVPR'09] Paul Schnitzspan, Mario Fritz, Stefan Roth and Bernt Schiele: Discriminative Structure Learning of Hierarchical Representations for Object Detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, 2009

[ECCV'08] Paul Schnitzspan, Mario Fritz and Bernt Schiele: Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features, Europoean Conference on Computer Vision (ECCV), Marseille, 2008