KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering

Context-based Reasoning for Object Detection and Object Pose Estimation

José Antonio Oramas Mogrovejo

Supervisor:
Prof. dr. ir. T. Tuytelaars
Prof. dr. L. De Raedt, co-supervisor

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

April 2015

Context-based Reasoning for Object Detection and Object Pose Estimation

José Antonio ORAMAS MOGROVEJO

Examination committee:
Prof. dr. ir. J. Vandewalle, chair
Prof. dr. ir. T. Tuytelaars, supervisor
Prof. dr. L. De Raedt, co-supervisor
Prof. dr. ir. L. Van Gool
Prof. dr. M.-F. Moens
Prof. dr. A. Leonardis
(University of Birmingham, UK)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

April 2015

Uitgegeven in eigen beheer, José Antonio Oramas Mogrovejo, Kasteelpark Arenberg 10 box 2441, B-3001 Heverlee (Belgium)

Acknowledgements

Some time ago I decided to go on the journey of working towards a doctoral degree. During this journey, I had the good fortune of meeting many wonderful people who motivated and supported me to fulfil this goal.

First of all, I would like to thank Prof. Tinne Tuytelaars and Prof. Luc De Raedt for giving me the opportunity to pursue a Ph.D. under their supervision. I want to express my gratitude to my main supervisor, Prof. Tuytelaars. She is the best supervisor I could ever have, possessing broad and deep scientific knowledge, while always being kind and sympathetic. Throughout my Ph.D., Prof. Tuytelaars granted me a good balance between supervision and freedom of research. Her guidance provided me space to learn from my mistakes and develop my research and teaching skills. I also had the good fortune of having Prof. De Raedt as my co-supervisor. I am very thankful for the insightful comments and the alternative perspectives that he always provided during our discussions. There are several lessons that I learned from Prof. De Raedt while I was his student during my Pre-Doctoral period and which were very valuable during my Ph.D.

I would like to thank the members of my Examination Committee for their motivating comments and for their useful remarks throughout my Ph.D. I am thankful to Prof. Aleš Leonardis for accepting to be my examiner and for his contributions to this thesis. It is a great honor to have the names of the members of my Examination Committee in this manuscript.

Research is a team sport. Therefore, I am happy that many people took the time to share their ideas with me. I very much enjoyed discussions and collaborations with different people. I want to give special thanks to my co-authors, Dr. Laura Antanas, with whom I collaborated for a significant part of my Ph.D., and to Dr. Basura Fernando, Dr. Efstratios Gavves and Amir Ghodrati, for the interesting discussions and all the good work related to *VideoDarwin*. In addition, I would also like to thank the rest of the team, the members of the VISICS lab. Thank

you guys for building such a nice working environment and for all the support during deadline times. I would like to give special thanks to my friends from the big office, especially to Basura, David, and Kostas, for all the awesome time we shared together. Guys, you are like brothers to me.

I thank the financial support from the DBOF Research Fund KUL 3E100864, the FP7 ERC grant 240530 COGNIMUND and the KULeuven OT Project VASI.

I want to give special thanks to my wonderful girlfriend, Jasmina. Thanks for being the sun that enlightens my life and for being my motivation to move forward. I really appreciate how positive, understanding and tolerative you were during deadline times. Te quiero mucho mijn princesitatje. In addition, I would like to thank her family for all their attention and support, and for making my stay in Belgium more comfortable and enjoyable.

Last but not least, I am very thankful to my parents, my sister and all my family for their unconditional support and for providing me the perfect environment to develop and succeed. Their love and actions have been invaluable to me. This thesis is dedicated to them.

Abstract

Computer vision algorithms have become very effective at detecting the occurrence of objects in images. Parallel to this, notable advancements have been achieved in estimating the orientation in which such objects occur. Different methods have been proposed during the last decade, ranging from methods that model the appearance of the object as it is projected on the 2D image space to methods that reason about physical properties of the objects in the 3D scene. These methods have proved to be effective at the tasks at hand. However, one weakness of these methods is their complete reliance on intrinsic features, e.g. color, size, texture, that define the objects of interest. This weakness becomes evident in difficult scenarios triggered by factors such as high inter-object occlusion; which affects perceived shape and size of objects, as well as drastic changes in illumination; which affects how the texture and color of objects are perceived by the camera.

There are additional, extrinsic, cues that can help under these scenarios. For example, some object categories tend to appear more often in some scene types than in others. For instance, it is more likely to find a computer in an indoor scene rather than in an outdoor setting. Likewise, in natural and man-made objects there are some, imposed or desired, rules that determine the configurations in which objects co-occur. For example, birds fly following a flocking behavior, keyboard and mouse are usually found below the computer screen, and so on. This thesis investigates the potential of these extrinsic cues to assist computer vision tasks such as object detection and object pose estimation.

Context cues have been used before for object detection. Here we show that they can also help in object pose estimation. This applies to both scene cues (e.g. the groundplane) as well as location and pose of other objects in the scene. Furthermore, we show that cautious inference on object relations brings improvements over traditional inference for object detection. Finally, we show how to use context cues not only to filter our false object detections but also to retrieve object instances missed in an initial detection step.

Beknopte samenvatting

Computervisiealgoritmes zijn zeer efficiënt geworden in het detecteren van voorwerpen in afbeeldingen. Tegelijk is er opmerkelijke vooruitgang geboekt op het vlak van de inschatting van de omgeving waarin dergelijke voorwerpen voorkomen. De voorbije tien jaar zijn diverse methodes voorgesteld, gaande van methodes die het uitzicht van het voorwerp modelleren zoals het wordt geprojecteerd in een tweedimensionale afbeeldingsruimte, tot methodes die redeneren over de fysieke eigenschappen van het voorwerp in een driedimensionale ruimte. Die methodes zijn efficiënt gebleken bij het uitvoeren van praktische taken. Toch is één van de zwakheden van deze methodes dat ze volledig vertrouwen op de intrinsieke kenmerken zoals de kleur, afmeting en textuur van die bewuste voorwerpen. Die zwakte komt tot uiting in moeilijke scenario's, veroorzaakt door factoren zoals grote occlusies tussen voorwerpen, die de gepercipieerde vorm en afmetingen van voorwerpen beïnvloeden, en ook bij drastische belichtingswijzigingen, die de manier waarop de camera de textuur en kleur van voorwerpen percipieert, beïnvloeden.

Er zijn bijkomende extrinsieke aanwijzingen die in deze omstandigheden hun bijdrage kunnen leveren. Zo hebben bijvoorbeeld bepaalde voorwerpcategorieën de neiging om vaker voor te komen in bepaalde omgevingen dan in andere. Een computer tref je bijvoorbeeld eerder binnenshuis dan buitenshuis aan. Ook gelden voor natuurlijke en artificiële voorwerpen vaak regels die de configuraties bepalen waarin voorwerpen samen voorkomen. Zo vliegen vogels bijvoorbeeld in een bepaalde formatie, en zo worden computertoetsenbord en muis meestal onder het computerscherm aangetroffen. Deze thesis onderzoekt het potentieel van die extrinsieke aanwijzingen om een bijdrage te leveren aan computerherkenningsopdrachten zoals de detectie van voorwerpen en het inschatten van hun oriëntatie.

Contextaanwijzigingen zijn vroeger al gebruikt bij voorwerpsdetectie. Hier willen we aantonen dat ze ook kunnen helpen bij de voorwerporiëntatieschatting. Dat geldt zowel voor omgevingsaanwijzingen (bijvoorbeeld het grondvlak) als

voor de plaats en de oriëntatie van andere voorwerpen in de omgeving. Voorts tonen we aan dat een voorzichtige gevolgtrekking uit voorwerpsverhoudingen de traditionele aanpak verbetert, voor het geval van voorwerpsdetectie. Tot slot tonen we aan hoe contextaanwijzingen niet alleen verkeerde voorwerpsdetectie kunnen wegfilteren, maar ook voorwerpen kunnen opsporen die in aanvankelijke detectiestappen over het hoofd werden gezien.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human visual system is very accurate when interpreting scenes depicted in images. It is not only able to identify intrinsic visual features that define the appearance of certain object categories but it is also able to identify the features that distinguish instances of a particular category from those of other categories. In addition, the human visual system is capable of considering, additional, contextual cues. These contextual cues may take different forms. One of those forms is as occurrence relations of object categories within specific types of scenes, e.g. cars are more likely to occur in urban scenes than in beach scenes. Likewise, chairs are more likely to occur in indoor scenes than in urban scenes. Another form of contextual cue is to consider typical arrangements in which objects tend to co-occur, e.g., chairs tend to be located around or below tables; cars park next-to or behind each other, etc. It is the combination of these aspects, plus other aspects related to visual perception that assists humans in interpreting images [12, 64, 77, 152].

Since the early years of visual object recognition, most attention has been given to the intrinsic aspects that define object categories, e.g. size, color, texture. Towards this goal, on one hand, several methods [101, 120, 133, 137, 161] have been proposed to model the 3D shape of the object categories of interest. On the other hand, there is a group of methods [26, 40, 42, 153] that focuses on modeling the 2D appearance of the objects when depicted on the images. Reports in the literature [94, 110, 116] show that these methods have a drop in performance in particular scenarios such as low image resolution, high occlusion between objects, changes in illumination, etc. Given these circumstances, subsequent research started to look beyond intrinsic features of the objects and explored different directions, e.g. inter-object occlusion reasoning [62, 94, 116, 162],

2D/3D spatial reasoning [3, 53, 125, 160, 165], and other sources of contextual information [59, 60, 68, 163], as a means to disambiguate confusing scenarios. This thesis focuses on the exploration of methods that exploit such contextual information to improve the performance of computer vision tasks.

In Section 1.1 we present the computer vision tasks that will be of interest in this thesis. Then, in Section 1.2 we position this thesis with respect to existing works. The motivation that drives this work is presented in Section 1.3. Section 1.4 gives a brief overview of contents and contributions of this thesis.

1.1 Tasks of Interest

There are a variety of problems addressed in the computer vision literature, e.g. object detection, image classification, single/multiple view 3D reconstruction, visual tracking, human pose estimation, gesture recognition, to name a few. In this thesis we focus our attention on two particular object-centric problems, i.e., object detection and object pose/viewpoint estimation. Furthermore, there are different modalities in which context cues can be obtained, e.g. text from image captions, audio or gps coordinates collected from cameras, metadata from digital image/video files, etc. In this thesis we focus on exploiting visual-context cues.

We perform our analysis in urban scenes focusing on cars as the object category of interest as there are many cars occurring in urban scenes which is necessary to model relations between objects. Furthermore, there are several datasets available, which focus on urban scenes and cars, that can be useful for experimentation. Finally, there is a variety of application-oriented problems that can benefit from our findings, e.g. obstacle detection, lane detection, or traffic pattern recognition.

1.1.1 Object Detection

Object detection is a task that deals with detecting, or localizing, instances of semantic objects of a certain category, e.g. pedestrians, buildings, or cars, in digital images or videos. Well-researched domains of object detection include face detection [93, 113, 153] and pedestrian detection [10, 11, 26, 61, 94]. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

In this thesis we focus on object detection, that is, the localization of object instances in still images. More formally, given an image (Figure 1.1) and an

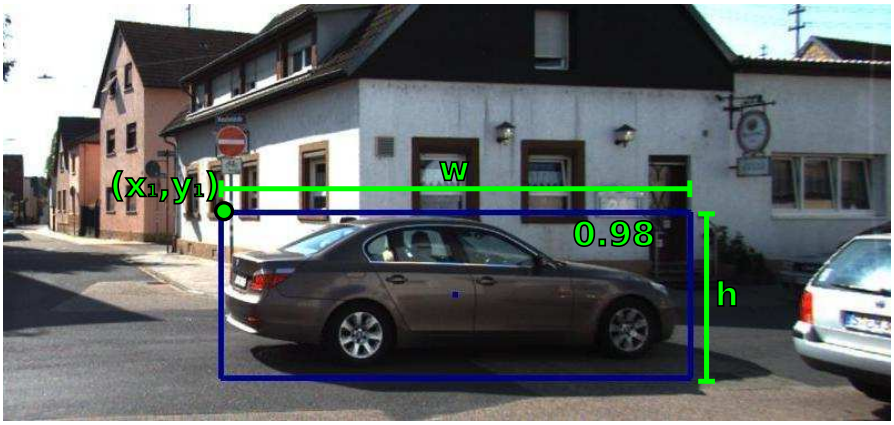

Figure 1.1: Car detection result output: coordinates of the top point (x_1, y_1), width w and height h of the bounding box accompanied with a detection score.

object category to be detected (e.g. cars), the goal is to provide the (x, y, w, h) data of each of the 2D bounding boxes that surround instances of the category of interest. An alternative output for the bounding box is to provide the (x, y) coordinates of the top-left and bottom-right corners that define it. In addition, most object detectors provide a score which represents the confidence on each of the predicted objects. In the literature [10, 11, 26, 61, 94], these predicted bounding boxes are usually referred to as "object hypotheses" or "detections". Furthermore, when an object hypothesis is correct, i.e. it overlaps an instance of the object of interest, it is referred to as a true positive. On the contrary when hypotheses that are incorrect are referred to as false positives.

It is important to note that object detection is different from image classification task. Image classification, usually referred to as "classification" in the computer vision literature, consists of predicting the presence/absence of an instance of an object category in a test image. Different from the object detection task, a bounding box localizing each object instance in the image is not required as output for classification. Thus, object detection is a more complex problem since it focus on measuring both occurrence and localization of the object instances of a given category.

1.1.2 Object Pose/Viewpoint Estimation

Object Pose Estimation is a typical task in computer vision which consists in identifying specific object instances in an image and determining each object's

position and orientation relative to some coordinate system. This information can then be used, for example, for object grasping by robots. The pose of an object in a 3D scene can be described by means of a rotation and translation transformation (see Figure 1.2.a), which brings the object from a reference 3D pose in the scene to the observed pose, or viewpoint, in the image. The azimuth orientation of the object in the 3D scene is usually referred to as "pose" and is denoted as θ. Similarly, the orientation of the object as is observed by the camera is referred to as "viewpoint" and is denoted as α (see Figure 1.2.b). In this thesis, specifically in Chapter 5, we introduce an alternative object orientation angle, the object elongation orientation ϵ. As its name suggests, ϵ measures the direction of the maximum physical extent, the elongation, of an object. In this thesis, the elongation orientation ϵ of an object is defined as perceived by the camera. As seen in Figure 1.2.c, the elongation orientation ϵ combines opposite viewpoints under a single label.

The image data from which the pose of an object is determined can be either a single image [47, 51, 81, 111], a stereo image pair [21, 83], or an image sequence [7, 6, 159] where, typically, the camera is moving with a known speed. The main focus of this thesis is on determining the pose/viewpoint from objects in still 2D images. The output of a pose/viewpoint estimation algorithm is similar to the one from object detection. However, in addition to the detection score and the bounding box, a value indicating the pose/viewpoint is predicted for each detected object instances (see Figure 1.3). In this thesis we focus on the prediction of discrete object poses/viewpoints. For this reason, the output value indicating the pose/viewpoint is a discrete variable. Furthermore, given that our experiments are focused on cars, it is reasonable to assume the objects to rest on the groundplane. For this reason, we focus on predicting the azimuth angle along the viewing circle projected over the groundplane and ignore the elevation angle (see Figure 1.2).

1.2 Related Work

Some object categories can be easily recognized based on their material, color and texture, while others are characterized predominantly by their shape or appearance. Based on this particularity, Forsyth et al. [43] introduced the division of object categories into *Things* and *Stuff*. In the works covered in this thesis we focus on the prediction of object categories with defined shape and appearance, the *Things*. We propose methods that exploit relations and configurations between them as well as cues derived from the scene in order to predict their occurrence and pose. In this section we describe relations based on *Things*, *Stuff* and the scene to position the content of this thesis w.r.t. existing

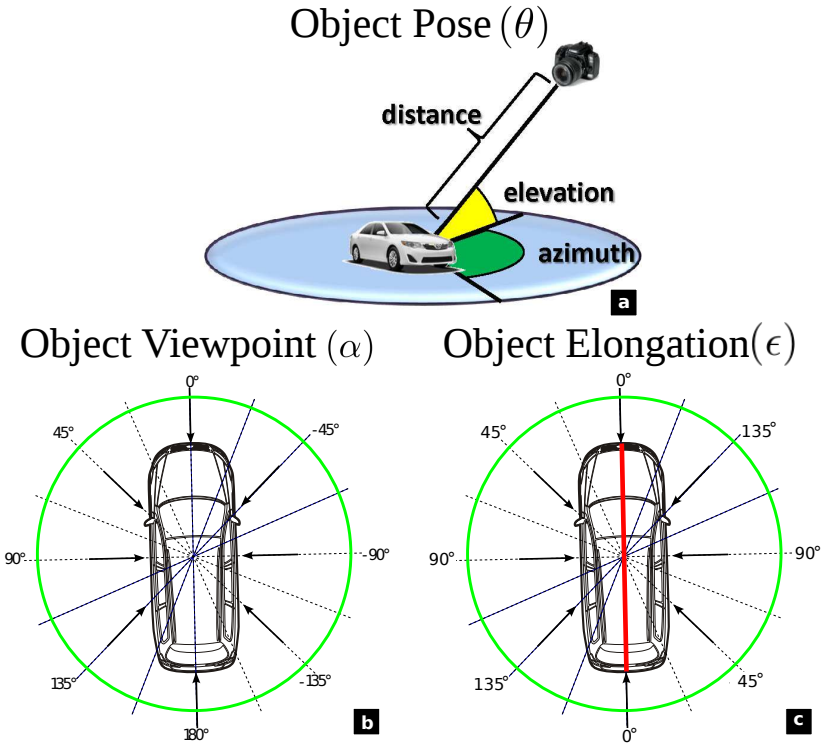

Figure 1.2: Object orientation angle. a) Azimuth and Elevation angles of the object in the 3D scene. b) Discrete viewpoint angles of the object in the viewing circle projected over the groundplane. c) Object Elongation angle as observed by the camera.

Figure 1.3: Car viewpoint estimation output. Note how in addition to the bounding box, a label indicating the predicted viewpoint is returned.

work. Moreover, this will serve as starting point to define the research questions addressed in this thesis.

1.2.1 *Scene* - *Things* Relations

This group of work starts from the observation that some object categories tend to occur more often in some particular environments than in others, e.g. a car is more likely to be found within an urban scene rather than in a nature scene. This observation provides a rich source of contextual associations that can be used as priors for the occurrence of particular object categories based on scene features [106]. Following this idea, [104, 105, 145] modeled the scene aiming to mimic the way in which humans can recognize a scene on a single glance. Towards this goal they proposed the *gist* descriptor which is a holistic representation of the scene that informs about its probable semantic category. This scene-centered representation was used as a proxy to constrain future feature analysis and provide a strong prior on the location of the object categories of interest. Parallel to this, [61] proposed a representation for urban scenes where the scene is defined by a common ground plane that supports all the object categories of interest. Consequently, object instances are considered to be consistent with the scene if they rest on the common ground plane. This served to constrain both the location and scale of the objects in the scene. Similarly, [9, 140] proposed a method to jointly reason about objects present in the image and the surfaces that support them. This relaxed the limitation of [61], where all the objects need to be supported by the same ground plane. More recently, [165] extended the previous ideas by shifting inference from the 2D image space to the 3D scene space obtaining more accurate results in the 3D scene.

As can be seen, work exploiting scene cues has mainly focused on the object detection task. In contrast, in this thesis we investigate several algorithms to compute scene-cues and use these cues to improve the performance of the object pose/viewpoint estimation task.

1.2.2 *Stuff* - *Things* Relations

In this group of work both *Things* and *Stuff* entities are assumed to be present in the images. Starting from this assumption, Heitz and Koller [59] proposed a method where they cluster image regions, defined by their color and texture, the *Stuff*, based on their ability to serve as contextual information for the detection of object categories, the *Things*. They showed that the obtained region clusters provide some level of interpretation while improving object detection. In [117, 119], the scene is modeled by computing the response of

texture and geometric features over the image. Polarity, contrast, and anisotropy are computed as texture features while the response of the geometric context classes [60] is considered as geometric cues of the scene. Then, in order to define *Stuff-Things* relations, for each of the object instances a log-polar sampling approach is applied on each of the response maps of the different cues. Finally, a classifier is trained using these relations. Recently, given the undefined shape characteristic of *Stuff* entities, a group of work [44, 92] opted to start from an initial image segmentation step to collect *Stuff* entities. Segments resulting from this initial step covered both *Stuff* and *Things* entities. From this point, [92] proposed an exemplar-based model that encoded both the relative appearance and 2D spatial relations between *Things* and *Stuff* entities. Parallel to this, [44] focused on the contextual interactions at different levels of the image. Pixel-level interactions were defined at the segments boundaries and aimed to separate *Things* entities from the background. Region-level interactions defined *Stuff-Things* relations within a bounding box defined for each *Thing* entity. Finally, object-level relations considered the co-occurrence along *Things* entities.

1.2.3 *Things* - *Things* Relations

In recent years, learning relations between *Things* entities has gained popularity in the computer vision community, particularly to assist the task of object detection. For the object detection problem, relations between object instances have been used to remove or reduce the uncertainty in hypotheses predicted by appearance-based detectors. A common pipeline in these works proceeds as follows: 1) an initial set of object hypotheses is obtained using an object detector; 2) for each hypothesis, a set of neighbor objects is selected as sources of contextual information; and 3) information from these neighboring objects is used to re-evaluate the initial object. Early work [30, 32, 118, 148, 156] represented objects as regions in the image. Then, by learning qualitative 2D spatial relations (e.g. top-left, far-left) between them, hypotheses in unlikely areas were filtered out. More recently, [78, 126] use discriminant relations between objects to learn the collective appearance of related objects in order to guide the detection of the individual objects. Reasoning about relations between "Thing" entities for the task of object pose estimation has received much less attention. Parallel to the work covered in this thesis, Xiang et al. [160] proposed a method that reasons about intrinsic features from the objects such as the appearance of patches taken from the object, and contextual cues such as 2D occlusion reasoning, to estimate the location and pose of the objects in the scene. Similarly, Zia et al. [165] used detailed shape representations based on CAD models. This representation improved model-object matching in the scene and, as a consequence, better reasoning about object support on

the ground-plane and mutual occlusion between objects. An important factor to consider, when reasoning about relations between objects, is the assumed nature of the relations that the algorithm is able to process. Until now, most existing work, e.g. [22, 30, 118], has assumed that relations between objects are pairwise in nature. As result, the proposed methods have mostly focused on reasoning about pairwise relations between objects. Recently, a small group of works [18, 162] that reason about higher order relations were proposed. A relation is considered to be higher order in nature if it is defined over more than two object instances, e.g. schooling in fishes, parked cars, queues, etc. In [18], a Pure-Dependency [65] framework is used to link groups of objects. In [108], objects are grouped by clustering pairwise relations between them. The work of [162] is able to reason about higher-order semantics in the form of traffic patterns. While slightly different to the previous group of work, this recent work still focuses on addressing the object detection task.

It is evident that there is a higher amount of work exploiting relations between objects for the task of object detection than for the task of object pose estimation. It is for that reason that in this thesis we will investigate methods to exploit relations between objects for the task of object pose estimation. Furthermore, within the group of work focused on detection, all relations between objects are considered despite the fact that some of these relations might be derived from noisy object hypotheses. This results in a rather crude, possibly inaccurate, way of reasoning about object relations. There is very little work [68] considering this aspect of the inference process, and it is mainly focused on the image segmentation task. Aiming to provide some insight to this matter, we will take steps to verify how the selection of related objects affects the relational inference process. Finally, we will present a method that shows the potential of reasoning about pairwise and higher-order relations between objects to recover object instances missed during the detection process. A more detailed look at related work will be provided in the respective incoming chapters.

1.3 Motivation and Research Questions

The past decade saw significant advances in the task of appearance-based object detection. One of the main driving forces of these advances is the Pascal Visual Object Classes Challenge [36], which over the lapse of ten years promoted the design of methods to tackle several computer vision-related problems, including object detection. It is important to notice, that despite these advances (and maybe in contrast to the easier problem of image/object classification), object detection is not a solved problem yet. As mentioned before, appearance-based methods have proved to perform well in some scenarios, e.g. when the object

of interest is fully visible and on relatively large scale, but this performance is affected by the presence of inter-object occlusions, low image resolution, etc. Nevertheless, in its current state, the available methods for appearance-based detection can serve to extract object hypotheses from which the top-scored ones are highly likely to be correct. These hypotheses can then serve to bootstrap the less certain ones.

Parallel to the advances on the object detection task, Collective and Relational Classification [134, 135]; the classification of networked data, has also achieved promising results in several applications. Examples of these applications include spam detection [76], suspicion scoring [90], social network analysis [87, 88], link prediction [28, 84], and web analysis [91]. In all these applications, collective classification proved to be effective at reasoning about interconnected nodes in large networks. Furthermore, by exploiting links, or relations, between nodes the methods for collective classification effectively identified ambiguous nodes. The ability of collective classification to perform reasoning on networked data in addition to its successful results, makes it a promising approach for the task of reasoning about detected object hypotheses.

In addition, recent work [22, 23, 30, 118] has shown that contextual information, in the form of relations between object instances can be exploited to improve object detection precision. This is achieved by using learned pairwise relations between objects in order to degrade object hypotheses that are out of context, and are likely to be false positive predictions. Inspired by these works, we experiment with relations between objects for improving object pose estimation. Furthermore, this type of contextual information from two different perspectives. On one hand, we verify the influence of selecting different sets of objects as sources of contextual information. On the other hand, we verify the effect of considering a different type of relations that associate the objects.

Last but not least, we can observe that there are higher order relations that play an important role in the occurrence of certain object categories. On the one hand, work in psychology and behavioral sciences has studied the influence of groups of animals in the behavior of new participants [25] and has showed that this herding behavior is even present in humans [121]. On the other hand, rules have been established of how certain man-made objects should be arranged in "desirable" or "permitted" configurations, i.e. cars should be parked aligned next or behind each other.

These four factors drive the work presented in this thesis where we evaluate the benefits brought by contextual information for the task of detecting and predicting object poses while exploring different factors that affect the performance of using relations between objects as contextual information.

1.3.1 Research Questions

As was evident in previous sections, the potential of some types of contextual information to improve object pose estimation performance has not been explored. Moreover, for the case of the object detection task, contextual information has mainly been used for filtering out false object hypotheses, thus, only providing improvements in terms of precision.

Taking these observations as starting point, the objective of this thesis is to investigate methods to reason about relations between objects and relations between the objects and the scene as means to improve object detection and object pose estimation performance. To this end, this thesis addresses the research question:

Can contextual information improve the performance of vision tasks?

There are several factors that can be considered when integrating contextual information in vision tasks, e.g. specific vision tasks to be addressed, sources of contextual information to be considered, methods for context-based reasoning, etc. For this reason, and for clarity of presentation, we split the main research question into three questions to address specifically some of these factors.

1. Is contextual information, in the form of relations between objects, useful for object pose estimation?
2. To what extent does the nature of the association between objects affect the performance of using relations between objects to improve object detection?
3. Is contextual information, in the form of scene-driven cues, useful for the task of object viewpoint estimation?

The journey aimed at answering these research questions resulted in the following contributions.

1.4 Overview and Contributions of the thesis

The line of work presented in this thesis is focused on reasoning about contextual information to assist computer vision tasks such as object detection and object pose estimation. The work covered in this thesis has been disseminated across

several papers. The content and contributions of these papers are presented in chapters 3 to 6. As a whole, the contents of these papers address the research questions introduced earlier with each chapter having specific contributions.

In Chapter 2, we present fundamental principles and tools used in the different methods presented in the thesis. In addition, we introduce some standard machinery from computer vision and machine learning for completeness. The objective of this chapter is to lay the foundations for the rest of this thesis.

In Chapter 3 we show that considering configurations between objects can be beneficial for pose estimation. To this end, we provide a more detailed literature review on context-based object pose classification. Then, we propose a novel approach based on a relational neighbor framework [91] that reasons about pairwise relations between objects with the objective of predicting object poses. Our experiments show that the proposed context-based method is able to complement state-of-the-art methods for local pose estimation. In addition, we provide an analysis of the effect of the frame of reference when defining relations in the 3D space. Our analysis showed that defining relations between objects from an object-centered perspective can increase the performance of object detection and object pose estimation. Furthermore, our results indicated that reasoning about object overlap is more effective when done in the 3D scene rather than in the 2D image space. This observation was later confirmed in [160, 165, 166]. These findings essentially show that the answer to Research Question 1 is positive. The contents of this chapter is based on the following publication:

- Oramas M, J., De Raedt, L., and Tuytelaars, T. *Allocentric pose estimation.* In IEEE International Conference on Computer Vision (ICCV) 2013.

Chapter 4 addresses the problem of object detection and bring in contextual information in the form of relations between objects [108]. In this chapter, we evaluate cautious algorithms that consider the most certain relational information first and use it to bootstrap less certain object hypotheses. In addition, we analyze the nature of the association between objects. To this aim we evaluate the changes in detection performance when assuming that objects are associated purely based on their category, and when assuming that objects are associated by underlying "relationships" that explain the object co-occurrence patterns that we see in images. As main contribution we show that cautious inference on object relations brings improvements over traditional, aggressive, inference for object detection. In addition, the proposed method for relationship-driven object association constitutes an early step towards reasoning about higher-order relations, since the "relationships" that drive object associations are not necessarily pairwise. This aspect partially answers

Research Question 2, which analyzes how the association between objects affects object detection performance. The contents of this chapter are closely related to the publication:

- Oramas M, J., De Raedt, L., and Tuytelaars, T. *Towards cautious collective inference for object verification.* In IEEE Winter Conference on Applications of Computer Vision (WACV) 2014.

Complementary to Chapters 3 and 4, in Chapter 5 we approach object pose estimation from an alternative perspective. Instead of bringing in contextual information in the form of relations between objects, we define context in the form of object-scene relations [109] (see Section 1.2.1). As contributions of this chapter, we introduce an intermediate step, towards viewpoint classification, which we refer to as object elongation orientation classification. To this end, as first step we classify the angle of the elongation orientation of an object, which can be estimated relatively accurately. Then, as second step, we use this elongation angle to classify the viewpoint of the object. In addition, we propose different top-down approaches to extract scene consistent cues that can be later integrated to appearance-based viewpoint classifiers. Our experiments show that considering the proposed scene-driven cues brings improvements on object viewpoint estimation. This chapter gives a positive answer to Research Question 3 by showing that considering contextual cues taken from the scene improves the performance of object viewpoint estimation. The content of this chapter is based on the publication:

- Oramas M, J., and Tuytelaars, T. *Scene-driven cues for viewpoint classification of elongated object categories.* In British Machine Vision Conference (BMVC) 2014.

In Chapter 6, we focus on improving object detection performance in terms of recall. We propose a post-detection stage during which we explore the image with the objective of recovering missed object instances. This exploration is performed by sampling object proposals on the image. Contributions of Chapter 6 focus on reasoning beyond pairwise relations between objects and effectively improving the object detection recall. First, we propose a novel way of using contextual information to improve object detection performance in terms of recall. To this end, we show that given a set of object hypotheses collected after an initial detection step, we can use relations between objects to recover missed object instances. Second, we propose a novel method based on Topic Models [15, 52] for discovering high-order relations between objects. We show that our method is able to discover object arrangements as those found on

traffic patterns on urban scenes. This, in consequence, provides some level of interpretation to the images. The findings of this chapter complement those of Chapter 4 and provide an answer to Research Question 2. This answer is that considering underlying relationships increase the performance of relations-based methods for object detection. Furthermore, due to the presence of noise in vision-based tasks, e.g. false hypotheses, caution should be exercised when using object relations as source of contextual information. The content of this chapter is based on the article:

- Oramas M, J., and Tuytelaars, T. *Recovering hard-to-find object instances by sampling context-based object proposals*. Submitted to IEEE International Conference on Computer Vision (ICCV) 2015.

Chapter 7 concludes the thesis. This chapter begins by presenting a summary of the results obtained and lessons learned during the execution of the work covered by the thesis. Then, we provide a critical analysis about the limitations of this work. Finally, we suggest directions for future research.

Chapter 2

Background

The work presented in this thesis is related to a variety of problems in computer vision and machine learning. In this chapter we present the concepts on top of which the methods proposed in the following chapters build. We start by introducing traditional methods for object detection based on intrinsic appearance features, e.g. color, texture, gradients, etc. Then, we present some related work that exploits contextual information to improve object detection. We conclude this chapter by introducing machinery used throughout this work to reason about relations between object instances.

2.1 Local Appearance-based Object Detection

2.1.1 Histograms of Oriented Gradients

Histogram of Oriented Gradients (HOG) [26] are feature descriptors used in computer vision to describe the appearance of objects and other content in images. It is usually computed on a dense grid of uniformly spaced cells over an image region.

The computation of this descriptor can be summarized in three steps. Given an input image, the first step consist in the computation of the gradient values. A classical procedure to achieve this is by filtering the color or intensity data of the image with 1D filter kernels, e.g. $[-1, 0, 1]$ and $[-1, 0, 1]^T$. The second step involves dividing the image, or image window, into smaller regions called cells. Then a histogram is assigned to each cell where each bin corresponds

to a discrete orientation within the cell. Each pixel within the cell casts a weighted vote for a bin using the values obtained in the gradient computation step. Finally, in the third step a normalization is performed to account for changes in illumination and contrast. This requires grouping the cells together into larger, spatially connected blocks. See Figure 2.1 for a visualization of HOG descriptors computed over an image. The HOG descriptor is then the vector resulting from the concatenation of the components of the normalized cell histograms from all of the block regions. These blocks typically overlap, meaning that each cell contributes more than once to the final descriptor.

Since their proposal in [26], HOG descriptors have become the backbone of other computer vision methods [5, 40, 99, 155, 160]. In this thesis, they are used within the object detectors to describe the appearance of the object categories of interest, in this case cars.

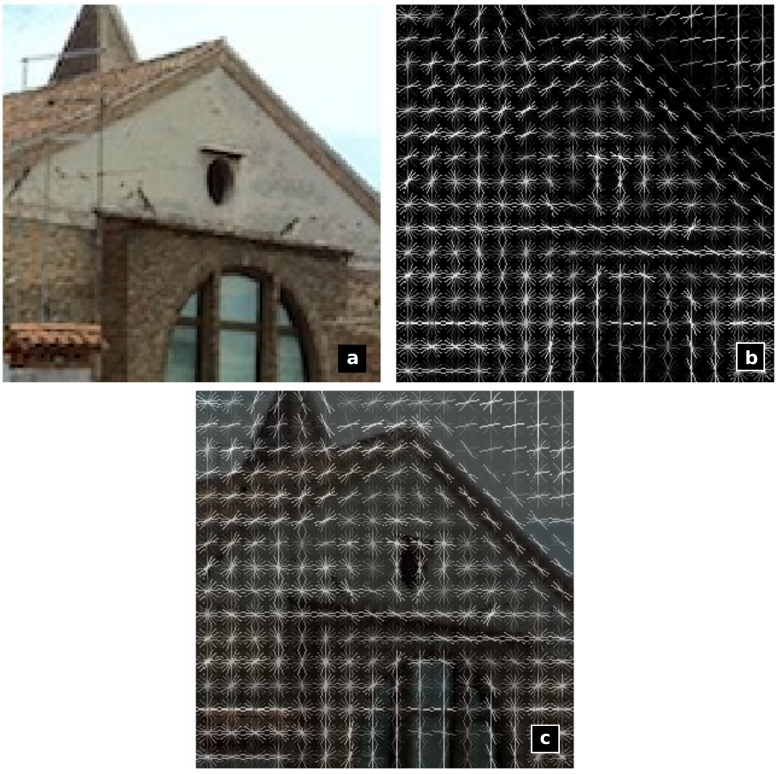

Figure 2.1: Histogram of Oriented Gradients (HOG) descriptor [26]. a) Input image. b) HOG descriptor visualization. c) HOG visualization overlapped on the input image.

2.1.2 Rigid Models for Object Detection

A very popular method to perform object detection is the *sliding window approach*. This approach consists in identifying the 2D regions of the image that are likely to contain instances of the categories of interest by the exhaustive evaluation of all windows in the image. These 2D windows are generated densely over the image in order to detect objects occurring at different locations within the image. Furthermore, these windows are generated at different sizes to cope with the different scales in which instances of the object categories of interest may occur. In addition, the windows may be generated with different aspect ratios which depend on the object category to be detected. During training, given a set of training images containing the objects of interest annotated with bounding boxes, the appearance of the object is modeled. In addition, information about common scales and aspect ratios is stored for later usage during the window-generation process. At test time, inference becomes a binary classification problem where the objective is to distinguish the object of interest from the background. See Figure 2.2 for a sliding window example for face detection.

Figure 2.2: Sliding window for face detection.

A landmark work following this approach, is the work from Dalal and Triggs [26] which used image gradients within the window to describe appearance. Specifically, Histograms of Oriented Gradients (HOG) (see Figure 2.3.b). During training, HOG descriptors are computed for all the bounding boxes of the annotated objects (Figure 2.3.a,b). These descriptors constitute the positive examples. Then, windows are randomly generated taking care not to overlap the annotated objects, HOG descriptors are computed on these windows and are considered the negative examples. As classifier, a binary Support Vector Machine (SVM) is trained using a linear kernel. As can be seen in Figure 2.3.c the linear SVM effectively learns how to weight the HOG descriptors (Figure 2.3.b) to model the appearance of a pedestrian (Figure 2.3.a). This method proved to be robust for (more or less) rigid object categories in which changes of appearance were mainly attributed to changes in illumination, viewpoint or scale. On the contrary, for the case when object instances showed larger changes in appearance due to changes in the shape of the objects, the performance of this method was suboptimal.

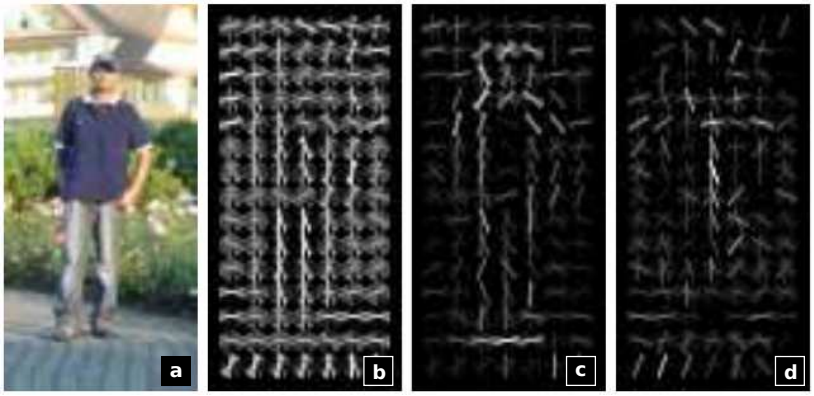

Figure 2.3: Dalal-Triggs, rigid model, detector [26]. a) input image patch, b) computed HOG descriptor over the image patch, c) and d) HOG descriptor weighted by the positive and negative SVM weights. Example taken from [26].

2.1.3 Deformable Models for Object Detection

Within the object detection literature, there are some object categories that are labeled as deformable categories. This covers object categories where the appearance of instances of that class changes across images due to changes on the shape of the object itself. This is a common feature of flexible or particles-based categories, e.g. fire, clouds, snakes, etc; or of limb-composed categories,

e.g. hands, human body, etc. As mentioned earlier, for this type of deformable objects, the performance of rigid models is suboptimal. To address this problem, a Pictorial Structure framework [42] was proposed to identify the parts of the object that produced the deformations. This was followed by learning the appearance and relative displacement of the parts with respect to a root location within the object window. The goal of this framework is to maximize the object's appearance likelihood by allowing some level of deformation on its parts. A landmark work based on the Pictorial Structures framework is the method from Felzenzswalb et al. [40, 112]. This work complements the Pictorial Structure framework by modeling appearance of the root template and parts of the object using HOG descriptors (see Figure 2.4.b). During training, root templates are modeled similar to the rigid model case with the addition that the common displacements of a pre-defined number of parts is learned (see Figure 2.4.c). In addition, several components are trained for each model in order to cope with drastic changes in appearance possibly caused by articulations, or changes in object viewpoint. The method from [40], and following releases, are commonly referred to in the computer vision literature as Deformable Parts Model (DPM) detectors.

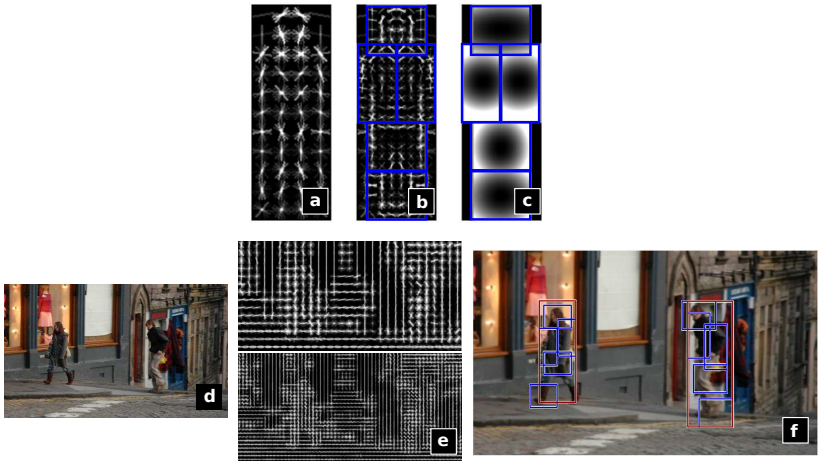

Figure 2.4: Deformable parts model detector [40]. Top row, model learned for pedestrian detection: a) HOG template of the root filter, b) identified parts, and c) part displacement. Bottom row, pipeline during testing: d) input image, e) HOG descriptors computed over the image at 2 different scales, and f) detected instances. Example taken from [40].

Since the focus of this thesis is on measuring the effect of contextual information to assist object detection and object pose estimation, instead of designing

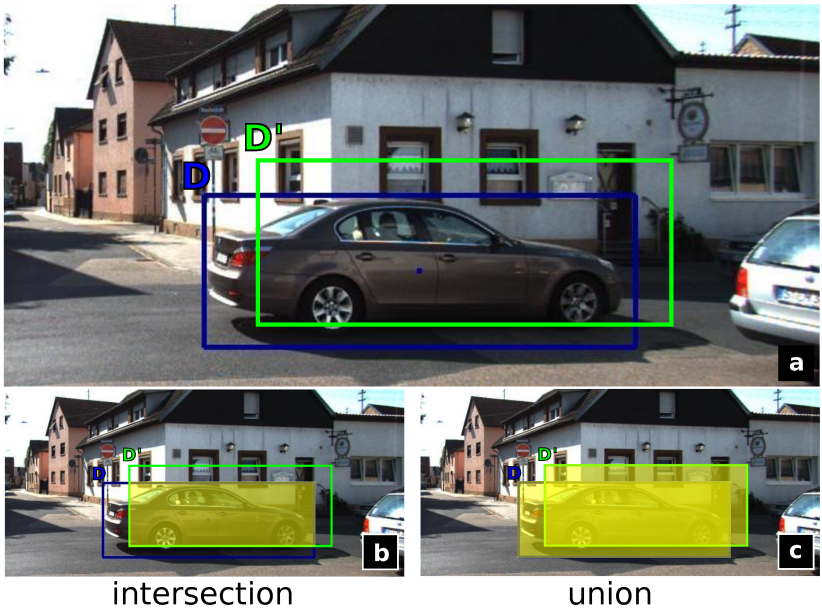

Figure 2.5: Pascal VOC intersection-over-union matching criterion. a) Object detection D and D'. b) Intersection area (yellow) of D and D'. b) Union area (yellow) of D and D'.

our own object detector we use standard off-the-shelf detectors. Specifically, throughout the methods proposed in this thesis we use DPM detectors [40] [47][81] to collect object hypotheses.

2.2 Non-Maximum Suppression

As can be noted in Figure 2.2, given the dense characteristic of the sliding window approach there is a very high possibility of obtaining multiple windows as candidate match for each instance of an object. In order to cope with this problem a greedy post-detection step is usually applied. This step is *non-maximum suppression (NMS)*. Given a set of detections, defined by a bounding box and a score, the detections are sorted by score, and the detections with highest score are greedily selected while skipping detections with bounding boxes that are "covered" by a bounding box of a previously selected detection. A common method to measure the level of "coverage" or matching between the bounding boxes of two detections is the intersection-over-union criterion from

the Pascal VOC Challenge [36]. Given the bounding boxes of two detections D and D' (Figure 2.5.a), these two detections are said to match or overlap if:

$$\frac{area(D \cap D')}{area(D \cup D')} > \tau \tag{2.1}$$

where the numerator of this fraction is the intersection of bounding boxes of detections D and D' (Figure 2.5.b). The denominator is the union of both bounding boxes (Figure 2.5.c). The threshold τ is usually set to 0.5 to define that two bounding boxes match if they overlap at least by 50%.

2.3 Multiview Object Recognition

The viewpoint α of an object is an important factor when modeling the appearance of a specific category of interest. As stated earlier, the 2D appearance of an object can be affected by the viewpoint from which the object of interest is observed. Based on this observation, taking into account the viewpoint of an object is of interest for two computer vision tasks. For the object detection task, reasoning about object viewpoints is useful to achieve robustness to possible changes in appearance caused by changes of viewpoint. These type of detectors are usually referred to as multiview (or viewpoint-invariant) detectors. Examples of this type of detector are [115, 143, 158]. The other computer vision task is object viewpoint estimation for which reasoning about viewpoint-specific appearance is essential.

The problem of object viewpoint estimation has been traditionally addressed under the assumption that the visual features, e.g. color, texture or gradients, projected by an object on an image differ between viewpoints. The problem is then to define proper descriptors to represent such visual features and to find a method to distinguish between the descriptors from each viewpoint. Based on this assumption, a classical paradigm towards object viewpoint estimation is to formulate the viewpoint estimation problem as an object detection problem, where each discrete viewpoint α to be predicted is considered as a separate category. To this end, an appearance model is trained for each of the discrete viewpoints (see Figure 2.6). During testing, each appearance model is used together with a sliding window to detect objects in different viewpoints. For the case of overlapping hypotheses, the viewpoint is classified by selecting the model with the highest response. In [47, 81, 115] we can find some examples of methods following this paradigm. The methods from [47, 81] are inspired by the DPM detector, where the 2D appearance of each viewpoint to be predicted is linked to a specific component within the model. During testing the viewpoint of the

component with maximum response is selected as predicted object viewpoint. The method from [114] takes this idea further and defines the object region and parts as 3D cuboids. Then, by aligning these 3D parts with 3D CAD models their model is able to maintain consistency between part positions across different viewpoints while being competitive with state-of-the-art methods for viewpoint estimation. Methods following this paradigm showed to be accurate at classifying object viewpoints in scenarios with low object occlusion where the object is clearly visible. However, on the downside, they inherit the weakness of their object detector nature in the sense that, in addition to the dense windows at multiple scales, they need to evaluate a higher number of models as the viewpoints to be classified get finer.

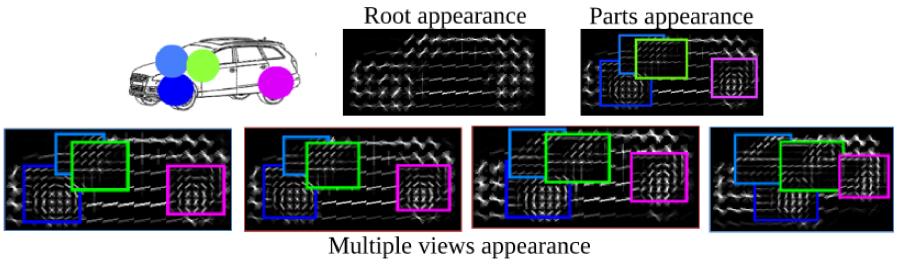

Figure 2.6: 3D2PM detector models [114]. Top row, model related to the root and part appearance of a specific viewpoint angle. Bottom row, models from different viewpoints. Note the displacement of the parts over the different viewpoints. Example taken from [114].

In this thesis we focus on the study of using contextual cues to improve object-centric tasks, i.e. object detection and object pose estimation. Furthermore, one of the contextual cues that we analyze is the one defined by relations between objects. It is for this reason that methods for object detection, like the ones presented above, are used to collect the initial set of object hypotheses that will be enriched with contextual cues. More specifically, we focus on viewpoint-aware object detectors, i.e. methods that provide information about the localization and viewpoint of object instances in images. In the next section we give more details about the specific object detectors used in next chapters.

2.4 Object Detection

In most of the work presented in this thesis, object hypotheses are obtained with detectors derived from the Deformable Parts Model (DPM) detector [40]

(see section 2.1.3). The DPM detector [40] is an object detector that models the appearance of an object category based on the intrinsic features, e.g. color and gradients, that define it. Since the main focus of this thesis is on reasoning about contextual features, we will give the task of describing objects by local, intrinsic, features to standard off-the-shelf detectors. Specifically, we use three detectors: a) The release 5 of the original DPM detector [40] (see section 2.1.3), b) the mDPM detector from [81], and c) the LSVM-MDPM-sv detector used in [47]. These detectors feed our methods with object hypotheses with their corresponding bounding boxes and confidence scores (see Figure 1.1). As stated earlier, theses hypotheses are predicted evaluating only intrinsic features of the object categories of interest. Furthermore, mDPM and LSVM-MDPM-sv are viewpoint-aware detectors, i.e. they predict the viewpoint of the object in addition to its bounding box and confidence score (see Figure 1.3). This is achieved by training a specific component of the DPM for each of the discrete object poses to be predicted. This effectively learns the appearance features of an object category w.r.t. its viewpoint label (Section 2.3). Although mDPM and LSVM-MDPM-sv are DPM detectors where each mixture component corresponds to a specific viewpoint there is a small difference in how they generate false examples when training their respective classifiers. LSVM-MDPM-sv follows the traditional DPM guideline [40] of sampling image patches from images that do not contain instances of the object category. mDPM extends this procedure by adding images of objects instances in an opposite viewpoint to the one of the component being trained. Given the characteristic of these detectors to learn from intrinsic object features, they constitute the local classifiers that we use in the work covered by subsequent chapters.

2.5 Kernel Density Estimation

Given a set of object instances in an image, we are free to define pairwise relations between all the instances. However, there are some relations that occur more often than others. This suggest that a mechanism to model the distribution of these relations is necessary for a proper reasoning about relations between objects. In this thesis we use Kernel density estimation (KDE) to achieve that objective.

Kernel density estimation is a non-parametric method to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. Formally speaking, given a set $(x_1, x_2, \ldots, x_n)$ of independent and identically distributed samples drawn from

some distribution with an unknown density f, we are interested in estimating the shape of this function f. Its kernel density estimator is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{2.2}$$

where $K(\cdot)$ is the kernel, a non-negative function that integrates to one and has mean zero and $h > 0$ is a smoothing parameter usually referred to as the bandwidth. A kernel with subscript h is called the scaled kernel and defined as $K_h(x) = \frac{1}{h}K(\frac{x}{h})$. A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, Gaussian, and others. In this work, we will adopt Gaussian kernels. Intuitively one wants to choose a bandwidth h as small as the data allow, however there is always a trade-off between the bias of the estimator and its variance.

Kernel density estimates are closely related to histograms, but can be endowed with properties such as smoothness or continuity by using a suitable kernel. For instance, in Figure 2.7 we show a histogram and a kernel density estimate computed from the same sample points x_i. As can be noted, there is a boundary effect introduced by the discontinuities between the bins of the histogram. In addition, the selection of number of bins, and their widths, of the histogram is not a trivial problem. Moreover, these parameters directly affect how the underlying distribution is perceived. On the contrary, by using a smoothing kernel at each sample point and estimating the bandwidth in a data-driven fashion, kernel density estimates converge faster to the true underlying density for continuous random variables [132] .

2.5.1 Bandwidth Selection

The bandwidth h of the kernel is a free parameter which exhibits a strong influence on the resulting estimate. When h is small, each training instance has a large effect in a small region and close to no effect on distant points. In the case when h is large, there is more overlap between neighboring kernels which results in smoother estimates. Over the years, several methods have been proposed for bandwidth selection, e.g. Maximum likelihood cross-validation [33, 55], Unbiased cross-validation [16, 124], Biased cross-validation [131], Silverman's Rule of Thumb [154]. Each method has its respective strengths and weaknesses.

In the work presented in this thesis we adopt Silverman's Rule of Thumb [154] for bandwidth selection. This is a plug-in method that defines the "rule-of-thumb" bandwidth $\hat{h}_{rot}$ as follows:

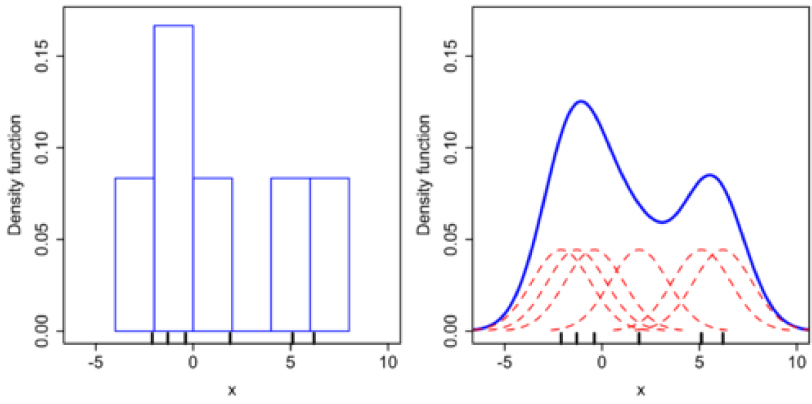

Figure 2.7: Comparison of a 1D histogram (left) and a kernel density estimate (right) constructed from the same data. The dashed curves represent the kernels while the continuous curve represent the kernel density estimate.

$$\hat{h}_{rot} = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \approx 1.06\hat{\sigma}n^{-1/5} \tag{2.3}$$

This is an explicit, applicable formula for bandwidth selection where $\hat{\sigma}$ is the standard deviation of the sample points x_i. In practice, if the underlying distribution is normally distributed, then Equation 2.3 gives the optimal bandwidth. If not, it will give a bandwidth not too far from the optimum if the distribution is not too different from the normal distribution (the "reference distribution").

A practical problem with the rule-of-thumb bandwidth is its sensitivity to outliers. A single outlier may cause a too large estimate of $\hat{\sigma}$ and hence implies a too large bandwidth. A more robust estimator is obtained from the interquartile range $R = x_{0.75n} - x_{0.25n}$ with $x_{0.75n}$ and $x_{0.25n}$ being the 75% and 25% quantiles, respectively. Based on the interquartile range R, the "rule-of-thumb" bandwidth $\hat{h}_{rot}$ (Eq. 2.3) is re-defined as [154] :

$$\hat{h}_{rot} = 1.06 \min\left\{\hat{\sigma}, \frac{R}{1.34}\right\} n^{-1/5} \tag{2.4}$$

Fixed vs. Variable bandwidth

A popular practice during the density estimation process is to use a fixed bandwidth value h for all the sample points x_i of Eq. 2.2. This setting is usually referred to as *fixed* kernel density estimation. Diverging from this fixed setup for kernel density estimation, a group of work [38, 56, 127] has pointed out the potential gains in modifying the shape of $\hat{f}_h$ to better adapt to local conditions of the data. This is achieved by modifying the bandwidth value h. Two variants have been proposed for modifying the bandwidth value h: the *balloon*, or local, estimator [150] and the *variable*, or sample point adaptive, kernel estimator [17]. For the case of the *balloon* estimator [150], the bandwidth depends directly upon the point x of estimation. This redefines Eq. 2.2 as:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h(x)}(x - x_i) = \frac{1}{nh(x)}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h(x)}\Big), \tag{2.5}$$

Several works, e.g [37, 58, 98], have proposed methods to compute the bandwidth $h(x)$. However, simulation studies [39] assessing the finite sample performance of balloon estimators revealed that fixed plug-in bandwidth selectors had a similar performance on when using up to $n = 1000$ samples.

For the case of *variable* kernel density estimation [17], a specific bandwidth value $h(x_i)$ is defined for each of the sample points x_i. The main goal of this method is to smooth less where there is more structure and vice versa. Applying this principle to the vanilla kernel density estimation results in Eq. 2.6.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h(x_i)}(x - x_i) = \frac{1}{nh(x_i)}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h(x_i)}\Big), \tag{2.6}$$

2.5.2 Multivariate Kernel Density Estimation

Kernel density estimation can easily be generalized from univariate to multivariate data. Formally speaking, the general form of the estimator is defined as:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H^{(d)}(x - x_i) = \frac{1}{n|H|}\sum_{i=1}^{n} K^{(d)}\Big(\frac{x - x_i}{H}\Big), \tag{2.7}$$

where $|H|$ is the absolute value of the determinant of the matrix H, which is a non-singular $d \times d$ bandwidth matrix. The kernel function $K^{(d)} : \mathbb{R}^d \to \mathbb{R}$ is often taken to be a d-variate probability density function. A common technique for generating $K^{(d)}$ from a univariate kernel K is by using product kernels,

$$K(u)^{(d)} = \prod_{j=1}^{d} K(u_j) = \prod_{j=1}^{d} K_{h_j}(x_j - x_{ij}) = \prod_{j=1}^{d} K\Big(\frac{x_j - x_{ij}}{h_j}\Big) \tag{2.8}$$

Here u is a d-dimensional argument of $K^{(d)}$ with a specific value u_j for each dimension j. Note how in the last part of Equation 2.8 we focus on the values of each dimension-specific estimator (sample points x_{ij}, the bandwidth h_j and point to be evaluated x_j). Following the product kernels from Equation 2.8 redefines Equation 2.7 as:

$$\hat{f}(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K^{(d)}\Big(\frac{x - x_i}{H}\Big) = \frac{1}{n|H|} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} K\Big(\frac{x_j - x_{ij}}{h_j}\Big) \right\}. \tag{2.9}$$

The matrix bandwidth H has a total of $d(d+1)/2$ entries (given that H is symmetric). This number becomes unmanageable very quickly, as the number of dimensions d in the data increases. It is for this reason that H is restricted to take a simpler form. Three common possibilities to define this matrix are:

1 $H = hI$. Setting the bandwidth to be constant for every variable, which results in the same amount of smoothing along each dimension. This is only applicable if the scales of the variables are comparable.

2 $H = diag(h_1, ..., h_d)$. This allows a different amount of smoothing for each dimension.

3 $H = hS^{\frac{1}{2}}$. Here S is an estimate of the covariance matrix of x. This can be seen as a multivariate generalization of coordinate-wise scaling.

Unfortunately, none of these approximate methods is rich enough to be able to handle all possible density shapes. Using a diagonal H matrix is often good enough, although sometimes a full matrix is necessary. In this thesis, we use the second method for defining the bandwidth matrix so $H = diag(h_1, ..., h_d)$.

2.5.3 Online Kernel Density Estimation

Online Kernel Density Estimation (oKDE) [73] is a variant of variable Kernel Density Estimation aimed at online estimation, i.e. the construction of models from continuous data streams. Due to its goal of performing online estimation, it has the characteristic of producing models with significantly lower complexity. This estimator is based on two key ideas. First, instead of building and keeping a model of the target distribution $f(x)$, a non-parametric model of the data itself (the sample points x_i) is stored in the form of a *sample distribution*. This model is used when estimation of the target distribution $f(x)$ is required. Second, each new observation (sample point x_i) is treated as a Dirac-delta function and is mixed with Gaussian functions to update the *sample distribution*. The update of the *sample distribution* can be summarized in three steps. Given a new observation x_i, the first step is to update the sample model with the new sample point. Then, during the second step, the updated sample model is used to recalculate optimal bandwidth values h. Finally, in the last step, the *sample distribution* is refined and compressed. This is achieved by replacing clusters of the components (sample points x_i within the estimator) by a single Gaussian component. This produces a good balance between complexity and generalization of the observed data points.

Later this method was extended to perform supervised online estimation of probabilistic discriminative models for classification tasks [72]. The idea is to be able to construct models on-the-fly in a supervised fashion so that the constructed models can be used later for classification. A straight forward step to achieve this online construction of classifiers would be to estimate the target distribution for each class using oKDE. Then, a Bayesian classifier can be constructed from these distributions. However, as pointed out in [72], oKDE is agnostic to the fact that target distributions are being estimated for the construction of classifiers. As a result, discriminative features of the data are not maintained during the estimation of the target distributions. To address this problem, oKDE is extended by constructing class-wise distributions in the form of Gaussian mixture models by taking into account all the classes jointly. When a new sample point x_i is observed, each of these distributions are updated independently. However, in order to retain the discriminative properties of each class, all the distributions are compressed jointly using a function to measure loss of interclass discrimination during compression. This extended model is referred to as the online discriminative Kernel Density Estimation (odKDE).

As said earlier, in this thesis, the main use of Kernel Density Estimation is to model the distribution of pairwise relations between objects (Chapters 3, 4 and 6). In Chapter 5, we use oKDE to model the distribution of object instances, with their corresponding attributes, over the scene. Since in Chapter 5 no

online estimation is required, we apply low compression and construct the initial estimator from the whole set of training examples. In consequence, we only keep its variable multivariate properties for kernel density estimation.

2.6 Topic Models

In machine learning and document analysis, a topic model [15] [52] is a type of statistical model for discovering the abstract topics that occur in a collection of documents. Intuitively, given that a document is about a particular topic one would expect particular words to appear in the document more or less frequently. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about topic A and 90% about topic B, there would probably be about 9 times more B-related words than A-related words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering the topics that might be covered in them. This discovery of topics, and their proportion, within the documents is achieved based on the statistics of words in each document.

Formally speaking, a document d_i can cover multiple topics t_k and the words w_j that appear in such a document reflect the set of topics t_k that it covers. From the perspective of statistical document analysis, a topic t_k can be viewed as a distribution over words w_j; likewise, a document d_i can be considered as a probabilistic mixture over the topics t_k. Based on this we can estimate the probability of a word w_j given document d_i as:

$$p(w_j|d_i) = \sum_{k=1}^{T} p(w_j|t_k)p(t_k|d_i) \tag{2.10}$$

where T is the total number of topics t_k, the term $p(w_j|t_k)$ represents the probability of word w_j appearing under topic t_k, and $p(t_k|d_i)$ is the probability of document d_i covering topic t_k. These terms are computed following Latent Dirichlet Allocation (LDA) [15] as presented in [52]:

$$\begin{aligned} p(w_j|t_k) &= \frac{n_k^{(w_j)} + \beta}{n_k^{(\cdot)} + W\beta} \\ p(t_k|d_i) &= \frac{n_k^{(d_i)} + \alpha}{n_{\cdot}^{(d_i)} + T\alpha_{\cdot}} \end{aligned} \tag{2.11}$$

where W is the total number of words w_j. The terms α and β are hyperparameters which specify the nature of the priors on $p(t_k|d_i)$ and $p(w_j|t_k)$, respectively. Note that in this formulation, the term α is not referring to the viewpoint angle of an object instance, as was stated in Section 1.1.2. Finally, $n_k^{(w_j)}$ represents the number of times word w_j has been assigned to the topic t_k. Similarly, $n_k^{(d_i)}$ is the number of times a word from document d_i has been assigned to topic t_k. The term $n_k^{(\cdot)}$ is the total number of words that has been assigned to the topic t_k while $n_{\cdot}^{(d)}$ is the number of words in document d_i. Please refer to [52] for more details about the approach that we employed for topic modeling.

In Chapter 6, we explore the assumption that the association between object instances is driven by underlying relationships that explain how objects co-occur in the scene. To this end, we use a topic model to discover these underlying relationships (topics) from a set of observed pairwise relations (words) between objects.

2.7 Collective Classification

At this point we have presented methods that can be used to localize and predict the viewpoint of object instances in images. Furthermore, in the previous sections, we proposed Kernel Density Estimation and topic models as means to model the distribution of relations between object pairs. However, there is the possibility that more than two objects could be present in an image. Moreover, some of these objects might be false hypotheses. For these reasons, a method is needed to classify each object instance by taking into account its local features as well as its relations with the other objects in the image. In this thesis, we adopt Collective Classification for such purpose.

Collective classification is a combinatorial optimization problem, in which, given a set of possibly connected nodes in a graph, the objective is to classify all the nodes. More formally, given a set of nodes $V = \{v_1, ..., v_n\}$ interconnected by links $L = \{l_{12}, l_{13}, ..., l_{n(n-1)}\}$, where the link l_{ij} goes from node v_i to v_j, our task is to assign to each of the nodes v_i a class label from the set $C = \{c_1, ..., c_k\}$. In the machine learning literature we can find several methods [28, 54, 71, 91, 123, 141] to address this objective. First, the classification problem can be addressed from a very local perspective where only the observed intrinsic features or attributes that define the class of the node are considered. Through the rest of this thesis, we refer to this classification, based on intrinsic features, as *local classification*. A second method to address the given classification problem, is to learn how to predict the class of a node by considering the attributes

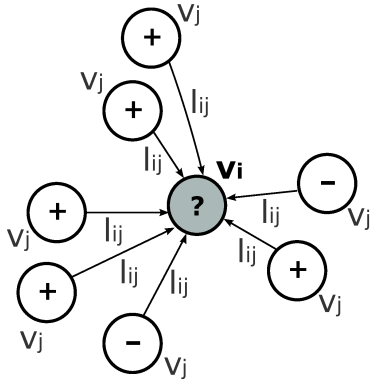

Figure 2.8: Collective Classification. The classification of a particular node v_i (in gray) is based on the links/relations l_{ij} with its neighboring nodes v_j.

of observed nodes in the neighborhood of the node to be classified. In the machine learning literature, the combination of the previous two methods for classification is sometimes referred to as *relational classification*. Furthermore, the problem of classifying unknown nodes in a partially labeled graph is referred to as the Within-Network classification problem [31]. Similarly to the second method, we can define a third approach in which a node is classified based on the attributes of unknown nodes on its neighborhood, without using intrinsic features. Collective classification refers to classification methods which combine the three methods presented above. As illustrated in Figure 2.8, the class label assignment step is performed in such a way that consistency is enforced between the node v_i being classified w.r.t. the set of nodes v_j on its neighborhood N_i.

This collective method for classification has proven to be beneficial in problems where links or relations rise naturally on the domains to be classified. For example in social networks [19, 85, 86], users define links between themselves based on common topics of interest, family affinity, friendship, etc. Similarly, in the domain of bibliographic data [24, 49, 91], authors position their work by referring to existing work that have some relationship of either similarity, dependence, or contrast with the work being presented. In both domains, the nodes, either contacts within a social network or related articles in bibliographic data, to which an unknown node is related provides a strong indication about the class or type of the unknown node.

Several methods have been proposed to provide exact inference for collective classification, e.g. the junction tree algorithm [66] and variable elimination [29] in probabilistic models. However, there is no guarantee that real-world networked data will satisfy the required conditions that make exact inference

tractable in these algorithms. For this reason, in this thesis we will focus on algorithms for approximate inference for collective classification. Following [135] we divide algorithms for approximate inference into two groups: algorithms based on local classifiers, and algorithms based on global formulations. As its name suggests, algorithms based on local classifiers, depend on a set of local conditional classifiers. Algorithms based on global formulations, define the collective classification problem as a global objective function to be optimized. In this thesis we aim at defining an initial simple baseline against which other more sophisticated methods for collective classification, e.g. Statistical Relational Learning (SRL) [50], can be compared. In addition, this simple baseline will provide initial evidence of the potential of contextual cues for improving the performance of object detection and object pose/viewpoint estimation. For the moment we will not consider more sophisticated methods for Statistical Relational Learning since, as suggested in [75, 144], these methods tend to be computationally expensive. Furthermore, an additional step to find a good trade-off between expressiveness and efficiency might me required. For these reasons, in this thesis we focus on methods from the first group, which are based on local conditional classifiers. Specifically, we follow a three-stage algorithm for collective classification, similar to the one proposed in [91] (see Algorithm 1). For a wider, more detailed, view of methods for Collective Classification please refer to [134, 135].

Algorithm 1 Collective Classification

Given

- A set of interrelated nodes v_i connected by links l_{ij} and some classified nodes.

Steps

1. **Local Classification:** classify each of the nodes v_i using the non-relational (local) model. This model focuses purely on attributes of the nodes.
2. **Relational Classification:** classify each of the nodes v_i using the relational classifier which takes into account nodes v_j, on the neighborhood N_i, connected via links l_{ij}.
3. **Collective Inference:** re-classify the nodes v_i together, taking into account the classification output obtained in steps 1 and 2, and possibly iterate.

This method has two main components, the first component is a local classifier. This local classifier should be able to learn the correlation between the intrinsic features from each node v_i with its corresponding class label c_i. Within the object recognition setting, on one hand, examples of intrinsic features are shape, color, gradients, or texture from the object instances. On the other hand, class labels refer to the specific object category of interest, e.g. cars, pedestrians, airplanes, sheep, etc., or viewpoints. The second component of the collective

classification algorithm is a method to perform relational classification, i.e. a classifier that learns the correlation between the class c_i from node v_i w.r.t. the nodes v_j present on its neighborhood N_i. A component of the relational classifier is a function to define the set of neighbor nodes N_i. In the object recognition setting, neighboring objects v_j have been defined in different ways e.g. objects within a maximum distance radius [119], objects with contact edges [44], etc. Finally, based on the responses of the local and the relational classification stages, we re-classify each of the nodes in such a way that a balance is obtained between the predictions of the local and relational classifiers.

2.7.1 Cautious Inference in Relational Data

An identifying characteristic of algorithms for collective classification, is that the simultaneous classification of nodes should maintain some level of consistency w.r.t. each other. It is for this reason that the classification of each node is dependent on the nodes on its neighborhood, see step 2 from Algorithm 1. However, as pointed out in [95, 96], this relational aspect introduces some problematic scenarios since the relational features are derived from nodes that are uncertain, hence possibly introducing noise in the inference process. Taking this observation as a starting point, [95, 96] labeled as *cautious*, the algorithms that seek to identify and exploit the most certain relational information first. In contrast, the algorithm is labeled as *aggressive* if it uses *all* the relational information ignoring the certainty on either the relational features or the neighboring nodes that produced them. The motivation behind cautious algorithms is that using reliable relational data for inference should, in consequence, produce more reliable predictions than when considering possibly noisy data. Furthermore, there is some evidence [48] that suggests that this type of iterative cautious algorithms are fairly robust to a number of simple ordering strategies, such as random ordering, visiting nodes in ascending order of diversity of its neighborhood class labels and labeling nodes in descending order of label confidences.

Controlling the degree of Caution

In their study [95], McDowell et al. identified three parameters to control the degree of *caution* of a relational method.

The first aspect favors the most reliable nodes or instances, i.e., the nodes predicted with the highest certainty. These nodes are then used to define relational features and re-classify the less certain ones. Following this initial step, this process is repeated, effectively propagating the predictions from the

most certain nodes to the rest of the nodes in the network. The motivation behind this aspect is that since class label assignment is done using only the most reliable information, subsequent assignments should also be more reliable. At the same time we are removing potential sources of noise by ignoring the less certain nodes.

The second aspect, favors known links or relations between nodes, which also include links previously seen on the training data. This is applicable on within-network classification problems, e.g. the "in-sample" task from [102], where some of the unknown nodes are linked to nodes with known, annotated, class labels. A current, realistic, example of this type of scenario is the Web. In the Web, new websites are constantly appearing and are usually linked to other, well identified, websites. Under this aspect, only the known links are used to compute the relational feature in the first iteration of the classification process. Then, based on predictions of the first iteration the rest of the nodes is classified.

Finally, the third aspect is in charge of cautiously handling missing links. As stressed by [95], the need for this aspect rises from the assumption made by many machine learning algorithms in that their input data has no missing values. For example, at a given iteration in the cautious inference setting, links to a particular, currently ignored, node class might be missing. However, this does not imply that links with such a node class do not exist at all. When considering this aspect, relational feature values computed from such unknown nodes are set to *unknown* as well. This allows the relational feature to make a distinction between nodes without links from nodes with invalid links. As presented above, the first scenario rises at the earlier iterations of cautious inference, when there are unknown nodes that have not been linked. On the opposite, the second scenario is more evident at the last iterations of cautious inference, when links between all the nodes have been defined.

2.7.2 Weighted-vote Relational Neighbor Classifier

In this thesis we investigate how different objects influence the occurrence and pose of each other. Therefore, we estimate the degree to which an object o_i fits in the scene based on its relations with other objects in the same scene. This can be seen as a *Collective Classification* problem (see Section 2.7) in which the class of an object influences that of another. In order to take into account the influence between objects, we estimate a response for each object o_i based on the relations with all the objects o_j in its context. This contextual response is obtained using the weighted-vote Relational Neighbor classifier (wvRN) [91]. This relational classifier, formally known as the probabilistic Relational Neighbor classifier (pRN) [89], is a simple, yet powerful classifier that is able to take advantage of

the underlying structure between networked data. This classifier operates in a node-centric fashion, that is, it processes one object o_i at a time taking into account a set of n objects in its neighborhood N_i. wvRN estimates $p(o_i|N_i)$, the probability that o_i occurs given its neighborhood N_i, as the weighted mean of the class-membership probabilities predicted by the entities in N_i. It is defined as follows:

$$wvRN(o_i|N_i) = \frac{1}{z} \sum_{o_j \in N_i} v(o_i, o_j).w_j \tag{2.12}$$

with z a normalization term, $v(o_i, o_j)$ a pairwise term measuring the likelihood of object o_i given its relation with o_j, and the weighting factor w_j modulating the effect of the neighbor o_j. In the methods proposed in this thesis we use wwRN to compute the relational, contextual, score $wvRN(o_i|N_i)$ of an object o_i given its neighborhood N_i.

The usage of wvRN is motivated by its performance during the last decade, where it has been successfully applied in work related to text mining [89, 91], web-analysis [91], suspicion scoring [90], link prediction [84], and social network analysis [87, 88].

As mentioned earlier, in this thesis, we perform Collective Classification as means to reason about interconnected object instances. This is done with the objective of predicting their location (Chapter 4) and viewpoint (Chapter 3) while taking into account both local and contextual features. Furthermore, in Chapter 4, we evaluate the effect of performing cautious inference for context-based object detection.

2.8 Evaluation Protocol

2.8.1 Datasets

As mentioned earlier, in this thesis we focus on urban scenes. For this reason, in the work presented in this thesis we conduct experiments on two datasets: the object detection set of the KITTI benchmark [45], and the MIT-StreetScenes (MITSS) dataset [13].

The *KITTI dataset* is collected from a car-mounted camera, resembling an autonomous navigation setting. We consider "car" as category of interest due to its multiple occurrences within each image of this dataset. This dataset presents a variety of difficult scenarios ranging from object instances with high

occlusions to object instances with very small size. Furthermore, it provides precise annotations of objects in the 2D image and in the 3D space, including their respective poses. Since this is a benchmark dataset, annotations are not available for the test set. For this reason, we run our experiments on the training set. Using the time stamps of the dataset, we divide the data into disjoint subsets. More detailed information about how the dataset is divided will be provided in subsequent chapters. Please see Figure 2.9 for some image examples from this dataset.

Figure 2.9: Example images from the KITTI object detection benchmark [45].

Different from the KITTI benchmark, the *MIT-StreetScenes dataset* was obtained using a consumer camera and offers more viewpoint variability (see Figure 2.10 for example images). In addition, this dataset only provides annotations for the 2D bounding box of the objects of interest. For our experiments we divided this dataset in 4 subsets. The first two quarters were used for training and validation while the third and fourth quarters were used for testing.

2.8.2 Performance Metrics

Methods presented in this thesis reason about contextual information with the final goal of improving the performance on the object detection and object pose estimation tasks. For this reason, we need a metric to measure the obtained performance on each of the two tasks.

Bounding Box Matching: Throughout this thesis we follow the *intersection-over-union* criterion, introduced in Section 2.2), in order to verify whether an

Figure 2.10: Example images from the MIT StreetScenes dataset [13].

object hypothesis matches an object annotation. This criterion considers an object hypothesis as a match if its Jaccard Similarity coefficient w.r.t. an object annotation is above 0.5. This coefficient is computed following Eq. 2.1. As presented in Figure 2.5, this takes into account the bounding boxes of the object hypothesis and the object annotations.

Measuring Object Detection Performance: We will further follow the evaluation protocol proposed in the Pascal VOC challenge for measuring performance on the object detection task. To this end, we report Average Precision (AP) as performance metric on the object detection task. AP is the interpolated average precision [128]. It is estimated based on the precision/recall curve computed from the score-ranked set of hypotheses produced by a method. Recall is defined as the proportion of all positive examples ranked above a given rank. Precision is the proportion of all examples above that rank which are from the positive class. The average precision summarizes the shape of the precision/recall curve, and is defined as the mean precision at a set of eleven equally spaced recall levels $[0, 0.1, ..., 1]$. This metric is used Chapters 3 and 4 which have evaluate methods for improving object detection.

In addition, in Chapter 6 we use an alternative metric that measures changes in recall as function of the number of object hypotheses predicted per image. For the computation of this metric object hypotheses are sorted based on the order in which they are generated, or the likelihood of their occurrence. Recall is computed in the same way as for average precision. In contrast to average precision, this new metric gives higher importance to recall. This type of

metric is commonly applied to measure the performance of methods to generate category-independent object proposals [2, 151, 168].

Measuring Pose/Viewpoint Estimation Performance: We adopt Mean Precision for Pose Estimation (MPPE) as performance metric, in order to measure performance on the object pose/viewpoint estimation task. This metric is traditionally used to measure pose classification performance [51, 81, 115, 130]. It is computed as the average of the diagonal of the class-normalized confusion matrix of the pose classifier. It is computed from hypotheses that are assumed correct based on the Pascal VOC matching criterion [36].

2.9 Conclusion

This chapter presented standard principles and machinery that constitute the foundations on top of which this thesis is constructed. Since the objective of this work is contextual reasoning for vision tasks, we start by presenting methods to acquire evidence from the images. To this end, we provide an overview of methods for object detection and pose estimation. Then, we introduce collective classification as the engine for reasoning about relations between groups of objects. Bridging the previous content, we present kernel density estimation and topic models as means to model the relations between object instances. Finally, the evaluation protocol followed in this thesis is explained.

Chapter 3

Allocentric Pose Estimation

In Chapter 1 we discussed on the emerging group of works that successfully exploit contextual information to improve object detection. On the contrary, works that exploit contextual information for the task of object pose estimation are almost nonexistent. Based in this observation, in this chapter we focus on measuring the effect of exploiting contextual information, in the form of pairwise relations between objects, to improve performance on object pose estimation. This chapter directly addresses Research Question 1 by exploring the potential effects of reasoning about object relations to improve object pose estimation performance.

Work covered in this chapter is based on:

- Oramas M, J., De Raedt, L., and Tuytelaars, T. *Allocentric pose estimation.* In IEEE International Conference on Computer Vision (ICCV) 2013.
- Oramas M, J., De Raedt, L., and Tuytelaars, T. *Reasoning about Object Relations for Object Pose Classification.* In Netherlands Conference on Computer Vision (NCCV) 2014 .

3.1 Introduction

Object pose or viewpoint estimation is an important problem for a wide range of applications, including robotics and road safety systems. Various methods for tackling this problem have been proposed [63, 80, 81, 115, 130], yet it is still far from being solved. Especially in 'real-world' scenarios, like the one

Figure 3.1: The natural or "desired" configurations in which objects occur in the world often provide strong cues of their pose. For instance, it is not difficult to guess the pose of the cars below the yellow circles by only looking at the rest.

depicted in the KITTI dataset [45], with lots of clutter, occlusions, etc. results are still relatively poor. Context information has been used successfully for object detection [30, 59, 104] in various forms (*Stuff*, *Things* and scene related cues). This has been effective in clarifying ambiguous scenarios. Yet, to the best of our knowledge, context information has not yet been exploited for pose estimation.

Imagine you are given the task of predicting the pose of the objects below the yellow circles in Fig. 3.1. Even when there is no access to intrinsic features of the objects, the overall configuration of surrounding objects provides a strong cue to predict their pose. This can be considered a Collective Classification problem [134] in which the class (pose) of one object influences that of another. We face two challenges towards solving this problem. First, we need a method to define informative relations between objects. These relations should be robust to viewpoint changes and general enough to be applicable to different categories of objects (i.e. not using category-specific features). Second, a method to discover and reason about configurations of objects should be adopted. In this chapter, we explore how information from other objects in the scene can be exploited for the task of pose estimation. In particular, we look at configurations of "Things". We show that, even when starting from a noisy pose estimator, results can be improved by looking at configurations. Considering the first challenge, robust and informative relations, we explore both a camera-centered and an object-centered representation for relations. Related to the second challenge, we use a simple, yet powerful, method to reason about the configuration of objects. We

capture statistics of typical object configurations using kernel density estimation, and combine this information using collective classification, more specifically a Relational Neighbor classifier [91]. We refer to the previous chapter for background information on object pose estimation, collective classification, and kernel density estimation.

The main contributions of this chapter are: First, we show that considering configurations between objects can be beneficial for pose estimation: the proposed collective classification method complements state-of-the-art local pose estimation methods. Second, we show the influence of the Frame of Reference (FoR) – i.e. object-centered or camera-centered, used to define relations between objects for object pose estimation and detection. To our knowledge this is the first attempt to exploit relations defined between object entities via collective classification for the task of pose estimation. Additionally, we show our scheme can also be used to improve object detection performance.

This chapter is organized as follows: section 3.2 presents related work. The following three sections show how we define and learn relations between objects in the scene, and how we combine them with the evidence from local detectors. In section 3.6 we provide implementation details, while section 3.7 describes the experimental results. Finally, we draw conclusions in section 3.8.

3.2 Related Work

Some object categories can be easily recognized based on their material, color and texture, while others are characterized predominantly by their shape or appearance. This led to the division of object categories into *Things* and *Stuff* [43]. Here we focus on object categories with defined shape and appearance, the *Things*, and methods exploiting relations and configurations between them to predict their pose.

Several pose estimation methods have been proposed in the literature. All of these rely on intrinsic characteristics of the object category. In the traditional processing pipeline for pose estimation, first, candidate regions to host object instances are proposed. Secondly, an appearance descriptor is computed in the area of each candidate region. Finally, based on a pre-trained model, each descriptor is classified as one of the possible poses the object may take. Following this pipeline, methods have evolved from modeling 2D views of the categories of interest (e.g. [81]) to reasoning about object parts in the 3D space [63, 80, 115, 130].

Recently, methods related to structure from motion such as [6, 7, 47, 159] aim

at understanding the full scene layout. They assume that correspondences between scene elements such as points, regions and objects across image views or sequences introduce constraints in the scene behind the images. These correspondences are exploited and among the different tasks these methods target, they also perform 3D pose estimation. These methods have shown impressive qualitative results. Yet they rely on the availability of image sequences or stereo pairs. Similar to these works we define relations between scene elements. However, instead of defining relations between different scene element types such as points, regions or objects, we focus on relations between object instances. Additionally, we drop the requirement of multiple images for the extraction of evidence - we only assume the ground plane to be known.

In recent years, learning relations between *Things* has gained popularity in the computer vision community, particularly to assist the task of object detection. Early work [30, 32, 118, 148, 156] represented objects as regions in the image. Then, by learning qualitative 2D spatial relations (e.g. top-left, far-left) between them, detections in unlikely areas were filtered out. Extending this idea, [44, 68, 92] went beyond object categories and also take the appearance of the objects into account. More recently, [78, 126] use discriminant relations between objects to learn the collective appearance of related objects in order to guide the detection of the individual objects. Similar to these works, we learn relations between object instances. Different from these works, in addition to predicting the occurrence of an object instance, we also predict its pose. Moreover, we reason in a 3D representation of the scene asuming we know the ground plane, not in the 2D image space. Additionally, instead of using symbolic spatial relations (e.g. *in-front-of, close, near, far*) we use continuous measures to define relations between entities as in [23, 118]. Finally, different from existing work, we explore the use of relations defined in an *object-centered* Frame of Reference.

3.3 Relations between Objects

We believe that the pose of an element is not only affected by its individual behavior but also by its behavior towards other elements in the scene. This idea is inspired by "Psychological Allocentrism" which states that elements tend to be interdependent, defining themselves in terms of the group they are part of, and behaving according to the norms of the group [67, 149]. Allocentric elements appear to see themselves as an extension of their in-group. Based on this description, our method takes into account the group consistency of each element relative to the group defined by the other elements in the scene.

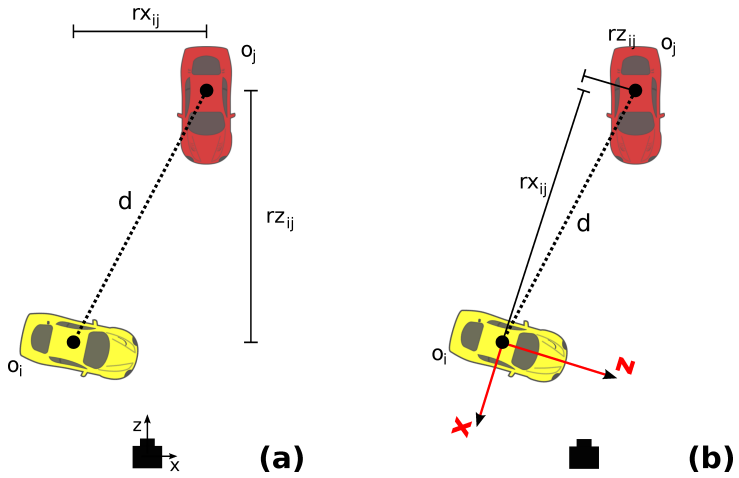

Figure 3.2: Spatial relations between objects. a) Camera-centered relations, b) object-centered relations. Note the difference between relative X and Z values.

In order to measure the level to which an object instance fits in a group of objects, first, we need to define *relations* between objects. Here, we limit ourselves to purely pairwise relations. Given a set of objects $o = \{o_1, ..., o_m\}$ in the 3D scene, each object o_i is defined by its location (x_i, y_i, z_i) and discrete pose label θ_i. We define pairwise relations between objects in two different ways, by changing the location and orientation of the frame of reference (FoR). This results in *camera-centered (CC)* and *object-centered (OC)* relations. We define *object-centered* relations between objects as follows. First an object o_i is selected and the frame of reference is centered on it with the Z-axis facing in the frontal direction of the object (see Figure 3.2b). Then, we measure the relative location and pose of each of the other objects o_j, one at a time, producing a relational descriptor $r_{ij} = (rx_{ij}, ry_{ij}, rz_{ij}, r\theta_{ij})$. For an image with m objects a total of $(m(m-1))$ pairwise relations are extracted. In practice we ignore ry_{ij} since all the objects we consider in our experiments are found on the ground plane, so $ry_{ij} = 0$ in all cases. As a baseline, we also perform experiments with *camera-centered* relations, as used traditionally. For these, we use the same relational descriptor as above, yet with everything measured relative to a frame of reference attached to the camera (see Figure 3.2a). Note that $rx_{ij}^{CC} \neq rx_{ij}^{OC}$ and $rz_{ij}^{CC} \neq rz_{ij}^{OC}$, but $r\theta_{ij}^{CC} = r\theta_{ij}^{OC}$.

3.4 Learning

3.4.1 Allocentric Pose Estimation

With *allocentric pose estimation*, we refer to the task of estimating the pose θ_i of an object o_i purely based on the objects in its neighborhood N_i. In our experiments, N_i is the set containing all the other objects o_j in the scene. This pose is defined as the pose value θ^* for object o_i that maximizes the occurrence of o_i given the neighborhood N_i. That is:

$$\theta_i^* = \arg \max_{(\theta_i \in o_i)} (wvRN(o_i|N_i)) \tag{3.1}$$

where $wvRN(o_i|N_i)$ is the weighted-vote Relational Neighbor classifier (wvRN) [91] as introduced in Section 2.7.2. Note that the pose θ_i is an attribute of the object o_i. In order to stress this, and for the sake of clarity, we will adopt the following notation. We will explicitly add the pose attribute θ_i of object o_i. In addition, we will refer as o^+ to the object hypotheses that are well localized, i.e. their bounding boxes cover valid object instances. On the contrary, we will refer as o^- to false object hypotheses. Similarly, we use θ^+ and θ^- to indicate whether the pose of the object is predicted correctly or not. Taking into account this new notation Eq. 3.1 is redefined as:

$$\theta_i^* = \arg \max_{(\theta_i)} (wvRN(\theta_i^+, o_i^+|N_i)) \tag{3.2}$$

Similarly, we take into account the new notation and define the terms of the $wvRN$ classifier in the following way:

$$\begin{aligned} wvRN(o_i|N_i) &= \frac{1}{Z} \sum_{o_j \in N_i} v(o_i, o_j).w_j \\ wvRN(\theta_i^+, o_i^+|N_i) &= \frac{1}{Z} \sum_{o_j \in N_i} p(\theta_i^+, o_i^+|r_{ij}).w_j \end{aligned} \tag{3.3}$$

where the weighting term w_j measures the effect of the neighboring object o_j on o_i. We use the weights w_j to bring into the model the uncertainty in object detection for objects o_j in the Neighborhood N_i. In a perfect scenario, where all the objects are accurately detected, the term $w_j = 1$, since we are certain of their occurrence, and the normalization term Z corresponds the number of neighboring objects. In our setting, the term w_j can be defined in different

ways depending on the task to be addressed. This will be explained in the next section. We define the pairwise term $v(o_i, o_j) = p(\theta^+, o_i^+|r_{ij})$ as the probability of object o_i occurring, with pose θ_i, given its relation r_{ij} with the neighboring object o_j. Using Bayes' rule we estimate $p(\theta_i^+, o_i^+|r_{ij})$ as the posterior:

$$p(\theta_i^+, o_i^+|r_{ij}) = \frac{p(r_{ij}|\theta_i^+, o_i^+)p(\theta_i^+, o_i^+)}{\sum_{o_i \in (o_i^+, o_i^-)} \sum_{\theta_i \in (\theta_i^+, \theta_i^-)} p(r_{ij}|\theta_i, o_i)p(\theta_i, o_i)} \tag{3.4}$$

The components of Eq. 3.4 are obtained through the following procedure. During the training stage, we compute pairwise relations r_{ij} between the annotated objects in the training images. Furthermore, we extend this set of objects and relations by running a local detector on the training set producing a set of object hypotheses per image. Then, we assign the flags θ_i^+, θ_i^-, o_i^+ and o_i^- to the hypotheses o considering both pose and spatial matching based on the Pascal VOC [36] matching criterion. Note that hypotheses with flags (θ_i+, o_i^-) do not occur since it is not possible to predict correctly the pose of a false hypothesis. In order to avoid repeated object instances, we replace the hypotheses with correct localization and pose (θ_i^+, o_i^+) by their corresponding annotations. Similarly, we replace the relations produced by the (θ_i^+, o_i^+) hypotheses by those produced by their corresponding annotations. This combination of annotations and hypotheses allows our method to model, up-to-some-level, the noise introduced by the local detector in the form of false detections. This produces a set of objects o_i with their corresponding pairwise relations $R = \{r_{ij}\}$ from the whole training set. Finally, during testing, $p(r_{ij}|\theta_i^+, o_i^+)$, $p(r_{ij}|\theta_i^-, o_i^+)$ and $p(r_{ij}|\theta_i^-, o_i^-)$ are estimated by performing Kernel Density Estimation using the set of relations $R = \{r_{ij}\}$ as samples.

This method captures the statistics of typical configurations. For instance, when applied on top of object-centered relations, it effectively encodes that cars with the same pose tend to be one behind the other - as when driving in the same lane, while cars with opposite poses are more likely to be driving on the left - as in opposite lanes (see figure 3.3). The priors $p(\theta_i^+, o_i^+)$, $p(\theta_i^-, o_i^+)$ and $p(\theta_i^-, o_i^-)$ of the object occurring or not at the given location with a given pose, are estimated based on their proportion on the training set, respectively.

3.4.2 Working with noisy detections

In practice, state-of-the-art object detectors are not perfect and produce many false hypotheses. Moreover, the location of true hypotheses are also noisy, while the viewpoint is often simply wrong. In addition, the score s provided as output by viewpoint-aware detectors is a stronger indicator of the confidence on the

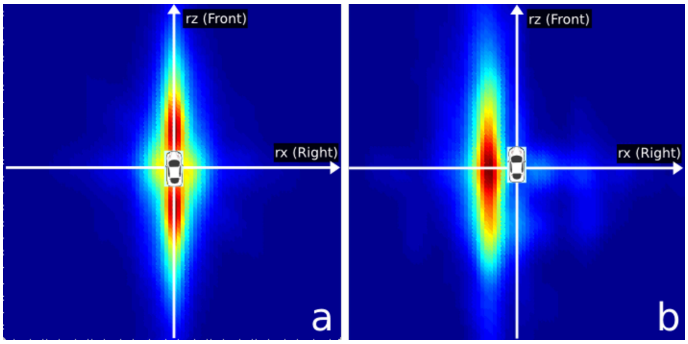

Figure 3.3: Top view of the distribution of object-centered relations for cars with the same pose (a) and opposite pose (b) respectively .

occurrence of the object rather than of its viewpoint. For these reasons the confidence on the hypotheses predicted by the local detector should be considered during the voting procedure of Eq. 3.3, via w_j. We define the weighting term w_j in two different ways depending on the objective of the classification. In this chapter we focus mainly on the task of object pose estimation. However, as a side experiment, we measure the effect of our method on the task of object detection (see Sec. 3.7.2).

Pose Estimation: For the task of object pose estimation, we define w_j as $w_j = p(\theta_j)$ aiming to compensate for the noise in the poses used to compute r_{ij}. This noise is introduced in greater extent by errors in the viewpoint predictions of the detector, e.g. opposite viewpoints are commonly confused. Moreover, these errors are further propagated when projecting the 2D hypotheses to the 3D space of the scene. As presented in Section 2.3, traditional viewpoint-aware object detectors predict the viewpoint angle α of the object as it is seen by the camera. Since the pose angle θ of the object, in the scene, is a function of its viewpoint α and its 3D location, errors committed during viewpoint estimation will have a strong effect in the estimated 3D poses θ. Furthermore, the bounding boxes predicted by the object detectors are not 100% accurate, i.e. the sides of the bounding boxes do not perfectly circumscribe the object detections. This introduces noise in localization when projecting the 2D object in the 3D space. In order to overcome these issues, we exploit the information from the confusion matrix of the pose estimator. Given a 3D object o_j with estimated *continuous* pose $\hat{\theta}_j$ (see Sec. 3.6), we define w_j by performing a linear interpolation to its nearby discrete poses θ_{low} and θ_{top} using their corresponding responses $p(\theta_{low})$ and $p(\theta_{top})$ from the diagonal of the confusion table.

$$w_j = p(\hat{\theta_j}) = p(\theta_{low}) + (p(\theta_{top}) - p(\theta_{low}))\frac{(\hat{\theta_j} - \theta_{low})}{(\theta_{top} - \theta_{low})} \tag{3.5}$$

Object detection: For this task, we need to put more emphasis on the occurrence of the object rather than its pose. For this reason, we estimate w_j through a *probabilistic local classifier* that takes into account the detection score s_j of the predicted object detection o_j. We consider the posterior of the object occurrence given its detection score as the output of the local classifier, $w_j = p(o_j^+|s_j)$. We compute this posterior following the procedure of [118]:

$$p(o_j^+|s_j) = \frac{p(s_j|o_j^+)p(o_j^+)}{p(s_j|o_j^+)p(o_j^+) + p(s_j|o_j^-)p(o_j^-)} \tag{3.6}$$

To obtain the components of this equation we perform a procedure similar to the one done for Eq. 3.4 up to the point were hypotheses are labeled as o^+ or o^-. Then, considering these hypotheses, we compute the conditionals $p(s|o^+)$ and $p(s|o^-)$ respectively based on KDE. Finally, the priors $p(o^+)$ and $p(o^-)$ of the detection occurring or not at the given location, are estimated as the percentage of o^+ and o^- hypotheses in the training set, respectively. As a result, $p(o_j^+|s_j)$ expresses the probability of an object hypothesis being properly localized given its detection score. This procedure allows us to plug-in any standard object detector in our method.

3.5 Modeling consistency between local appearance and allocentric behavior

At this point, we have two methods to estimate the probability of a certain pose for an object hypothesis o_i: based on its intrinsic features, as evaluated by a traditional pose estimator, and based on its neighborhood N_i, respectively. The reader should note the "competitive" behavior of these two methods. While the local classifier (LC) pulls the decision towards individual features, the relational classifier (RC) (Eq. 3.3) pulls it towards the collective feature of group fitting. Given the "competitive" nature of these classifiers, local and relational, we need to find a method to reconcile them.

To achieve this we follow a method similar to [118]. First, we collect the responses of the local (Eq. 3.6) and relational (Eq. 3.3) classifiers on a validation set, giving us score pairs $S = (s_{\text{LC}}, s_{\text{RC}})$ for each object hypothesis o. Then we group these score pairs for true detections as well as for false detections. At test

time, we estimate $p(S|o^+)$ and $p(S|o^-)$ via KDE. These terms are used in the equation $p(o|S) = p(S|o^+)p(o^+)/(\sum_{(o^+,o^-)} p(S|o)p(o))$ to estimate the desired posterior.

3.6 Implementation Details

The focus of this chapter is on the study of how relations between objects can assist the task of object pose estimation. For this reason rather than proposing our own object detector and pose estimator we use state-of-the-art detectors to acquire evidence of objects in the scene. To show the generality of our method, we build on two different detectors / pose estimators, namely those proposed in [81] and [47]. Both methods are based on the popular deformable parts model of [112], and both of them jointly tackle the problems of object detection and pose estimation. We use them as off-the-shelf detectors with default parameters. These detectors, separately, feed our framework with confidence scores, locations (2D bounding box) and poses of object hypotheses discretized into 8 and 16 partitions respectively. Then, using a stereo pair and the algorithm for efficient large-scale stereo matching proposed in [46] we obtain a 3D point cloud of the scene. To obtain the 3D location of the object, we project the point cloud into the image plane and take as location the 3D point at the bottom center of the bounding box predicted by the detector. For the 3D size of the object (used purely for visualization purposes), we use the mean width, length and height of 3D annotations in the training data. Though this is not very accurate, it is an approximation that worked well in practice. Reasoning about the relative location of objects permits the usage of alternative methods (e.g. [9], [61] and [82]) that focus on building 2.5D-3D scene representations from still images in cases where stereo pairs are not available. It should be noted that the stereo pairs are used solely to estimate the 3D location of the objects and not to derive information (e.g. 3D shape) that can be used to estimate the pose of the object. Additionally, this dependence on relative location rather than shape/volume, regions or scene type, sets our work in the middle between works based on 2.5D and works from Holistic Scene Understanding.

For the pose, the detectors provide a discrete viewpoint angle α of the object as seen by the camera. From this angle we obtain a continuous azimuth angle θ in the world coordinate frame by back projecting the object o on the ground plane. To measure the certainty of this estimation during testing, we perform a linear interpolation of the estimated azimuth angle using the closest discrete pose angles and the confusion table of the local pose estimator as discussed in Sec. 3.4.2. Since one of our objectives is to evaluate the influence of the frame of reference for defining informative relations, we define relations using both

camera-centered and *object-centered* FoRs. The procedure is directly applied for the case of *camera-centered relations*. For *object-centered relations* an additional step is required where the FoR should be centered in the trajector object before any relation attribute can be measured (see Section 3.3).

We perform Kernel Density Estimation, $f(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$, where K is a Gaussian kernel. Furthermore, the estimation is performed in a fixed modality selecting the bandwidth values h in a data-driven fashion using Silverman's Rule of Thumb [154] (see Section 2.5). The samples x_i that compose the estimator depend on the function to be modeled. For the case of $p(r_{ij}|\theta_i^+, o^+)$, $p(r_{ij}|\theta_i^-, o^+)$ and $p(r_{ij}|\theta_i^-, o^-)$, in Eq. 3.4, the samples x_i correspond to the pairwise relations computed between the object instances in the training data. For the case of the terms $p(s_j|o_j^+)$ and $p(s_j|o_j^-)$, in Eq. 3.6, the samples x_i are the detection scores from detections in the training images. Finally, for the case of multivariate data, e.g. the relations between objects r_{ij}, we follow the multivariate approach presented in Section 2.5 (Eq. 2.9) which relies on product kernels (Eq. 2.8).

3.7 Evaluation

3.7.1 Dataset

Most pose estimation datasets do not include groups of objects in images. Usually, there is just a single object in the main focus of the picture. This is a side effect due to the way in which datasets are collected [32]. This reduces the datasets in which the proposed method could be evaluated to one, the KITTI benchmark [45], which was introduced in Section 2.8.1. We evaluate the influence of the FoR when defining relations between objects in both ideal (annotated) and real (estimated) world settings. For the ideal setting, the dataset provides 3D location and pose vectors for the objects. For the real setting, it provides stereo pairs for each scene and object annotations that allow us to build methods to learn and evaluate the configurations between object instances. Additionally, the multiple cars occurring in each image provide a challenging realistic scenario with occlusions and clutter that will be useful to evaluate our proposed allocentric pose estimator. We evaluate against all the object annotations despite their occlusion level and considered images with more than two objects. We split the training set of the KITTI dataset [45] into four subsets. The first quarter of the set is used for training the relational classifier and estimating the pose estimator confusion matrix. The second is used for validation and learning the combination of the local and the relational classifier.

The third and fourth quarters are used for testing. We run experiments in 5 different splits of data.

3.7.2 Experiments

We report results in four sets of experiments: The first experiment measures the performance of the proposed method considering an *ideal scenario* where all the neighboring objects are predicted accurately. The second experiment focuses on a *real scenario* in which neighboring objects are obtained using an object detector. Thus, evaluating the performance of the proposed method under the presence of noise. The third, *object verification*, experiment aims to analyze the potential of the proposed relations-based scheme for the task of object detection. Finally, the fourth experiment measures the changes in performance when defining relations between objects from a *camera-centered and an object-centered* frame of reference.

Pose Estimation

Similar to prior work [51, 80, 81, 115, 130], in this chapter we adopt the MPPE metric, introduced in Section 2.8.2, in order to evaluate pose estimation performance. Both the baseline detectors and our method start from the same initial set of hypotheses.

Ideal Scenario Experiment: The first experiment aims at answering the question: "How much information about the object's pose can be obtained based on the locations and poses of objects in its neighborhood?". To this end, we consider the ideal scenario, where the local object detector and pose estimator are 100% accurate for the objects in the neighborhood. In this scenario all the objects of interest in the scene have been detected and their pose has been accurately predicted. For this experiment we use ground-truth annotations from the dataset. The pose of each object is then predicted based on the ground truth locations and poses in its neighborhood. The objective of this experiment is to present the upper limit of the performance that the Relational Classifier (RC) used for allocentric pose estimation can achieve in an ideal setting on the current dataset. We compare 2 ideal allocentric pose estimators that are able to predict 8 and 16 poses respectively.

Discussion: Table 3.1 shows that, in an ideal scenario, the allocentric pose estimator takes advantage of finer discretization of object poses. While the absolute number is lower for the 16 poses classifier, with twice as many output labels this is a significantly harder problem. This experiment shows the upper

Method	testSet
Ideal Local Classifier (8 poses)	0.47
Ideal Local Classifier (16 poses)	0.37

Table 3.1: Allocentric Pose Estimation Performance in the Ideal Setting (MPPE values per method).

limits in performance that can be expected from allocentric pose estimation using local detectors [47, 81]. Based only on context information, it is not possible to accurately estimate the object's pose. At the same time, this upper bound is similar or even higher than what current state-of-the-art local detectors can obtain (see below), and therefore using context information to improve pose estimation results seems promising.

Real Scenario Experiment: This experiment starts from the local detectors [47, 81] introduced in Section 3.6. We define object-centered relations between the 3D hypotheses in the scene (i.e. the 2D object detection back projected onto the ground plane) and perform pose estimation based on the method proposed in Section 3.3. The objective of this experiment is to evaluate: a) the performance of the local pose estimators, b) the performance of pose estimation based on object relations alone, and c) the changes in performance brought by the method proposed in Sec. 3.5 for modeling the consistency of local and relational classifiers. We report results on two sets. The first set runs on the raw output of the baseline detectors, while the second set adds a 3D Non-Maximum Suppression (3DNMS) pre-processing step to remove overlapping hypotheses. Given a set of 3D hypotheses o_i we suppress all the hypotheses that are closer than a threshold value t. This value is heuristically estimated from the training set, by estimating the mean width of the objects of interest. Any object closer than a factor of 0.8 is assumed to overlap and is suppressed.

Discussion: The results of this experiment (see table 3.2) show it is possible, at least to some extent, to estimate the pose of objects by looking at the poses and locations of other objects – even if these poses and locations are noisy themselves. While the performance of the relational classifier alone is lower than the one obtained by the local classifier, it is above the chance level (12.5% for the 8-poses [81] setting and 6.25% for the 16-poses [47] setting). The combination of both local and relational classifiers brings a mean absolute improvement, over the local classifier, of 2.5% and 1.7%, on [81] and [47] respectively. This improvement seems marginal, however, encouraging given that we only tried the most basic collective classification schemes. Furthermore, this improvement is consistent given a standard deviation on the improvement of 0.7% and 0.6%, respectively. In addition, as depicted in Fig. 3.4 the inclusion of object configurations helps to fix some of the, initially, wrongly estimated poses. This example also shows, that

testSet			testSet (3DNMS)		
LC (Lopez et al.[81])	RC	LC+RC	LC (Lopez et al.[81])	RC	LC+RC
0.27	0.20	**0.30**	0.29	0.20	**0.31**

testSet			testSet (3DNMS)		
LC (Geiger et al.[47])	RC	LC+RC	LC (Geiger et al.[47])	RC	LC+RC
0.55	0.27	**0.57**	0.57	0.24	**0.58**

Table 3.2: Mean Pose Estimation Performance in the Real Scenario (MPPE values per method) using the detectors from [81] (8 poses) and [47] (16 poses) to collect object hypotheses. LC (Local Classifier), for their respective pose-aware detectors. RC (Relational Classifier). LC+RC (Combination of the responses of the Local and Relational classifier). Top: results from detections collected with the pose-aware detector from [81], and Bottom: results from detections collected with the pose-aware detector from [47]. Left Column: results reported considering all the detections reported by the detectors. Right Column: results reported considering only the detections that remained after performing a 3D Non-Maximum Suppression (3DNMS) pre-processing step.

even for the case of false hypotheses, our method predicts poses for objects that could have occurred in such locations. We can notice that the local classifiers directly benefit from the application of the 3DNMS step. As presented in Table 3.2, local classifiers have an average performance improvement of 2% by applying 3DNMS. As a result, the relational classifiers now have less room for improvement. This can be verified in Table 3.2 by the reduced improvement of 2% and 1%, over [81] and [47] respectively. Future work will focus on a proper integration of heuristic strategies, e.g. 3DNMS, with relations-based methods like the one proposed in this chapter.

We additionally tried a variation of this setting where pose information is ignored when defining relations between object instances. As a result, reasoning will be performed based purely on relative locations between objects. As expected, allocentric pose estimation in this setting has lower performance. In fact, its performance is close to chance level and is 15% lower than the setting where relations include pose information. Given these observations, we conclude that object pose information plays an important role when modeling configurations between object instances and that it is a feature that must be considered in future algorithms that take into account contextual features for reasoning. This also provides evidence that we are dealing with a true collective classification problem as the pose of one object depends on the pose of the other ones. This motivates the use of wvRN.

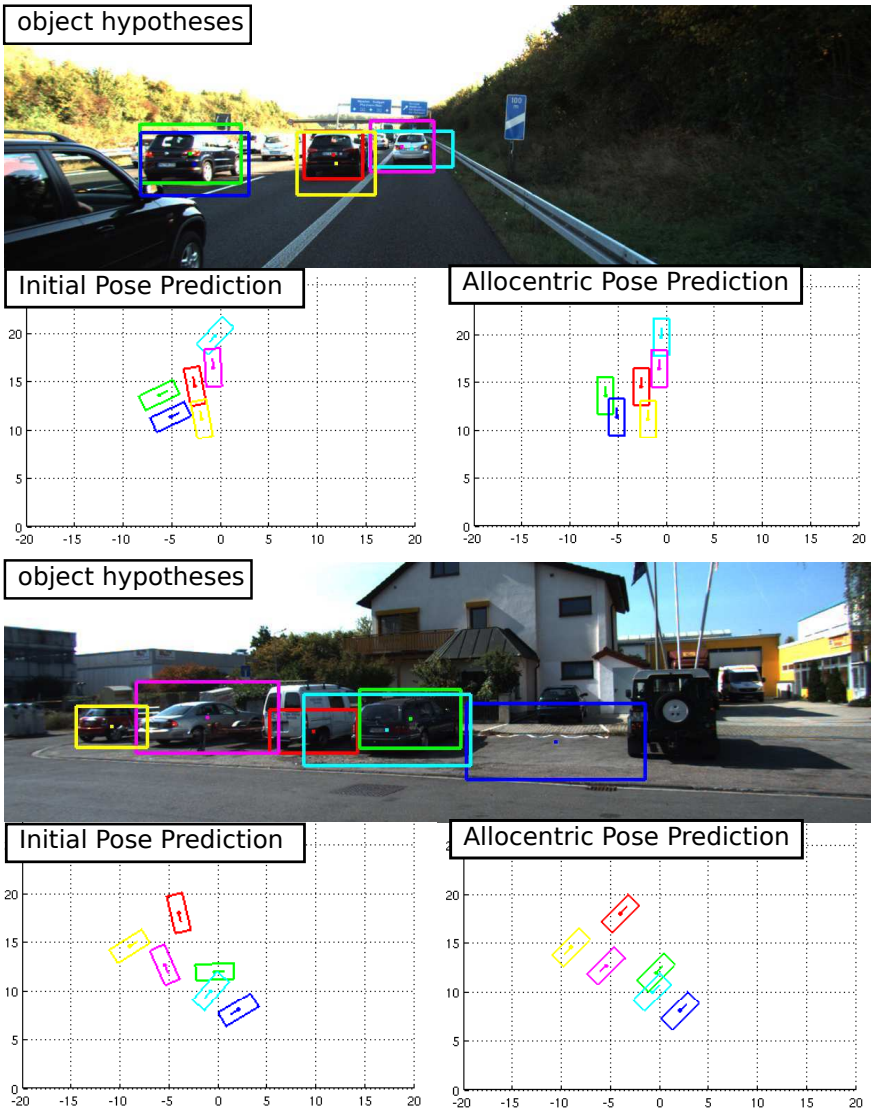

Figure 3.4: Effect of considering object configurations for pose estimation. Per set: Top image, hypotheses reported by the detector; bottom left, in top view, initial pose prediction given by the standard pose estimator; bottom right, in top view, pose prediction when considering object configurations.

Object Verification

While in this chapter we focus on the task of pose estimation, the configuration of objects and their poses in a neighborhood around a given object can also be exploited for object verification, i.e. to correct errors of the object detector. This is tested in the next experiment. We define *Object Verification* as the task of re-ranking the set of hypotheses given by a detector in such a way that the most likely hypotheses get a higher score. For this task we need a relational classifier that focuses on the prediction of the occurrence of an object o_i given the objects in its neighborhood N_i. We define this classifier as:

$$\begin{aligned} wvRN(o_i|N_i) &= \frac{1}{Z}\sum_{o_j \in N_i} v(o_i, o_j).w_j \\ wvRN(\theta_i^+, o_i^+|N_i) &= \frac{1}{Z}\sum_{o_j \in N_i} p(\theta_i^+, o_i^+|r_{ij}).p(o_j^+|s_j) \end{aligned} \tag{3.7}$$

where the weighting factor is assumed to be equal to $w_j = p(o_j^+|s_j)$ as presented in Section 3.4.2. The weighting term w_j is computed using Eq. 3.6. while the term $p(\theta_i^+, o_i^+|r_{ij})$ is computed as in Eq. 3.4. Since the pose of the predicted hypothesis is not evaluated, just its location, we will not take the response over possible poses as in Eq. 3.2. The task of object verification is evaluated based on the criterion used in Pascal VOC [36] (see Section 2.8.2). We report results using Average Precision (AP) as performance metric on the testing set described before. Additionally, we report the performance of using traditional camera-centered relations and our proposed object-centered Relations. Again we show results for the two selected object detectors, relational classifiers based on them, and the combination of the two (Table 3.3). Considering the fact that we are reasoning in 3D Space, we repeat the previous object verification experiment adding a pre-processing 3DNMS step applied on the 3D hypotheses (Table 3.4).

Discussion: The change in performance brought by the combination of local and relational classifiers, over the local classifier alone, confirms that indeed the proposed relations can also assist the task of object verification. In our experiments we obtained mean improvements of 3% and 4% for [81] and [47] baselines respectively. Furthermore, it is remarkable how the relational classifiers (RC) are clearly above their respective chance levels, 24% and 14%, by 10% and 16% respectively. These chance levels correspond to the true detection - false detection ratio of the baselines [81] and [47], respectively. Table 3.4 shows how using the "heuristic" 3DNMS step improves all the baselines by 7%. However, the improvement brought by contextual information in that case is

RC / LC	none	CCRel.	OCRel
None	-	0.342	0.347
Lopez et al.[81]	0.600	0.622	0.629
None	-	0.300	0.314
Geiger et al.[47]	0.637	0.666	0.671

Table 3.3: Object Verification Performance (AP) related to the baseline [81] and [47]. LC (Local Classifier), RC (Relational Classifier), CCRel (Camera-Centered Relations), OCRel (Object-Centered Relations).

reduced to 1% for both detectors. This can be explained by the fact that the increase in performance given by the 3DNMS makes the local classifier better, leaving less room for improvement. One might argue that our distance based 3DNMS is sub-optimal when compared with methods used in Holistic Scene Understanding for NMS based on volumetric overlap. However, our experiments presenting 3DNMS results should be considered as just a hint of additional advantages that can be obtained from reasoning in a 3D rather than a 2D space. Future work will address reasoning about the volumetric properties of objects and the effect of the re-estimated poses on the aspect ratios of the hypotheses initially predicted by the detector.

RC / LC	none	CCRel.	OCRel
None	-	0.396	0.399
Lopez et al.[81]	0.676	0.685	0.682
None	-	0.353	0.364
Geiger et al.[47]	0.717	0.724	0.725

Table 3.4: Object Verification Performance (AP) related to the baseline [81] and [47] using 3DNMS. LC (Local Classifier), RC (Relational Classifier), CCRel (Camera-Centered Relations), OCRel (Object-Centered Relations).

Object-centered or Camera-centered

To analyze the effect of the FoR when defining relations between objects, we evaluated the performance of the relational classifier with camera-centered relations and object-centered relations respectively (Sec. 3.3). As in the previous experiments, we present results in an ideal and realistic scenario. Furthermore,

we add an experiment on the realistic scenario where we apply 3DNMS as a preprocessing step. This complements the previous experiment involving these types of relations and will provide an overview of their effect in such tasks.

Detector	Relations	Real	Real (3DNMS)	Ideal
Lopez et al.[81]	CCRel.	0.20	0.19	0.44
	OCRel.	0.20	0.20	0.47
Geiger et al.[47]	CCRel	0.24	0.22	0.32
	OCRel.	0.27	0.24	0.37

Table 3.5: Effect of the Frame of Reference when defining relations for pose estimation (MPPE values per method). CCRel (Camera-Centered Relations), OCRel (Object-Centered Relations). Columns 3 and 4 present the performance when starting from the detections reported by the detectors [81] (8 poses) and [47] (16 poses) without and after 3D Non-maximum suppression. Finally, column 5 reports the performance in the ideal scenario.

Discussion: On the KITTI dataset, the difference between the object-centered and camera-centered settings seems to be minimal for object detection (Table 3.3). While the object-centered setting does not depend on the camera viewpoint and therefore can be expected to generalize better to different camera setups (e.g. surveillance cameras as opposed to cameras mounted on a vehicle), the camera viewpoints in the KITTI dataset are consistent, and therefore the camera-centered setting works equally well than the object-centered one. On the pose estimation problem, previous experiments proved that pose information plays an important role when defining relations. Here object-centered relations bring an improvement of $\sim$2% over their camera-centered counterparts (Table 3.5).

3.8 Conclusions

In this chapter we presented an initial attempt to reason about object configurations to estimate and refine object poses. Even when, in isolation, allocentric pose estimation does not solve the object pose estimation problem, its performance (clearly above chance levels) suggest that the proposed relations-based method is able to encode information about the pose of the related objects. Again, this is achieved by only looking at the relative location and pose of other objects in the scene. This makes it a good alternative for cases where local, appearance, information about the unknown object is unavailable (i.e. when augmenting a scene with virtual objects). Though there is room for improvement, our results support our hypothesis that there is something to gain

from object configurations when predicting object poses. These results directly answer Research Question 1, showing that contextual information, in the form of relations between objects, is useful for object pose estimation. Complementing this, the experiments show how defining relations from an object-centered perspective can increase performance in object pose estimation and detection. In this regard, increased performance gains are expected by using object-centered relations in datasets acquired with more arbitrary camera setups. The findings made in this chapter serve as starting point for following two directions of future work: first, the combination of allocentric pose estimation with more advanced local pose estimators that can reason about the 3D geometry of the object; and second, the use of more advanced relational classifiers and collective classification methods to reason about object configurations.

Chapter 4

Towards Cautious Collective Inference for Object Verification

From the experiments conducted in the previous chapter we observed that when the ratio between true and false positive object hypotheses was low, the proposed method for Allocentric Pose Estimation decreased its performance. This indicates that, initially high-ranked, false positive hypotheses introduce noise to the pose estimation process. To address this problem, we investigate methods to "clean" the initial hypotheses prior to any estimation processes. To this end, in this chapter we propose a method that re-ranks object hypotheses based on relations derived from their 2D bounding boxes. To achieve this re-ranking step, we define a cautious algorithm where each object hypothesis is re-ranked based on the most certain/reliable neighboring hypotheses. In addition, we explore the assumption that objects are not related by their categories, but by underlying relationships that explain how groups of objects co-occur. This chapter partially addresses Research Question 2 by analyzing the effect that different types of object associations have on object detection performance.

Work and findings corresponding to this chapter were published in:

- Oramas M, J., De Raedt, L., and Tuytelaars, T. *Towards cautious collective inference for object verification.* In IEEE Winter Conference on Applications of Computer Vision (WACV) 2014.

4.1 Introduction

Recently, contextual information has been used in several computer vision tasks including segmentation [44, 59, 68] and object detection [30, 78, 126]. As presented in Section 1.2.3, for the object detection problem, relations between object instances have been used to remove or reduce the uncertainty in object hypotheses predicted by appearance-based detectors. The underlying methods differ in the way they define neighboring objects. Some works (e.g. [30, 118]) use all the other objects as neighbors, while others (e.g. [44]) use only the objects in a spatial vicinity. We refer to these two types as "global" and "near" neighborhoods, and empirically evaluate which setting yields best results.

For inference, the neighboring object hypotheses are commonly considered without taking into account the certainty of their prediction. As as result, *all* the neighbors participate for the classification of each object [118]. Following the literature [95, 96] on Collective Classification [134], instead, we propose an iterative scheme where we first classify the objects with certain relational information, and then use these to bootstrap the predictions of the other objects. This is useful in collective classification tasks, like object detection, where multiple possibly related objects all need to be classified. Following the terminology of [95], we refer to these two inference variants as "aggressive" and "cautious" inference. Again, we empirically evaluate the added value of cautious vs. aggressive inference.

Furthermore, probabilities or likelihoods are typically computed based on the frequency of occurrences of object relations in the training data. Usually, this is computed relative to *all* the relations involving two objects of the same category. This is an example of classical *homophily*-based relational classification. *Homophily* is the tendency of individuals to associate with others of the *same* category. This homophily-based model is inspired by observations in a vast array of network studies, e.g. [97], in both explicitly defined and latent-assumed networks. In *homophily*-based relational classification, objects are expected to give higher support to hypotheses belonging to the same category [91] independent of the relation between them. Here we also investigate an alternative definition for *homophily*, based on the relation between object instances rather than strictly focused on the categories of the objects. Following this idea, we assume that the observed pairwise relations between objects belong to a set of underlying relationships that determine how the different objects are associated with each other. In this setting, during inference, only a subset of the relations (those covered by the same relationship) are involved in the estimation of probabilities or likelihoods. We refer to these two cases as "category-based homophily" and "relation-based homophily", and empirically evaluate their respective merits. Let us illustrate these ideas by an example. Imagine you are

given the task of predicting whether the green box in Fig. 4.1 (corresponding to an object hypothesis) contains a car or not, based on the context given by the objects in the other bounding boxes (Fig. 4.1a). Shouldn't the true hypotheses, in blue, have a higher influence on the prediction than the false hypotheses, in red? Furthermore, focusing on the true hypotheses (Figure 4.1b), wouldn't it be more intuitive to take into account also the color-codes of the objects (defined by their relations with the object under the green box)?. These two figures, Fig. 4.1a and Fig. 4.1b , depict examples of cautious inference and relation-based homophily respectively.

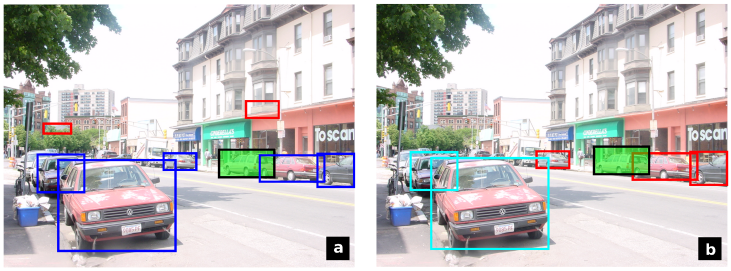

Figure 4.1: Different ways in which neighboring objects can cast a vote for the query object (green box): a) cautious inference, selecting the more reliable neighbors, b) relation-based homophily, considering the relations between objects. (See text for details.)

The main contributions of this chapter are: a) we show that cautious inference on object relations brings improvements over traditional inference for object detection, and b) we investigate a notion of relational homophily, which specifies how neighboring objects influence unknown objects based on underlying relationships. In addition, we evaluate these techniques using three different representations for object relations. These contributions are very related to Research Question 2, which focuses on the study of how the way in which objects are linked, or related, affects object detection performance when applying relational reasoning.

This chapter is organized as follows: Section 4.2 presents existing work on using relations between objects for object detection. Section 4.3 introduces how we define and learn relations between object hypotheses provided by a detector and emphasizes the principles to be considered when performing cautious inference with such relations. Section 4.4 presents how the output of the local and relational classifiers are combined. In section 4.5 we provide implementation details, while section 4.6 describes the experimental results. Finally, we draw conclusions in section 4.7.

4.2 Related Work

In Section 1.2 we provided an overview of sources of contextual cues that are commonly used to assist the object detection problem. Similar to chapter 3, in this chapter we focus on the contextual cues provided by *Things-Things* relations. Using this type of contextual cue, [30] represented objects as regions in the image. Then, by learning qualitative spatial relations (e.g. top-left, far-left) between them, hypotheses in unlikely areas were filtered out. Following this trend, Felzenszwalb et al. [40], Perko & Leonardis [118] and Choi et al. [22] defined continuous spatial relations between the centers of object bounding boxes and used the learned relations to filter out the out of context objects. Song et al. [138] suggested a joint detection-classification scheme to identify ambiguous ranked hypotheses and use an adaptive method to exploit context information on these hypotheses. This method favored the prediction of the local detector on top and low ranked hypotheses and considered the context information for the ambiguous cases. In [23] relations between objects considered additional features such as relative scales, bounding box overlap ratio and scores. Furthermore through a set-based formulation this method is able to reason about object spatial configurations that go beyond pairwise interactions. Similar to these works, we learn relations between object instances. In particular, we build on the work of [118] and their framework based on Kernel Density Estimation (KDE) to estimate the probabilities of certain relations. Different from these works, which follow an "aggressive" approach considering all the neighboring objects as sources of contextual support, we explore an alternative "cautious" inference scheme in which the set of neighboring objects is defined based on the level of certainty of their occurrence. Our method is inspired by *Cautious Inference* [96, 103], a method used in collective classification that seeks to identify and exploit the more certain relational information. Such a cautious approach was used in [68] for the problem of labeling object superpixels. In [68], discriminative relations were mined between object regions and discriminative attributes were discovered per relation. Our approach to object detection differs from that in [68] in that in object detection the object regions (the bounding boxes) can overlap, which increases the complexity of the problem. To the best of our knowledge, this is the first attempt to use *cautious* collective inference to improve object detection.

An important issue in collective classification problems is the way in which each of the neighboring objects casts a vote on the unknown object. Experiments [89, 91] with Collective Classification algorithms on text data used "category-based" homophily models, that is, an object is only associated with objects of the same category. These models have the limitation that the association of objects is strictly category-driven, and the weakness that all the neighboring

objects influence the occurrence of others of the same category in a direct and rather crude fashion. Recent work in object detection that exploits object relations takes into account inter/intra category object relations, however they still perform inference using the relations between objects directly. We follow the suggestion of [102], which aims to uncover underlying groups, which represent the cause of the frequently seen pairwise relations. However, different from [102], we do not require fully visible explicit relations at both training and testing stages. Furthermore, since in our work we define attribute-based relations, there is a relation between any pair of objects, linked to its underlying group. This removes the requirement of group membership for objects that are not explicitly related. Summarizing, our work differs from the category-based homophily models since we assume that objects are not linked by their category, but by the relationships underlying the visible pairwise relation between objects of the same category. Focusing on the relational aspect of the problem, we formulate our object classification problem as a Within-Network classification problem, which consists of making a prediction about an object based on the neighboring objects.

A recent group of works [78, 126] promotes the use of groups of objects with consistent relations among them. Following this idea, [126] exploits explicitly defined pairwise relations to learn the collective appearance of object pairs. Taking this idea further, [78] removes the requirement of explicitly defined relations and discovers composite relations to learn the appearance of the group of objects. Our procedure of recovering the underlying *relationships* and their corresponding densities is quite similar to the Hough transform and mode finding approach in [78]. However, different from [78] and [126], which use underlying groups towards learning the collective appearance of the groups, we discover the underlying groups as a means to improve object detection accuracy.

4.3 Object Relations as Source of Context

Before we discuss how relations between objects can be used as a source of contextual information, we introduce the representations for objects and relations used in this chapter. Given an image, we use an object detector to collect a set of object hypotheses $O = \{o_1, o_2, ..., o_n\}$ of the category of interest. Each object hypothesis o_i is represented as a tuple $o_i = (x_i, y_i, f_i, s_i)$ where (x_i, y_i) represents the location of the center of the bounding box of the object, f_i represents additional object-related features (e.g. aspect ratio or scale of the bounding box), and s_i the detection score reported by the detector. In addition, we will refer as o^+ to the object hypotheses that are well localized, i.e. their bounding boxes cover valid object instances. On the contrary, we will refer as

o^- to false object hypotheses. Given the set of hypotheses O, we define pairwise relations r_{ij} between each pair of objects o_i and o_j. Here, the relations r_{ij} are defined by relative attributes such as relative location, relative size, etc. Note that no attribute regarding the category of the participating objects is specified. In section 4.5 we describe how we compute the relative attributes that define the relations r_{ij} .

4.3.1 Inference

In this chapter we follow the principle proposed in [95] that stresses that instances are not independent, on the contrary, "in some classification tasks they are implicitly or explicitly related". Therefore, we estimate the degree to which an object o_i fits into the scene based on its relations with the other objects in the scene. This is a *Collective Classification* [134] problem in which the occurrence (class) of an object influences that of another (see Algorithm 1). For simplicity we focus on the case of a single object category for now. To take into account the interdependencies between objects based on their relations we re-rank the predicted object hypotheses using the Weighted Vote Relational Neighbor Classifier (wvRN) [91] introduced in Section 2.7.2. It is defined similarly as in the previous chapter:

$$\begin{aligned} wvRN(o_i|N_i) &= \frac{1}{Z}\sum_{o_j \in N_i} v(o_i, o_j).w_j \\ wvRN(o_i^+|N_i) &= \frac{1}{Z}\sum_{o_j \in N_i} p(o_i^+|r_{ij}).p(o_j^+|s_j) \end{aligned} \tag{4.1}$$

where $wvRN(o_i^+|N_i)$ represents the likelihood of an object o_i occurring, i.e. of being a true positive o_i^+, given its neighborhood N_i. The term Z is a normalization factor and w_j is a weighting term that takes into account the noise in the object detector. It is computed as $Z = \sum w_j = \sum p(o_j^+|s_j)$. The conditional $p(o_i^+|r_{ij})$ represents the probability of object o_i occurring, i.e. of being a true positive o_i^+, given its relation r_{ij} with object o_j. Using Bayes' Rule we estimate $p(o_i^+|r_{ij})$ as the posterior:

$$p(o_i^+|r_{ij}) = \frac{p(r_{ij}|o_i^+)p(o_i^+)}{p(r_{ij}|o_i^+)p(o_i^+) + p(r_{ij}|o_i^-)p(o_i^-)} \tag{4.2}$$

The components of Eq. 4.2 are obtained through the following procedure. First, we run the local detector on a training set with annotated objects producing a

set of hypotheses per image. Then we label the hypotheses as true positives o_i^+ or false positives o_i^- based on the Pascal VOC [36] matching criterion. In order to avoid repeated object instances, we replace true positive hypotheses by their corresponding annotations. We define pairwise relations r_{ij} between the hypotheses reported for each image generating a set of relations $R = \{r_{ij}\}$ for the whole training set. Finally, during testing, $p(r_{ij}|o_i^+)$ and $p(r_{ij}|o_i^-)$ are estimated via multivariate Kernel Density Estimation (KDE) using the pairwise relations r_{ij} as sample points. This method captures the statistics of typical configurations between objects. The priors $p(o_i^+)$ and $p(o_i^-)$ are estimated as the proportion of true positive and false positive hypotheses in the training set, respectively.

The weighting factor w_j of equation 4.1 takes into account the noise that is introduced by the object detector in the neighboring objects o_j. We estimate w_j using a *Probabilistic Local Classifier* that takes into account the score s_j provided by the object detector for its respective hypothesis o_j. The output of this classifier will be the posterior $p(o_j^+|s_j)$ of the occurrence of the object o_j given its detection score s_j. We compute this posterior following the procedure presented in [118]:

$$w_j = p(o_j^+|s_j) = \frac{p(s_j|o_j^+)p(o_j^+)}{p(s_j|o_j^+)p(o_j^+) + p(s_j|o_j^-)p(o_j^-)} \tag{4.3}$$

The components of this equation are obtained following a procedure similar to that for Eq. 4.2 up to the point where hypotheses are labeled as true or false positives. Then, based on the true and false positives we compute the conditionals $p(s|o_j^+)$ and $p(s|o_j^-)$ respectively via KDE. Finally, the priors $p(o_j^+)$ and $p(o_j^-)$ are estimated in the same way as $p(o_i^+)$ and $p(o_i^-)$ in Eq. 4.2. As a result, $p(o_j^+|s_j)$ expresses the probability of a hypothesis being a true positive given its detection score. This procedure allows us to plug-in any standard object detector in our method.

Finally, the condition $o_j \in N_i$ of the sum in Eq. 4.1 is of relevance for wvRN inference since wvRN estimates class-membership probabilities based on two assumptions: First, the class (occurrence) of an object depends on its neighbors, and second, the entities exhibit *homophily* in their behavior, i.e. they tend to associate with other objects of the same category. This implies that the posteriors $p(o_i^+|r_{ij})$ and $p(o_i^-|r_{ij})$ are estimated considering all the pairwise relations r_{ij} between objects of the same category. This suggests that objects of different categories should not associate or influence each other. Here, we explore an alternative idea, and assume that objects are not associated by their category, but by underlying "relationships". As a result, the posteriors $p(o_i^+|r_{ij})$ and $p(o_i^-|r_{ij})$ are computed by considering only the pairwise relations

r_{ij} covered by specific "relationships". Next, we will specify how to integrate "cautious" inference in this model and propose an alternative idea to define the associations between objects.

4.3.2 Cautious Inference

An algorithm is considered "cautious" if it seeks to identify and employ the more certain or reliable relational information [95]. We focus on two factors that [95] introduces to control the degree of caution in an algorithm. The first factor dictates to use only objects for which the prediction is confident enough. The second factor increases caution by favoring already-known relations. These are relations between objects that have been seen in the training images. See Section 2.7.1 for more details about the factors that control the degree of caution of a relational classifier.

For the aggressive version of our relational classifier, we use wvRN as described in Eq. 4.1. For each object hypothesis, it considers *all* the other objects o_j in its neighborhood N_i during the inference. For the *cautious* version of our relational classifier, we enforce the above principles in the following fashion. For the first principle, giving relevance to certain objects, we perform an iterative approach inspired by [103]. Given a set of hypotheses $O = \{o_1, o_2, ..., o_n\}$, we define the disjoint sets O^k and O^u as the known and unknown objects, respectively, with $O = O^k \cup O^u$ at all times. We initialize $O^k = \{\}$ and $O^u = O$ and flag as *known* object, the hypothesis with the highest score based on the probabilistic local classifier (Eq. 4.3). This hypothesis is moved to the set of known objects O^k. Then, the wvRN score for the unknown objects o_i^u is re-estimated considering *only* the known objects o_j^k in their neighborhood N_i. This re-defines Eq. 4.1 in the following way:

$$wvRN(o_i^{u+}|N_i) = \frac{1}{Z} \sum_{o_j^k \in (N_i \cap O^k)} p(o_i^{u+}|r_{ij}).p(o_j^{k+}|s_j) \tag{4.4}$$

We flag the hypothesis with highest wvRN score as *known* and move it to the set of known objects O^k. We repeat this procedure promoting one hypothesis o_i^u at a time until the set of unknown objects O^u is empty. Finally, for the sake of similarity in the ranking of the new scores, we re-estimate the score of the first promoted object using Eq. 4.4 with the second promoted object as known neighbor.

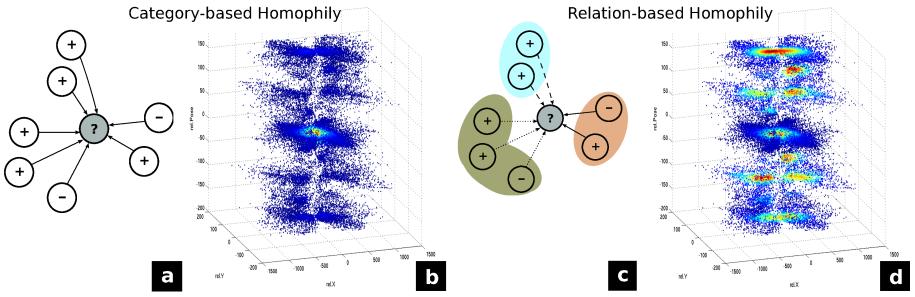

Figure 4.2: Cautious definition of Homophily. Category-based homophily: a) voting, b) density distribution; and Relation-based homophily c) voting, d) density distribution. Density distributions from RF1 relations on the KITTI dataset. See Sec. 4.5 and Sec. 4.6)

For the second principle of *cautious* inference, "favoring relations already seen on training data", our use of KDE for estimating the vote $p(o_i^{u+}|r_{ij})$ from each neighbor object o_j^k implicitly introduces this characteristic in the inference.

Relation-based Homophily: We add and explore an alternative principle to define homophily to use present relations. Our additional principle emphasizes the homophily exhibited in the behavior of the objects. We assume that the pairwise relations estimated from object pairs belong to a set of underlying relationships $Q = \{q_1, q_2, ..., q_m\}$ and that objects tend to associate with each other based on these relationships. As a consequence, homophily will shift from being defined by the object category (as in Eq. 4.4) to being defined by these relationships. To take into account this principle we compute the neighbor vote $p(o_i^{u+}|r_{ij})$ of Eq. 4.4 using an intermediate clustering step. First, we extract all the pairwise relations r_{ij} between the annotated objects from a training set of images. Then, we run an unsupervised clustering algorithm on them, producing a set of clusters. The centers of these clusters then represent the relationships $\{q_1, q_2, ..., q_m\}$. At test time, each of the pairwise relations r_{ij} is assigned to its closest relationship q_{ij} and the relations computed from the object hypotheses in the training set for the cluster q_{ij} are used to perform kernel density estimation. In summary, as Fig. 4.2c presents, all the known neighboring nodes participate as in Eq. 4.4 but their vote depends on the relationship that links them. Additionally, comparing the relation-based density distribution (Fig. 4.2d) with its category-based equivalent (Fig. 4.2b), one can see that considering underlying relationships has the effect of removing the bias towards the most frequent pairwise relation that is introduced when all the pairwise relations are used for inference (Fig. 4.2b).

4.4 Combining Information Cues

At this point, we have two methods to estimate the probability of the occurrence of an object hypothesis o_i: the local classifier, based on appearance, as evaluated by the object detector, and the relational classifier, based on its neighborhood N_i. The reader should note that while the local classifier pulls the decision towards individual features, the relational classifier (Eq. 4.4) pulls it towards the collective feature of group fitting. Given this opposite behavior of our classifiers, local and relational, we need a method to combine them. We follow a method similar to [118]. We use a validation set of images on which we run the object detector. After defining pairwise relations between object hypotheses, we label them as true and false positives using the annotations. Then, for each object hypothesis, we compute the score pair $(s_{\text{LC}}, s_{\text{RC}})$ of the local and relational classifier for each image. For the local classifier, we use the output of Eq. 4.3, applied on o. For the relational classifier we use the response of Eq. 4.4. Using these pairs we estimate the conditionals $p(s_{\text{LC}}, s_{\text{RC}}|o^+)$ and $p(s_{\text{LC}}, s_{\text{RC}}|o^-)$ via Kernel Density Estimation. Finally, the probabilistic score with enforced consistency is estimated as the posterior $p(o+|s_{\text{LC}}, s_{\text{RC}})$ using Bayes' Rule (see Eq. 4.5) with $p(o^+)$ and $p(o^-)$ determined as for Eq. 4.2.

$$p(o^+|s_{\text{LC}}, s_{\text{RC}}) = \frac{p(s_{\text{LC}}, s_{\text{RC}}|o^+)p(o^+)}{p(s_{\text{LC}}, s_{\text{RC}}|o^+)p(o^+) + p(s_{\text{LC}}, s_{\text{RC}}|o^-)p(o^-)} \tag{4.5}$$

4.5 Implementation Details

This chapter studies the impact of *cautious inference*, when reasoning about object relations, for object detection. For this reason rather than proposing our own object detector we use a state-of-the-art detector to acquire evidence of objects in the scene. We build on top of the DPM-based viewpoint-aware detector proposed in [81][1], and which was previously introduced in Section 2.4. This detector feeds our framework with confidence scores, locations (2D bounding box) and the orientation angle of object hypotheses discretized into 8 viewpoints.

We define relations between objects in three formats. The first format (RF1) considers differences in x- and y-coordinates $(\Delta x_{ij}, \Delta y_{ij})$ in the 2D image space and the relative viewpoint $\Delta\alpha_{ij}$ of the viewpoint α predicted by the object detector producing a triplet $r_{ij}^{(RF1)} = (\Delta x_{ij}, \Delta y_{ij}, \Delta\alpha_{ij})$. The second format (RF2) is based on [78]. In this work relations are represented as a tuple

[1] http://agamenon.tsc.uah.es/Personales/rlopez/data/pose-estimation/

$r_{ij}^{(RF2)} = (rx_{ij}, ry_{ij}, r\rho_{ij}, ra_{ij})$ where $rx_{ij} = x_i - x_j \frac{\rho_i}{\rho_j}$ and $ry_{ij} = y_i - y_j \frac{\rho_i}{\rho_j}$. The factor $\frac{\rho_i}{\rho_j}$ normalizes the translation by object size and is used as a proxy for handling the global scale of the scene. $r\rho_{ij} = \frac{\rho_i}{\rho_j}$ denotes the relative scale ρ_i (the scale of object o_i) and is computed as the square root of the bounding box area of the object. Finally, $ra_{ij} = \frac{a_i}{a_j}$ represents the relative viewpoint, where the viewpoint a_i is encoded by the aspect ratio of the bounding box. The third format (RF3), is purely spatial and considers differences in x- and y-coordinates $(\Delta x_{ij}, \Delta y_{ij})$ only in the 2D image space. This is used in cases where object viewpoint annotations are not available.

In our experiments relationships are discovered using the XMeans [27] clustering algorithm. XMeans is an iterative version of an accelerated KMeans in which the user only provides the range of values in which K may be located. In our experiments, we provide the range $K \in [4, 64]$ to the XMeans algorithm.

Kernel Density Estimation (KDE), with $f(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h})$ is performed using publicly available code[2]. We use multivariate KDE for estimating $p(r_{ij}|o_i^+)$ and $p(r_{ij}|o_i^-)$ in Eq. 4.2 using the relations $R = \{r_{ij}\}$ as sample points x_i. For the case of Eq. 4.3 we use univariate KDE to estimate $p(s_j|o_j^+)$ and $p(s_j|o_j^-)$ using the detection scores s_j as sample points x_i. For both cases we use Gaussian kernels for K, and fixed bandwidth values h. These h values are obtained in a data-driven fashion using Silverman's Rule of Thumb [154]. See Section 2.5 for more details about Kernel Density Estimation

4.6 Evaluation

Datasets: We run experiments in the object detection set of the *KITTI benchmark* [45] introduced in Section 2.8.1. We evaluate against all the object annotations independent of their occlusion level. In our experiments, the training set is used for extracting the pairwise relations used to perform KDE in the relational classifier and to discover relationships. The validation set is used for learning the combination of the local and the relational classifier. This dataset was obtained using a car-mounted camera and resembles the settings used for autonomous navigation. Additionally, we run experiments on the *MIT-StreetScenes* (MITSS) dataset [13]. As described in Section 2.8.1 this dataset was obtained using a consumer camera and offers more viewpoint variability. We run experiments on three splits of each dataset and report mean performance results. In addition, in order to check the behavior of the object detector, the relational classifier and the combination of the two, we split the

[2] http://www.ics.uci.edu/∼ihler/code/kde.html

test set in three subsets. These subsets are defined based on the number of hypotheses available for the inference stage. The subsets contain images with [2,3],[4,7] and [8,∞) hypotheses respectively.

Experiment: We reason about object relations as means for object verification, i.e. to correct errors of the object detector. We define *Object Verification* as the task of re-ranking the set of hypotheses given by a detector in such a way that the most likely hypotheses get a higher score. The task of object verification is evaluated following the protocol used in Pascal VOC [36] (see Section 2.8.2). We report results using the mean Average Precision (mAP) computed over the data splits.

We report experiments with eight baselines defined by the combination of three parameters. The first parameter *Neighborhood Scope*, indicates how the neighborhood N_i of a particular object o_i is defined based on their relative location. It is set to "global" if it considers *all* the objects despite their location. It is set to "near" if it only considers the objects within a relative distance t where t is defined as the median distance of all the spatial relations in the training set. This parameter represents the Markovian assumption that some relational methods enforce by only considering neighboring objects in their spatial vicinity. The second parameter indicates the type of inference to use which can be "aggressive" (Eq. 4.1) or "cautious" (Eq. 4.4). The last parameter *Homophily drive* covers the possible causes that relate entities. It can be driven by the *category* of the object, as in traditional homophily, or by the *relationships* that we propose in this work. We present results using the relation representations RF1 and RF2 (Sec. 4.5) for the KITTI dataset. For the case of MITSS we use representations RF2 and replace RF1 with RF3 (Sec. 4.5) due to the lack of annotated object viewpoints. Table 4.1 shows the performance of the different baselines when only the relational classifier is used, that is, only considering contextual information. Table 4.2 shows the performance of the combination of local and relational classifiers for the top performing baselines. Note that the baseline defined by *aggressive* inference with RF3 relations, assuming *Category-based Homophily* in a *Global* Neighborhood, is a *Things*-based version of [118].

Discussion: Overall, based on the parameters previously mentioned, the performance of the evaluated algorithms present the following trend: First, and maybe somewhat surprisingly, on average, *global* neighborhoods provide higher performance than the *near* option. Second, on the scope of a *global* neighborhood, *cautious* methods outperform their *aggressive* counterparts. Third, dataset-wise, *Relation-based Homophily* performs better in the KITTI dataset, where camera settings are more constrained. This may suggest that the method to uncover relationships may be sensible to changes in viewpoint. Finally, the proposed *cautious* scheme boosts the performance of the baselines [40, 81, 118]. Now we

Dataset **KITTI benchmark**	**Relations Representation : RF1**							
	Category-based Homophily				Relation-based Homophily			
	Global		Near		Global		Near	
Set	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.
2-3	0.33	0.35	0.32	0.33	0.33	0.35	0.27	0.34
4-7	0.31	0.40	0.30	0.29	0.30	0.40	0.28	0.35
8+	0.28	0.36	0.27	0.23	0.26	0.36	0.26	0.30
all	0.29	***0.38***	0.29	0.26	0.28	***0.37***	0.27	0.32
	Relations Representation : RF2							
	Category-based Homophily				Relation-based Homophily			
	Global		Near		Global		Near	
Set	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.
2-3	0.32	0.32	0.36	0.34	0.40	0.39	0.37	0.36
4-7	0.34	0.43	0.38	0.34	0.44	0.53	0.41	0.49
8+	0.30	0.39	0.37	0.29	0.40	0.51	0.40	0.44
all	0.32	0.40	0.37	0.31	0.41	***0.50***	0.40	0.45

Dataset **MIT StreetScenes**	**Relations Representation : RF1**							
	Category-based Homophily				Relation-based Homophily			
	Global		Near		Global		Near	
Set	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.
2-3	0.71	0.74	0.65	0.58	0.70	0.71	0.62	0.60
4-7	0.54	0.65	0.46	0.42	0.49	0.59	0.48	0.50
8+	0.35	0.46	0.30	0.29	0.33	0.45	0.29	0.39
all	0.54	***0.63***	0.47	0.43	0.51	0.59	0.47	0.50
	Relations Representation : RF2							
	Category-based Homophily				Relation-based Homophily			
	Global		Near		Global		Near	
Set	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.
2-3	0.70	0.69	0.68	0.61	0.67	0.68	0.68	0.66
4-7	0.49	0.56	0.50	0.48	0.47	0.55	0.52	0.57
8+	0.33	0.42	0.38	0.34	0.33	0.43	0.39	0.47
all	0.51	***0.56***	0.52	0.48	0.49	0.55	0.53	***0.57***

Table 4.1: Mean Average Precision of the Relational Classifier for object detection on the KITTI and MITSS datasets. (Only using context to predict the object presence)

discuss the results in more detail.

Regarding the *relations format*, the difference in performance of RF2 on the different datasets in Table 4.1 suggests that RF2 is better suited for working on constrained camera settings, as in the KITTI dataset. Furthermore, the difference in performance between RF1 and RF3, shows a weakness of relational methods when relations are defined from, possibly, unstable attributes. In this case, the relative viewpoint information used in RF1 may be the cause of its relatively lower performance.

Regarding the *type of inference* to use, both Tables 4.1 and 4.2, show that *cautious* reasoning with object relations always outperforms its aggressive

Dataset **KITTI benchmark**		**RF1**		**RF2**	
		Category-based Homophily		Relation-based Homophily	
		Global		Global	
Set	Detector [81]	aggre.	caut.	aggre.	caut.
2-3	0.65±0.027	0.65±0.022	0.66±0.020	0.66±0.033	0.64±0.017
4-7	0.63±0.010	0.64±0.009	0.66±0.016	0.67±0.017	0.71±0.019
8+	0.60±0.011	0.59±0.007	0.61±0.004	0.63±0.004	0.68±0.009
all	0.61±0.011	0.61±0.009	0.63±0.007	0.65±0.011	***0.68±0.003***

Dataset **MIT StreetScenes**		**RF3**		**RF2**	
		Category-based Homophily		Category-based Homophily	
		Global		Global	
Set	Detector [81]	aggre.	caut.	aggre.	caut.
2-3	0.74±0.005	0.83±0.007	0.86±0.002	0.79±0.009	0.80±0.011
4-7	0.68±0.005	0.77±0.001	0.81±0.031	0.73±0.004	0.77±0.016
8+	0.68±0.033	0.69±0.003	0.71±0.044	0.68±0.043	0.70±0.030
all	0.69±0.006	0.77±0.001	***0.80±0.028***	0.73±0.011	0.76±0.014

Table 4.2: Mean Average Precision of the top performing baselines of the combination of Local [81] and Relational Classifiers for object detection on the KITTI and MITSS datasets. Note that the baseline defined by *aggressive* inference with RF3 relations, assuming *Category-based Homophily* in a *Global* Neighborhood, is a *Things*-based version of [118].

counterpart when exercised on a *global* neighborhood. This is supported by mean improvements, over traditional *aggressive* inference, of 8%, on the Relational Classifiers (Table 4.1), and 2.5% on the combination of Local and Relational classifiers (Table4.2). In addition, there is an improvement of 5% and 3% over the baselines [81] and [118], respectively. It may seem that the proposed method for cautious inference has a significant weakness in that it relies on identifying a true positive object hypothesis to seed the rest of the process. In the current implementation, such hypothesis is found by taking the object hypothesis with the highest detection score (Sec. 4.3.2). This method may fail for other object categories which are more difficult to detect. However, as mentioned in Sec. 2.7.1, there is some evidence [48] that suggests that this type of iterative cautious algorithms are fairly robust to a number of simple ordering strategies. This suggests that the inference method might be able to recover from possible mistakes while visiting and promoting object hypotheses (Sec. 4.3.2). Furthermore, additional alternatives can be tested where not just the top one but a few object hypotheses are tried out as seed objects.

Related to the alternative notion of *homophily*, Relation-based homophily outperforms class-based homophily on a *global* neighborhood when using RF2. This is opposite to what is seen with the related RF1 and RF3 where

category-based homophily performs better. It seems that, similar to RF2, Relation-based homophily performs better in constrained settings, with lower viewpoint variability as in KITTI. In this context, the representation used for the relations plays a relevant role since the clustering method used to discover the underlying relationships operates directly on the attributes of the pairwise relations. Likewise, the method to discover these underlying relations affects the inference process, i.e. boundary effects that can be introduced by hard clustering methods as the one employed in this work. Future work will focus on analyzing the influence that the selected method for discovering the relationships has on relation-based homophily. The mean boost in performance of 8.5% on the relational classifier makes relation-based homophily an appropriate principle in scenarios where no local information is available on the unknown object. Indeed, it is remarkable that the cautious relational classifiers, only using context information, can get as low as 8% behind the local detector for their top performing cases. Note in Fig. 4.3 how the baselines based on cautious inference effectively promote hypotheses that had been ranked low by the detector.

The change in performance obtained by the local classifier, the object detector, and the relational classifier in the different subsets of images hints at the scenarios for which each classifier is better. For the local classifier, its performance is at its highest point when a low number of hypotheses is reported and decreases as the number of hypotheses increases. This represents the scenario with few, possibly non-overlapping, hypotheses (see Fig. 4.3 top row). On the other hand, the relational classifier performs better as the number of hypotheses increases (see Fig. 4.3 and Fig. 4.4). This proves their "competitive" behavior. However, it should be noted that the increase in performance of the relational classifier has a peak in the second subset of images. It should be noted that the following subset, where performance drops, is the one with higher number of hypotheses, thus, more likely to contain a larger proportion of false hypotheses. This is confirmed by the performance of the object detectors for each of the subsets. The true positive - false positive ratio is of importance if we consider that the number of relations grows almost exponentially with the number of object hypotheses, hence introducing a significant amount of noise in the context-based classification process. This ratio is known as *class skewness*, or *labeled proportion*, in the collective classification literature [20, 91, 96, 141], more specifically when focusing on within network classification tasks where predictions about some nodes are based on other nodes. In this type of tasks, class skewness measures the proportion of data that is known, or predicted, with high certainty w.r.t. the whole data. In scenarios where class skewness is low there is not enough certain information to guide the inference process, e.g. our experiments on the third subset of images. In scenarios with high class skewness, the performance of collective classification is comparable to that of local classification. Based on these observations we stress that class skewness is a factor that can give

Dataset **KITTI benchmark**		**RF3**		**RF2**	
		Category-based Homophily		Relation-based Homophily	
		Global		Global	
Set	Detector [40]	aggre.	caut.	aggre.	caut.
all	0.65±0.003	0.68±0.007	0.71±0.007	0.72±0.009	**0.75±0.003**
Dataset **MIT StreetScenes**		**RF3**		**RF2**	
		Category-based Homophily		Category-based Homophily	
		Global		Global	
Set	Detector [40]	aggre.	caut.	aggre.	caut.
all	0.62±0.004	0.66±0.011	**0.71±0.012**	0.65±0.026	0.69±0.014

Table 4.3: Mean Average Precision of the top performing baselines of the combination of Local [40] and Relational Classifiers for object detection on the KITTI and MITSS datasets. Note that the baseline defined by *aggressive* inference with RF3 relations, assuming *Category-based Homophily* in a *Global* Neighborhood, is a *Things*-based version of [118].

an indication of whether the method proposed on this chapter will perform successfully on a specific problem. For the combination of the two classifiers, there is a similar trend along the different image subsets. We see that the combination of the responses of the local and relational classifiers produces an average and maximum improvement of 5% and 9%, respectively over the baseline from [81].

For object detection the use of a neighborhood with reduced spatial scope is discouraged since it has relatively lower improvement of 1.3% than when reasoning in *global* neighborhoods where a mean improvement of 4.7% was obtained over different relations representations. A possible cause for this difference in performance could be that most of the object hypotheses occurring within a *near* neighborhood might be false hypotheses. As a result, the contextual model learns that high-overlapping hypotheses are very likely to be false. This type of context model is helpful for performing non-maximum suppression (see Section 2.2), but has little effect on removing out-of-context hypotheses located at larger distances. Note that for the case of object-centric datasets, i.e. datasets where a single object instance covers a significant part of the image, *global* and *near* neighborhoods are expected to have a similar coverage and performance.

Finally, we ran experiments using the detector from [40] to generate the initial object hypotheses and defined RF2 and RF3 relations between objects. Table 4.3 shows how results follow a similar trend as the ones obtained in the other experiments.

Taking a brief look outside the field of computer vision, similar findings were

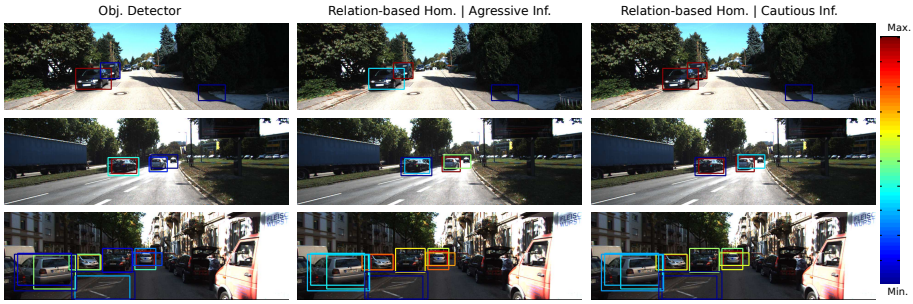

Figure 4.3: Qualitative results in a Global Neighborhood setting. Confidence scores color coded in jet scale. Note how Cautious Inference promotes hypotheses with initial low score. (Best viewed in color)

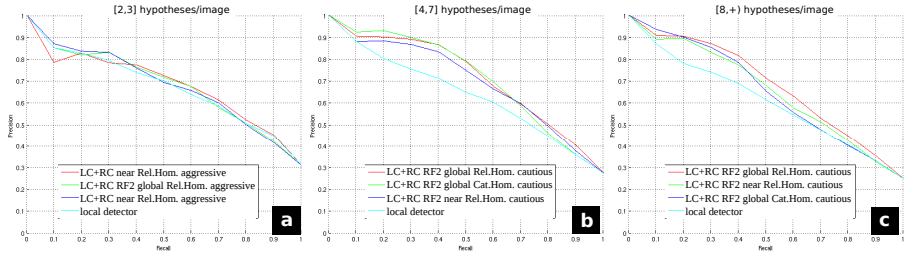

Figure 4.4: Precision-Recall curves for the top 3 ranking baselines on the KITTI dataset for the different image sub-sets based on their respective number of hypotheses: a) [2,3], b) [4,7] , and c) [8,∞) hypotheses/image respectively.

obtained in [157] for the task of node classification in multi-relational networks. Similar to our setting, the method from [157] starts from a graph with some local information about its nodes. Furthermore, an edge clustering method is used to extract "social context features" from each node. This is very similar to the clustering of the relational space (Section 4.3.2) that we perform for Relation-based homophily. Finally, the results from experiments on large graphs derived from text data showed that use of social features boosts node classification performance. This further supports our hypothesis that the object co-occurrence patterns that we see in images are driven by underlying relationships.

4.7 Conclusions

In this chapter we showed that *cautious* inference about object relations outperforms traditional *aggressive* inference methods for object detection. In this regard, *Cautious* methods empirically provided mean improvements of 8% and 2.5% on relational and combined classifiers, respectively, over its *aggressive* counterparts. Furthermore, in this chapter we introduced a notion of *relational-homophily* that recovers underlying structures from the observed relations aiming to better understand the behavior of the related object categories of interest and improve inference. Improvements of 8.5% on purely relational methods makes *relational-homophily* a promising principle to use when local information about the unknown instances is not available (e.g. in an inpainting scenario). Our experiments suggest that performing cautious inference paired with Relation-based homophily with relations in RF2 representation is beneficial for more camera constrained settings such as the ones found in systems for autonomous navigation. In addition, following the observations made by the Collective Classification community, we propose class skew as an indicator flag. This flag can be used to measure whether the method proposed in this chapter can be applied to a specific problem. These results suggest three possible directions for future work: the exploration of better representations for reasoning in 3D space, which typically outperform methods that operate in 2D; the evaluation of better methods to recover the underlying structures of the relational space defined by the object relations and investigating the generality of these observations in the context of other object categories or other application scenarios.

Chapter 5

Scene-driven Cues for Viewpoint Classification of Elongated Object Categories

Until now the focus of this thesis have been purely on exploiting relations between objects in order to improve the precision of computer vision-related tasks such as object detection and object pose estimation. These relation-based methods are inspired by observations of the usual arrangements in which the objects of interest tend to co-occur (*Things*-based context). However, as found in several studies [106, 145, 146], in addition to object instances the scene has an important impact on the occurrence of the objects of interest. In this chapter, we put aside the inter-object relational aspect of context. We give focus on how cues from the scene can be exploited to improve the estimation of the viewpoint of objects in the scene.

Contents of this chapter are based on the publication:

- Oramas M, J., and Tuytelaars, T. *Scene-driven cues for viewpoint classification of elongated object classes.* In British Machine Vision Conference (BMVC) 2014.

5.1 Introduction

In this work we explore the use of non-local, scene-driven cues for object viewpoint estimation. In particular, we exploit a particular feature of elongated objects in that their physical extent provides a strong cue about their orientation. For example, consider the object in Fig. 5.1a. Even when we have no direct access to the local features of the object itself, we are able to predict, up to some level, the orientation of the underlying object (Fig. 5.1b). In this work we use the bounding boxes covering the objects, as a proxy to classify their viewpoint. Particularly, we are interested in how objects in specific orientations in the scene, project bounding boxes in the image space and use this as an intermediate step towards viewpoint classification. To this aim, we use the elongation orientation, i.e. the direction of maximum physical length, of the object, as a cue to estimate its viewpoint. For the sake of brevity, in the rest of the chapter we drop the term "orientation" and refer to elongation orientation purely as *elongation*. Moreover, in order to enforce scene-consistency in the viewpoint classification process, we define the scene not only as a space in which the objects of interest occur, but rather as a space with specific regions that are more likely to host certain objects with particular features such as object category, orientation, or size. For example, note how the orientation of the objects in Fig. 5.1c is closely related to the regions of the scene in which they occur. Combining these two ideas, here we propose a Top-Down approach in which we first generate scene-driven object proposals in the image and select the closest ones to object hypotheses gathered with an object detector. Then, we define a correspondence descriptor between each hypothesis and its closest object proposal and perform classification to predict the elongation orientation of the object. Finally, the object viewpoint is determined by a late fusion of the elongation prediction and the intrinsic prediction of the object detector. We explore four means to produce scene-driven object proposals, based on: a) the scene ground-plane, b) previously seen 3D annotated objects in the scene, c) previously seen 2D annotated objects in the image, and d) previously seen 2D object hypotheses in the image, obtained by a detector.

The contributions of this chapter are: a) the introduction of an intermediate step, towards viewpoint classification, object elongation orientation classification, and b) a top-down approach that produces scene consistent results for viewpoint classification outperforming results that are obtained in a purely local fashion. This chapter is organized as follows: Section 5.2 positions the contents of this chapter with respect to similar work. In Section 5.3 we present the details of our method. Section 5.4 introduces the evaluation protocol used, followed by experimental results and discussion (Section 5.5). In Section 5.6 we conclude this chapter.

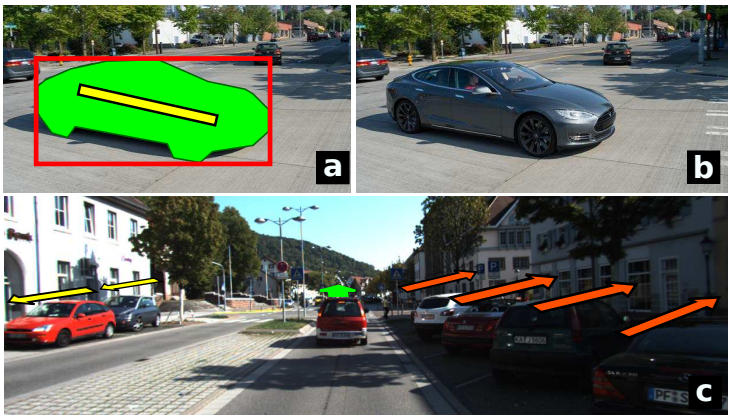

Figure 5.1: Note how the shape (a,b) and the location (c) of the bounding box of an object is related to its viewpoint.

5.2 Related Work

There exists a significant amount of work addressing the problem of object viewpoint classification. For the sake of brevity, we position our work based on two, closely related, groups of work:

Local viewpoint classification: As presented in Section 2.3, over the years, methods for viewpoint estimation have evolved from reasoning about features of the object in the image space [81, 111] to reasoning about object parts in the 3D space [63, 80, 115]. Similar to the recent group of work, we explore how the scene and the features of the 3D objects in the scene shape the 2D object evidence we perceive in an image. We focus on the size feature of the objects of interest. Specifically, we use features derived from the bounding boxes that circumscribe the objects in the image. This removes the requirement of, more detailed, CAD models at the cost of producing a coarse object representation.

Scene-driven viewpoint classification: This group covers work that exploits the full scene to estimate the viewpoint of the objects of interest. Methods from this group enforce geometric consistency of the objects with the scene and/or consistency between objects in the scene. Aiming to enforce object-scene consistency, [41] proposed a deformable 3D cuboid model composed of faces and parts that can deform with respect to their anchors in the 3D bounding box. Then, after learning a viewpoint-invariant appearance of each face, a sliding 3D bounding box approach is used for localization. In Chapter 3, we introduced a complementary approach which exploits pairwise relations between 3D objects

in the scene from an object-centered perspective. We show how reasoning about relations between objects in the scene could serve as a cue for the classification of the object pose. Parallel to this, [160] presented a spatial layout model that enforced scene consistency based on the 3D aspectlets of individual objects with object-object consistency in the form of occlusion reasoning. This combination not only improved 3D object detection but also produced accurate oriented object hypotheses. Very recently, and in parallel to our work, [165] uses a fine detail shape representation based on CAD models. This representation improved the reasoning about object support on the ground-plane and mutual occlusion. Similar to this group of works, we derive our object representation from features from the 3D object in the scene. However, we employ much simpler features derived from the bounding boxes of the objects and not from, more complex, CAD models as in [165]. We enforce scene consistency by either assuming that the objects of interest are located on the scene-ground plane as proposed in [61] in the context of object detection; or by assuming that the object evidence extracted by the detector should align with objects previously seen with the same camera setup. Finally, different from [107, 160, 165], in this chapter we do not reason about object relations. Furthermore, our method operates in still images and does not require image sequences as in some SfM-based approaches [6, 7].

5.3 Proposed Method

In a nutshell our method can be summarized in five steps (see Fig. 5.2): First, we run a viewpoint-aware object detector in order to collect a set of hypotheses $o = \{o_1, o_2, ..., o_n\}$. Then, based on a proposal generation function Ω we generate a set of scene-driven object proposals $o' = \{o'_1, o'_2, ..., o'_n\}$. In the next step we estimate a correspondence descriptor d_i between each object hypothesis o_i and its closest scene-driven object proposal o'_i. Then, we estimate the elongation of the initial object hypothesis o_i via multiclass classification of the descriptor d_i. Finally, the viewpoint of the objects is estimated by the fusion of the responses of the viewpoint-aware local object detector and the scene-driven elongation classifier. In some of the following methods we perform reasoning about physical objects in the 3D scene. For this reason, we will adopt the following notation. We will use the term O_i (in upper case) to refer to 3D objects, either hypotheses or proposals, located in the 3D scene. Likewise, we will use the term o_i (in lower case) to refer to 2D objects located in the image space. Now we take a deeper look in the different stages of the proposed method.

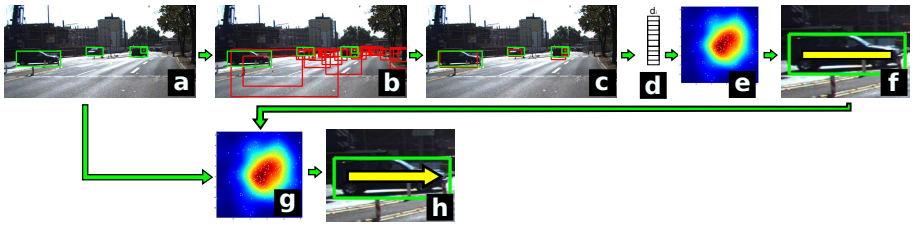

Figure 5.2: Algorithm Pipeline: a) Object Detection, b) Scene-driven Object Proposal Generation, c) Object-hypotheses - Object-Proposal Matching, d) Correspondence descriptor extraction, e) Elongation Classification, f) Elongation Estimate, g) Viewpoint Classification, and h) Viewpoint Estimate.

5.3.1 Scene Representation and Object Detection

This work is inspired by the work of Hoiem et al. [61] in the sense that we consider the idea of a scene concept behind the image. Whereas [61] defined the scene as a ground-plane and focused on the task of object detection, we explore different scene representations and exploit these ideas for the problem of object viewpoint estimation. Here we define a scene as the area in which the objects of interest occur. This scene can be defined either in 3D or in 2D. Furthermore, we explore an extension of the original idea of [61] where the scene can serve as a prior for the location of the objects of interest, to consider the scene as a space with specific regions that are more likely to hold certain objects with particular features like category, orientation and/or size.

To guide the process, we locate regions of the image that appear to host the objects of interest based on appearance. To this end, we run a standard object detector which produces a set of 2D object hypotheses $o = \{o_1, o_2, ..., o_n\}$, where each object hypothesis $o_i = (s_i, b_i, \alpha_i)$ is defined by its confidence score s_i, its bounding box $b_i = (x_i, y_i, w_i, h_i)$, and its viewpoint α_i, for the case of viewpoint-aware detectors such as the ones presented in [47] and [81]. Finally, as mentioned in Section 1.1, we will focus on urban scenes and cars as the object category of interest.

5.3.2 Scene-driven Object Proposal Generation

Once we have spotted a set of regions in the image, i.e. the object hypotheses o_i, which are likely to host the objects of interest the next step is to recover scene-driven object proposals o_i' (i.e. Fig. 5.2.b), that will serve to validate the evidence o_i collected by the detector. We generate a set of scene-driven object

proposals as $o' = \Omega(scene)$, where Ω is an object proposal generation function defined over the scene. We define Ω using one of the following methods:

a) Ground-Plane: This approach is heavily inspired by the work of Hoiem et al. [61]. We model the geometry of the scene by assuming the existence of a ground plane that supports the objects of interest. Given the ground plane, we densely generate a set of 3D object proposals $O' = \{O'_1, O'_2, ..., O'_m\}$ resting on the ground plane for each of the discrete orientations $\theta = \{\theta_1, \theta_2, ..., \theta_K\}$. Each 3D object proposal, $O'_i = (X_i, Y_i, Z_i, L_i, W_i, H_i, \theta_i)$, is defined by its 3D location (X, Y, Z), its physical length, width and height (L, W, H) and its orientation θ in the scene. In our work we define the length, width and height (L, W, H) of the proposed 3D object proposals O' based on statistics from real world objects. We drop the 3D location coordinate Y since all the 3D object proposals are assumed to be supported by the ground plane, hence $Y = 0$ for all the proposals. As stated in Section 1.1.2, θ is a discrete one-dimensional value which represents the azimuth angle of the object in the 3D scene.

Then, once we have generated all the 3D objects that can physically be in the scene, using the camera parameters we project each of the 3D object proposals O' to the image space, assuming a perspective camera model, producing a set of 2D object proposals o'. Specifically, each 2D proposal o' is obtained by projecting each of the corners of the 3D proposal O', using a perspective camera model and selecting the 2D points that enclose the rest. Given a calibrated camera, the viewpoint α of each 2D object proposal o' is estimated as a function of the 3D orientation θ and location (X, Y, Z) of the 3D object proposal O' that generated it.

b) History of 3D objects: This approach is an extension of the previous scenario, in which we assume that there are some regions of the 3D space that are more likely to support objects with particular features. For this purpose, we start from a set of previously seen ground truth 3D objects and sample from the distribution defined by $p(X, Z, \theta)$ a subset of $O' = \{O'_1, O'_2, ..., O'_m\}$ 3D object proposals. Following the same procedure as before, assuming a perspective camera model, we project the 3D objects O' to the image space producing a set of 2D object proposals o'. Note that for history-based methods to be informative, the same camera setup is required for training and testing.

c) History of 2D objects: This is the 2D counterpart of the previous approach where we assume that the scene is defined over the image space, hence, we start from a set of ground truth 2D objects from a training set. Here we obtain the set of 2D object proposals o' by sampling the distribution defined by $p(x, y, w, h, \alpha)$, where x and y are the 2D location coordinates of the object in the image, w and h define the bounding box size and α its viewpoint. Note that since this approach operates in the image space, its annotation cost is relatively lower

than its 3D counterpart.

d) History of 2D object hypotheses: This approach is very similar to the previous in the sense that the scene is defined over the image space. It differs in that it starts from the output of an object detector instead of relying on ground-truth annotations. The set of object proposals o' is obtained by sampling the distribution defined by a set of previously detected object hypotheses.

In order to model the history-based distributions we gather a specific set of features from the object instances present in the training set. Depending on the scene representation, we collect either (X, Z, θ) for 3D objects or (x, y, w, h, α) for the 2D objects. Then, based on the oKDE [73] method presented in Section 2.5.3, we model each of these distributions, respectively. We uniformly sample this distribution producing a total of 1000 object proposals o'.

As final step of this stage (see Fig 5.2.c), for each of the hypotheses o_i spotted by the object detector, we select its closest scene-driven object proposal o'_i using the Pascal VOC criterion [36] (see *bounding box matching* in Section 2.8.2). Note that due to the box representation, objects with opposite orientations (orientation difference=π) will project the same 2D bounding boxes. For this reason we will focus on a smaller set of $K/2$ discrete viewpoints. This subset of viewpoints measures the orientation of the maximum extent of the objects (i.e. Fig. 5.2.f), to which we refer in this thesis as the *elongation* orientation ϵ of the object. Given an object o_i its elongation ϵ_i is defined as $\epsilon_i = (\alpha_i \bmod \pi)$.

5.3.3 Elongation Classification

So far, for each of the object hypotheses o_i reported by the detector, we have spotted its closest 2D object proposal o'_i. For each hypothesis-proposal pair (o_i, o'_i) we compute the correspondence descriptor $d_i = (rw, rh, rx, ry, \alpha')$ where $rw = |\frac{w'}{w}|$, $rh = |\frac{h'}{h}|$, $rx = |\frac{x'-x}{w}|$, $ry = |\frac{y'-y}{h}|$ and the viewpoint α' of the closest object proposal.

In this work, we take advantage of the physical extent of elongated objects and how the elongation of an object is related to its viewpoint perceived in the image space. To this end, we first classify the elongation $\hat{\epsilon}$ of an object, and then use this prediction to improve the viewpoint prediction $\hat{\alpha}$ given by the detector based on appearance features.

Training: Given a set of training images containing objects with annotated bounding boxes b_i and elongation ϵ_i, for each annotated object o_i we obtain its matching object proposal o'_i and compute its correspondence descriptor d_i following the procedure described above. Then, using the pair (d_i, ϵ_i) we model

each discrete elongation ϵ by a probability density function (pdf). We use the odKDE method from [72] (Section 2.5.3) to model these pdfs. This method has the advantage of providing compression mechanisms in which each pdf is approximated by a Gaussian Mixture Model (GMM), reducing the number of components required to model each distribution.

Inference: Given a new image, we compute the set of correspondence descriptors d between the object hypotheses o and the object proposals o' generated from $\Omega(scene)$. Then, the elongation $\hat{\epsilon_i}$ of each object o_i is the MAP estimate by applying the Bayes rule:

$$\hat{\epsilon_i} = \arg\max_{(\epsilon_k)} p(\epsilon_k|d_i) = \arg\max_{(\epsilon_k)} p(d_i|\epsilon_k)p(\epsilon_k), \tag{5.1}$$

where the class likelihoods $p(d_i|\epsilon_k)$ are computed using Kernel Density Estimation (odKDE) and the priors $p(\epsilon_k)$ are obtained from the occurrence of the elongation ϵ_k on the training data.

5.3.4 Viewpoint Classification

We estimate the viewpoint of the object o_i by late fusion of the response of the viewpoint-aware object detector and the response of our object elongation classifier (Fig. 5.2.g). Given the two responses, we define the coupled response $r_i = [\alpha_i, s_i, \epsilon_i, \lambda_i]$ where (α_i, s_i) are the responses of viewpoint and score from the object detector, and (ϵ_i, λ_i) are the responses of elongation and score of our elongation classifier. In this case, the score λ_i is the posterior $p(\epsilon_i|d_i)$ estimated in Eq. 5.1. Finally, to classify the viewpoint $\hat{\alpha_i}$ of an object o_i we perform MAP inference for the coupled response r_i over the discrete viewpoint classes in a similar fashion as Eq. 5.1:

$$\hat{\alpha_i} = \arg\max_{(\alpha_k)} p(r_i|\alpha_k)p(\alpha_k). \tag{5.2}$$

Here, the class likelihoods $p(r_i|\alpha_k)$ are computed performing Kernel Density Estimation (odKDE) considering the response of the object detector and the elongation classifier on a validation set. The priors $p(\alpha_k)$ are obtained from the occurrence of the viewpoint α_k on the validation data.

5.4 Evaluation

5.4.1 Experimental Settings

We perform experiments on the KITTI benchmark [45], specifically, on the object detection dataset. See Section 2.8.1 for more details of this dataset. Following the methodology from [165, 166], we report results in two subsets of ground truth objects from the testing set. The first set, coined *fullSet*, contains all the objects from the testing set while the second set, coined *easySet*, contains all the objects whose bounding box height is larger than 50 pixels. We run experiments starting from one of three off-the-shelf standard detectors used to collect the initial object hypotheses. We employ the deformable parts model (DPM) detector (release 5) [112] and two extensions of DPM, [47] and [81], modified to predict 8 viewpoints. These detectors were trained on the Pascal VOC 2007 [35], the Karlsruhe urban [47], and the EPFL cars [111] datasets, respectively. We present results for five methods: 1) *local detector*, the isolated output of one of the object detectors ([47, 81, 112]), i.e. not using any scene-level information; 2) *GroundPlane*, when assuming the scene is defined by its ground-plane (Sec. 5.3.2(a)); 3) *Hist3DObjects*, when considering the 3D regions of the scene that are more likely to host objects with particular features (Sec. 5.3.2(b)), 4) *Hist2DObjects*, the 2D counterpart of *Hist3DObjects*, where we start from a history of 2D ground-truth objects (Sec. 5.3.2(c)), and 5) *Hist2DHypotheses*, where we start from a history of 2D hypotheses collected with an object detector (Sec. 5.3.2(d)).

5.4.2 Experiment: Object Evidence Extraction

This experiment aims to show the performance of the object detectors for recovering the object bounding boxes. We evaluate detection of 2D objects following the Pascal VOC protocol [36] (introduced in Section 2.8.2) and report Average Precision (AP) as performance metric. The local detectors [47, 81, 112] produced a total of 3986, 9713 and 10525 hypotheses, respectively, during the detection stage. On the *easySet*, the selected detectors achieved a performance of 52%, 35% and, 44% AP, respectively. This performance dropped to 34%, 20% and 32% AP, respectively, when the *fullSet* was considered. The superior performance of the detector from [47] can be attributed to the fact that it was trained on the Karlsruhe urban dataset [47], which more closely resembles the settings from the KITTI dataset used for evaluation. We provide these performance measurements to indicate the volume of data processed in the following experiments.

	Geiger et al. [47]		Lopez et al. [81]		Felzenswalb et al. [112]	
Method	**easySet**	**fullSet**	**easySet**	**fullSet**	**easySet**	**fullSet**
Local detector	**0.69**	**0.68**	0.42	0.46	—	—
GroundPlane	0.50	0.52	**0.69**	**0.57**	**0.72**	**0.64**
Hist3DObjects	0.39	0.41	0.57	0.53	0.54	0.53
Hist2DObjects	0.40	0.39	0.46	0.43	0.53	0.51
Hist2DHypotheses	0.41	0.42	0.61	0.49	0.34	0.34

Table 5.1: Object Elongation Classification Performance. Mean Precision on Pose Estimation (MPPE).

5.4.3 Experiment: Elongation Classification

The elongation of an object is a feature closely related to its viewpoint. For this reason, we use a performance metric that has been traditionally used in previous work for measuring the performance of pose/viewpoint estimation. In particular, we adopt the Mean Precision in Pose Estimation (MPPE) as performance metric. MPPE is computed as the average of the class-normalized confusion matrix of the pose/viewpoint classifier (see Section 2.8.2). It is computed from hypotheses that are assumed correct based on the Pascal VOC criterion [36], as in prior work [51, 81, 115]. For the evaluation of this experiment, we derived elongation annotations, from the original viewpoint annotations of the dataset, producing four possible discrete elongation values in the range $[0,\pi)$. Furthermore, we run experiments using the object detector from [112] to collect object hypotheses, and using our method we extend its output to provide elongation estimates. Note that [112] is purely an object detector and does not provide viewpoint estimates.

Discussion: We can see in Table 5.1 that all the proposed methods for elongation classification have a performance clearly above chance levels (25% for the case of 4 elongation orientations). This shows that indeed our methods are encoding some useful cue for elongation classification. It is also remarkable that this can be achieved without having direct access to the local data of the objects, the pixels inside the bounding boxes. Furthermore, the difference in performance between detectors is more evident. For the case of [81], there is more room for improvement and our elongation classifiers achieve a mean improvement of 16.3 and 4.5 percentage points (pp) for the *easySet* and the *fullSet*, respectively. In addition, starting from this particular detector, the method defined by *GroundPlane* leads with improvements of 23 and 11 pp in the respective image sets. On the opposite, for the case of [47], none of the proposed methods improves over the local detector for the task of elongation classification. For all cases, it is notable how the methods based on a 2D scene representation (*Hist2DObjects*, *Hist2DHypotheses*), which require significant lower annotation effort, have a comparable performance to some of the 3D-based

Method	Geiger et al. [47]		Lopez et al. [81]	
	easySet	**fullSet**	**easySet**	**fullSet**
Local detector	0.38	0.43	0.38	0.36
GroundPlane	0.44	**0.45**	0.50	0.42
Hist3DObjects	**0.46**	0.43	**0.55**	**0.48**
Hist2DObjects	0.42	0.39	0.43	0.38
Hist2DHypotheses	0.45	0.42	0.47	0.39

Table 5.2: Object Viewpoint Classification Performance. Mean Precision on Pose Estimation (MPPE).

methods. Also, we can see that for the pure detector [112], we are able to predict its elongation to a significant level, clearly above chance levels.

5.4.4 Experiment: Viewpoint Classification

We now measure the performance of the proposed method for the task of viewpoint classification, that is, the classification of 8 discrete viewpoints. For this evaluation we will again use MPPE. Similarly, we report performance results on the same five methods: *local detector*, *GroundPlane*, *Hist3DObjects*, *Hist2DObjects*, *Hist2DHypotheses*. However, notice that since the viewpoint of an object is defined by the combination of the responses of the detector and the elongation classifier (Sec. 5.3.4), we can only report viewpoint classification results on the detectors from [47] and [81], which provide viewpoint-related information in their response. Please see Fig. 5.3 for some qualitative results.

Discussion: For the task of viewpoint classification we notice a drop in most of the performance values (Table 5.2). This is to be expected as it involves more classes than the one of elongation. However, all the proposed methods are again well above chance levels (12.5% for 8 viewpoint classification). For the case of [81] our proposed methods achieve mean improvements of 10.75 and 5.75 pp, for the *easySet* and *fullSet*, respectively. Experiments using this detector are lead by the *Hist3DObjects* method, with the respective improvement of 17 and 12 pp. Different to the elongation classification task, for viewpoint classification our methods do bring an improvement also over [47] namely, 6.25 pp on the *easySet*. This may hint that some of the failure cases for the detector from [47] are caused by opposite viewpoints. In addition, for this detector, best results are obtained by the methods based on a 3D scene representation (*GroundPlane*, *Hist3DObjects*), producing mean improvements of 6 and 1 pp on the corresponding image sets. The mean improvement of the (*GroundPlane* and *Hist3DObjects*) methods on the *fullSet* seem to be marginal, when compared to those on the *easySet*. This low mean improvement, seems to be affected by the *Hist3DObjects*) which brings no improvement. Now we will look into

a possible cause for this low improvement. For the case of the *Hist3DObjects* method, 3D Object proposals O' are sampled from the distribution of annotated 3D objects in training images. During the annotation stage, these 3D objects are sensed using a laser scanner. Since the scanner has a defined effective range, it introduces an effect on how 3D objects are distributed in the scene. More specifically, objects within this effective range have higher occurrence likelihood. In consequence, during the proposal generation stage, objects within this range are sampled first. Furthermore, some of these 3D proposals may generate 2D proposals that match the smaller objects in the images with incorrect viewpoints. This suggests that a higher number of proposals should be sampled in order to recover 3D objects at larger distance which are projected as the smaller 2D objects in the images. Similarly to the elongation experiments (Sec. 5.4.3), the methods defined on a 2D scene representation have a comparable performance to the methods starting from a 3D scene representation. In addition, for the case of viewpoint classification, at least for the relatively larger objects of the *easySet*, the cues from the 2D scene always bring improvements over the purely local methods ([47, 81]). Finally, by looking at the performance of the proposed methods to extract scene-driven cues, we can see that superior performance is achieved by methods that define the scene in the 3D space, i.e. *GroundPlane* and *Hist3DObjects*. However, these 3D-based methods come with the additional cost of requiring 3D object annotations which are more difficult to obtain. For the case when no such 3D annotations are available, the proposed 2D methods, *Hist2DObjects* and *Hist2DHypotheses*, offer a good trade-off between performance and annotation cost. In this regard, compared to the *local detector*, the proposed methods that operate in the 2D space have superior performance in objects with larger sizes (*easySet*). For the case of the smaller objects that are included in the *fullSet*, the proposed 2D methods (*Hist2DObjects* and *Hist2DHypotheses*) have comparable performance to the *local detector*.

5.5 Discussion

In this chapter we have explored several ways to extract cues from the scene with the objective of estimating the viewpoint of the objects of interest while enforcing scene consistency. We have seen that by taking into account scene-driven cues, viewpoint classification results can be improved relative to those obtained when using only local information. Recently, other methods that perform scene-driven viewpoint estimation have been proposed, [160] presented a few months ago and [165] developed parallel to this work. However, both of these methods focus purely on the ground plane assumption as means to enforce consistency with the scene. Furthermore, in addition to the traditional object annotations on the image set and calibrated camera, they required CAD

Figure 5.3: Viewpoint classification results encoded in jet scale. Continuous line, local detector prediction; Dashed line, scene-driven object proposals. Circle, ground-truth viewpoint. (Best viewed in color).

models. Both methods depend on fine-part detection during the object detection stage, which makes them inappropriate for low resolution images. Finally, they have a strong link between the methods to enforce scene consistency and to perform object detection. This complicates the integration of future, possibly improved, object detectors in their methods for enforcing scene consistency. To their advantage, by learning from CAD models, they are able to predict object polygons or wireframe models that are more pleasing to the eye and closer to the original object shape than our bounding box predictions. Furthermore, they are able to predict continuous object orientation values.

On the opposite, we have presented several ways to enforce scene consistency with different levels of annotation cost. Additionally, as demonstrated in our experiments, we are able to integrate any object detector in our method. This last feature allows our method to improve the box representation of its predictions by integrating more advanced detectors, e.g. the ones used in [160] and [165], as long as they produce viewpoint information in their responses. In its current state, our method does not have the requirement of high resolution images for proper performance. This is again handled by the flexibility of the method for the integration of any object detector. This flexibility also makes our method useful to extend pure detectors, such as [112], to produce elongation estimates. Note that for some applications, such as obstacle detection, object elongation prediction might be enough. Finally, the similar camera setup requirement of our history-based methods may be seen as a strong constraint. However, there are many scenarios that resemble this setting, e.g. dashcams and backup cameras attached on cars; inspection in manufacturing, and fixed security cameras found on streets, airports, shopping centers and several areas of interest where human activity takes place.

5.6 Conclusions

In this chapter we have introduced scene-driven object elongation orientation classification as an intermediate step prior to viewpoint classification. Our experiments show how considering object elongation estimates based on scene-cues brings improvements over purely appearance-based viewpoint-aware object detectors. In addition, we have presented several approaches to perform scene-driven object viewpoint classification at different levels of annotation cost. The proposed method is flexible enough to allow the integration of future, more advanced, viewpoint-aware detectors. To conclude, this work complements very recent work, by sending the message that there are relatively simple cues in the scene that can bring improvements for the task of object viewpoint classification.

Chapter 6

Recovering Missed Detections by Sampling Context-based Object Proposals

Exploiting contextual information in the form of relations between object instances, is becoming an accepted practice when trying to disambiguate uncertain object hypotheses with the goal of improving object detection precision. Moreover, as we have seen in most of the related work presented through this thesis, existing methods rely mostly on pairwise relations between objects. In contrast to this trend, in this chapter we focus on means to improve the recall in the object detection process. Towards this goal, we propose a method to sample context-based object proposals after an initial object detection stage. In addition, we take early steps towards reasoning beyond pairwise relations between object instances.

The content of this chapter is based on the article:

- Oramas M, J., and Tuytelaars, T. *Recovering hard-to-find object instances by sampling context-based object proposals. Submitted to* IEEE International Conference on Computer Vision (ICCV) 2015.

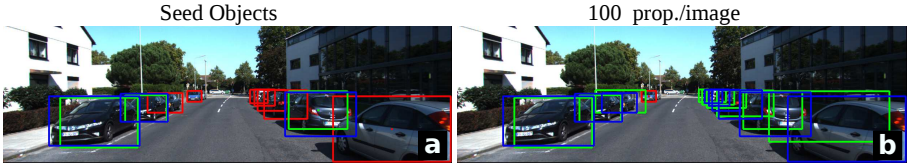

Figure 6.1: Object detections collected: a) after running a standard appearance-based detector, b) after sampling only 100 context-based object proposals post detection. Notice how we manage to recover many of the initial missed detections. For the sake of clarity in the visualization, we removed the false positive proposals and only show the bounding boxes of the matched object annotations (in blue), missed detections (in red) and matching object proposals (in green). (Best viewed in color).

6.1 Introduction

The literature on object detection to date is very *precision-focused* [4, 30, 79, 118, 122]. It is generally acknowledged that precision and recall should be considered simultaneously, e.g. in the form of precision-recall curves, as one can be traded for the other. Combined measures such as mean Average Precision are common practice. At the same time, the curves typically drop steeply at some point, leaving lots of objects undetected. The high-precision low-recall area is often considered the more interesting part of the curve [79, 108, 122]. Methods are optimized and typically perform well in this region. The high-recall low-precision area, on the other hand, receives little attention – as if we all have come to accept there is some percentage of object instances that are just too hard to be found.

A notable exception to this view is the work on category-independent object detection (e.g. objectness [1, 2], selective search regions [151], edge boxes [168]). When the object category is unknown, no-one expects a high precision, and it is only natural to focus on recall instead. A common evaluation protocol in this context is the obtained recall as a function of the number of window proposals per image. Here, we adopt the same evaluation scheme, but now for standard supervised object category detection.

At the same time, most methods in the literature, be it for category-specific or category-independent object detection, are also explicitly *appearance-focused*. Other cues like context are sometimes added (e.g.[18, 30, 108, 118]), but only as secondary cues and mostly to filter out false detections, i.e. to further improve precision, rather than improving recall. Appearance, indeed, may be the most reliable cue available when the goal is to detect objects with high

confidence. If the object is clearly visible in the image, appearance cues can be very strong. Unfortunately, appearance-based approaches cannot cope well with more difficult cases, such as small object instances or highly occluded ones. In spite of some efforts in this direction (e.g. Frankenclassifier [94], Occlusion patterns [116]), these mostly remain undetected, resulting in lower recall. In a real world setting, highly cluttered scenes and therefore small and occluded objects are actually quite common - probably more common than in typical benchmark datasets, which are often object-focused (e.g. because they have been collected by searching images that have the object name mentioned in the tags).

In this work, we focus on recall instead of precision. The goal is to find as many object instances as possible, even if this comes at a cost, in the form of many false detections (low precision). Because of the lower precision, we refer to the detections as "object proposals" as in the category-independent object detection work. This reflects the idea that further verification (e.g. using other modalities, other viewpoints or higher resolution imagery) may be required to separate the true detections from the many false detections – a process which may be application dependent and is out-of-scope of this work. We have selected the KITTI dataset [45] for our analysis, since it provides real world, challenging imagery and high quality ground truth annotations, including object instances that are small or highly occluded. On this dataset, we compare various strategies to generate category-specific object proposals: i) a sliding-window baseline, ii) a method for category-independent object proposals (selective search regions [151]), and iii) two category-specific context-based schemes. In particular, we focus on context cues from other objects in the scene. Indeed, multiple objects in a scene often appear in particular spatial configurations. Detecting one object then also provides information about possible locations of other objects. We start from a few high-confidence appearance-based detections and use these as seeds based on which other likely object locations are identified. We explore one method that uses pairwise relations, and propose a new topic-based method that builds on higher-order spatial relations between groups of objects. We have found that despite its dependence on simple features, relative location and orientation, our method is able to discover arrangements between objects that resemble those found in the real world. Furthermore, it does not enforce restrictions on the number of objects participating in each of the higher-order relations. For simplicity, we assume the ground plane to be known, both for the baselines as for the newly proposed context-based schemes. Despite its simplicity, our method is able to bring significant improvement to standard object detectors. For example, notice how in Figure 6.1(b) we manage to recover many of the initial missed detections (Figure 6.1(a)). Furthermore, this is achieved at the low cost of just 100 additional object proposals.

The remainder of this chapter is organized as follows: Section 6.2 presents existing work that inspired the method proposed in this chapter. In Section 6.3 we present the details of the analysis and of the methods for generating object proposals. Experiments, results and discussions are presented in Section 6.4. Finally, Section 6.6 concludes this chapter.

6.2 Related Work

The analysis presented in this chapter lies at the intersection of category-independent and context-based category-specific object detection. These two groups of work constitute the axes along which we position our work.

Context-based category-specific object detection: In recent years, contextual information, in the form of relations between objects, has been successfully exploited to improve object detection performance [22, 30, 40, 118]. These works typically follow a two-stage approach where a set of detections is collected using an appearance-based detector. Then, using pre-learned relations between objects, out-of-context detections are degraded. Following this methodology, [30] learns qualitative spatial relations between object bounding boxes (e.g. top-left, far-left). Using these relations, detections in unlikely areas are filtered out. Similarly, [22, 40, 118] define continuous spatial relations between the centers of object bounding boxes and use the learned relations to filter out the out-of-context objects. Works belonging to this group have proven successful in improving object detection, specifically, in terms of increasing the precision. However, objects missed by the object detector are not recovered. This, in consequence, leaves no room for improvement in terms of recall.

One work that tries to increase recall is the co-detection work of Bao et al. [8]. They exploit detections of the same object instances in multiple images to generate bounding boxes. Our work, on the other hand, operates purely on a single image.

Additionally, our work differs from [22, 30, 40, 118] in that we consider higher-order relations whereas most of the methods that exploit relations between objects focus on the pairwise case. Recently, a small group of works [18, 108, 162] that consider higher order relations have been proposed. In [18], a Pure-Dependency [65] framework is used to link groups of objects. In [108], objects are grouped by clustering pairwise relations between them. The work of [162] is able to reason about higher-order semantics in the form of traffic patterns. Different from these works, our topic-based method to discover higher-order relations does not require the number of participating objects to be predefined

[65]. Furthermore, objects do not need to be "near" in the space defined by pairwise relations in order to be covered by the same higher-order relation [108]. Finally, our method does not require scene-specific cues (e.g. lane presence, lane width or intersection type), or motion information [162].

Category-independent object detection: Another group of work operates under the assumption that there are regions of the image that are more likely to contain objects than others. Based on this assumption, the problem is then to design an algorithm to find these regions. Following this trend, Alexe et al. [1] proposed a method where windows were randomly sampled over the image. Following the sampling, a "general" classifier was applied to each of the windows. This classifier relied on simple features such as appearance difference w.r.t. the surrounding or having a closed contour and was used to measure the objectness of a window. In [1], windows with high objectness are considered to be more likely to host objects. Endres and Hoiem [34] proposed a similar method with the difference that their method generated object proposals from an initial segmentation step. This produced more detailed object proposals. Similarly, [151] proposed a selective search method which exploits the image structure, in terms of segments, to guide the sampling process. In addition, their method imposes diversity by considering segment grouping criteria and color spaces with complementary properties. More recently, [168] proposed a novel objectness score measure, where the likelihood of a window to contain an object is proportional to the number of contours entirely enclosed by it. A common feature of this group of work is that their precision is less critical since the number of generated proposals is a small percentage of the windows considered by traditional sliding window approaches. On the contrary, these methods focus on improving detection recall by guiding the order in which windows are evaluated by later category-specific processes. Inspired by these methods we propose to complement a traditional object detector with an object proposal generation step. The objective of this additional step is to improve detection recall even at the cost of more false detections.

6.3 Proposed Method

The proposed method can be summarized in two steps: In the first stage, we run a traditional object detector which produces a set of object detections. Then, in a second stage, we sample a set of object proposals aiming to recover object instances possibly missed during the first stage. similar to the previous chapter, in some of the following methods we perform reasoning about physical objects in the 3D scene. For this reason, we will adopt the following notation. We will

use the term O (in upper case) to refer to 3D objects, located in the 3D scene. Likewise, we will use the term o (in lower case) to refer to 2D objects located in the image space. For simplicity, we will focus on a single object category with car as the category of interest. Furthermore, for clarity, we will adopt a similar notation as in Chapter 5 to distinguish between object hypotheses (or detections) and object proposals. We will use a prime accent $'$ to refer to object proposals, i.e. objects sampled to complement the initial detection step. This will be applied to both 2D proposals o' in the image, as well as to 3D object proposals O' in the scene.

6.3.1 Category-specific object detection

The main goal of this work is to recover missed object instances after the initial detection stage has taken place. Given this focus on the post-detection stage, for the object detection stage we start from an off-the-shelf detector. In practice, given a viewpoint-aware object detector, we collect a set of 2D object detections $o = \{o_1, o_2, ..., o_n\}$ where each object detection $o_i = (s_i, \alpha_i, b_i)$ is defined by its detection score s_i, its predicted viewpoint α_i and its 2D bounding box coordinates $b_i = (x_i, y_i, w_i, h_i)$ (see Section 1.1.2).

6.3.2 Object Proposal Generation

Traditional appearance-based object detectors have proven to be effective to detect objects o with high confidence when the objects of interest are clearly visible. On the contrary, for small or highly-occluded object instances its predictions are less reliable resulting in a significant number of object instances being missed. To overcome this weakness we propose, as a post-detection step, to sample (category-specific) object proposals o' with the goal of recovering missed detections. We analyze four strategies to generate these proposals, as discussed in the next four sections.

Relaxed Detector

A first, rather straightforward method to recover missed detections consists of further reducing the threshold τ used as cutoff in the object detector. This is a widely used strategy, even though it usually does not increase recall that much. Here, we give it a different twist, by completely removing the non-maximum suppression step present in most object detectors, including the one used in our experiments. The removal of non-maximum suppression may sound counter-intuitive, though it is important to notice that non-maximum suppression is

beneficial in cases where objects are well separated. For cases of high occlusion, as in Figure 6.1, non-maximum suppression completely removes overlapping, still valid, detections. As a consequence, these suppressed detections become unrecoverable. For a given threshold value τ, removing the non-maximum suppression step results in many more object proposals being generated. We refer to this strategy as *Relaxed Detector*.

3D Sliding Window

This is a 3D counterpart of the traditional 2D sliding window approach used by traditional detectors (e.g. [40]). This approach is inspired by the work of Hoiem et al. [61]. We assume the existence of a ground plane that supports the objects of interest. Given the ground plane, we densely generate a set of 3D object proposals $O' = \{O'_1, O'_2, ..., O'_m\}$ resting on the ground plane for each of the discrete poses $\theta = \{\theta_1, \theta_2, ..., \theta_K\}$. Each 3D object proposal, $O' = (X, Y, Z, L, W, H, \theta)$, is defined by its 3D location (X, Y, Z), its physical length, width and height (L, W, H) and its pose θ in the scene. We define the length, width and height (L, W, H) of the proposed 3D object proposals O' as the mean length, width and height values of annotated 3D objects in the training set. Since in our analysis we focus on cars as the category of interest, we drop the 3D location coordinate Y since all the 3D object proposals are assumed to be supported by the ground plane, hence $Y = 0$ for all the proposals. The pose θ is a discrete one-dimensional value which represents the azimuth angle of the object in the 3D scene (see Section 1.1.2). Then, once we have generated all the 3D objects that can physically be in the scene, using the camera parameters we project each of the 3D object proposals O' to the image space, assuming a perspective camera model, producing a set of 2D object proposals o'. Specifically, each 2D proposal o' is obtained by projecting each of the corners of the 3D proposal O, and selecting the 2D points that enclose the rest. Note that due to the box representation, objects with opposite poses (pose difference=π) will project the same 2D bounding boxes. For this reason we only generate proposals from a smaller set of $K/2$ discrete poses.

Category Independent Object Proposals

Here we follow the strategy of generic, category-independent, object proposal generators. A crucial part of this strategy is to define a proper objectness measure to be able to estimate how likely is a window defined over an image to contain an object of any category. In this analysis we evaluate the effectiveness of using this type of strategy to recover missed detections. Particularly, we use the popular Selective Search method from Uijlings et al. [151].

Category-specific Context-based Object Proposals

In this strategy we generate a set of object proposals o' as a function $o' = f^{\eta}(o)$ of the object detections o predicted by the appearance-based detector. The function f^{η} enforces contextual information in the form of relations between object instances. This way, all the proposals o' sampled from f^{η} follow a distribution of relations previously seen in the training data where η is the number of object instances participating in the relation. This produces a relation-driven search where given a seed object o object proposals o' are sampled at locations and with poses that satisfy these relations. In this chapter we propose two relation-driven functions: f^2 for the case of objects being associated by pairwise relations, and f^+ for the case when objects are associated by higher-order relations. More details on how we define $f^{\eta}(o)$ will be given below.

From 2D object detections to 3D objects in the scene In this work, reasoning about relations between objects is performed in the 3D scene. For this reason, we first need to project the object detections used as seeds on the 3D scene. We define the objects $O = \{O_1, O_2, ..., O_n\}$ as 3D volumes that lie within this 3D space. Each object $O_i = (X_i, Y_i, Z_i, L_i, W_i, H_i, \theta_i, s_i)$, is defined by its 3D location coordinates (X_i, Y_i, Z_i), its physical size (length, width and height) (L_i, W_i, H_i), its pose θ_i in the 3D scene and its confidence score s_i. We assume that all the objects rest on a common ground plane, so $Y = 0$ for all the objects. For brevity, we drop the Y term, then each object is defined as $O_i = (X_i, Z_i, L_i, W_i, H_i, \theta_i, s_i)$. In order to define the set of 3D objects O from the set of 2D objects o, we execute the following procedure: first, given a set of annotated 3D objects, we obtain the mean size (length,width and height) of the objects in the dataset. Second, similar to Chapter 5, based on a calibrated camera we densely generate a set of 3D object proposals O' over the ground plane. Third, each of the 3D object proposals from O' is projected in the image plane producing a set of 2D proposals o'. Then, for each object detection o_i we find its corresponding proposal o'_i by taking the proposal with highest matching score. We use the matching criterion for 2D detections from the Pascal VOC Challenge [36] (see *Bounding Box Matching* from Section 2.8.2). Finally, we use the 3D location (X_i, Z_i) from the 3D proposal O'_i from which o'_i was derived and the viewpoint angle α_i, predicted by the detector, to estimate the pose angle θ_i of the object O_i in the scene. This step might look redundant, since we can directly assume the pose θ_i of the 3D object O_i to be equal to the pose θ' of its matching proposal O'_i. However, this step is required since, as mentioned earlier, we generate proposals for a reduced number of $K/2$ discrete symmetric poses. As result of this procedure, we obtain a set of 3D objects defined as $O_i = (X_i, Z_i, L_i, W_i, H_i, \theta_i, s_i)$.

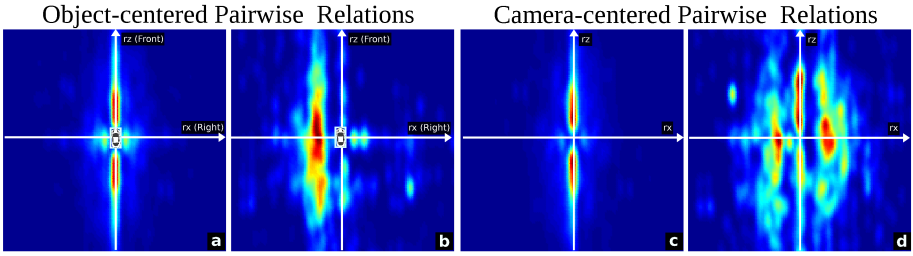

Figure 6.2: Top view of the distribution of pairwise relations for cars in the KITTI dataset [45]. Pairwise relations of cars with the same pose (a,c) and opposite pose (b,d), respectively. Images (a,b) corresponds to object-centered relations while (c,d) corresponds to camera-centered relations. Note how, for the object-centered case, cars with the same pose (a) are expected to be behind or in-front of the reference car, as commonly driving on the same lane. Similarly, note how cars in opposite poses (b) are expected to be on the left of the reference car, as is the case in opposite lanes. (Best viewed in color).

Pairwise Relations (f^2) : Pairwise relations between 3D objects are computed following the procedure introduced in Section 3.3. Following this procedure we define *camera-centered (CC)* pairwise relations between objects. Alternatively, we define *object-centered (OC)* relations. As a result, we obtain a set of pairwise relations $R = \{r_{ij}\}$ for each image, where r_{ij} is a pairwise relation between the 3D objects O_i and O_j. Each pairwise relation r_{ij} is defined as $r_{ij} = (r_X, r_Z, r_\theta)$, where (r_X, r_Z) represent the relative location of the objects. It is computed as $r_X = (X_j - X_i)$ and $r_Z = (Z_j - Z_i)$. The attribute r_θ represents the relative orientation between the object instances and is computed as the angular difference $r_\theta = (\theta_j - \theta_i)$. In order to model relations between pairs of objects, we compute pairwise relations r_{ij} between each pair of annotated 3D objects within each image of the training set. Then, we use the relations r_{ij} as sample points of a multivariate kernel density estimator to model the distribution $p(r_{ij})$. Please see Section 2.5.2 for more details about kernel density estimation. This is a simple method that manages to find some common arrangements in which pairs of objects co-occur. For example, in Figure 6.2 we show top view of the distribution of pairwise relations from cars in the KITTI dataset [45]. For the object-centered case, note how cars with the same pose (Figure 6.2.a) are expected to occur behind or in-front of the reference car. This is a common behavior when driving on the same lane. Similarly, note how cars in opposite poses (Figure 6.2.b) are expected to be on the left of the reference car, as is the case when driving in opposite lanes.

During the proposal generation stage, we sample a set of relations r' from this distribution. Then, for each seed object O we generate object proposals O' following the sampled relations r'. Finally, the 3D object proposals O' are projected into the image plane producing the 2D object proposals o'.

Higher-order Relations Discovery (f^+) : Given a set of training images containing objects occurring in a scene, our goal is to discover the underlying higher-order relations that influence the location and pose in which each object instance occurs w.r.t. each other. The intuition behind this idea is that some object categories tend to arrange themselves in such a way that a specific goal is achieved. For example, in the urban setting, cars tend to drive behind each other in the same direction following lanes. For parking, cars park in specific arrangements, outside of the lane area. All this is done in order to promote circulation of cars. Likewise, in office settings, computer screens, keyboard and mouse, are arranged in a way that provides an interactive experience for the user. Based on these observations, our goal is to discover these underlying higher-order relations from annotated images. Then, during testing, we generate object proposals in such a way that they are consistent with these higher-order relations. A very similar problem, of discovering abstract topics $t = \{t_1, t_2, ..., t_T\}$ that influence the occurrence of words w within a document d, is addressed by Topic Models [15] [52]. Motivated by this similarity we formulate our higher-order relation discovery problem as a topic discovery problem. According to the topic model formulation, a document d_i can cover multiple topics t_k and the words w that appear in the document reflect the set of topics t_k that it covers. From the perspective of statistical text analysis and document analysis, a topic t_k can be viewed as a distribution over words w; likewise, a document d can be considered as a probabilistic mixture over the topics t (see Section 2.6 for more details).

In order to meet this formulation in our particular setting, given a set of training images, we compute pairwise relations r_{ij} between all the objects O_i within each image as before. Then, for each object O_i we define a document d_i where the words w^i are defined by the pairwise relations r_{ij} that have the object O_i as the source object. Additionally, we experiment with an alternative way to compute the pairwise relations between objects. Specifically, we run tests with a variant of the relative orientation attribute of the relation where instead of considering the pose of the target object we consider the orientation of its elongation ϵ (similar to Chapter 5). This orientation is less affected by errors during prediction, since traditional pose estimators tend to make mistakes by confusing opposite orientations, e.g front-back, left-right, etc.

In order to make the set of extracted pairwise relations R applicable within the topic model formulation we need to quantize them into words. To this end, we

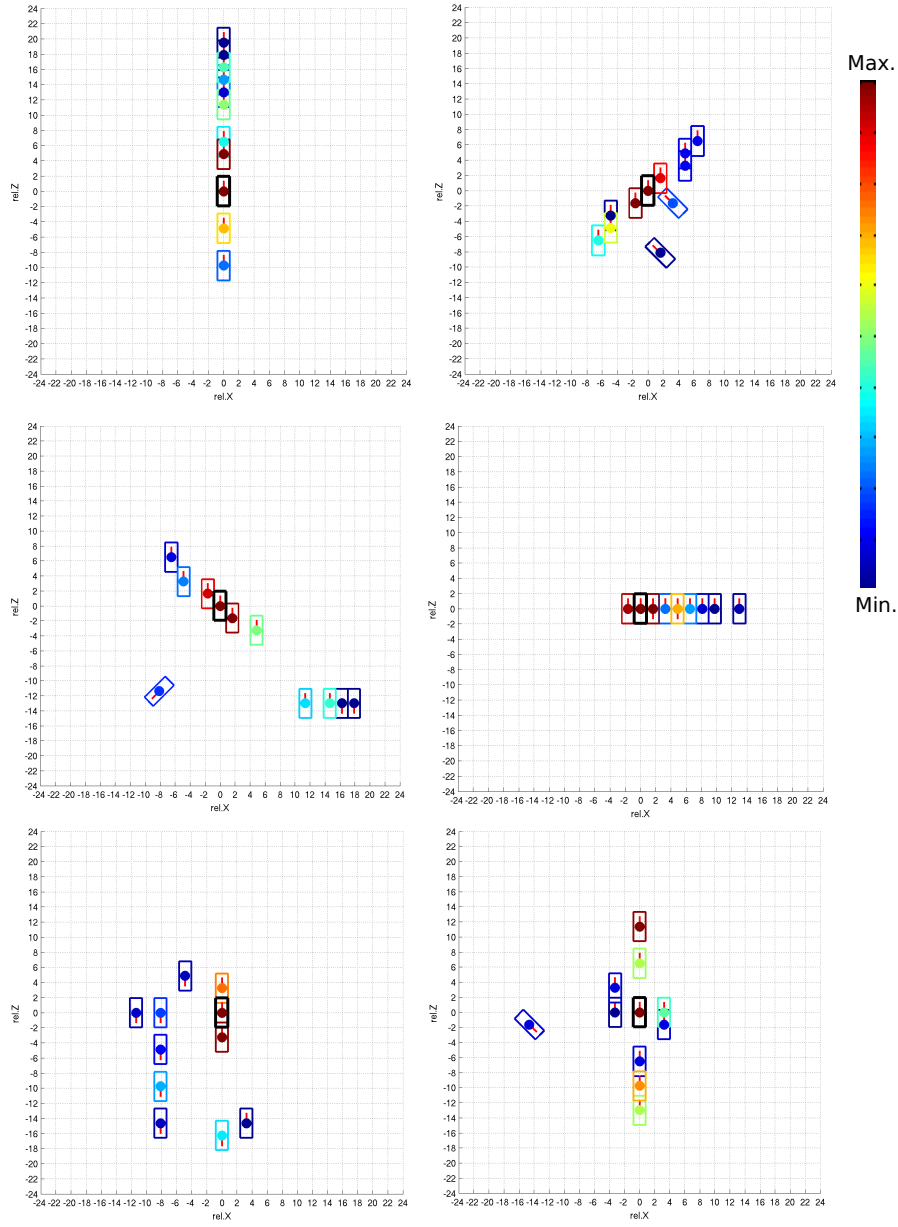

Figure 6.3: Top view of the discovered relational Topics from an object-centered perspective from cars in the KITTI dataset [45]. For each topic, the reference object is in the center and colored in black. The related objects are presented with their occurrence likelihood color-coded in jet scale. Notice how the discovered topics resemble traffic scenarios from urban scenes. For visualization purposes, each object is being plotted with average size of the annotations in the training set of images. We only show the top 10 most likely words per topic. (Best viewed in color).

discretize the space defined by the relations R by $(W/2, W/2, K)$ where W is the average width of the annotated 3D objects in the training set, and K is a predefined number of discrete poses of the object, 8 in our experiments. At this point, we are ready to perform topic modeling in our data. Here we use Latent Dirichlet Allocation [15] for topic modeling. For inference, we follow a Gibbs sampling method as in [52]. Specifically we use the implementation from the Matlab Topic Modeling toolbox 1.4 [1] released by Steyvers and Griffiths.

Our main goal is to identify the set of topics t that define higher-order arrangements between objects O in the scene. In our experiments we extract 16 topics from our documents d. Fig 6.3 shows a top view of a subset of the discovered topics when considering object-centered pairwise relations as words. These relations were computed from cars in the KITTI dataset [45]. Notice how some of the topics resemble common traffic patterns of cars in urban scenes. These topics represent the underlying higher-order relations that we claim influence the way in which objects tend to co-occur.

During the object proposal generation stage, we assume that each 3D object O_i, estimated from the seed object detection o_i, is related with the object proposals O' under higher-order relations. For simplicity, we assume that all the higher-order relations are equally likely to occur. Object proposals O' are then generated by sampling the word distributions $p(w|t)$ given each of the topics t. Finally the sampled 3D object proposals O' are projected to the image plane, yielding o'. The assumptions made at this stage have three desirable effects. First, object proposals are sampled in such a way that they follow the higher-order relations between objects. Second, the exploration process gives priority to the most likely proposals from each of the discovered higher-order relations, see Figure 6.3. Third, we are able to reason about higher-order relations even for the scenario when just one object detection o_i was collected by the detector.

6.4 Evaluation

Experiment Details
We perform experiments in the KITTI object detection benchmark [45]. This dataset constitutes a perfect testbed for our analysis since it covers a wide variety of difficult scenarios ranging from object instances with high occlusions to object instances with very small size. Furthermore, it provides precise annotations from objects in the 2D image and in the 3D space, including their respective viewpoints and poses. We focus on cars as the category of interest given its

[1] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

high occurrence within this dataset which makes it appropriate for reasoning about relations between objects. We focus our evaluation on images with two or more objects, where it is possible to define such relations. Matching between annotated objects and object proposals is estimated based on the intersection over union criterion from the Pascal VOC [36]. We report as evaluation metric the recall as a function of the number of object proposals generated per image as is often used for evaluating object proposal methods. In this analysis we use the LSVM-MDPM-sv detector from [47] to collect the initial set of object detections. See Chapter 2 for more details regarding the KITTI dataset, the hypothesis matching criterion and the detector used during this evaluation.

As baselines we use the methods introduced in Section 6.3.2: *Relaxed Detector*, *3D Sliding Window* proposals, and the proposals generated by *Selective Search* [151]. For the case of category-specific context-based proposals, we evaluate one method based on pairwise relations, *Pairwise*, and two methods based on higher-order relations, *HOR* and *HOR-Elongation*, where the latter is the variant based on object elongation orientation instead of object pose. For the special case when no seed objects are available, i.e. images where the object detector was unable to find detections above the threshold, we fallback to the *3D Sliding Window* strategy and consider the proposals proposed by this strategy. Similar to Chapter 3, we evaluate the changes in performance when considering camera-centered (CC) relations vs. considering object-centered (OC) relations.

Exp.1: Relations-based Object Proposals

In this experiment we focus on evaluating the strategies based on relations between objects (pairwise and higher-order). We consider as seed objects for our strategies the object detections collected with the detector [47]. Figure 6.4 presents performance on the range of [0,1000] generated object proposals.

Discussion: Strategies based on camera-centered higher-order relations seem to dominate the results. They achieve around 10% higher recall than all other methods over a wide range of the curve. This can be attributed to the fact that higher-order relations consider object arrangements with more than two participating objects. This allows them to spot a larger number of areas that are likely to contain objects. In addition, higher-order relations cover a wider neighborhood, whereas the pairwise relations have a more "local" coverage (i.e. they explore mostly a small neighborhood around the seed detections). As a result, strategies based on higher-order relations are able to explore highly likely regions to contain the objects within a larger neighborhood. This is more visible in the range [0,500] of the sampled proposals, where recall from methods based

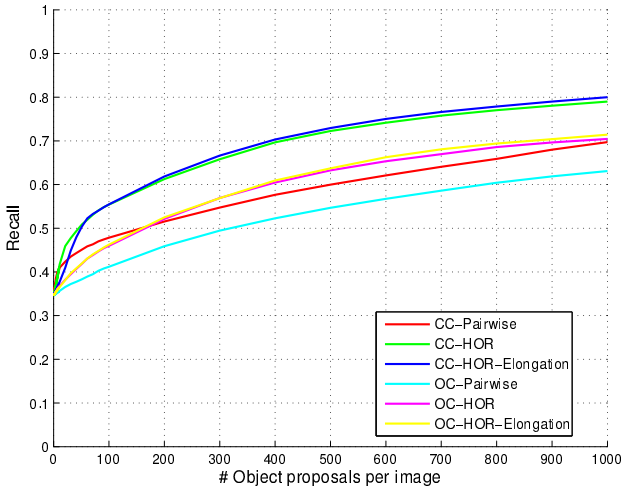

Figure 6.4: Recall vs. number of generated proposals for the scenario when all the object hypotheses reported by the detector are used as seed objects. We report performance when considering pairwise and higher-order relations (HOR). In addition, we report results when defining both camera-centered (CC) and object-centered (OC) relations. (Best viewed in color).

on higher-order relations increases faster than for pairwise relations. This can be further verified in Figure 6.5 (second row).

For the case of relation-based strategies, it is visible that strategies based on camera-centered relations have superior performance than their object-centered counterparts. This can be partly attributed to the fact that object proposals sampled following object-centered relations are affected by errors during the prediction of the pose of the seed object. Moreover, the camera setup in the KITTI dataset is fixed, introducing low variability in the camera-centered relations. In a scenario with higher variability on camera viewpoints we expect object-centered relations to have superior performance over camera-centered relations.

In addition, for the case of camera-centered relations, the higher-order relations where the elongation orientation is considered are slightly better, albeit only marginally so. This can be attributed to the fact that the orientation of the elongation of an object is less affected to errors in the pose estimation. Moreover, by defining camera-centered relations we also avoid the noise introduced in the pose of the seed objects.

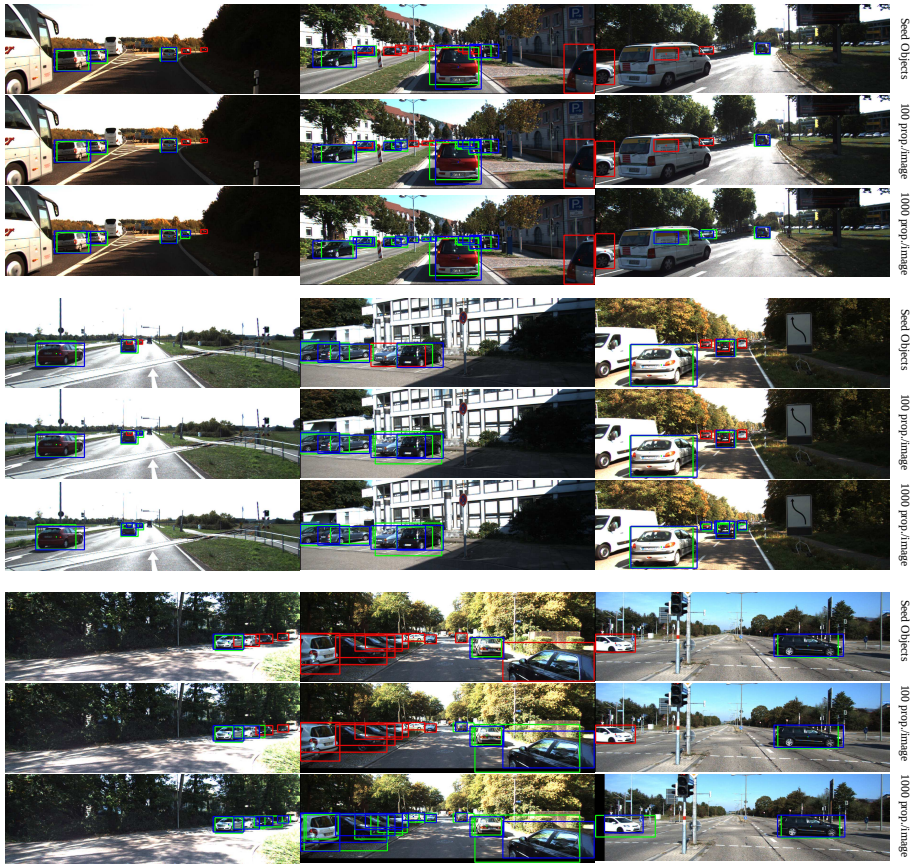

Figure 6.5: Object proposals generated in chronological order using the context-based strategy based on camera-centered higher-order relations. For the sake of clarity in the visualization, we removed the false positive proposals and only show the bounding boxes of the matched object annotations (in blue), missed detections (in red) and matching object proposals (in green). First row, seed objects collected with the object detector [47]; second row, results after sampling 100 object proposals; and third row, results after sampling 1000 object proposals. (Best viewed in color).

Despite the difference on performance between the proposed strategies, it is remarkable that we are able, on average, to double the initial recall obtained by the object detector by following relatively simple strategies. This suggests that object proposal generation should not be employed solely as a pre-detection step as it is commonly found in the literature [1, 34, 151, 168]. Furthermore, this suggests that there should be some level of interoperability between object detection and object proposal generation methods.

Exp.2: Starting from a Single Object Seed

This experiment is similar to the previous experiment with the difference that for each image we only consider the top scoring object detection as seed object. The objective of this experiment is to measure what performance can be achieved if we start from the top scoring object detection. This single seed scenario is of interest for several reasons. As stated earlier, appearance-based detectors can be very reliable at levels of high precision and low recall. As a result, by starting from the top scoring object detection we will be seeding our search with the most reliable object detection. Thus, introducing less noise in the search. Additionally, for real-time applications there may be either hardware or time constraints that restrict the possibility of performing an exhaustive search of objects. In some situations, locating multiple object instances may not be possible without searching densely over different locations, scales and aspect ratios. Finally, there are difficult scenarios, e.g. high inter-object occlusion, objects in very low scale, and drastic changes in illumination, where appearance-based detectors just manage to produce a single detection as output, if any. Similar to the previous experiment, Figure 6.6 shows performance on the range of [0,1000] generated object proposals.

Discussion: A quick inspection of Figure 6.6 shows similar trends as the ones observed in the previous experiment. However, different from the previous experiment, recall is relatively lower in the range of [0,100] proposals. This is to be expected since we start from a smaller pool of seed objects. However, it is surprising to see how we can achieve nearly similar performance around the range of 400 proposals by just starting from a single seed object. This further supports the idea of interoperability between object detectors and object proposal generators.

Exp.3: Comparison with non-contextual strategies

The objective of this experiment is to compare the performance of the relations-based strategies w.r.t. the non-contextual ones. For the relation-based strategies

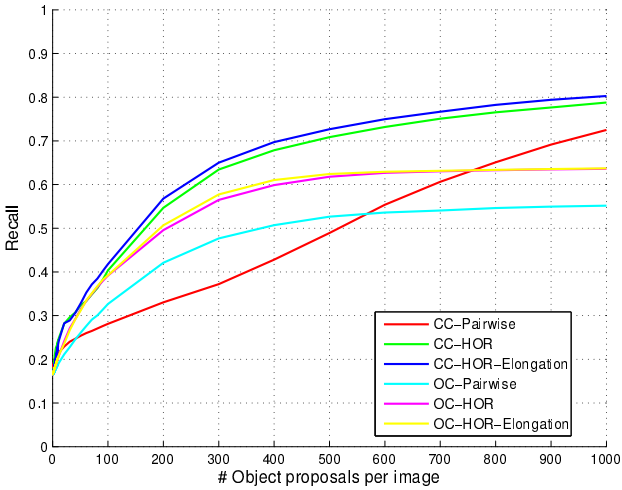

Figure 6.6: Recall vs. number of generated proposals for the scenario when only the top scoring object detection reported by the detector is used as seed object. We report performance when considering pairwise and higher-order relations (HOR). In addition, we report results when defining both camera-centered (CC) and object-centered (OC) relations. (Best viewed in color).

we consider the camera-centered variants only since, in the previous experiments, they achieved higher performance than their object-centered counterparts. As non contextual strategies we consider *Relaxed Detector*, the *3D Sliding Window*, and the *Selective Search* method from [151]. We report results considering all the detections as seed objects in Figure 6.7.

Discussion: We notice that the contextual strategies based on higher-order relations have a superior performance than all the other strategies. In addition, the performance of the strategy based on pairwise relations is similar to that of non-contextual strategies, except for the range of [0,200] where the strategy based on pairwise relations has higher performance. Focusing on the group non-contextual strategies, it can be noted that, their performance is relatively comparable. From this group, the *Relaxed Detector* has superior performance. Interestingly, a clear difference can be noted between the performance of contextual and non-contextual strategies. In the the range of [0,200], all the contextual strategies achieve superior performance than the non-contextual counterparts. This suggests that indeed contextual information is useful for an early exploration of regions of the image that are likely to host instances of the

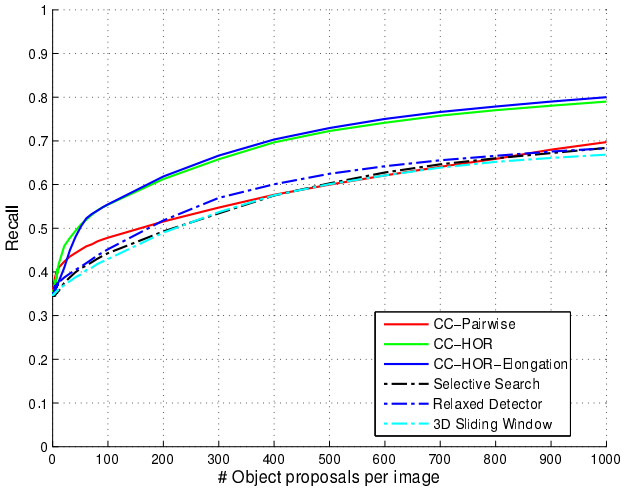

Figure 6.7: Comparison with non-contextual strategies. Recall vs. number of generated proposals. Here we report results on pairwise and higher-order relations (HOR) defined in a camera-centered (CC) fashion. (Best viewed in color).

objects of interest.

Exp.4: Proposal Localization/Fitting Quality

Finally, we add an experiment to measure the quality of the object proposal to localize and fit the region of the recovered object instance. For this purpose, in this experiment we employ a stricter matching criterion [36] of at least 0.75 intersection over union between the bounding boxes of the object annotations and the object proposals, respectively. We evaluate the performance of the same, contextual and non-contextual, strategies from experiment 3. In Figure 6.8 we report results considering all the detections as seed objects.

Discussion: Recall values obtained in this experiment are significantly reduced now that matching an object is a more complicated task. The performance of the *Relaxed Detector* is surprisingly high. This can be attributed to the fact that with the non-maximum suppression step removed in the *Relaxed Detector*, the detector is able to exhaustively explore the areas where appearance have triggered a detection. In addition, we notice that pairwise relations are now

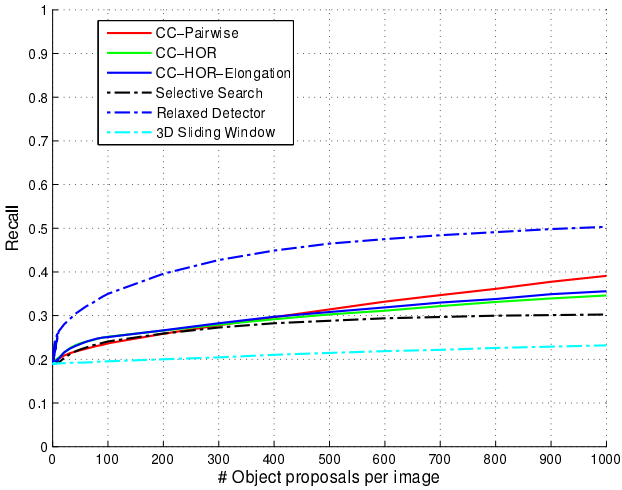

Figure 6.8: Recall vs. number of generated proposals considering a stricter overlap threshold. Here we report results on pairwise and higher-order relations (HOR) defined in a camera-centered (CC) fashion. (Best viewed in color).

outperforming the higher-order alternatives in the range of [500,1000] proposals per image. This may be caused by the discrete nature of the words in the topic models which are used to discover higher-order relations. As a result, the proposals generated from higher-order relations are spatially sparser than the ones produced by pairwise relations. The strategies based on pairwise relations tend to first concentrate on regions of high density before exploring other areas. This is why we notice improvements in the range [500,1000] and not earlier. These observations hint at a possible weakness of our relation-based strategies to generate object proposals. On one hand, relation-based proposals have some level of sparsity embedded, in our case, either by vector quantization of the relational space or by assuming mean physical sizes for the objects in the scene, when reasoning in 3D. This can be a weakness compared to the exhaustive *Relaxed Detector* strategy, when the objective is to have good localization/fitting. On the other hand, relation-based strategies seem to be better suited for "spotting" the regions where the objects of interest might be. This is supported by their superior recall on Experiment 1. This further motivates our idea of a joint work of object detectors and object proposal generators.

6.5 Discussion

The objective of this chapter is to explore an alternative way to exploit contextual information to recover object instances missed in an initial detection step. Note that we do not claim our method to be a full pipeline for object detection. That is why no standard detection metric/evaluation is presented. We see our method as an intermediate step that a subsequent stage can feed from towards improving detection. In this regard, a conditional random field that operates over the set of detections and proposals, produced by our method, might be a good way to re-score the proposals and improve precision.

In addition, it is important to clarify that we do not claim that recall is more important than precision. However, methods for context-based detection, e.g. [30, 108, 118], mostly focus on using context for filtering false detections, thus only improving precision. Different from them, we exploit contextual information to recover missed detections. That is why our evaluation is based on recall. From the results obtained in the previous experiments, it is arguable that a method with high recall and low-precision might not be optimal. However, in systems with various sources of information (multimodal sensors, multi-cameras or image sequences) it is desirable to detect the majority of the objects since the pool of detections can be further reduced by imposing consistency along the different sources.

A current limitation of the presented relations-based methods could be their focus on the car category on a groundplane. However, it is important to note that our relations-based methods can be extended to cover other object categories not necessarily on the groundplane. This can be achieved by adding the relative Y location (r_Y) and related object category (r_C) as relation attributes.

A deeper inspection of the qualitative results (see Figure 6.5) produced by our methods reveals a particular trend on how it addresses object instances of different sizes. As a starting point, our method generates proposals using detections as seeds. Since detectors are better at detecting objects at larger sizes, our method first focuses on objects on their vicinity, which have similar size. Eventually, objects with smaller sizes at larger distances to the seeds are explored. Furthermore, we have noted that our method consistently misses object instances close to the camera. This is due the mismatch between the proposal and annotations bounding boxes which seems to be caused by high level of truncation (see Figure 6.5, columns 2 and 3 of row 1). This can be solved by detecting truncation when projecting the 3D proposal O' to the image and rescaling the 2D proposal o' bounding box accordingly.

Finally, a future extension to our method based on higher-order relations is to adapt it depending on the number of seed objects. For images with a single seed

object, we follow the current method, i.e. we assume that all the higher-order relations are equally likely and sample proposals accordingly. For images with two or more seeds, first, we assign higher-order relations to those seeds. Then we sample the proposals from the new distributions of higher-order relations. Note that in this second case, the higher-order relations are not equally likely. This would provide a better balance between the enforcement of higher-order relations between objects and the ability to reason about higher-order relations when just a single seed is available.

6.6 Conclusions

In this chapter we have taken the first steps to demonstrate that sampling object proposals is an effective way to recover missed detections. Our experimental results suggest that object proposal generation should not be employed solely as a pre-detection step as it is commonly found in the literature. Furthermore, our results suggest relations-based strategies are better suited for spotting regions that may contain objects of interest rather than achieving high localization/fitting. In addition, our novel method to discover higher-order relations between objects is able to recover semantic patterns as those traffic patterns found in urban scenes. Future work should focus on investigating the complementarity of the proposed strategies as well as proper ways to integrate them.

Chapter 7

Conclusion

Identifying the context in which objects occur in images and using this context to clarify ambiguous cases is a relatively trivial task for humans. However, there are some obstacles that should be overcome before the perception and reasoning skills of computers match those of humans. While significant progress has been achieved in modeling and using contextual information for the task of verifying the occurrence of specific objects of interest, e.g. [30, 59, 118], existing methods have mostly focused on promoting true detections and reducing the effect of false predictions. This produces improvements on object detection performance which are limited to a boost on precision. In addition, the use of contextual information has not been analyzed for assisting tasks such as object pose/viewpoint estimation. In this thesis, we have taken as starting point those two weak points on the application of contextual information for computer vision problems. This chapter summarizes the contributions made in this thesis. Following this summary, we present the lessons learned during the execution of the work that are part of this thesis. Then, we discuss some limitations of the proposed methods and provide some directions for future research.

7.1 Summary of Contributions

In this thesis we have explored means to exploit contextual information in order to improve the performance of computer vision tasks, namely, object detection and object viewpoint/pose estimation. As a result, the contributions resulting from this work cover both tasks in addition to some remarks regarding the integration of contextual information for reasoning.

7.1.1 Context-based Object Pose Estimation

A significant amount of work presented in this thesis, Chapters 3 and 5, addresses the task of object pose/viewpoint estimation. In Chapter 3, we showed that it is possible to extract some information about the pose/viewpoint of an object instance by looking at relations defined with other object instances in the scene. Furthermore, our experiments suggest that, for this particular task, pairwise relations defined from an object-centered perspective provide stronger cues about the pose of the object instance to be classified.

In addition, in Chapter 5, we showed that there are cues that can be drawn from the scene which provide information about the pose/viewpoint of object instances. We proposed four methods to collect these cues. These methods are suited for reasoning on the 2D image space or in the 3D scene.

7.1.2 Reasoning about Object Relations for Object Detection

As it was mentioned before, exploiting relations between objects to improve object detection performance has received much more attention in the computer vision literature. However, most of the existing work uses contextual information from other object instances in a rather crude way. Different from traditional methods, in Chapter 4, we proposed a cautious approach that iteratively uses the most certain relational information to predict less certain object instances. We showed that our cautious approach is able to improve the performance of both local appearance-based detectors, e.g. [40, 81], and aggressive context-based methods such as [118].

In addition, in Chapter 6 we showed how models based on object relations can be used within an object proposal generation framework for effectively recovering missed detected object instances. This is clearly complementary to what is found in the literature which is focused purely on improving the precision of object detection by re-ranking the hypotheses initially collected by the object detector.

7.1.3 Relational Reasoning between Object Instances

Chapters 3, 4 and 6 of this thesis cover different aspects of relational reasoning between object instances. In Chapter 4, we showed that reasoning about object relations cautiously improves object detection precision, specially in the purely-contextual case where there is no access to appearance features from the object being processed. Later in Chapter 6, we proposed a method based on topic

models to discover higher-order relations between object instances. We showed that assuming that objects are associated to each other by relationships that are pairwise in nature, as is commonly done in existing work, might not be the best way to model object relations. In fact, the quantitative results obtained in our experiments suggest that modeling higher-order relations between objects is beneficial, especially to improve the recall of a detection process. Moreover, on the qualitative aspect, our method to discover higher-order relations between objects is able to discover object arrangements similar to those found in traffic patterns, providing some level of interpretation to the image.

7.2 Observations

Having presented the contributions made in this thesis now we will make some general observations regarding the proposed methods.

Complexity: In this thesis we have presented methods that take into account contextual cues to improve the performance of object detection and object pose/viewpoint estimation. As was stated in the previous section, the proposed methods do bring some level of improvement for such tasks. However, not much have been said about the extra computational cost required for bringing in such contextual cues. In the methods we have presented in this thesis, there are two possible scenarios that may cause processing bottlenecks. The first bottleneck scenario occurs in methods that reason about object relations (Chapter 3 & 4). In these methods, the number of pairwise relations between objects increases exponentially w.r.t. the number of object instances. This represents a problem for the case when we want to process large amounts of object hypotheses as those produced by high-recall detectors. In our methods, relational inference is performed via the weighted-vote relational neighbor (wvRN) classifier. This classifier has a marginal computational cost when compared to the cost involved in acquiring the object hypotheses. In addition, previous works [87, 90, 91] have demonstrated the ability of wvRN to scale to massive networks. Based on this evidence, we do not expect this to be a a big issue in practice. The second bottleneck scenario may be triggered if the current method to predict object orientations is used densely towards the prediction of orientations with high granularity. In this regard, our methods have been tested for the prediction of 8 and 16 discrete orientation values. However, they may not be suitable for the prediction of 360 discrete orientations. To this aim, we propose to follow the method from [142] in which the object orientation distribution is approximated by a curve fitted to the responses of a reduced number of discrete orientations. Then, the peak of such curve is considered as the orientation of the object. We will provide more details of this method in Section 7.6.

Reasoning in the 3D scene space vs. in the 2D image plane: In this thesis we showed that relations between objects can be defined in the 2D image space (Chapter 4) as well as in the 3D scene space (Chapter 3 & 6). Similarly, in Chapter 5 we showed that scene-driven cues can be extracted from the scene defined in both the 2D image plane as well as in the 3D space. During the execution of these works we have observed that there is a trade-off between performance and annotation cost, that should be defined when deciding about the space in which reasoning takes place. As presented in [165], there are several advantages that can be obtained from reasoning on the 3D scene. For example, improved object localization accuracy can be obtained by using detailed 3D representations of the objects of interest. In addition, scene-level constraints can be applied to reduce the possible locations, poses or configurations in which the objects of interest may occur. However, all these improvements come with the cost of relatively more "expensive" 3D annotations. Indeed, especially in some applications, 3D data may be difficult to acquire. Even though with the recent popularity of consumer-level depth cameras, e.g. MS kinect, datasets including 3D annotations are starting to appear [74, 136, 139], the number of datasets with 2D annotations is still a lot higher. On the downside, the appearance of elements depicted in images are commonly affected by illumination changes, motion, perspective effects and other factors that influence the imaging process. As a result, 2D annotations in the image space show higher variability which later affects performance. In addition, for the case of object relations in the 3D space, there is the possibility of defining relations either from a camera-centered or from an object-centered frame of reference. This provides the flexibility of defining relations between objects in such a way that it is suitable to our applications. For example, camera-centered relations can be defined when the camera setup is fixed while object-centered relations can be adopted for images acquired with more arbitrary camera setups. For these reasons, a trade-off between performance and annotation cost should be defined depending on the task at hand.

7.3 Lessons Learned

The execution of the work covered in this thesis has taught us some lessons that we believe are worth sharing since they serve as a motivation for the adoption of ideas presented in this thesis. Furthermore, they can be taken as guidelines to be considered when planning to integrate the ideas and methods proposed in this thesis.

Collective classification should be used cautiously in vision problems: Methods based on collective classification and relational inference have proven

to be effective when handling networks derived from documents, e.g. websites, bibliographic data, emails, see Section 2.7. However, in these networks the features, the text, that describe the nodes, the documents, are almost noise free and there are large corpora describing possible class labels for them. Noise on this type of data usually arises on unsupervised user-generated data, e.g. emails, forum posts, tweets, etc. in the form of typos. On the contrary, for vision based applications, especially in the networks like the ones defined in this work, some of the nodes might be produced by false hypotheses and in consequence produce false links. For this reason, for this specific application, caution should be exercised when building and reasoning on networks constructed from object hypotheses. This is also supported by our observations in Chapter 4 where, in the absence of local information, defining relations with neighboring objects cautiously always performed better than its aggressive counterpart where all the objects in the neighborhood where selected despite its certainty.

Based on our experience, and supported by the findings from the Collective Classification and the Link-based classification communities, we highlight two factors that should serve as indicators of whether the relational methods proposed in this thesis will bring benefits to a specific problem. The first factor measures the *link density* of the network. Findings made by the previously mentioned communities [14, 69, 102, 135] state that methods that reason about links, or relations, between objects have an improved performance when operating on a setting with high *link density*. However, in vision-related tasks *link density* is not a strong indicator on its own since, as said before, a lot of links could originate from false nodes, i.e. be produced by false detections. For networks similar to the ones presented in this thesis there is a second indicator that we believe should be considered in parallel with *link density* since it clarifies the scenario when high *link density* is produced by false nodes. In the collective classification literature this factor is known as *class-skewness* or *labeled proportion* [20, 91, 96, 141]. This indicator is of interest, specifically, for within-network classification tasks where predictions about some nodes are based on other nodes. In this type of tasks, *class-skewness* measures the proportion of data that is known, or predicted, with certainty w.r.t. the whole data. In scenarios where *class-skewness* is low there is not enough certain information to guide the inference process. In scenarios with high *class-skewness*, the performance of collective classification is comparable to that of local classification. In our setting, *class-skewness* refers to the ratio between true hypotheses and false hypotheses predicted by the local detector.

Object pose/viewpoint estimation is not purely a local intrinsic problem: In Chapter 3 we showed that it is possible to predict, up to some level, the pose of an object by only looking at objects in its neighborhood. Likewise, in Chapter 5 we showed that the pose/viewpoint of an object can also be predicted

by drawing cues from the scene. There is a clear underlying message if we put together the observations of these chapters: *Object pose/viewpoint estimation is not a purely-local problem.* As has been stressed in previous chapters, object pose/viewpoint estimation is a problem that has been traditionally addressed from a very local perspective. As result, recent work [81, 115] exploiting intrinsic object features such as color, texture, geometry, etc. have presented impressive results, even in the presence of high object occlusion [165]. This message should motivate future work on object pose estimation to fuse both intrinsic cues from the object category as well as external cues drawn from the context in which the object occurs.

Object relations can also be used to improve object detection recall: Exploiting contextual information to assist object detection has proven to be useful to filter out false object hypotheses. However, as pointed out by [164], some methods [107, 108, 118, 165] that exploit contextual information suffer from an early commitment to object hypotheses. This commitment refers to the fact that only a fixed set of hypotheses whose detection scores are above a particular threshold are processed. As a consequence, improvements in object detection are mostly in terms of precision since object instances missed during the detection step cannot be recovered. In Chapter 6, we showed that models based on object relations can be used to sample likely locations to host the miss-detected objects. Empirical results from our experiments comparing object instances collected by the appearance-based detector and those recovered using the relations-based models suggest a complementary behavior between the two. On the one hand, relation-based models are good for spotting regions that are likely to contain the objects of interest, similar to the root filters from the DPM detectors [40]. On the other hand, appearance-based models are better suited for refining the location of the objects.

7.4 Revisiting the Research Questions

At the beginning of this thesis we set the objective of investigating the potential of contextual information for improving the performance of computer vision tasks. To this end, we formulated three research questions which were presented in the introduction. From then on, we went on a journey aimed at answering these questions which resulted in a set of contributions and observations. Based on these contributions and observations, we will revisit each of the research questions and address them accordingly.

1. Is contextual information, in the form of relations between objects, useful for object pose estimation?

The results obtained in Chapter 3, when using a purely contextual baseline, suggest that our models based on relations between objects are able to encode information about the orientation of the participant objects. Moreover, additional experimentation showed that these models can be combined with methods that model local information to improve the performance in object pose estimation. These observations were further confirmed by other authors, namely Xiang et al. [160] and Zia et al. [165, 166]. These findings essentially show that the answer to this research question is positive.

2. To what extent does the nature of the association between objects affect the performance of using relations between objects to improve object detection?

This research question was explored from two different perspectives. From the first perspective, we analyzed to what extent the selection of related neighboring objects affects the object detection performance. From the second perspective we explored to what extent the assumption on the origin of the relations between objects has an impact on object detection performance. Regarding the first perspective, our results in Chapter 4 confirm that following specific strategies, like our cautious iterative approach, to select the objects that provide contextual information brings improvements to object detection performance. In particular, to improve precision on the detection process. Regarding the second perspective, our results in Chapters 4 and 6 suggest that considering that objects are associated via underlying relationships increases the performance of relations-based methods for object detection. Furthermore, results from 6 show that considering relationship-driven association between objects is effective to improve object detection performance in terms of recall.

3. Is contextual information, in the form of scene-driven cues, useful for the task of object viewpoint estimation?

Aiming at answering this research question, in Chapter 5 we proposed four methods to extract contextual cues from the scene. Our experiments using these cues showed that their performance for object viewpoint estimation, was well above chance levels. This suggests that the scene can effectively serve as a source of contextual information that is useful for object viewpoint estimation. Additional experiments showed that improvements on object viewpoint estimation performance can be obtained by combining our scene-driven contextual cues with methods that reason about intrinsic object features. These results give a positive answer to this Research Question by showing that considering contextual cues taken from the scene improves the performance of object viewpoint estimation.

7.5 Limitations

In previous sections we have presented that improvements in vision-related tasks such as object detection and object pose/viewpoint estimation are obtained when considering contextual cues. In this regard, this thesis proposed methods to reason about different contextual cues for object pose estimation as well as methods to recover missed detections and improve object detection recall. However, as can be expected, there is room for further improvement on these methods. The content of this section highlights weak points of the work presented in this thesis.

Data Quality: In this thesis we have presented several methods that reason about contextual information to complement local models and improve vision tasks. However, in order to be informative, models based on contextual information should be trained from data that is representative to the context setting to be modeled. In this regard, dataset bias is a problem that affects models based on local appearance features [70, 147], as well as models based on contextual features. For models that reason about pairwise relations (Chapters 3 & 4), their characteristic of exponential increase of pairwise relations w.r.t. the number of object instances becomes a double-edged sword in the presence of dataset bias. On the one hand, when local data is representative, a significant amount of meaningful contextual information can be extracted. Then, from this information, we can build contextual models that are effective at complementing methods that only reason about local information. On the other hand, when local data is not representative i.e. in the presence of dataset bias, contextual information produces exponentially biased context models. In consequence, the initial effect the dataset bias is further increased by integrating the biased context models. In this regard, the collective classification community has proposed some factors such as *link density* [14, 69, 102, 135] and *class-skewness* (or *labeled proportion*) [20, 91, 96, 141], that can be used to identify whether the relational methods proposed in this thesis will indeed improve the methods that reason about local information.

Reduced number of object categories: Even when the models proposed in this thesis can be extended to cover a wide variety of object categories, our experiments only consider *cars* as the category of interest. Further experiments considering additional object categories should provide insights on the generality of the proposed methods w.r.t. other object categories. Furthermore, experiments with different additional object categories might point out additional aspects that should be considered for a proper reasoning about contextual information.

Relational nature: An essential requirement of the methods that reason about relations between objects (Chapters 3, 4 & 6) is that at least two object instances should be available in order to define a relation. This restricts the applicability of the proposed relational methods on scenes where only a single object is present. In addition, in this single object scenario, the presence of multiple false object hypotheses may give the false impression that more than one object is present, thus representing a problem. In this regard, we proposed the integration of scene context (Chapter 5) to alleviate the problems arising on this scenario. Scene context cues do not have the requirement of at least two objects present on the scene. In addition, scene context can assist the relations-based model to reduce the effect of false object hypotheses during relational reasoning.

Focus on discrete pose/viewpoint estimation: In Chapters 3 and 5 we presented two different methods to reason about contextual information to improve the performance of object pose/viewpoint estimation. Although the proposed methods proved to be effective at improving pose/viewpoint estimation, they have the weakness of only predicting discrete orientation angles. This may represent a significant factor in applications that require more precise, continuous, predictions. In this regard, mechanisms to extend the methods proposed in this thesis should be explored.

Integration of weak intrinsic features: The main focus of this thesis is on reasoning about contextual information to assist computer vision problems. A direct consequence of focusing on the contextual aspect of the problem is that very little attention has been given to local appearance features of the objects of interest. To this end, we have limited ourselves to only using the size, orientation, and detection score as local features of the objects but have completely ignored more intrinsic features such as appearance, shape, etc. This is complementary to the work from [160] and [165] whose findings suggest that reasoning about fine intrinsic object features can boost object detection and pose estimation. Motivated by those findings we propose to integrate reasoning about fine intrinsic object features to complement the context-based methods proposed in this thesis.

In the next section we will use these limitations as starting point to draw directions for future work.

7.6 Directions for Future Research

In this section we draw some directions for future research based on the findings and limitations of the work presented in this thesis.

The first direction for future work lies in the integration of models that reason about relations between objects and models that reason about contextual cues extracted from the scene. As mentioned in the previous section, in addition to reducing the effect of false object hypotheses, the new joint-model would be applicable in any scene irrespective of the number of objects occurring in the scene. Furthermore, there are other sources of contextual information that have been proposed in the computer vision literature, e.g. geometric context [60], inter-object occlusions [165], geographic context [57] and others [32], that can be integrated in order to perform more informed predictions.

The second direction for future research is related to the integration of finer intrinsic features of the objects at the inference stage. In the current methods, the only local cues that are considered are derived from the output of the object detectors. However as suggested by [160] and [165], reasoning about finer features, such as those captured by detailed 3D object models, are beneficial for object detection even in the presence of severe occlusion. We believe that features derived from detailed 3D object models proposed in [160] and [165] can extend the methods proposed in this thesis in two ways. First, detectors based on detailed 3D models can better cope with the detection of highly occluded objects. This will provide the context-based methods with a larger set of true detections which, for the relational case, are beneficial [20, 91, 96, 141]. Second, we can compute more informative relation attributes based on features derived from the detailed 3D object models. We expect this to produce more informative relations than those produced from the weak local features extracted from the response of the detectors.

The third direction, addresses the limitation of discrete pose/viewpoint predictions. A straight forward strategy to obtain a continuous orientation output from the discrete output of our methods, is by following the method proposed in [142]. The idea is to use the set of discrete pose/viewpoint values considered by our method, together with their predicted scores, as approximations of the likelihood function for the pose/viewpoint at some "probing" points. Then a distribution (Gaussian and von Mises-Fisher distributions) is locally fitted considering those points. Finally, the mean of the fitted distribution is retained as the continuous pose/viewpoint prediction. In addition, in the recent years, an increasing number of methods [114, 142, 167] have been proposed to perform continuous object pose/viewpoint estimation based on intrinsic object features. Future work should evaluate the performance of our methods for pose/viewpoint when combined with these methods.

Fourth, as presented in Chapter 6, there is a strong potential on using context-based models for recovering object instances missed after an initial detection stage. In this regard, we propose to avoid early commitment to high-scoring object hypotheses and disposal of low-scoring hypotheses. In addition,

we propose to investigate mechanisms to iteratively reconsider low-scoring hypotheses by using context-based models.

The fifth direction is related to the evaluation of more advanced methods to perform Collective Classification. To this end, we propose to explore the potential of Statistical Relational Learning (SRL) [50] when applied in the methods proposed in this thesis. Statistical Relational Learning is a field derived from Inductive Logic Programming [100] and is concerned with the modelling of domains that exhibit both uncertainty and complex, relational structure. Knowledge in SRL models is usually represented by using first-order logic which is quite useful to describe relational properties of a domain in a general manner. In recent years, several methods have been proposed to integrate probabilistic, logical and relational representations in machine learning. Prominent examples of these methods include: Markov Logic Networks (MLNs) [123], PRISM [129] and ProbLog [54]. In addition to sophisticated machinery for Collective Classification, SRL methods will also enrich our methods by providing declarative means to integrate background knowledge as well as data from other sources of information, e.g. image captions, audio, etc.

Bibliography

[1] Alexe, B., Deselaers, T., and Ferrari, V. What is an object? In *CVPR* (2010). pages 92, 95, 106

[2] Alexe, B., Deselaers, T., and Ferrari, V. Measuring the objectness of image windows. *TPAMI* (2012). pages 38, 92

[3] Antanas, L., van Otterlo, M., Oramas M, J., Tuytelaars, T., and Raedt, L. D. There are plenty of places like home: Using relational representations in hierarchies for distance-based image understanding. *Neurocomputing* (2014). pages 2

[4] Atanasoaei, C., McCool, C., and Marcel, S. A principled approach to remove false alarms by modelling the context of a face detector. In *BMVC* (2010). pages 92

[5] Avidan, S., and Shamir, A. Seam carving for content-aware image resizing. In *ACM Transactions on graphics* (2007). pages 16

[6] Bao, S. Y., Bagra, M., Chao, Y.-W., and Savarese, S. Semantic structure from motion with points, regions, and objects. In *CVPR* (2012). pages 4, 41, 80

[7] Bao, S. Y., and Savarese, S. Semantic structure from motion. In *CVPR* (2011). pages 4, 41, 80

[8] Bao, S. Y., Xiang, Y., and Savarese, S. Object co-detection. In *ECCV* (2012). pages 94

[9] Bao, S. Y.-Z., Sun, M., and Savarese, S. Toward coherent object detection and scene layout understanding. In *CVPR* (2010). pages 6, 48

[10] Benenson, R., Mathias, M., Timofte, R., and Van Gool, L. Pedestrian detection at 100 frames per second. In *CVPR* (2012). pages 2, 3

[11] Benenson, R., Mathias, M., Tuytelaars, T., and Van Gool, L. Seeking the strongest rigid detector. In *CVPR* (2013). pages 2, 3

[12] Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology* (1982). pages 1

[13] Bileschi, S. M. Streetscenes: Towards scene understanding in still images. In *PhD Dissertation, MIT* (2006). pages 35, 37, 69

[14] Bilgic, M., Namata, G. M., and Getoor, L. Combining collective classification and link prediction. In *Workshop on Mining Graphs and Complex Structures at ICDM* (2007). pages 117, 120

[15] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *JMLR* (2003). pages 12, 29, 100, 102

[16] Bowman, A. W. An alternative method of cross-validation for the smoothing of kernel density estimates. In *Biometrika* (1984). pages 24

[17] Breiman, L., Meisel, W., and Purcell, E. Variable kernel estimates of multivariate densities. *Technometrics* (1977). pages 26

[18] Cao, X., Wei, X., Han, Y., and Chen, X. *IEEE Transactions on Cybernetics*. pages 8, 92, 94

[19] Chakrabarti, D., Funiak, S., Chang, J., and Macskassy, S. A. Joint inference of multiple label types in large networks. In *ICML* (2014). pages 31

[20] Chakrabarti, S., Dom, B., and Indyk, P. Enhanced hypertext categorization using hyperlinks. In *SIGMOD* (1998). pages 73, 117, 120, 122

[21] Chavez-Aragon, A., Macknojia, R., Payeur, P., and Laganiere, R. Rapid 3d modeling and parts recognition on automotive vehicles using a network of rgb-d sensors for robot guidance. *Sensors* (2013). pages 4

[22] Choi, M., Lim, J. J., Torralba, A., and Willsky, A. S. Exploiting hierarchical context on a large database of object categories. In *CVPR* (2010). pages 8, 9, 62, 94

[23] Cinbis, R. G., and Sclaroff, S. Contextual object detection using set-based classification. In *ECCV* (2012). pages 9, 42, 62

[24] Cohn, D., and Hofmann, T. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS* (2000). pages 31

[25] Couzin, I. D. Collective cognition in animal groups. *Trends in Cognitive Sciences* (2009). pages 9

[26] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *CVPR* (2005). pages 1, 2, 3, 15, 16, 18

[27] Dan Pelleg, A. M. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML* (2000). pages 69

[28] De Raedt, L., Kimmig, A., and Toivonen, H. Problog: A probabilistic prolog and its application in link discovery. In *IJCAI* (2007). pages 9, 30

[29] Dechter, R. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence* (1999). pages 31

[30] Desai, C., Ramanan, D., and Fowlkes, C. C. Discriminative models for multi-class object layout. *IJCV* (2011). pages 7, 8, 9, 40, 42, 60, 62, 92, 94, 110, 113

[31] Desrosiers, C., and Karypis, G. *Machine Learning and Knowledge Discovery in Databases*. pages 31

[32] Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., and Hebert, M. An empirical study of context in object detection. In *CVPR* (2009). pages 7, 42, 49, 122

[33] Duin, R. P. W. On the choice of smoothing parameters of parzen estimators of probability density functions. *IEEE Transactions on Computers* (1976). pages 24

[34] Endres, I., and Hoiem, D. Category-independent object proposals with diverse ranking. *TPAMI* (2014). pages 95, 106

[35] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. pages 85

[36] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. pages 8, 21, 38, 45, 54, 65, 70, 83, 85, 86, 98, 103, 108

[37] Fan, J., Hall, P., Martin, M. A., and Patil, P. On local smoothing of nonparametric curve estimators. *Journal of the American Statistical Association* (1996). pages 26

[38] Farmen, M., and Marron, J. S. An assessment of finite sample performance of adaptive methods in density estimation. *Comput. Stat. Data Anal.* (1999). pages 26

[39] Farmen, M., and Marron, J. S. An assessment of finite sample performance of adaptive methods in density estimation. *Comput. Stat. Data Anal.* (1999). pages 26

[40] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI* (2010). pages 1, 16, 19, 20, 22, 23, 62, 70, 74, 94, 97, 114, 118

[41] Fidler, S., Dickinson, S. J., and Urtasun, R. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS* (2012). pages 79

[42] Fischler, M. A., and Elschlager, R. The representation and matching of pictorial structures. *IEEE Transactions on Computers* (1973). pages 1, 19

[43] Forsyth, D. A., Malik, J., Fleck, M. M., Greenspan, H., Leung, T., Belongie, S., Carson, C., and Bregler, C. Finding pictures of objects in large collections of images. In *ECCV* (1996). pages 4, 41

[44] Galleguillos, C., McFee, B., Belongie, S., and Lanckriet, G. Multi-class object localization by combining local contextual interactions. In *CVPR* (2010). pages 7, 33, 42, 60

[45] Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR* (2012). pages 35, 36, 40, 49, 69, 85, 93, 99, 101, 102

[46] Geiger, A., Roser, M., and Urtasun, R. Efficient large-scale stereo matching. In *ACCV* (2010). pages 48

[47] Geiger, A., Wojek, C., and Urtasun, R. Joint 3d estimation of objects and scene layout. In *NIPS* (2011). pages 4, 20, 21, 23, 41, 48, 51, 52, 54, 55, 56, 81, 85, 86, 87, 88, 103, 105

[48] Getoor, L. Link-based classification. In *Advanced Methods for Knowledge Discovery from Complex Data*. 2005. pages 33, 72

[49] Getoor, L., Friedman, N., Koller, D., and Taskar, B. Learning probabilistic models of relational structure. In *ICML* (2001). pages 31

[50] Getoor, L., and Taskar, B. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007. pages 32, 123

[51] Glasner, D., Galun, M., Alpert, S., Basri, R., and Shakhnarovich, G. Viewpoint-aware object detection and pose estimation. In *CVPR* (2011). pages 4, 38, 50, 86

[52] Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences* (2004). pages 12, 29, 30, 100, 102

[53] Gupta, A., Efros, A. A., and Hebert, M. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV* (2010). pages 2

[54] Gutmann, B., Thon, I., and De Raedt, L. Learning the parameters of probabilistic logic programs from interpretations. In *ECML PKDD* (2011). pages 30, 123

[55] Habbema, J. D. F., H. J., and Van den Broek, K. A stepwise discrimination analysis program using density estimation. In *Compstat* (1974). pages 24

[56] Hall, P., Hu, T. C., and Marron, J. S. Improved variable window kernel estimates of probability densities. *The Annals of Statistics* (1995). pages 26

[57] Hays, J., and Efros, A. A. im2gps: estimating geographic information from a single image. In *CVPR* (2008). pages 122

[58] Hazelton, M. L. An optimal local bandwidth selector for kernel density estimation. *Journal of Statistical Planning and Inference* (1999). pages 26

[59] Heitz, G., and Koller, D. Learning spatial context: Using stuff to find things. In *ECCV* (2008). pages 2, 6, 40, 60, 113

[60] Hoiem, D., Efros, A. A., and Hebert, M. Geometric context from a single image. In *ICCV* (2005). pages 2, 7, 122

[61] Hoiem, D., Efros, A. A., and Hebert, M. Putting objects in perspective. In *CVPR* (2006). pages 2, 3, 6, 48, 80, 81, 82, 97

[62] Hoiem, D., Efros, A. A., and Hebert, M. Recovering occlusion boundaries from an image. *IJCV* (2011). pages 1

[63] Hoiem, D., Rother, C., and Winn, J. M. 3d layoutcrf for multi-view object class recognition and segmentation. In *CVPR* (2007). pages 39, 41, 79

[64] Hollingworth, A. Object-position binding in visual memory for natural scenes and object arrays. *Experimental Psychology. Human Perception & Performance* (2007). pages 1

[65] Hou, Y., He, L., Zhao, X., and Song, D. Pure high-order word dependence mining via information geometry. *Advances in Information Retrieval Theory* (2011). pages 8, 94, 95

[66] Huang, C., and Darwiche, A. Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning* (1996). pages 31

[67] Hulbert, L. G., Correa da Silva, M. L., and Adegboyega, G. Cooperation in social dilemmas and allocentrism: a social values approach. *European Journal of Social Psychology* (2001). pages 42

[68] Jain, A., Gupta, A., and Davis, L. S. Learning what and how of contextual models for scene labeling. In *ECCV* (2010). pages 2, 8, 42, 60, 62

[69] Jensen, D., and Neville, J. Autocorrelation and linkage cause bias in evaluation of relational learners. In *ILP*. 2003. pages 117, 120

[70] Khosla, A., Zhou, T., Malisiewicz, T., Efros, A., and Torralba, A. Undoing the damage of dataset bias. In *ECCV* (2012). pages 120

[71] Kinderman, R., and Snell, S. *Markov random fields and their applications*. American mathematical society, 1980. pages 30

[72] Kristan, M., and Leonardis, A. Online discriminative kernel density estimator with gaussian kernels. *Transactions on Cybernetics* (2014). pages 28, 84

[73] Kristan, M., Leonardis, A., and Skočaj, D. Multivariate online kernel density estimation with gaussian kernels. *Pattern Recognition* (2011). pages 28, 83

[74] Lai, K., Bo, L., and Fox, D. Unsupervised feature learning for 3d scene labeling. In *ICRA* (2014). pages 116

[75] Landwehr, N. Trading expressivity for efficiency in statistical relational learning. In *Ph.D. thesis, KU Leuven, Department of Computer Science* (2009). pages 32

[76] Laorden, C., Sanz, B., Santos, I., Galán-García, P., and Bringas, P. Collective classification for spam filtering. In *Computational Intelligence in Security for Information Systems*. 2011. pages 9

[77] Leushina, L., and Nevskaya, A. Perception of spatial relations between objects in early ontogeny. *Human Physiology* (2000). pages 1

[78] Li, C., Parikh, D., and Chen, T. Automatic discovery of groups of objects for scene understanding. In *CVPR* (2012). pages 7, 42, 60, 63, 68

[79] Li, H., and Chen, L. Removal of false positive in object detection with contour-based classifiers. In *ICIP* (2010). pages 92

[80] Liebelt, J., and Schmid, C. Multi-view object class detection with a 3d geometric model. In *CVPR* (2010). pages 39, 41, 50, 79

[81] Lopez-Sastre, R. J., Tuytelaars, T., and Savarese, S. Deformable part models revisited: A performance evaluation for object category pose estimation. In *ICCV WS* (2011). pages 4, 20, 21, 23, 38, 39, 41, 48, 50, 51, 52, 54, 55, 56, 68, 70, 72, 74, 79, 81, 85, 86, 87, 88, 114, 118

[82] Lv, F., Zhao, T., and Nevatia, R. Camera calibration from video of a walking human. *TPAMI* (2006). pages 48

[83] Lysenkov, I., and Eruhimov, V. Pose refinement of transparent rigid objects with a stereo camera. In *Transactions on Computational Science*. 2013. pages 4

[84] Macskassy, S. A. Leveraging contextual information to explore posting and linking behaviors of bloggers. In *ASONAM* (2010). pages 9, 35

[85] Macskassy, S. A. Contextual linking behavior of bloggers: leveraging text mining to enable topic-based analysis. *Social Netw. Analys. Mining* (2011). pages 31

[86] Macskassy, S. A. Definition and multi-dimensional comparative analysis of ad hoc communities in twitter. In *Proceedings of the Sixth International Conference on Weblogs and Social* (2012). pages 31

[87] Macskassy, S. A. On the study of social interactions in twitter. In *ICWSM* (2012). pages 9, 35, 115

[88] Macskassy, S. A., and Michelson, M. Why do people retweet? anti-homophily wins the day! In *ICWSM* (2011). pages 9, 35

[89] Macskassy, S. A., and Provost, F. A simple relational classifier. In *MRDM* (2003). pages 34, 35, 62

[90] Macskassy, S. A., Provost, F., and Macskassy, S. A. Suspicion scoring of networked entities based on guilt-by-association, collective inference, and focused data access. In *NAACSOS* (2005). pages 9, 35, 115

[91] Macskassy, S. A., and Provost, F. J. Classification in networked data: A toolkit and a univariate case study. *JMLR* (2007). pages 9, 11, 30, 31, 32, 34, 35, 41, 44, 60, 62, 64, 73, 115, 117, 120, 122

[92] Malisiewicz, T., and Efros, A. A. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS* (2009). pages 7, 42

[93] Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L. Face detection without bells and whistles. In *ECCV* (2014). pages 2

[94] Mathias, M., Benenson, R., Timofte, R., and Van Gool, L. Handling occlusions with franken-classifiers. In *ICCV* (2013). pages 1, 2, 3, 93

[95] McDowell, L., Gupta, K. M., and Aha, D. W. Cautious inference in collective classification. In *AAAI* (2007). pages 33, 34, 60, 64, 66

[96] McDowell, L., Gupta, K. M., and Aha, D. W. Cautious collective classification. *JMLR* (2009). pages 33, 60, 62, 73, 117, 120, 122

[97] McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* (2001). pages 60

[98] Mielniczuk, J., Sarda, P., and Vieu, P. Local data-driven bandwidth choice for density estimation. *Journal of Statistical Planning and Inference* (1989). pages 26

[99] Moeslund, T. B., Hilton, A., and Krüger, V. A survey of advances in vision-based human motion capture and analysis. *CVIU* (2006). pages 16

[100] Muggleton, S., and Raedt, L. D. Inductive logic programming: Theory and methods. *Journal of Logic Programming* (1994). pages 123

[101] Murase, H., and Nayar, S. K. Visual learning and recognition of 3-d objects from appearance. *IJCV* (1995). pages 1

[102] Neville, J., and Jensen, D. Leveraging relational autocorrelation with latent group models. In *ICDM* (2005). pages 34, 63, 117, 120

[103] Neville, J., and Jensen, D. D. Iterative classification in relational data. In *Workshop on SRL at AAAI* (2000). pages 62, 66

[104] Oliva, A., and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV* (2001). pages 6, 40

[105] Oliva, A., and Torralba, A. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research* (2006). pages 6

[106] Oliva, A., and Torralba, A. The role of context in object recognition. *Trends in Cognitive Sciences* (2007). pages 6, 77

[107] Oramas M, J., De Raedt, L., and Tuytelaars, T. Allocentric pose estimation. In *ICCV* (2013). pages 80, 118

[108] Oramas M, J., De Raedt, L., and Tuytelaars, T. Towards cautious collective inference for object verification. In *WACV* (2014). pages 8, 11, 92, 94, 95, 110, 118

[109] Oramas M, J., and Tuytelaars, T. Scene-driven cues for viewpoint classification of elongated object classes. In *BMVC* (2014). pages 12

[110] Osadchy, M., and Keren, D. Efficient detection under varying illumination conditions and image plane rotations. *CVIU* (2004). pages 1

[111] Ozuysal, M., Lepetit, V., and Fua, P. Pose estimation for category specific multiview object localization. In *CVPR* (2009). pages 4, 79, 85

[112] P. Felzenszwalb, R. Girshick, D. M. Cascade object detection with deformable part models. In *CVPR* (2010). pages 19, 48, 85, 86, 87, 90

[113] Pedersoli, M., Timofte, R., Tuytelaars, T., and Gool, L. V. Using a deformation field model for localizing faces and facial points under weak supervision. In *CVPR* (2014). pages 2

[114] Pepik, B., Gehler, P. V., Stark, M., and Schiele, B. 3d2pm - 3d deformable part models. In *ECCV* (2012). pages 22, 122

[115] Pepik, B., Stark, M., Gehler, P., and Schiele, B. Teaching 3d geometry to deformable part models. In *CVPR* (2012). pages 21, 38, 39, 41, 50, 79, 86, 118

[116] Pepik, B., Stark, M., Gehler, P. V., and Schiele, B. Occlusion patterns for object class detection. In *CVPR* (2013). pages 1, 93

[117] Perko, R., and Leonardis, A. Context awareness for object detection. In *Workshop of the Austrian Association for Pattern Recognition* (2007). pages 6

[118] Perko, R., and Leonardis, A. A framework for visual-context-aware object detection in still images. *CVIU* (2010). pages 7, 8, 9, 42, 47, 60, 62, 65, 68, 70, 72, 74, 92, 94, 110, 113, 114, 118

[119] Perko, R., Wojek, C., Schiele, B., and Leonardis, A. Probabilistic combination of visual context based attention and object detection. In *International Workshop on Attention in Cognitive Systems (WAPCV)* (2008). pages 6, 33

[120] Ponce, J., Lazebnik, S., Rothganger, F., and Schmid, C. Toward true 3d object recognition. In *In Reconnaisance de Formes et Intelligence Artificelle* (2004). pages 1

[121] Raafat, R. M., Chater, N., and Frith, C. Herding in humans. *Trends in Cognitive Sciences* (2009). pages 9

[122] Ramanan, D. Using segmentation to verify object hypotheses. In *CVPR* (2007). pages 92

[123] Richardson, M., and Domingos, P. Markov logic networks. *Machine Learning* (2006). pages 30, 123

[124] Rudemo, M. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* (1982). pages 24

[125] Russell, B. C., and Torralba, A. Building a database of 3d scenes from user annotations. In *CVPR* (2009). pages 2

[126] Sadeghi, M. A., and Farhadi, A. Recognition using visual phrases. In *CVPR* (2011). pages 7, 42, 60, 63

[127] Sain, S. R., and Scott, D. W. On locally adaptive density estimation. *Journal of the American Statistical Association* (1996). pages 26

[128] Salton, G., and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986. pages 37

[129] Sato, T., and Kameya, Y. Parameter learning of logic programs for symbolic-statistical modeling. *JAIR* (2001). pages 123

[130] Savarese, S., and Fei-Fei, L. 3d generic object categorization, localization and pose estimation. In *ICCV* (2007). pages 38, 39, 41, 50

[131] Scott, D., and George, R. T. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association* (1987). pages 24

[132] SCOTT, D. W. On optimal and data-based histograms. In *Biometrika* (1979). pages 24

[133] Selinger, A., and Nelson, R. C. A perceptual grouping hierarchy for appearance-based 3d object recognition. *CVIU* (1999). pages 1

[134] Sen, P., Namata, G., Bilgic, M., and Getoor, L. Collective classification. In *Encyclopedia of Machine Learning*. 2010. pages 9, 32, 40, 60, 64

[135] Sen, P., Namata, G. M., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. Collective classification in network data. *AI Magazine* (2008). pages 9, 32, 117, 120

[136] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *ECCV* (2012). pages 116

[137] Song, S., and Xiao, J. Sliding shapes for 3d object detection in depth images. In *ECCV* (2014). pages 1

[138] Song, Z., Chen, Q., Huang, Z., Hua, Y., and Yan, S. Contextualizing object detection and classification. In *CVPR* (2011). pages 62

[139] Spinello, L., and Arras, K. O. People detection in rgb-d data. In *IROS* (2011). pages 116

[140] Sun, M., Bao, S. Y.-Z., and Savarese, S. Object detection using geometrical context feedback. *IJCV* (2012). pages 6

[141] Taskar, B., Abbeel, P., and Koller, D. Discriminative probabilistic models for relational data. In *UAI* (2002). pages 30, 73, 117, 120, 122

[142] Teney, D., and Piater, J. Continuous pose estimation in 2d images at instance and category levels. In *CRV* (2013). pages 115, 122

[143] Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Schiel, B., and Van Gool, L. Towards multi-view object class detection. In *CVPR* (2006). pages 21

[144] Thon, I., Landwehr, N., and De Raedt, L. Stochastic relational processes: Efficient inference and applications. *Machine Learning* (2011). pages 32

[145] Torralba, A. Contextual priming for object detection. *IJCV* (2003). pages 6, 77

[146] Torralba, A., Castelhano, M. S., Oliva, A., and Henderson, J. M. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review* (2006). pages 77

[147] Torralba, A., and Efros, A. A. Unbiased look at dataset bias. In *CVPR* (2011). pages 120

[148] Torralba, A., Murphy, K. P., and Freeman, W. T. Contextual models for object detection using boosted random fields. In *NIPS* (2004). pages 7, 42

[149] Triandis, H. C., and Suh, E. M. Cultural influences on personality. *Annual Review of Psychology* (2002). pages 42

[150] Tukey, P., and Tukey, J. Data driven view selection, agglomeration, and sharpening. *Interpreting Multivariate Data* (1989). pages 26

[151] Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. Selective search for object recognition. *IJCV* (2013). pages 38, 92, 93, 95, 97, 103, 106, 107

[152] V, M. J. R., , K, O. J., and C, T.-B. Objects and inter-object relations in visual working memory. In *Perception 32 ECVP (Abstract Supplement)* (2003). pages 1

[153] Viola, P. A., and Jones, M. J. Rapid object detection using a boosted cascade of simple features. In *CVPR* (2001). pages 1, 2

[154] Wand, M., and Jones, M. Kernel smoothing, 1995. Chapman & Hall CRC. pages 24, 25, 49, 69

[155] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. Locality-constrained linear coding for image classification. In *CVPR* (2010). pages 16

[156] Wang, X., and Grimson, E. Spatial latent dirichlet allocation. In *NIPS* (2007). pages 7, 42

[157] Wang, X., and Sukthankar, G. Multi-label relational neighbor classification using social context features. In *KDD* (2013). pages 75

[158] Weber, M., Einhäuser, W., Welling, M., and Perona, P. Viewpoint-invariant learning and detection of human heads. In *FG* (2000). pages 21

[159] Wojek, C., Walk, S., Roth, S., Schindler, K., and Schiele, B. Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes. *TPAMI* (2013). pages 4, 41

[160] Xiang, Y., and Savarese, S. Object detection by 3d aspectlets and occlusion reasoning. In *3ddr workshop at ICCV* (2013). pages 2, 7, 11, 16, 80, 88, 90, 119, 121, 122

[161] Yan, P., Khan, S. M., and Shah, M. 3d model based object class detection in an arbitrary view. In *ICCV* (2007). pages 1

[162] Zhang, H., Geiger, A., and Urtasun, R. Understanding high-level semantics by modeling traffic patterns. In *ICCV* (2013). pages 1, 8, 94, 95

[163] Zhang, Y., Song, S., Tan, P., and Xiao, J. PanoContext: A whole-room 3D context model for panoramic scene understanding. In *ECCV* (2014). pages 2

[164] Zia, M. Z. High-resolution 3d layout from a single view. In *Ph.D. thesis, Swiss Federal Institute of Technology (ETH-Zurich)* (2014). pages 118

[165] Zia, M. Z., Stark, M., and Schindler, K. Are cars just 3d boxes? - jointly estimating the 3d shape of multiple objects. In *CVPR* (2014). pages 2, 6, 7, 11, 80, 85, 88, 90, 116, 118, 119, 121, 122

[166] Zia, M. Z., Stark, M., and Schindler, K. Towards scene understanding with detailed 3d object representations. In *IJCV* (2014). pages 11, 85, 119

[167] Zia, Z., Stark, M., Schiele, B., and Schindler, K. Detailed 3d representations for object recognition and modeling. *TPAMI* (2013). pages 122

[168] Zitnick, C., and Dollár, P. Edge boxes: Locating object proposals from edges. In *ECCV*. 2014. pages 38, 92, 95, 106

Curriculum

José Antonio Oramas Mogrovejo was born on February 10^{th} 1985, in Guayaquil, Ecuador. He received the degree of Computer Engineering with a major in Multimedia Systems from the Escuela Superior Politécnica del Litoral (ESPOL) in Ecuador. During his studies at ESPOL, he co-founded KOKOA, the free software community of ESPOL. He was a junior researcher at CTI-ESPOL where he worked on projects related to iTV, gestural interfaces and assistive technology under the supervision of Prof. dr. Xavier Ochoa and Prof. dr. Katherine Chiluiza. During his time at CTI-ESPOL, in 2008, he won a VLIR-ESPOL competitive project grant for the project "A Hand Gesture Interface for Ecuadorian Sign Language". In 2008, he was a visiting scholar at AVNET (Belgium) where he worked on technology related to media streaming and video learning objects based on free software. In April 2009, he visited the TELIN-IPI group from the UGent (Belgium) where he worked on algorithms for skin segmentation applied to gesture recognition. In 2010, he joined the PSI-VISICS computer vision lab at KU Leuven (Belgium) for pursuing his Ph.D. under the advise of Prof. dr. Tinne Tuytelaars and Prof. dr. Luc de Raedt. His research focuses on reasoning about contextual information to improve the performance of computer vision tasks.

List of publications

Journal Articles

- Antanas, L., van Otterlo, M., Oramas M., J., Tuytelaars, T., De Raedt, L. (2014). *There are plenty of places like home: Using relational representations in hierarchies for distance-based image understanding.* Neurocomputing, 123, 75-85.

Conference Articles

- Oramas M., J., Tuytelaars, T. *Recovering hard-to-find object instances by sampling context-based object proposals. Submitted to* IEEE International Conference on Computer Vision - (ICCV).
- Martinez-Camarena, M., Oramas M., J., Tuytelaars, T. *Towards sign language recognition based on body parts relations. Submitted to* IEEE International Conference on Image Processing - (ICIP).
- Fernando B., Gavves, E., Oramas M., J., Ghodrati, A., Tuytelaars, T. (2015). *Modeling video evolution for action recognition.* Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition - (CVPR). Boston, MA, USA, 7-10 June 2015 .
- Oramas M., J., Tuytelaars, T. (2014). *Scene-driven Cues for Viewpoint Classification of Elongated Object Classes.* Proceedings of the British Machine Vision Conference - (BMVC). Nottingham, UK, 1-5 September 2014 (pp. 1-11).

- Oramas M., J., De Raedt, L., Tuytelaars, T. (2014). *Towards cautious collective inference for object verification.* Proceedings of the IEEE Winter Conference on Applications of Computer Vision - (WACV). Steamboat Springs, CO, USA, 24-26 March 2014 (pp. 1-8).

- Billiet, L., Oramas M., J., Hoffmann, M., Meert, W., Antanas, L. (2013). *Rule-based hand posture recognition using qualitative finger configurations acquired with the Kinect.* Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods. International conference on pattern recognition applications and methods - (ICPRAM). Barcelona, Spain, 15-18 February 2013 (pp. 1-4).

- Oramas M., J., De Raedt, L., Tuytelaars, T. (2013). *Allocentric pose estimation.* Proceedings of the IEEE International Conference on Computer Vision - (ICCV). Sydney, Australia, 3-6 December 2013 (pp. 289-296).

- Antanas, L., van Otterlo, M., Oramas M., J., Tuytelaars, T., De Raedt, L. (2012). *A relational distance-based framework for hierarchical image understanding.* Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods - (ICPRAM). Algarve, Portugal, 6-8 Feb 2012 (art.nr. 150) (pp. 206 -218).

- Antanas, L., van Otterlo, M., Oramas M., J., Tuytelaars, T., De Raedt, L. (2011). *Not far away from home: A relational distance-based approach to understand images of houses.* In Lecture Notes in Computer Science: Vol. 6489. International Conference on Inductive Logic Programming - (ILP). Firenze, Italy, 27-30 June 2010 (pp. 22-29) Springer.

Extended Abstracts

- Oramas M., J., De Raedt, L., Tuytelaars, T. (2014). *Reasoning about object relations for object pose classification.* Proceedings of the Netherlands Conference on Computer Vision - (NCCV). Ermelo, Netherlands, 24-25 April 2014.

FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL ENGINEERING
PSI-VISICS
Kasteelpark Arenberg 10 box 2441
B-3001 Heverlee
Jose.Oramas@esat.kuleuven.be
http://homes.esat.kuleuven.be/∼joramas