Explore Objects and Categories in Unexplored Environments Based on Multimodal Data

Dissertation zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und
Naturwissenschaften,
Fachbereich Informatik
der Universität Hamburg
vorgelegt von

Jianhua Zhang

Hamburg, 2012

Tag der mündlichen Prüfung: 12. July 2012

Gutachter:

Prof. Dr. Peer Stelldinger
Prof. Dr. Jianwei Zhang

Abstract

In deterministic environments object detection and recognition are based on the assumption that object categories are known. However, in unexplored environments these assumptions cannot be fulfilled since there is not enough prior information about what kinds of objects and how many objects there are. Thus the execution of computer vision tasks in such environments requires the ability of detecting category-independent objects and discovering novel object categories.

In this thesis, a set of novel methods is presented to explore objects and categories in unexplored environments. The first step to achieve this is to detect objects, then to recognize objects belonging to known categories. If there are unknown objects, these object should be clustered as new categories, and be described and be related to known categories. Thus the proposed methods in this thesis can be separated into two parts that concern the problems of category-independent object detection and novel object category discovery, respectively. When humans explore an environment, 3D information is indispensable in addition to 2D information. Therefore, the presented methods are all based on multimodal data (i.e. the 2D images and 3D point clouds).

Concerning the first problem, most existing methods either can only detect one object per image or need to sample a large number of regions to cover multiple object instances. This thesis first proposes a set of novel category-independent object features that describe an object on a stand-alone instance regardless of its category. Based on these features, a cross-modal co-segmentation method is proposed to simultaneously segment paired images and 3D point clouds which are obtained by RGB+D cameras, and to detect and localize multiple category-independent object instances without sampling extra regions. A new discriminative model is designed, namely cross-modal higher-order Conditional Random Field model, which consists of unimodal and cross-modal terms. Unimodal terms include unary, pair, and higher order potentials, which are computed from the new category-independent features. Cross-modal terms add global constraints that keep the cross-modal spatial consistency in both 2D and 3D space. The category-independent object detection is treated as a labeling process with three kinds of labels (i.e. the object, the background and the boundary). Taking advantage of these labels, single object instances can be separated efficiently from a resulting labeled map. By comparison with state-of-the-art methods, experimental results on a public RGB+D dataset show that the proposed method yields a promising performance.

After localizing objects in an unexplored environment, a dynamic category hierarchy is proposed to improve object recognition and discover novel categories for the second problem. First, multimodal object attributes are extended from 2D ones to describe objects since they have excellent generalizability across categories, by which novel categories can also be depicted. Then a supervised hierarchical latent Dirichlet allocation model (shLDA) is presented to organize a large scale number of categories. A category hierarchy is an inherent structure in the human mind, and more importantly it can dynamically change. However, existing methods concern building static category hierarchies. In this thesis, a novel framework is presented to build such a dynamic hierarchy based on the multimodal attributes and the shLDA model. The framework can effectively recognize objects belonging to known categories and can detect and distinguish objects belonging to unknown categories. After discovering novel categories, the framework can integrate them into the hierarchy and construct a new one, thus forming a dynamic category hierarchy. Experiments first demonstrate the improvement of multimodal attributes with respect to 2D ones. Then they show the promising performance of object recognition and novel category discovery by comparing with state-of-the-art methods. Moreover, this novel framework can find the most representative object attributes to compactly describe objects.

Finally we draw some conclusions, and discuss limitations of the presented work and suggest the directions for future work.

Kurzfassung

In abgeschlossenen Umgebungen basieren Objektdetektion und Identifizierung oft auf der Annahme, dass Objektkategorien vorab bekannt sind. Allerdings kann diese Annahme in unbekannten Umgebungen nicht erfüllt werden, da die Art und Anzahl der Objekte unbekannt ist. Aus diesem Grund wäre für die maschinelle Bildverarbeitung in solchen Umgebungen die Fähigkeit wichtig, unabhängig von bereits bekannten Kategorien Objekte im Bild zu detektieren und neue Objektkategorien zu entdecken.

In dieser Arbeit wird eine Reihe neuer Methoden vorgestellt, um Objekte und Kategorien in unbekannten Umgebungen zu erforschen. Den ersten Schritt stellt hierbei die Detektion der Objekte dar. Es folgt die Klassifikation derjenigen Objekte, die zu den bekannten Kategorien gehören. Wenn unbekannte Objekte existieren, sollen für diese neue Kategorien entdeckt und mit den bereits bekannten Kategorien in Verbindung gebracht werden.

Somit lassen sich die in dieser Arbeit behandelten Methoden in zwei Klassen unterteilen, zum Einen mit dem Ziel der Kategorie-unabhängigen Objekterkennung und zum Anderen mit dem Ziel der Entdeckung neuartiger Objektkategorien. Wenn eine Umgebung erkundet wird, sind neben den 2D-Informationen die 3D-Informationen unverzichtbar. Daher basieren die vorgestellten Methoden auf multimodalten Daten (2D-Bildern und 3D-Punktwolken).

Im Hinblick auf die erste Problemstellung können die meisten bekannten Verfahren entweder nur ein Objekt pro Bild erkennen oder mehrere Objektinstanzen nur beim Erproben einer großen Anzahl von Regionen bestimmen. Diese Arbeit führt zuerst eine Reihe von neuen Kategorie-unabhängigen Objekteigenschaften ein, die ein Objekt unabhängig von dessen Kategorie als eine eigenständige Instanz beschreiben. Basierend auf diesen Merkmalen wird eine ,,intermodale“ Segmentierungs-Methode vorgestellt, um gleichzeitig Bilddaten und 3D-Punktwolken zu verarbeiten. Diese werden durch RGB+D-Kameras erzeugt. Somit können mehrere Kategorie-unabhängige Objekt-Instanzen ohne die Erprobung zusätzlicher Regionen erkannt und zu lokalisiert werden. Es wird ein neues Entscheidungs-Modell entwickelt, das ,,Cross-Modal Higher-Order Conditional Random Field Model“. Dieses verwendet sowohl ,,unimodale“ als auch ,,intermodale“ Merkmale. ,,Uni-modale“ Merkmale beschreiben Potentiale verschiedener Ordnung, die von den entwickelten Kategorie-unabhängigen Merkmalen berechnet werden. ,,Intermodale“ Merkmale definieren globale Bedingungen, um die Integrität der Daten im 2D-und 3D-Raum zu sichern. Die Kategorie-

unabhängige Objekterkennung wird als Klassifizierungsvorgang der Regionen in drei Klassen (Objekt, Hintergrund und Grenze) behandelt. Unter Ausnutzung dieser Kennzeichnung können einzelne Objektinstanzen effizient aus der resultierenden Karte isoliert werden. Ein Vergleich mit den gängigen Methoden für diese Problemstellung zeigt die Leistungsfähigkeit des entwickelten Verfahrens. Dieser Vergleich erfolgt unter Verwendung eines öffentlich zugänglichen RGB + D Datensatzes.

Im Hinblick auf die zweite Problemstellung wird nach der Lokalisierung von Objekten in einer unbekannten Umgebung eine dynamische Kategorie-Hierarchie zur Verbesserung Objekterkennung und zur Entdeckung neuer Kategorien eingeführt. Die 2D Merkmale werden zur Objektbeschreibung zu multimodalen Objektattributen erweitert, da diese eine sehr gute Generalisierbarkeit versprechen und somit auch neuartige Kategorien formuliert werden können. Die große Anzahl an Kategorien wird in einem „supervised hierarchical latent Dirichlet allocation model (shLDA)“ organisiert. Eine Kategorie-Hierarchie ist eine inhärente Struktur des menschlichen Gehirns, die sich dynamisch ändert. Allerdings implementieren die bisherigen Methoden den Aufbau von statischer Kategorie-Hierarchien. In dieser Arbeit wird ein neuartiges Framework vorgestellt, um eine dynamische Hierarchie basiert auf den multimodalen Attributen und dem shLDA Modell zu erzeugen. Das Framework kann die zu bekannten Kategorien gehörenden Objekte effektiv erkennen und kann auch die zu unbekannten Kategorien gehörenden Objekte erkennen und unterscheiden. Nach der Entdeckung neuer Kategorien kann das Framework diese in die bestehende Hierarchie integrieren und eine neue erzeugen, wodurch eine dynamische Kategorien-Hierarchie entsteht.

Experimente demonstrieren zuerst die Verbesserung der multimodalen Attribute gegenüber 2D-Merkmalen. Die Leistung der Objekterkennung und Entdeckung neuartiger Kategorien wird durch den Vergleich mit gängigen Methoden gezeigt. Darüber hinaus kann dieses neuartige Framwork die relevanten Objektattribute in einer kompakten Form beschreiben.

In einem Fazit werden die Einschränkungen der beschriebenen Verfahren diskutiert und es wird ein Ausblick auf mögliche zukünftige Forschungsrichtungen gegeben.

Acknowledgements

This dissertation would not have been possible without the help and support of my family, and many friends and colleagues.

Firstly, I would like to thank my supervisor, Prof. Dr. Jianwei Zhang. I have learnt a huge amount from Jianwei, his inspiration, insight and feedback have been invaluable. I am also very grateful to him for his tolerance in allowing me to pursue my own academic path.

I would also like to thank Prof. Dr. Shengyong Chen and Prof. Dr. Houxiang Zhang. I am grateful to their support and encouragement, as well as many great conversations, form which I obtained many inspiration to work on this dissertation.

I am also grateful to my colleagues from the TAMS group for providing professional support and personal advice in all the time of research for and writing of this dissertation. Specifically I would like to thank Lu, Norman, Bernd, Hannes and Dominik. I thank all my friends in Hamburg, Junhao, Hansong, Gang Cheng, Guoyuan and Bo Sun for their support.

I would like to thank my wife for her understanding and support, without which it is impossible for me to devote all my energies to research work. Finally, I thank to my parents who give me a lifetime of love and care.

This work is funded by the DFG German Research Foundation (grant 1247) International Research Training Group CINACS (Cross-modal Interactions in Natural and Artificial Cognitive Systems).

Dedicated to my wife, Jie Wan.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1. Motivation

This thesis presents novel methods for exploring objects and categories in unexplored environments based on multimodal data, which include object detection and category discovery. Human beings have an amazing ability to quickly find unfamiliar objects and relate them with similar object categories, by which people can explore our world. This ability is also important for artificial intelligent systems. For example, when robots work in an unexplored environment, it is not guaranteed that all objects in such an environment have been learned by the robots. Hence, robots need to learn new objects belonging to novel categories based on their knowledge of known object categories. As shown in Fig. 1.1, one can image such a scene where a robot is asked to deliver a cup of coffee from the kitchen to the meeting room. However, the robot has never been in this kitchen before and consequently does not know any objects in it. Actually, there is no cup but many mugs. The robot only has knowledge about the cup category, so it cannot find the object that it needs and therefore cannot continue its tasks. If the robot is equipped with an intelligent method which endows it with the ability of automatically learning novel categories from known knowledge, it can learn this new mug category from the knowledge about the cup category. Furthermore, if this intelligent method can further find the relationship between the mug category and the cup category, which have the same superordinates as container, the robot would know the mug can also be used to fill coffee in and then it can continue with the following steps.

However, to develop such an intelligent method is an extremely challenging task. From this example, we found several key points to execute such a complex task. First the robot should detect and localize all objects in which it is interested. Then it should describe these objects to recognize objects it knows and cognize objects it does not know. The new objects should be connected to the knowledge of the robot to discover their characteristics and functions for the following steps. Thus this complex task can be divided into three sub-problems.

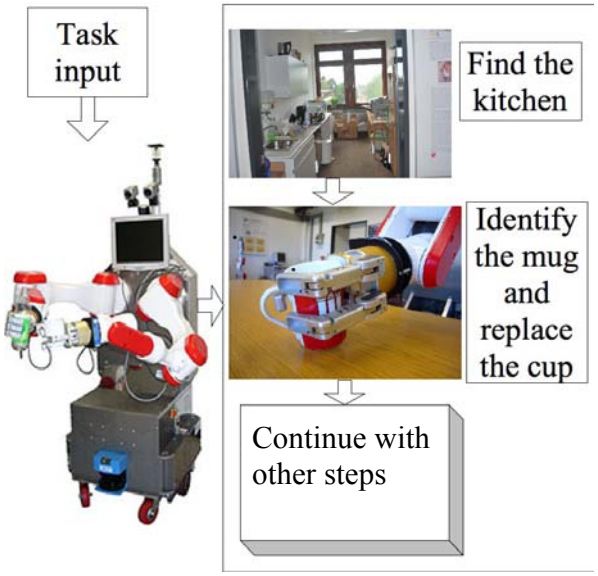


Figure 1.1.: An example of the necessity of learning to discover novel categories in unexplored environments.

In general, a complex environment contains more than one object which increases the complexity of object recognition. Furthermore, when these objects are placed messily, one object may be occluded by others, which also obstructs object recognition. Therefore, the first step of categorizing more than one object is usually to segment a scene into several parts. Each part only involves one object. However, this segmentation problem is also a challenging and open problem due to similar objects, clutter backgrounds and occlusion. Although many methods have been proposed to effectively segment images (e.g. Stein et al. (2008), Bagon et al. (2008)), the segmentation based on only images cannot achieve a promising performance for complicated scenes yet, which motivates this thesis to resort to multimodal data. Therefore in this thesis, the presented work is based on multimodal data (i.e. 2D images and 3D point cloud data). The 3D data can complement 2D data well. When one object is occluded by others and is difficult to segment it in an image, its 3D point cloud may be easily segmented. If two objects look similar, their positions in 3D are different and are easily distinguished.

Since the aim of this thesis is to explore novel categories, it is necessary to develop efficiently generalizable descriptions for objects, which can transfer the knowledge from known categories to those unknown categories. This is also a challenging problem. The object attribute is a good description which was originally proposed by Farhadi et al. (2009) and Lampert et al. (2009b) and is widely used in the literature. However, in current literature object attributes are based on images. For some attributes, extracting them from 3D space is more efficient; especially in case of shape attributes.

How to organize categories in an efficient manner is also a key problem for presenting the relationship of object categories. Besides describing novel categories using generalizable object attributes, determining category relationships is also useful for deriving object functions for further work.

The aforementioned three sub-problems simulate the human cognitive procedure. Let us recall the human cognitive procedure of discovering novel categories from a mass of objects. For example, when someone enters a strange kitchen to find something, he will focus on different objects and then recognize which known object he needs or find some new objects as substitutes. According to this procedure, it is natural to divide these problems into two sub-tasks, detecting and localizing category-independent objects, and distinguishing and categorizing novel objects. In this thesis, the presented work corresponds to these two sub-tasks. The first sub-task is to solve the first problem of segmentation, and the second sub-task is to solve the last two problems, generalized object description and category organization.

1.2. Related Work

In this section we consider the aforementioned problems as a whole and introduce the common way of how to solve them. This problem is a key issue for many computer vision tasks, so a lot of research has carried out on different methods trying to solve this problem. These methods have similar procedures, which is illustrated by the block diagram shown in Fig. 1.2.

At first, most methods separately present objects by unsupervised segmentation (e.g. Lee and Grauman (2012), Russell et al. (2006), Sivic et al. (2008)), or given bounded boxes (e.g. Fritz and Schiele (2008)), or saliency detection (e.g. Li et al. (2010e)). Subsequently, a set of features is extracted from these objects, followed by the next step of clustering similar objects into one category. Besides these common steps, some methods can refine the segmentation of objects by the discovered results (e.g. Lee and Grauman (2009), Russell et al. (2006)), and some methods also output corresponding classifiers for different object categories (e.g. Grauman and Darrell (2006)).

These methods can be categorized into different classes from different aspects. At the first step, most methods employed different technologies to present objects separately, while others omit this step by directly extracting features from the whole image (e.g. Kim et al. (2008), Liu and Chen (2007b)). By presenting objects separately, several advantages can achieved. The extracted features can be confined to single objects, avoiding the influence of other objects or background, which can improve the performance of clustering results. As shown in Lee and Grauman (2012), segmenting an image into different parts, where each part contains one object, can model the interaction between different parts. Through this, the relationship between known categories and unknown categories can be obtained, which forms an object graph leading to practical improvements of category discovery. In Li et al. (2010e), the foreground objects to be discovered are detected by saliency, which

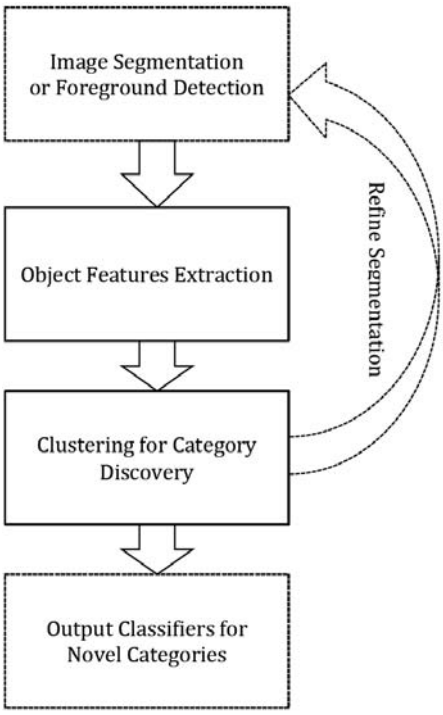


Figure 1.2.: A classical block diagram of novel category discovery.

causes the method to focus on those parts with higher objectiveness. In Bodesheim (2011), the author employed a generic object detector to find regions of interest to avoid the influence of background. However, the results of these methods rely on the performance of segmentation or saliency detection. The inaccurate segments and salient regions may cause some outliers at the next clustering step, which may reduce the accuracy of novel category discovery. On the other hand, those methods that directly extract different features, such as SIFT from a whole image (e.g. Liu and Chen (2007b), Lou et al. (2010)), HoG (e.g. Winn and Jojic (2005)), match features across images and subsequently filter out the parts with lower correspondence of features. Then the clustering is based on those well-matched features to discover object categories. These methods rely entirely on features matching, therefore the mismatched features may result in failed object discovery.

Different methods also extract different features to describe objects. Many excellent features have been proposed for describing visual appearance, such as SIFT (Lowe (2004)), MSER (Matas et al. (2004)) and HoG (Dalal and Triggs (2005)). Many methods employ one or more kinds of features to represent objects. For example, Liu and Chen (2007b), Dueck and Frey (2007) and Grauman and Darrell (2006) only utilize SIFT features, while Cho et al. (2010) consider MSER and SIFT simultaneously. Although these methods achieve some good results, the mismatched features still decrease the accuracy of object category discovery. Recently, a new 'bag of feature' has been proposed in Fei-Fei and Perona (2005). The bag of feature

constructs a set of visual words which forms a vocabulary, then an image or object is composited by a subset of this vocabulary. Since the bag of feature can integrate different kinds of features and be represented in a concise manner, many methods employ it to describe objects, such as (Pineda et al. (2010), Lee and Grauman (2012), Sivic et al. (2008), Kim et al. (2008), Lou et al. (2010), Russell et al. (2006), Fritz and Schiele (2008)). However, the bag of feature ignores the spatial relationship among features, which is an important aspect for the presence of objects in an image. Therefore in Lee and Grauman (2012) the proposed method also takes into account the spatial relationship among objects. The aforementioned features are all non-semantic, which cannot endow the discovered object categories with specific semantic meanings. Therefore, more recently some methods have been using object semantic attributes to describe objects. For example, in Rohrbach et al. (2010) the authors employed the algorithm proposed in Lampert et al. (2009b) to generate object semantic attributes and then made use of linguistic knowledge to provide the semantic links between known and unknown object categories for discovering novel object categories.

Different clustering techniques are employed by different methods in the clustering step. The two most popular clustering techniques for unsupervised object discovery are the latent variables methods (Russell et al. (2006), Liu and Chen (2007b), Fritz and Schiele (2008), Li et al. (2010e), Bodesheim (2011)) and spectral clustering schemes (Lee and Grauman (2012), Grauman and Darrell (2006), Lee and Grauman (2009), Kim et al. (2008), Pineda et al. (2010), Triebel et al. (2010)). The latent variables methods are also referred as discrete independent component analysis, including Latent Dirichlet Allocation (Blei et al. (2003)), Probabilistic Latent Semantic Analysis (Hofmann (1999)), and simple Gaussian Mixture Models (Reynolds (2008)). A good introduction to this topic can be found in Buntine and Jakulin (2006). The spectral clustering is a family of techniques relying on the eigen-decomposition of a modified similarity matrix, which can be roughly divided into two categories, the Laplacian eigenmap (Belkin and Niyogi (2003)) and Kernel Principle Components Aanlysis (Bengio et al. (2004)). In Tuytelaars et al. (2010) the authors gave a comprehensive review of unsupervised object discovery based on these two major clustering techniques. Besides them, other clustering techniques are also used. In Dueck and Frey (2007), the author employed the affinity propagation clustering (Frey and Dueck (2007)) to find a better exemplars to represent object categories. The information bottleneck (Slonim and Tishby (1999)) and locality-sensitive hashing (Cohen et al. (2001)) methods are also used for clustering novel object categories in Lou et al. (2010) and Pineda et al. (2010) respectively. The most relevant clustering method used in this thesis was also used in Sivic et al. (2008) where the authors employed a hierarchical model to cluster and organize object categories which can yield relationships among categories.

After clustering and discovering novel object categories, some methods explicitly provide a classifier for each category. For example, in Grauman and Darrell (2006), the algorithm outputs a set of classifiers trained from the clustered objects to recog-

nize the novel categories from images. Some methods provide feedback for refining object segmentation. For example, in Russell et al. (2006), the segments are sorted by outputs of the topic model algorithm and the better object segmentation can be obtained by choosing the segments with high scores from the pool of segments obtained by multiple segmentations.

Most of these methods are based on 2D images. However, in 3D scene unsupervised object category discovery is also important for many tasks, such as object mapping (e.g. Herbst et al. (2011)) and robot navigation (e.g. Modayil and Kuipers (2004)). To discover novel object categories, it is necessary to provide multiple object instances in different space or time. Most of the aforementioned methods use different images containing the same object category to discover this category. But the temporal difference is also useful for discovering novel categories, especially in dynamic environments (e.g. Modayil and Kuipers (2004)) or in a visual tracking system (e.g. Liu and Chen (2007a)).

The methods mentioned here are related to the method proposed in this thesis, however, there are significant differences between them. First, the proposed method employs a more elaborate algorithm to efficiently and accurately segment objects from the background. Note that the segmented objects are not all foreground objects but those objects to be discovered, which are interesting to the whole method and users. By this, the object descriptions are more precise and will yield more accurate results for novel category discovery. Second, the proposed method employs both non-semantic features and semantic object attributes to describe objects, which provide not only accurate descriptions but also semantic meanings for those discovered novel categories, which is extremely helpful for the next reasoning tasks. Last but not least, the presented work also constructs an efficient hierarchical category structure which is similar to human cognition for learning novel categories and improves the performance of category discovery.

1.3. Category-independent Objects Detection

Similarly to most methods mentioned above, the presented work in this thesis also detects objects before discovering novel object categories. In a complex environment, objects are various and background is cluttered. Thus directly extracting features to describe each object is inevitably influenced by background due to similar color, shape and texture. Although many excellent features can be extracted to match objects across images, there are still some mismatched results lowering the performance of object discovery. By segmenting objects to different parts from an image, the extracted features avoid being affected by other objects or background. However, the performance of object category discovery relies heavily on the accuracy of segmentation. Therefore, an efficient and precise segmentation method is necessary. Moreover, if the segments used by category discovery methods are distinct from the background segments, the algorithm of novel category discovery can

focus further on objects themselves and will have a better performance. Thus the presented method in this thesis will detect the generic objects first.

Segmentation cannot be done based on the assumption that object categories are known, since the aim of object discovery is to find novel object categories in unexplored environments. In such an environment, there must be some objects that belong to unknown categories, except objects belonging to known categories. Thus a category-independent object detector is really needed. On the one hand, although there are many excellent category-specific detectors that achieve promising results (e.g. Lampert et al. (2009a), Lehmann et al. (2009), Torralba et al. (2007)), they are not suitable for the problem in this thesis. Because these methods need to train a specific detector for each category, they cannot be applied to those unknown objects. Consequently, when there are objects of unknown categories, these methods cannot correctly detect and localize them. On the other hand, the unsupervised segmentation methods are also not suitable for category-independent object detection. Although a lot of unsupervised segmentation methods are widely used in the computer vision community (e.g. Shi and Malik (2000), Felzenszwalb and Huttenlocher (2004)), they cannot guarantee that each segment contains an entire object because an object may be broken by oversegments.

In a complex environment, the influence comes not only from the cluttered background, but also from other foreground objects since there may be too many objects. To obtain good object segments for object category discovery, the category-independent detector should only detect objects of interest. Recently, a set of category-independent object detectors have been proposed based on various methods, such as object ranking (e.g. Alexe et al. (2010), Rahtu et al. (2011)), saliency detection (e.g. Feng et al. (2011), Liu et al. (2011b)), and structure segmentation(e.g. Endres and Hoiem (2010), Collet et al. (2011)). However, due to the limitations that object ranking methods cannot achieve sufficient accuracy, saliency detection can only detect a single object per image, and the results of structure segmentation heavily relys on initial segmentation, they need to be improved significantly before applying them to novel category discovery.

In this thesis, the proposed object detection aims at detecting and localizing generic objects which are interesting for novel object category discovery. The proposed method uses the multimodal data (i.e. 2D images and 3D point cloud data) to enhance the performance of object detection and localization. The basic idea is illustrated in Fig. 1.3.

This method is a kind of novel conditional random fields, namely Cross-Modal Higher-order Conditional Random Field (CMH-CRF) model, which simultaneously utilizes 2D and 3D oversegments as basic nodes for decreasing computational costs. A set of new multimodal category-independent features are developed first. Then, the CMH-CRF model is designed to integrate the 2D and 3D potentials and the cross-modal potentials into one uniform model. After inferencing the CMH-CRF model, the 2D and 3D labeling results are simultaneously obtained. The pixel-wise results can then be produced by combining 2D and 3D results at pixel level. The

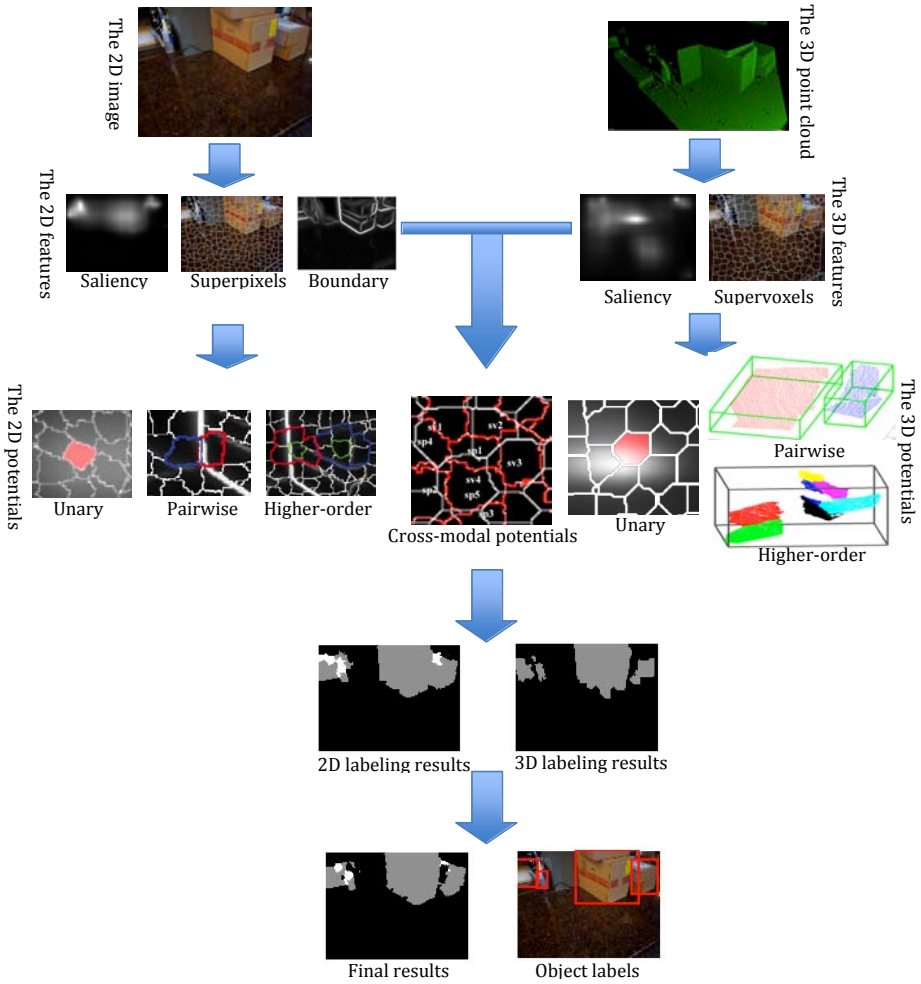


Figure 1.3.: Overview of the proposed generic object detection method. From the first row to the third row, there are the original image and point cloud, their features and their potentials used in the proposed model. In the fourth row, 2D and 3D labeling results are shown. By combining the 2D and 3D results at the pixel level, the final results are obtained and shown in the fifth row.

final results that each region only contains a single object instance can be achieved by a simple algorithm.

1.4. Novel Categories Detection and Clustering

After detecting and localizing generic objects, the task of novel category discovery can be executed based on these segmented regions. To discover novel object categories means no prior knowledge about the novel categories, but the knowledge about known categories can be taken into account. Like with human beings, novel knowledge is always derived from experience. Naturally, humans always use attributes to describe objects. For example, an apple may be portrayed by a set of attributes, such as red color, ball shape and sweet taste. Furthermore, humans will construct a dynamic structure for all known categories to facilitate memorizing and utilizing objects. For instance, a vase on a dining table is always looked at as a container for arranging flowers, while it will often be regarded as an artwork when people see it in a museum. Thus, given single object instances, there are three key issues to the problem of category discovery, which are (a) object description, (b) learning method and (c) organization of categories.

For the first issue, a generalizable description should be chosen since it must be generalized from known categories to unknown ones. In this thesis, object attributes are chosen because of their excellent generalizability and natural consistency to human cognition. Based on the algorithm proposed in Farhadi et al. (2009), a new set of attributes are developed. From the semantic aspect, attributes can be divided into two classes, the semantic attributes and the non-semantic attributes.Through the semantic attributes, the novel categories can obtain a certain number of semantic tags which can help us to deduce the new categories' functions. Besides semantic attributes, to accurately describe an object requires non-semantic attributes as well. On the one hand, because of the limited knowledge one the object's semantic attributes and the limitation of the ability of the human language to accurately describe an object by compact statements, the difference between objects cannot be described only by semantic attributes, at least not by simple semantic attributes. On the other hand, because of the limitation of computer vision, it is not easy to obtain an attribute classifier to represent complex semantic attributes.

From the modal aspect, attributes can also be divided into two classes, the unimodal attributes and the multimodal attributes. In current literature, the existing attributes are all trained from unimodal data, such as color and texture, which are referred to as unimodal attributes in this thesis. However, some attributes cannot be trained from only one modality. For example, training an attribute 'cup handle' needs texture and 3D shape features. Therefore, opposite unimodal attributes, in this work the new multimodal attributes are first proposed, which are trained from multimodal data.

For the second and third issue, a novel supervised hierarchical topic model is developed to learn to discover novel categories and organize them in a dynamic

hierarchy. The topic models (e.g. Hofmann (2001), Blei et al. (2003)) were originally proposed for document analysis, and then they were widely extended to the field of computer vision for unsupervised object discovery. The supervised hierarchical topic model was developed to integrate the advantages of the supervised topic model (Hannah et al. (2011)) and the hierarchical topic model (Blei et al. (2010)). By the supervised method a more precise relationship among the known categories can be built than that built by the general unsupervised topic model. By the hierarchical method, a concise structure can be built to keep consistent with natural human cognition of learning novel categories.

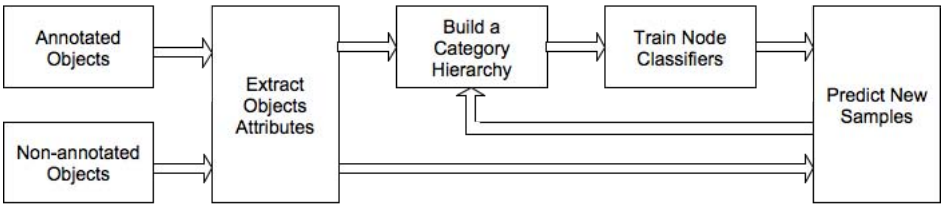


Figure 1.4.: Block diagram of the proposed method for discovering novel object categories from dynamic category hierarchies.

A block diagram of the proposed novel category discovery method is illustrated in Fig. 1.4. Attributes of annotated objects in the training dataset are computed based on the extracted base features. Given these attributes, a hierarchy representing the relationship among current categories can be built by using a supervised hierarchical topic model. For each node in the built hierarchy a classifier can then be trained based on those object samples assigned to this node. For new object samples, their attributes are also computed, then node classifiers are used to determine if they belong to known concrete categories or unseen categories. If new object samples are corresponding to unseen categories, the proposed method can indicate their superordinates and the hierarchy will branch off at appropriate nodes to generate new paths to represent new categories. Therefore the proposed methods feed predictive results back to the built hierarchy, by which the hierarchy can change dynamically.

1.5. Contributions

The main contribution of the presented work is a set of novel methods towards unsupervised novel category discovery in unexplored environments based on multimodal data. First a set of multimodal category-independent object features is developed for the novel multimodal co-segmentation method, which is extended from the robust higher order conditional random field (Kohli et al. (2009)). Then the object attributes are extended to more comprehensive version by integrating multimodal data. A novel supervised hierarchical latent Dirichlet allocation is developed to dis-

cover novel categories and construct a hierarchical topological structure for category relationship. The particular research contributions can be summarized as follows.

- A set of novel category-independent object features is developed based on multimodal data. The methods computing 2D saliency and oversegments are extended to be executed on 3D data, which yield effective 3D saliency and oversegments to complement 2D ones. The 2D and 3D unary features are implemented on 2D and 3D saliency and oversegments. The novel 2D pairwise and clique features are developed based on 2D oversegments and 2D object boundaries computed by global probability of boundary (Maire et al. (2008)). Since there is no efficient algorithm to obtain 3D boundary, 3D pairwise and clique features are developed by using the statistics of 3D point cloud among adjacent 3D oversegments.

- A novel multimodal co-segmentation method, called Cross-Modal Higher order Conditional Random Field (CMH-CRF) model is developed to efficiently detect and localize category-independent objects. To the best of our knowledge, this is the first time that co-segmentation is extended to multimodality. It is also the first time that higher order CRF model is extended to cross-modality. By integrating novel cross-modal potentials, the proposed method can simultaneously label 2D and 3D oversegments. Different from general figure/background segmentation methods, we use three kinds of labels to make the object instances easily distinguishable from the final labeling results.

- We extend object attributes to describe objects more comprehensively. First the new intra-class non-semantic object attributes is developed to improve the original object attributes to describe objects more accurately. Furthermore, since some object attributes are multimodal, which cannot be computed from only one modality, the new multimodal object attributes are implemented based on the original object attribute algorithm.

- A new supervised hierarchical topic model is developed to efficiently build category hierarchies from known categories. The proposed topic model has the supervised topic model's and hierarchical topic model's advantages. The more accurate clusters can be obtained in the supervised manner and the more compact structure of clusters can be obtained in the hierarchical manner. Furthermore, this new topic model organizes categories in a manner similarly to the organization of categories in the human mind.

- A new framework is implemented to precisely build dynamic category hierarchies and efficiently discover novel categories. The new category will be identified by the classifiers associated with each node in the built category hierarchy, and the new node will be branched off in this hierarchy to form a dynamic category hierarchy. When novel objects belonging to more than one new category are identified, the new node will automatically divide into several nodes to represent all new categories.

1.6. Thesis Outline

This thesis is organized in nine chapters. In the following chapter, the theoretical foundation of the graphical model is introduced, which is strongly related to the proposed novel methods. The basic concepts of probabilistic graphical model will be explained first, including Bayesian networks and Markov networks. Next, we introduce two of the most important models in this area, the conditional random fields and topic models. Two important extensions of these are presented. Finally, the relevant state-of-the-art methods are also reviewed here.

The state-of-the-art on the problems in this thesis will be presented in Chapter 3. We state them through two parts. In the first part, we introduce methods of category-independent object detection, by dividing them into four categories. The category-independent object features used in these methods are also briefly introduced. We analyze the characteristics of these methods and point out their limitations. Next the co-segmentation frameworks are presented which are also related to the topic in this part. In the second part, the object attributes and related work are introduced, followed by hierarchical object model and novelty detection.

In Chapter 4, how to implement the category-independent object features is described. We first extend 2D saliency and oversegments to 3D version. Then how to compute unary features, pairwise features, higher order features and cross-modal features that are used in Chapter 5 will be presented. Finally, extensive experiments are carried out. In these experiments, parameters for 2D and 3D saliency and oversegments are first tested, from which the best parameters are chosen to compute category-independent features. Then the distinctiveness of unary and pairwise features are also evaluated in experiments.

The CMH-CRF model is presented in Chapter 5. This model includes not only the unary and pairwise potentials, but also higher-order and cross-modal potentials. All of these potentials are defined on 2D and 3D modalities, respectively, which will be introduced in turn. Finally, extensive experiments will be carried out on a public RGB+D (which is referred to as red, green, blue and depth respectively) dataset to show state-of-the-art performance of the proposed model.

The extension of object attributes will be introduced in Chapter 6. We first introduce how to learn semantic attributes. Then the non-semantic attributes are extended by adding the intra-class ones. Next the algorithm of learning multimodal attributes is presented. Extensive experiments are executed and show the improvements achieved by our extensions.

In Chapter 7, a supervised hierarchical latent Dirichlet allocation is proposed. We give its graphical model, generative process, and probabilistic inference. Why this model can represent relationships among categories well is also discussed. Finally, the proposed model is evaluated on several public datasets, including well-known image datasets and one public RGB+D dataset. The experiments show the satisfactory performance of the proposed model.

The method for novel category discovery given single object instances is involved in Chapter 8. First the proposed framework is introduced. How to train the classifiers for each node in the built category hierarchy by the method in Chapter 7 is described subsequently. Then the algorithm of identifying novel category objects and distinguishing novel categories is presented. Finally, we execute extensive experiments on several public datasets and show the promising performance of the proposed method.

In the final chapter 9, we conclude this thesis with a brief summary. It summarizes the main achievements of this thesis, discusses limitations of the presented work and suggests the directions for future work.

Chapter 2

Probabilistic Graphical Models

Probabilistic models have a dominant position in the field of modern artificial intelligence. As a diagrammatic representation of probabilistic models, the probabilistic graphical models have several advantages (Bishop (2006)): 1) visualizing the structure of a probabilistic model in a simple way, 2) simplifying the analysis of conditional independence of a probabilistic model, and 3) simplifying the inference and learning of a complicated probabilistic model. In this chapter, we will first introduce the general concepts of probabilistic graphical models, then two concrete models which are extensively used in this thesis.

2.1. Probabilistic Graphical Model

A probabilistic graphical model can be denoted by a graph $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$, where $\mathcal{V}$ is the set of vertices (also called nodes) connected by edges in $\mathcal{E}$. Each node in $\mathcal{V}$ corresponds to a random variable (or group of random variables) and each edge in $\mathcal{E}$ denotes the probabilistic relationship between connected random variables. An example of a graphical model is shown in Fig. 2.1. The left one is called undirected graphical model where the edges have no directions and only express the soft constraints between variables. The middle one is called directed graphical model (also called Bayesian networks) where the edges carry arrows and denote the causality between connected variables. Directed graphs and undirected graphs can be converted to the factor graphs (shown in Fig. 2.1(c)), which are useful for solving inference problems.

2.1.1. Bayesian Networks

The Bayesian networks (BNs) are among of the most popular probabilistic graphical models and were originally proposed by Pearl (1985). Let $\{\mathbf{X}_v : v \in \mathcal{V}\}$ denote the set of variables associated with all nodes in the graph. For each node $v \in \mathcal{V}$, there is a set of parent nodes, denoted by $pa(v)$. Using $\mathbf{U}_v$ to represent a vector consisting

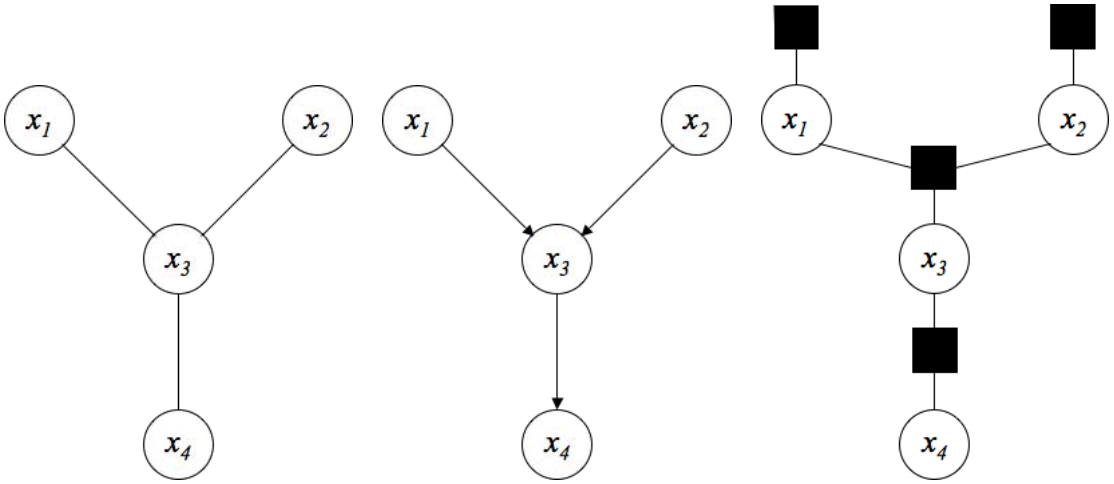


Figure 2.1.: An example of graphical models. From left to right, they are a undirected graph, a directed graph and a factor graph.

of a set of variables corresponding to $pa(v)$, if the conditional probability $p(\mathbf{X}_v|\mathbf{U}_v)$ exists, the joint probability distribution of the probabilistic model can be expressed as:

$$p(\mathbf{X}_v) = \prod_{v \in \mathcal{V}} p(\mathbf{X}_v|\mathbf{U}_v) \tag{2.1}$$

where $p(\mathbf{X}_v|\mathbf{U}_v)$ is also called the local conditional probability of the node v. If the node v does not have parents (i.e. $pa(v) = \varnothing$), then $p(\mathbf{X}_v|\mathbf{U}_v) = p(\mathbf{X}_v)$. Thus the BNs can efficiently decompose the probability $p(\mathbf{X}_v)$. For example, the directed graph of Fig. 2.1(b) describes the following conditional distribution:

$$p(\mathbf{X}) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3) \tag{2.2}$$

where the right side is the product of a set of local conditional probabilities. The rules of conditional independence can be expressed by "D-separation" (Jordan and Weiss (2002), Gallager (1962)).

Inference

The most common function of BNs is for the inference of uncertainty, which includes two aspects. If all distributions of parent nodes are determined, the distribution of a child node could be computed from local conditional distributions. If all distributions of child nodes are known, the distributions of parent nodes can be yielded by the Bayesian rules (Hartigan (1983)) which computes the posterior via local conditional distributions. The first inference is called causal reasoning, and

the second is called diagnostic reasoning. Several methods can serve as exact inference for small BNs, such as variable-elimination (Zhang and Poole (1996), Cozman (2000)), elimination-tree (Dechter (1999)) and junction tree (Murphy and Paskin (2002)). However, in most computer vision problems, there are thousands of nodes in graphical models; the exact inference is intractable since it is NP hard. Hence it is necessary to employ approximate inference methods, such as loopy belief propagation (Murphy et al. (1999), Ihler et al. (2006)), variational methods (Jaakkola and Jordan (2000), Bishop (2006)) and Markov chain Monte Carlo sampling methods (MCMC, Fruehwirth-Schnatter (2001)).

Here we only briefly introduce the Gibbs sampling, an important MCMC method (Bishop (2006)) which will be used in this thesis. For other approximate inference methods, readers can refer to corresponding literature. The MCMC method is a family of methods iteratively drawing samples from an intractable target distribution $p(x)$. Given an initial global configuration $x^{(0)} \in \mathbf{X}$, we can obtain subsequent states by a first-order Markov process:

$$x^{(t)} \sim q(x|x^{(t-1)}) \qquad t = 1, 2, \cdots . \tag{2.3}$$

where $q(x|x^{(t-1)})$ is transition distribution which describes the probabilities of transforming the state $x^{(t-1)}$ to current state x. After sufficient iterations, the state will be approximately distributed as $p(x)$.

As a special case of MCMC, the Gibbs sampling originally proposed by Geman and Geman (1984) is well suited to state spaces with internal structure. Consider a multivariate random variable $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ and a distribution over this $p(\mathbf{x}) = p(x_1, x_2, \cdots, x_N)$. Assuming that some initial states for the Markov chain are chosen, each step of the Gibbs sampling is to replace the value of one variable by a value drawn from the distribution of this variable conditioned on the given values of other variables. Formally, the algorithm of Gibbs sampling can be listed as (Bishop (2006)):

Algorithm 2.1: Gibbs sampling

Initialize $\{x_i : i = 1, \cdots, N\}$;
for $(\tau = 1, \cdots, T)$ **do**
  Sample $x_1^{\tau+1} \sim p(x_1|x_2^{\tau}, x_3^{\tau}, \cdots, x_N^{\tau})$;
  Sample $x_2^{\tau+1} \sim p(x_2|x_1^{\tau}, x_3^{\tau}, \cdots, x_N^{\tau})$;
  $\vdots$
  Sample $x_j^{\tau+1} \sim p(x_j|x_1^{\tau}, \cdots, x_j^{\tau}, \cdots, x_N^{\tau})$;
  $\vdots$
  Sample $x_N^{\tau+1} \sim p(x_N|x_1^{\tau}, x_2^{\tau}, \cdots, x_{N-1}^{\tau})$;

Learning

The learning of BNs involves two aspects: learning parameters and learning network structures. Here we are only concerned with the first one. When the network structure is known and the data is complete, the maximum likelihood estimate (MLE) can learn parameters in BNs. If the data is incomplete or the network involves hidden nodes, the gradient ascent algorithm (Binder et al. (1997)) can be employed for learning parameters. The expectation maximization (EM, Dempster et al. (1977)) is also an important algorithm to learn model parameters given incomplete data.

2.1.2. Markov Networks

The Markov Networks(also called Markov Random Fields (MRFs)) are an undirected graphical model which can also be denoted by $\mathcal{G}(\mathcal{V}, \mathcal{E})$. The joint distribution of MRFs can be formulated as:

$$P(\mathbf{X}_v) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c) \tag{2.4}$$

where c is a clique which denotes a set of nodes that are connected to each other and are conditionally dependent on other cliques, $\psi_c(x_c)$ denotes a non-negative potential function, Z is a normalizing constant known as the partition function, and $\mathcal{C}$ is the set of all cliques.

For example, the joint distribution of a Markov network as shown in Fig. 2.1(a) can be written as:

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi(x_1, x_3) \psi(x_2, x_3) \psi(x_3, x_4) \tag{2.5}$$

where the right side are the products of all potentials. For a different number of variables in a clique, there are different MRFs. For example, if the number of variables in every clique is not more than 2, an MRF model is called pairwise MRF, which can be written as:

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{x_i \in \mathcal{V}} \psi_i(x_i) \prod_{(x_i, x_j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \tag{2.6}$$

where $\psi_i(x_i)$ denotes the unary potential, ψ_{ij} is the pairwise potential and (x_i, x_j) denotes an edge in the edge set $\mathcal{E}$.

Inference

The basic principle of the inference of MRFs is the same as that of BNs. Actually, BNs and MRFs can be represented by the same graph via moralization. Therefore the methods of inference mentioned in section 2.1.1 are derived from MRFs. Two algorithms, loopy brief propagation (LPB, Murphy et al. (1999), Ihler et al. (2006))

and graph-cut (GC, Kohli and Torr (2007)) are most popularly used for the inference of MRFs.

Taking eq. 2.6 as example, we assume that $m_{ij}(x_j)$ is a "message" passed from node j to neighbor node i. Given the initial messages, the LBP algorithm propagates and updates the message $m_{ij}(x_j)$ iteratively till them converge. Two approaches can be used for message updating, the Sum-product:

$$m_{ij}(x_j) \leftarrow \sum_{x_i} \psi_i(x_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}_i \backslash j} m_{ki}(x_i) \tag{2.7}$$

and the Max-product:

$$m_{ij}(x_j) \leftarrow \max_{x_i} \psi_i(x_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}_i \backslash j} m_{ki}(x_i) \tag{2.8}$$

where $\mathcal{N}_i \backslash j$ denotes all nodes in the neighboring system of node i except node j. When all messages are converge, the belief $b(x_i)$ of node i is defined as:

$$b(x_i) \propto \psi_i(x_i) \prod_{j \in \mathcal{N}_i} m_{ji}(x_i) \tag{2.9}$$

The results that maximizes the belief $b(x_i)$ are used as the final ones.

The GC algorithm (Kumar and Hebert (2003b),Arora et al. (2007)) can be used to exactly infer MRF models if the energy function of MRFs is submodular. Otherwise, the MRF model can be approximately inferred by using two approximate GC approaches. The first approach is to use approximate submodular functions instead of the distribution of MRFs and then infers them by the GC algorithm (e.g. Kumar and Hebert (2003a)). Another approach is to employ some extended GC algorithms for approximate inference (e.g. Boykov et al. (2001)).

Learning

Due to the normalizing constant Z, the learning of MRFs is much more difficult than that of BNs. In BNs, all local factors take the probabilistic forms where the normalizing constant is unnecessary for computing the joint distribution. In other words, the normalizing constant of BNs is always fixed to 1. However, the local factor of MRFs does not possess the probabilistic form, and it should be converted to this form by dividing the normalizing constant after computing the joint distribution or marginal distribution. When learning the parameters of MRFs, the normalizing constant is related to each parameter. That is to say, changing any parameter will influence the normalizing constant. We cannot directly use MLE or the gradient ascent algorithm to learn these parameters. Thus, in general the exact learning of parameters for MRFs is intractable. Currently, only little research work involves the approximate learning of MRFs. In Welling and Hinton (2002), the authors proposed

the contrastive divergence algorithm based on the Mean Field to learn parameters for MRFs. A Bayesian learning algorithm based on MCMC has been discussed for learning parameters of the undirected graphical model in Murray and Ghahramani (2004).

2.2. Conditional Random Field Model

In this section we first introduce the definition of the general Conditional Random Fields (CRF) model via an image labeling task. Subsequently an extension of the CRF model, the Robust Higher-order CRF model and its inference are presented since the first sub-task in this thesis is solved based on it. Finally we will briefly review other extensions of the CRF model.

2.2.1. General CRF model

Generally, a CRF model is defined on unary and pairwise cliques. Consider a task of labeling an image. Let $\mathcal{V}$ denote the set of all nodes where each node corresponds to one pixel in an image. Given two sets of variables, $\mathbf{x} = \{x_i, i \in \mathcal{V}\}$ representing the labels of all pixels to be assigned, and $\mathbf{y} = \{y_i, i \in \mathcal{V}\}$ denoting the observed features of all pixels. The label set is denoted by $\mathcal{L} = \{l_1, l_2, \cdots, l_k\}$. The labeling task is to assign a label to each random variable x_i, and the configuration of labels is denoted by $\mathbf{x}$ which takes values from the set $\mathbf{L} = \mathcal{L}^N$. The most common way of labeling an image is to compute the maximum a posteriori, i.e. $\mathbf{x}^* = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})$. By applying the Bayesian rule, it can be written as:

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}, \mathbf{y}) \propto p(\mathbf{x}) p(\mathbf{y}|\mathbf{x}) \tag{2.10}$$

A model defined on the posteriori presenting by the term in the left side of this equation is called discriminative model. If the posteriori is directly modeled by Gibbs distribution, $(\mathbf{x}, \mathbf{y})$ is named Conditional Random Field (CRFs) and the model is thereby called CRF model. The CRF model is first proposed in Lafferty et al. (2001). Formally, the CRF model is defined as:

Definition 2.1. For random fields $\mathbf{x}$ and $\mathbf{y}$, if $\mathbf{x}$ obeys the Markov property when conditioned on $\mathbf{y}$

$$p(x_i|\mathbf{y}, x_j, j \neq i, j \in \mathcal{V}) = p(x_i|\mathbf{y}, x_j, j \in \mathcal{N}_i), \tag{2.11}$$

$(\mathbf{x}, \mathbf{y})$ is a Conditional Random Field.

By the Hammersley-Clifford theorem (Li (2009)), the posteriori of this CRF obeys the Gibbs distribution:

$$p(\mathbf{x}|\mathbf{y}, \theta) = \frac{1}{Z(\mathbf{y}, \theta)} \exp\left(-\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c, \mathbf{y}, \theta)\right) \tag{2.12}$$

where $Z(\mathbf{y},\theta) = \sum_{\mathbf{x}} \exp\left(-\sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c,\mathbf{y},\theta)\right)$ is the normalizing constant, ψ_c is the potential function with parameter θ defined on the clique c and $\mathcal{C}$ is the set of all cliques. The corresponding Gibbs energy is defined as:

$$E(\mathbf{x}) = -\log p(\mathbf{x}|\mathbf{y}) - \log Z = \sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c). \tag{2.13}$$

where we omit $\mathbf{y}$ and θ in the right items for concise representation. The task of labeling an image is to find a label configuration that maximizes a posteriori of $p(\mathbf{x}|\mathbf{y},\theta)$, which is equal to minimizing the energy function.

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{2.14}$$

Thus the most important issue for a CRF model is to define an appropriate Gibbs energy function. In the classic CRF model, the potential is defined as the linear combination of multiple features. For example, the energy of a linear CRF model with unary and pairwise potentials can be written as (Kumar and Hebert (2003b)):

$$\begin{aligned} E(\mathbf{x}) &= \sum_{i\in\mathcal{V}} \psi_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i,x_j) \\ &= \sum_{i\in\mathcal{V}} \sum_{k} \theta_{1k} f_k(x_i,\mathbf{y}) + \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{N}_i} \sum_{d} \theta_{2d} g_d(x_i,x_j,\mathbf{y}) \end{aligned} \tag{2.15}$$

where $\psi_i(x_i)$ and $\psi_{ij}(x_i,x_j)$ are the unary and pairwise potentials. $f_k(x_i,\mathbf{y})$ denotes the k-th item in the D dimensional unary feature $\mathbf{f}(x_i,y)$, $g_d(x_i,x_j,\mathbf{y})$ is the d-th item in the B dimensional pairwise feature $\mathbf{g}(x_i,x_j,\mathbf{y})$, $\theta_1 = \{\theta_{1k}, k = 1,\cdots,D\}$ and $\theta_2 = \{\theta_{2d}, d = 1,\cdots,B\}$ are the set of parameters for unary and pairwise potentials respectively. This energy function can be efficiently optimized by the LBP algorithm (Murphy et al. (1999)).

Another energy function often used takes the form of the contrast sensitive Potts model (Boykov et al. (2001)), where the unary potential is defined as $\psi_i(x_i) = -\log(p(x_i|y_i))$, the negative log of the likelihood of a label being assigned to variable i, and the pairwise potential is defined as:

$$\psi_{ij}(x_i,x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ \theta_p + \theta_v \exp(-\theta_\beta ||I_i - I_j||^2) & \text{otherwise,} \end{cases} \tag{2.16}$$

where I_i and I_j are the color vectors of pixel i and j respectively, θ_p, θ_v and θ_β are model parameters. This energy function can be efficiently optimized by a graph-cut algorithm (Boykov et al. (2001)) since its potentials are submodular.

2.2.2. Robust Higher-order CRFs

The pairwise potential in a CRF model is a kind of hard constraint. It encourages two variables taking the same label and smoothes object boundaries. For most

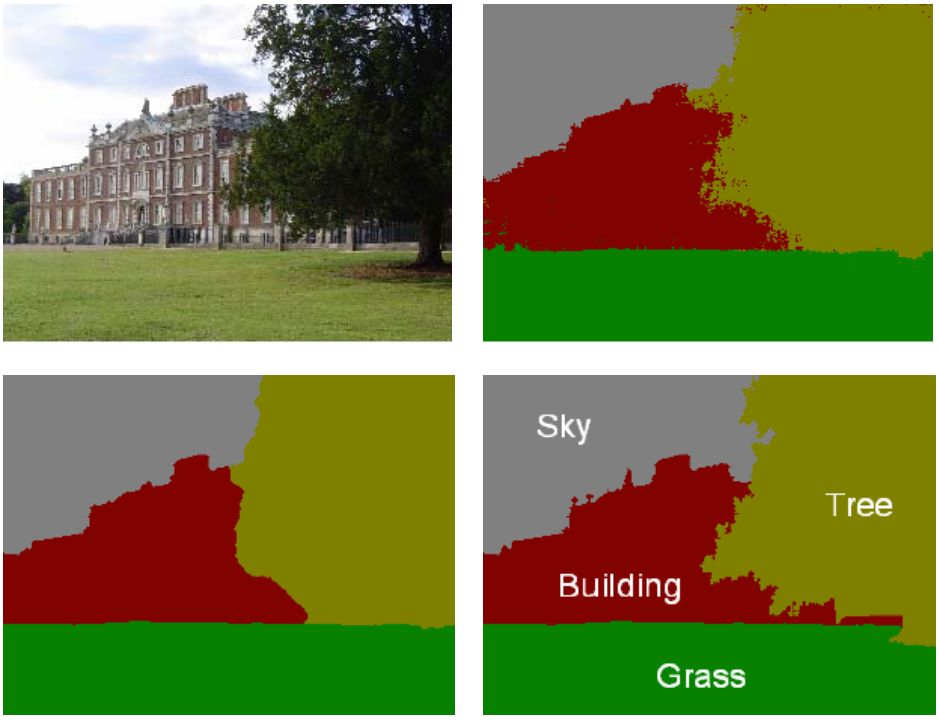


Figure 2.2.: Comparison between the pairwise CRF model and the higher-order CRF model (Kohli et al. (2009)). The left-top is the original image. The object segmentation in the right-top is obtained using the unary likelihood potentials. The left-bottom is the result of performing inference in the pairwise CRF. The right-bottom is the segmentation obtained by the RH-CRF model proposed in Kohli et al. (2009).

objects this leads to better results, however, it may cause oversmoothness of object boundaries. Thus it prevents the model finding the fine contours for certain object classes such as trees and bushes. Kohli et al. proposed to add robust higher-order potentials to improve the results of segmenting complex objects (Kohli et al. (2009)). As shown in Fig. 2.2, the resulting object contours are indeed improved by the Robust Higher-order CRF model (RH-CRF).

The Gibbs energy of the RH-CRF model is defined as:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) + \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) \tag{2.17}$$

where ψ_c is the higher order potential defined on a clique c. This robust higher order potential in Kohli et al. (2009) is defined as:

$$\psi_c(\mathbf{x}_c) = \begin{cases} N_i(\mathbf{x}_c)\frac{1}{Q}\gamma_{max} & \text{if } N_i(\mathbf{x}_c) \leqslant Q, \\ \gamma_{max} & \text{otherwise,} \end{cases} \tag{2.18}$$

where $N_i(\mathbf{x}_c)$ denotes the number of nodes which have different labels to the dominant label in the clique c, and can be calculated by $N_i(\mathbf{x}_c) = \min_k(|c| - n_k(\mathbf{x}_c))$. Here $|c|$ is the number of all nodes in clique c and $n_k(\mathbf{x}_c)$ is the number of nodes taking label l_k. The truncation parameter denoted by Q is used to control the rigidity of this higher order potential. The γ_{max} is defined as:

$$\gamma_{max} = |c|^{\theta_\alpha}(\theta_{hp} + \theta_{hv} \exp(-\theta_{h\beta} \frac{||\sum_{i \in c} (f(i) - \mu)^2||}{|c|})) \tag{2.19}$$

where $f(i)$ is the color vector of pixel i, μ is the mean color vector computed from all pixels in the clique c and parameters θ_α, θ_{hp}, θ_{hv}, and $\theta_{h\beta}$ are also learned from training data.

Through the robust higher order potential, the cost penalizing that not all pixels in a clique take the same label is a linear truncated function of the number of inconsistent variables, as shown in Fig. 2.3 (Kohli et al. (2009)).

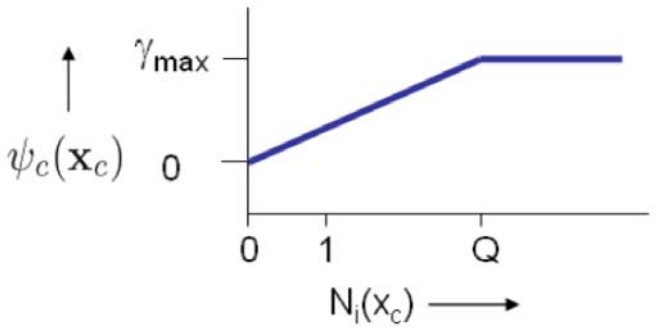


Figure 2.3.: Behavior of the robust higher order model potential (Kohli et al. (2009)). This figure shows how the cost changes with the number of variables not taking the dominant label in the clique.

The RH-CRF model can be inferred by the GC-based α-expansion and $\alpha\beta$-swap algorithm (Boykov et al. (2001)) since the higher order potential can be transformed to the submodular quadratic pseudo-boolean function. Note that eq. 2.18 can be reformulated as:

$$\psi_c(\mathbf{x}_c) = \min\{\min_{k\in\mathcal{L}} \mathcal{F}_c(|c| - n_k(\mathbf{x}_c)), \gamma_{max}\} \tag{2.20}$$

where $\mathcal{F}_c$ is a non-decreasing concave function. This formulation can be further generalized to the form:

$$\psi_c(\mathbf{x}_c) = \min\{\min_{k\in\mathcal{L}} ((P - f_k(\mathbf{x}_c))\theta_k + \gamma_k), \gamma_{max}\} \tag{2.21}$$

where parameter P and function $f_k(\mathbf{x}_c)$ are defined as:

$$P = \sum_{i\in c} w_i^k, \forall k \in \mathcal{L} \tag{2.22}$$

$$f_k(\mathbf{x}_c) = \sum_{i\in c} w_i^k \delta_k(x_i) \tag{2.23}$$

where

$$\delta_k(x_i) = \begin{cases} 0 & \text{if } x_i = k, \\ 1 & \text{otherwise,} \end{cases} \tag{2.24}$$

The weights $w_i^k \geqslant 0$ ($i \in c$, $k \in \mathcal{L}$) control the relative importance of different variables when preserving consistency of the labeling of the clique. Parameters γ_k, θ_k and γ_{max} satisfy the constraints which are $\theta_k = \frac{\gamma_{max}-\gamma_k}{Q_k}$ and $\gamma_k \leqslant \gamma_{max}$ for $\forall k \in \mathcal{L}$.

Note that eq. 2.21 is equivalent to the form of pseudo-boolean function:

$$f(\mathbf{t}_c) = \min(\theta_0 + \sum_{i\in c} w_i^0(1 - t_i), \theta_1 + \sum_{i\in c} w_i^1 t_i, \theta_{max}) \tag{2.25}$$

where $\mathbf{t}_c = \{t_i \in \{0, 1\}, i \in c\}$ denotes a set of binary random variables included in the clique c, and w_i^0, w_i^1, θ_0, θ_1, θ_{max} satisfy the constraints:

$$\begin{aligned} & w_i^0 \geqslant 0 \qquad w_i^1 \geqslant 0, \\ & \theta_{max} \geqslant \theta_0 \qquad \theta_{max} \geqslant \theta_1, \\ & ((\theta_0 + \sum_{i\in c} w_i^0(1 - t_i) \geqslant \theta_{max}) \\ & \vee (\theta_1 + \sum_{i\in c} w_i^1 t_i \geqslant \theta_{max})) = 1, \qquad \forall \mathbf{t}_c \in \{0, 1\}^{|c|} \end{aligned} \tag{2.26}$$

where $\vee$ is a boolean OR operator.

Kohli et al. proved that eq. 2.25 can be transformed to the submodular quadratic pseudo-boolean function (Kohli et al. (2009)). Therefore the higher order potentials can be optimized by the GC-based α-expansion and $\alpha\beta$-swap algorithm. Kohli et

al. also proved that the robust higher order potential possesses high efficiency, since only two extra auxiliary variables need to be added for each higher order potential and the complexity of the algorithm increases linearly with the size of the clique. For the detailed proof, readers can refer to Kohli et al. (2009) and Boykov et al. (2001).

2.2.3. Other Extensions of CRFs

Besides the RH-CRF model, there are a lot of different extensions of the CRF model. These extensions introduced different structures of models, utilizing the higher level and more forms of context information. Roughly, these extensions can be divided into four classes.

Dynamic Conditional Random Field Model

The dynamic CRF (DCRF) models (Sutton and McCallum (2005a), Wang and Mori (2009), Sutton et al. (2007), Shimosaka et al. (2007), Wang and Suter (2007), Wu et al. (2007)) simultaneously assign different kinds of labels to variables, by which the correlation among different kinds of labels can improve the labeling results. For example, a common way for object recognition via CRF models is to segment images first, then extract features from segments, and finally classify features to obtain the recognitive results. Actually, the segmenting and classifying are closely correlated. On the one hand, the segmenting results determine the accuracy of feature extracting and further effect the performance of classifiers. On the other hand, the class of objects also provides top-down information for image segmentation. Thus simultaneously segmenting and recognizing images, and making use of correlation between these two kinds of labels can improve the performance of object recognition.

Denoting by $\mathbf{x} = \{\mathbf{x}^m, m = 1, \cdots, M\}$ all the assignments of total M labeling tasks, where $\mathbf{x}^m = \{x_i^m, i \in \mathcal{V}\}$ is the assignment of the m-th labeling task given an image. The intra-clique that only considers the variables in one labeling task is denoted by c^1, and the inter-clique that considers the neighbor variables among multiple labeling tasks is denoted by c^2. Let $\mathcal{C}^1$ and $\mathcal{C}^2$ denote the set of all c^1s and c^2s respectively. The DCRF can be formulated as:

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z}\exp(-\sum_{c^1 \in \mathcal{C}^1}\sum_{c^2 \in \mathcal{C}^2}\psi_{c^1c^2}(\mathbf{x}_{c^1}^{c^2}, \mathbf{y}, \theta)) \tag{2.27}$$

where $\mathbf{x}_{c^1}^{c^2}$ denotes the collection of all assignments in cliques c^1 and c^2, and $\psi_{c^1c^2}$ is the potential defined on cliques c^1 and c^2. An illustration is shown in Fig. 2.4. It can be seen from this graph that the assignment of each variable of one labeling task relies on not only the whole observed image, but also the assignments of adjacent labeling tasks.

If the correlation of variables located at the same spatial position is only considered, the DCRF can be simplified to an FCRF model (Wang and Suter (2007), Wu

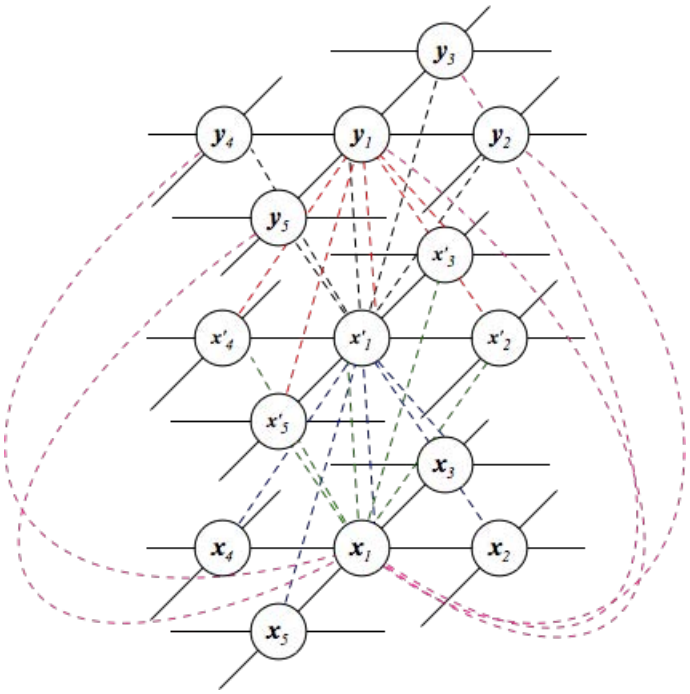


Figure 2.4.: A general graphical representation of DCRF.

et al. (2007)), which can be treated as a special case of DCRF. Moreover, there are several other kinds of DCRF, such as the edge DCRF model (Sutton et al. (2007)).

Hidden Conditional Random Field Model

The hidden CRF model (HCRF, Quattoni et al. (2007), Sung et al. (2007), Welling and Sutton (2005), Wang et al. (2006), Chu et al. (2007), Morency et al. (2007)) employs a middle layer called hidden variable layer between the observation and the common variable layer, as shown in Fig. 2.5. The HCRF augments the representability of the CRF model by adding the capability of modeling substructures of objects. For example, the hidden variables can be used to describe parts of objects, such as the head, foot for animals, wing and tail for airplanes. A HCRF model can be defined as:

$$p(\mathbf{x}|\mathbf{y}) = \sum_{\mathbf{h}} p(\mathbf{x}, \mathbf{h}|\mathbf{y}, \theta) = \frac{1}{Z} \exp\left(-\sum_{c \in \mathcal{C}} \psi_c(\mathbf{h}_c, \mathbf{x}, \mathbf{y}, \theta)\right) \tag{2.28}$$

where c is the clique defined on the hidden layer. It can be seen from Fig. 2.5 that the hidden layer and labeling layer rely on the whole observation, and the assignments for variables in the labeling layer rely on both the hidden layer and the observation. This is similar to the DCRF, but the difference is that the hidden layer is only considered as middle variables and need not be labeled.

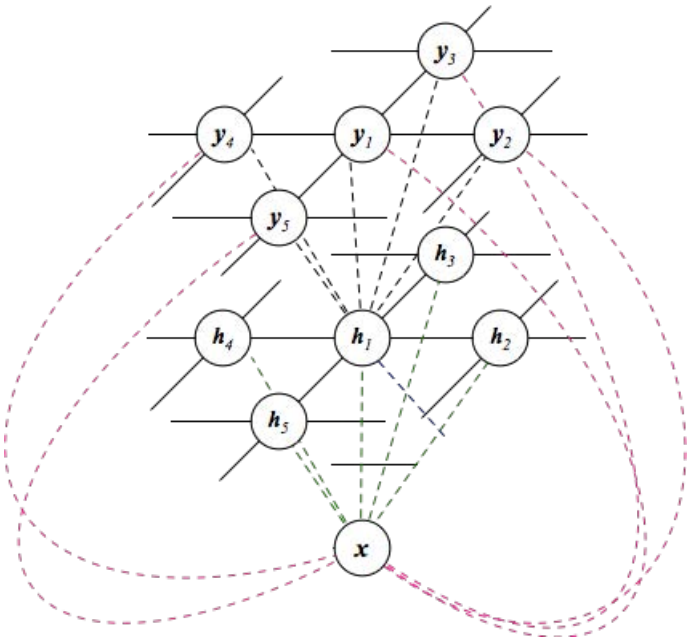


Figure 2.5.: A general graphical representation of HCRF.

Moreover, special HCRFs can be defined by adding some constraints on the hidden layer with special physical meanings. For example, when the hidden variables are used to describe object parts for object recognition, these hidden variables can be restricted to special locations, which yields the Layout Consistent CRF model (Winn and Shotton (2006),Zouhar et al. (2010)).

Tree Structured Conditional Random Field Model

In a standard CRF model the edges often take the lattice form, which can conveniently and accurately express the local context information. However, it is difficult to represent the large scale and global context information with this structure. One solution to represent large scale context is to transform the lattice structure to the tree structure (Bradley and Guestrin (2010), Lu et al. (2009), Duvenaud et al. (2011), Huang et al. (2011), Ladicky et al. (2009)), by which the directly connecting variables in the lattice structure are indirectly connected by the same parent nodes in a tree structure. Thus the tree structure CRF can model large scale context through transferring data via multiple layers. Furthermore, we can accurately and efficiently infer the tree structure graphical model and learn its parameters, since the tree is an acyclic structure. By introducing multiple hidden layers $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T - 1\}$, the tree structure CRF model can be defined as:

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z}\exp\Big(\sum_{c^1\in\mathcal{C}^1}\psi_{c^1}(\mathbf{x}_{c^1},\mathbf{y},\theta_1) + \sum_{c^2\in\mathcal{C}^2}\psi_{c^2}(\mathbf{h}_{T-1,c^2},\mathbf{x},\theta_2) + \sum_{t=1}^{T-2}\sum_{i,j}\xi(h_{t,j},h_{t+1,j},\theta_3)\Big) \tag{2.29}$$

where c^1 and c^2 denote the intra-clique in layer T and $T-1$ respectively, $\mathbf{h}_{T-1,c^2}$ is the collection of all variables in clique c^2 and $h_{t,j}$ is the label of variable i in the t-th layer.

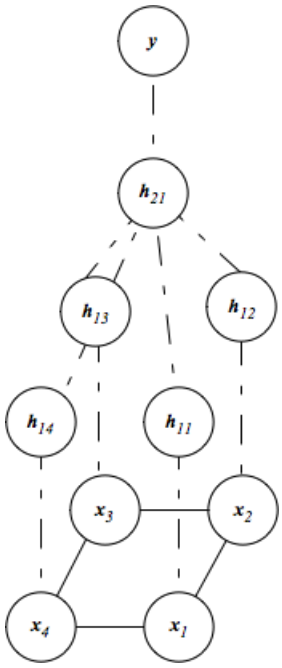


Figure 2.6.: A general graphical representation of tree CRF.

A quadtree CRF model is illustrated in Fig. 2.6 corresponding to this equation. Similar to the HCRF model, the tree structure CRF model is implemented by introducing the hidden layer, while the difference is that the tree structure CRF model introduces the hidden variables layer by layer where the observed image locates at the bottom.

Mixed Condition Random Field Model

The mixed CRF model is usually defined as the wighted sum of multiple CRF models, which can be formulated as:

$$p(\mathbf{x}|\mathbf{y}) = \sum_m w_m p_m(\mathbf{x}|\mathbf{y}) \tag{2.30}$$

where $p_m(\mathbf{x}|\mathbf{y})$ is the m-th CRF unit, w_m may be some simple weight coefficient (Sutton et al. (2006)), or some probabilistic model with special physical meanings (He et al. (2006)).

Furthermore, there are other kinds of extensions of CRF with respect to different application background, such as self-Markov CRF (Do and Artières (2005), Ngo et al. (2010)), segmented CRF (Liu et al. (2007)) and author related CRF (Mao and Lebanon (2007)).

2.3. The Topic Model

In this section, the general idea of topic models will be introduced first. Then we focus on the hierarchical Latent Dirichlet Allocation (hLDA), including the generative process and inference. This model will be extended for solving other sub-problems in this thesis. Finally, state-of-the-art work will be briefly reviewed, including the application of topic models in the fields of document analysis and computer vision.

2.3.1. Basic Concepts of the Topic Model

The concept of the topic model was proposed in recent year for document analysis. The topic model is a kind of probabilistic model. It introduces the topic space where a document can be represented by topics. Each topic is a probabilistic distribution over the word space. Thus the topic model has two advantages, as it can: 1) present documents in a lower dimension space, and 2) extract the latent semantics for a collection of documents.

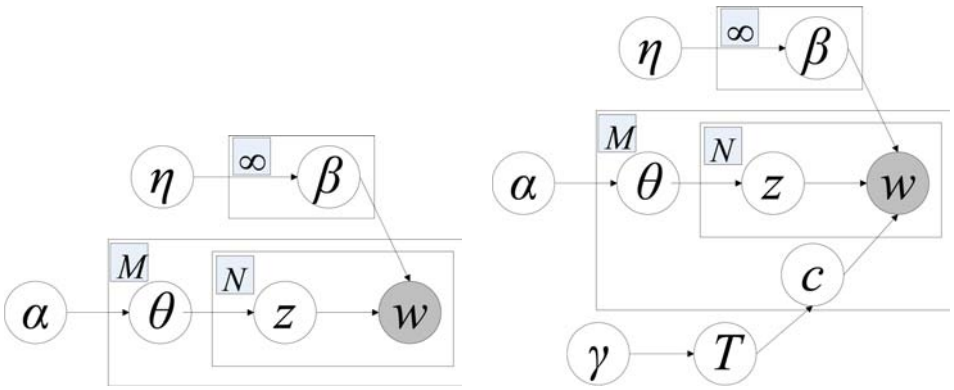
Given a collection of documents denoted by $D = \{d_1, d_2, \cdots, d_M\}$, where each document consists of a collection of words $d_i = \{w_1, w_2, \cdots, w_{n_d}\}$ sampling from a vocabulary containing V terms, and we assume that the document collection comprises K topics. Thus a classic topic model generates a document by the following process:

1. For a document d, draw a multinomial distribution over all K topics, $\theta(d) \sim Dir(\alpha)$,
2. For a word w_n in the document d:
   a) Draw an assignment of topic, $z_n \sim Multi(\theta)$
   b) Generate the word w_n, $w_n \sim Multi(w_n|z_n, \theta_z)$

where $\theta_z \sim Dir(\beta)$. This generative process corresponds to the Latent Dirichlet Allocation (LDA) model proposed in Blei et al. (2003). The $Dir(\alpha)$ is the Dirichlet priori distribution for sampling topics. $Multi(\cdot)$ denotes the multinomial probability of which the Dirichlet distribution is the conjunct distribution. A graphical model of LDA is illustrated in Fig. 2.7(a).

Now we introduce some basic concepts of topic models based on LDA.

- **Bayesian Hierarchical Model**. LDA is the two-levels hierarchical model where θ and θ_z are not the parameters, but random variables sampled from



(a) (b)

Figure 2.7.: The graphical model for LDA and HLDA.

Dirichlet priori distribution which is controlled by the hyper-parameters α and β as shown in the figure. The second level is z_n and w_n which are drawn from the multinomial distribution controlled by θ and θ_z.

- **Exchangeability**. Consider a set of N random variables $\{w_1, w_2, \cdots, w_N\}$. For any permutation, or reordering of these variables, if the probability satisfies:

$$p(w_1, w_2, \cdots, w_N) = p(w_{\tau(1)}, w_{\tau(2)}, \cdots, w_{\tau(N)}) \tag{2.31}$$

  where $\tau(\cdot)$ denotes any permutation, then these variables are exchangeable. The topic models assume that all words in a document satisfy the exchangeability and do not consider the orders of words, which simplifies the complexity of models.

- **Latent Variable**. In a latent variable model, some latent variables will be introduced, such as topics in LDA. These hidden variables are invisible, meaning they do not exist in the real data. But they can make the description and inference of the model more clear and simple. For example, by using topics, a document can be described in a lower dimensional topic space.

- **Conjunction**. In the Bayesian rule $p(\theta|x) \propto p(x|\theta)p(\theta)$, the conjunction distributions mean that posteriori and priori have the same probability density and only have different parameters. The conjunction distributions significantly simplify the model inference. For example, the Dirichlet distribution and multinomial distribution are the conjunction distributions.

The topic model concerns two distributions: 1) the "document $\sim$ topic" distribution and 2) the "topic $\sim$ word" distribution. In the LDA model, the priori of these two distributions are both symmetrical Dirichlet distributions, which come from $Dir(\alpha)$ and $Dir(\beta)$. In Griffiths and Steyvers (2004), the author gave a simple

approach to choose the values of α and β. Moreover, some research projects employed some nonparametric methods to introduce more complex priori, such as the Dirichlet process (Teh (2010)) and the Pitman-Yor process (Teh (2006)).

The inference of the LDA model is a difficult optimizing problem, since the exact inference is intractable. Usually, three approximate approaches are used for topic model inference, the Gibbs sampling (Griffiths and Steyvers (2004), He et al. (2009)), the variational inference (Blei et al. (2003)) and the expectation propagation (Minka (2001)). Generally, the Gibbs sampling is the most simple algorithm to infer topic models with promising performance. Particularly, the collapsed Gibbs sampling was employed for LDA inference (Griffiths and Steyvers (2004)). This sampling method integrates the hidden variables $\theta(\cdot)$ and $\theta_z(\cdot)$ and only samples the assignments of topics for all words, which significantly simplifies the algorithm complexity.

2.3.2. Hierarchical Latent Dirichlet Allocation

The hierarchical LDA (hLDA, Blei et al. (2010)) is extended from the LDA by transforming the flat topic structure to a tree-like topic structure. It adds the nested Chinese restaurant process into the model. Its graphical model is shown in Fig. 2.7(b). Compared to LDA, the hLDA has several advantages. First, it has a hierarchical, tree-like structure. Second, nodes (i.e. the topics) are generated directly from data. Third, the structure of the hLDA model can change dynamically as more and more samples are input.

The hLDA assumes that words in a document are generated according to a mixture model consisting of topics. The mixing proportions of topics are random and document-specific. Topics in hLDA are organized as a tree with fixed depth L. Each node in the tree has an associated topic. The document is then generated from L topics that form a path from the root to a leaf in the tree. To build a dynamic tree with a changeable structure and fixed depth, the nested Chinese restaurant process (nCRP) is used to generate paths.The first document generates an initial single branch tree with L nodes, which forms a single L-level path. The nCRP is then used to determine that subsequent documents are assigned either to one of the existing paths, or to a novel path branching off at any existing non-leaf node of the tree. The probability of assigning documents to a novel path is controlled by the parameter of nCRP, γ. A smaller value of γ results in a tree with fewer branches. Supposing that the infinite tree defined by nCRP is obtained and to use $\mathbf{c}_d$ denotes the path for the d-th document, the formally generative process of hLDA is as follows:

1. For each node $k \in \mathbf{T}$ in the infinite tree, draw a topic $\beta_k \sim Dir(\eta)$,
2. For each document, $d \in \{1, 2, ..., D\}$,
   a) Draw $\mathbf{c}_d \sim nCRP(\gamma)$,
   b) Draw a distribution over levels in the tree, $\theta_d \mid \{m, \pi\} \sim GEM(m, \pi)$,
   c) For each word,

i. Choose level $Z_{d,n} \mid \theta_d \sim Multi(\theta_d)$,

ii. Choose word $W_{d,n} \mid \{z_{d,n}, \mathbf{c}_d, \beta\} \sim Multi(\beta_{\mathbf{c}_d}[z_{d,n}])$, which is parameterized by the topic in position $z_{d,n}$ on the path $\mathbf{c}_d$.

Thus for a particular document W, the generative model of this sampling process can be represented by a joint distribution of observed and hidden variables given the hyper parameters:

$$p(\mathbf{w}, \mathbf{z}, c, \theta, \beta | \alpha, \eta, T) = \prod_{i=1}^{N} p(w_i|z_i, c, \beta)p(z_i|\theta)p(\theta|\alpha)p(\beta|\eta)p(c|T) \tag{2.32}$$

Two main steps are carried out for hLDA inference: the sampling of level allocations and the sampling of path assignments. Given current path assignments and the current values of all other variables, the level allocation of variable $z_{d,n}$ for word n in document d needs to be sampled from its distribution:

$$p(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta) \propto p(z_{d,n} \mid \mathbf{z}_{d,-n}, m, \pi)p(w_{d,n} \mid \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) \tag{2.33}$$

where $\mathbf{z}_{-(d,n)}$ and $\mathbf{w}_{-(d,n)}$ denote the vectors of level allocations and observed words leaving out $\mathbf{z}_{(d,n)}$ and $\mathbf{w}_{(d,n)}$ respectively. And $\mathbf{z}_{d,-n}$ is the level allocations excluding $\mathbf{z}_{d,n}$ in document d.

The first term in eq. 2.33 defines a stick-breaking distribution over levels which has an infinite number of components. If $k \leqslant \max(\mathbf{z}_{d,-n})$, its first component is computed as:

$$\begin{aligned} p(z_{d,n} \mid \mathbf{z}_{d,-n}, m, \pi) &= E[V_k \prod_{j=1}^{k-1}(1 - V_j)|\mathbf{z}_{d,-n}, m, \pi] \\ &= E[V_k|\mathbf{z}_{d,-n}, m, \pi] \prod_{j=1}^{k-1} E[1 - V_j|\mathbf{z}_{d,-n}, m, \pi] \\ &= \frac{m\pi + \#[\mathbf{z}_{d,-n} = k]}{\pi + \#[\mathbf{z}_{d,-n} \geqslant k]} \prod_{j=1}^{k-1} \frac{(1-m)\pi + \#[\mathbf{z}_{d,-n} > j]}{\pi + \#[\mathbf{z}_{d,-n} \geqslant j]} \end{aligned} \tag{2.34}$$

where $\#[\cdot]$ denotes the number of elements of an array that satisfy conditions in the bracket.

The second term in eq. 2.33 defines the distribution of a given word based on a possible assignment. The parameter β_i is generated from a symmetric Dirichlet distribution with hyperparameter η. Thus this distribution can be computed as:

$$p(w_{d,n}|\mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) \propto \#[\mathbf{z}_{-(d,n)} = z_{d,n}, \mathbf{c}_{z_{d,n}}, \mathbf{w}_{-(d,n)} = w_{d,n}] + \eta \tag{2.35}$$

which describes the smoothed frequency that a word $w_{d,n}$ is sampled to the topic at level $z_{d,n}$ of the path $\mathbf{c}_d$.

The first term in eq. 2.33 has an infinite number of components, thus the distribution of a new component emerging over topic assignments is formulated differently:

$$\begin{aligned} & p(z_{d,n} > \max \mathbf{z}_{d,-n} | \mathbf{z}_{d,-n}, \mathbf{w}, m, \pi, \eta) \\ & = 1 - \sum_{j=1}^{\max \mathbf{z}_{d,-n}} p(z_{d,n} = j | \mathbf{z}_{d,-n}, \mathbf{w}, m, \pi, \eta) \end{aligned} \tag{2.36}$$

When the level allocation variables are given, the path associated with each document conditioned on all other paths needs to be sampled:

$$p(\mathbf{c}_d \mid \mathbf{w}, \mathbf{c}_{-d}, \mathbf{Y}, \mathbf{z}, \eta, \gamma, \varphi) \propto p(\mathbf{c}_d \mid \mathbf{c}_{-d}, \gamma) p(w_d \mid \mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}, \eta) \tag{2.37}$$

where the first term in this equation is the prior of paths implied by the nested CRP, and the second term is the probability of the data given a particular choice of path, which can be integrated over the multinomial parameters:

$$\begin{aligned} p(\mathbf{w}_d | \mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}, \eta) = & \prod_{l=1}^{\max(\mathbf{z}_d)} \frac{\Gamma(\sum_w \#[\mathbf{z}_{-d} = l, \mathbf{c}_{-d,l} = c_{d,l}, \mathbf{w}_{-d} = w] + V_\eta)}{\prod_w \Gamma(\#[\mathbf{z}_{-d} = l, \mathbf{c}_{-d,l} = c_{d,l}, \mathbf{w}_{-d} = w] + V)} \\ & \frac{\prod_w \Gamma(\#[\mathbf{z} = l, \mathbf{c}_l = c_{d,l}, \mathbf{w} = w] + \eta)}{\Gamma(\sum_w \#[\mathbf{z} = l, \mathbf{c}_l = c_{d,l}, \mathbf{w} = w] + V_\eta)} \end{aligned} \tag{2.38}$$

2.3.3. Improvement and Application of Topic Model

Since the LDA model was proposed, a lot of research work has improved this topic model from different aspects to obtain better performance. Some improved topic models are not based on the assumption of exchangeability of the topic model, since this assumption limits the representative ability. For example, Blei and Lafferty (2007) and Li and McCallum (2006) introduced the correlation among topics, Chang and Blei (2009) considered the correlation among documents and Wang et al. (2007) took word order into account. However, the complexity of these models is significantly enlarged. Therefore, some literature compromises the complexity and the representation. For example, in Lu and Zhai (2008) the authors added more priori to distinguish between different topics.

Some novel topic models employ more complicated nonparametric Bayesian methods to describe more complex problems. For example, in Teh (2010) and Teh (2006) the authors introduced the Dirichlet process to automatically determine the number of topics. However, the parameters of these topic models need to be set more carefully, and the complexities of algorithms are higher than the LDA model.

Furthermore, the topic models have been widely used in not only the fields of document analysis and web information retrieval, but the field of computer vision. In the literature of Lu and Zhai (2008), Jo and Oh (2011), Li et al. (2010c), Lin and He (2009) and Mei et al. (2007), researchers used topic models to extract the authors

opinions and advice, then to generate the sentiment abstraction, and to construct a sentiment dictionary. The author-topic model (Rosen-Zvi et al. (2004)) considers the generation of topics from the authors' aspect, where the document-topic distribution is replaced by the author-topic distribution. McCallum et al. (2007) further extended the author-topic model by taking the receivers into account. In Zhao et al. (2011) the topic model is applied to analyze the data in Twitter.

In the field of computer vision, the topic models are used for automatical novel object discovery, object recognition, and object segmentation. A set of methods used the topic model as clustering algorithm to discover novel object categories. For example, in Russell et al. (2006) authors utilized the topic model to cluster object categories from segmented image parts. More literature integrated spatial information and topic models for simultaneously segmenting images and recognizing objects. In Sudderth et al. (2008) authors proposed a transformed Dirichlet process (TDP) model to simultaneously segment and recognize objects. The TDP model integrates a topic model and spatial transformations, which can discover contextual relationships, and better exploit partially labeled training images. In Sivic et al. (2005), authors employed the topic model to discover object categories, as it uses visual features to simulate words and treats object categories as topics. In Fergus et al. (2005), the authors extended the topic model to include spatial information in a translation and scale invariant manner, which can automatically learns an object category model from the training images obtained from the Google search. Sudderth and Jordan developed a novel statistical framework in which the object frequencies and segment sizes are modeled by the Pitman-Yor process (Sudderth and Jordan (2009)). This nonparametric prior distribution yields the statistical framework where learning algorithms discover an unknown set of objects, and segmentation methods automatically adapt their resolution to each image. Ghosh *et al.* proposed a novel hierarchical extension of the spatial distance dependent Chinese restaurant process (ddCRP) model for unsupervised image segmentation (Ghosh et al. (2011)). The ddCRP exploits spatial non-exchangeable data to enhance the representation of the topic model, which leads to better segmenting results. In Wang and Grimson (2007) authors also integrated the spatial information with the topic model (LDA) to discover object categories. Cao and Fei-Fei proposed a spatially coherent latent topic model for simultaneously recognizing and segmenting object and scene classes (Cao and Fei-Fei (2007)).

2.4. Summary

In this chapter, we first introduce the basic concepts of probabilistic graphical model. Then two widely used graphical models, the CRF model and the topic model, are presented. We explain two important probabilistic graphical models, the robust higher order CRF model and the hierarchical latent Dirichlet allocation in detail. They are important extensions with respect to the classical CRF models and topic models, respectively. Finally, we extensively review the related work about the two models, including their improvements and different application cases.

Chapter 3

The State-of-the-Art

In this chapter state-of-the-art methods concerning the techniques of solving the concrete computer vision problem related to this thesis are introduced. This introduction is divided into two parts. The first part is about category-independent object detection, in which we will extensively introduce existing methods. The second part discusses the problem of novel category detection and discovery, in which we will introduce object descriptions and different models about category discovery.

3.1. Category-Independent Object Detection

Before reviewing category-independent object detection, we first briefly introduce some state-of-the-art approaches about category-specific object detection, and point out why they cannot be used for category-independent object detection. Currently, most work on object detection is focused on category-dependent detectors. A usual technique is to employ effective object descriptions to represent objects (Dalal and Triggs (2005)) or object parts (Felzenszwalb et al. (2010)), and utilizes sliding windows to search objects in an image. However, this sliding window approach is extremely inefficient since it will search all windows over the whole image. Some researchers therefore improved the searching efficiency by using a proportion of all windows (Vedaldi et al. (2009)). A more effective searching method is to use the branch and bound optimization to find global optimal windows, which is called effective subwindow search (ESS) (Lampert et al. (2009a), Lehmann et al. (2009)).

Most work only uses local image features as mentioned above. Little work takes global features into account. In Torralba et al. (2010), the gist feature (Torralba (2003), Torralba et al. (2006)) is combined with effective local features (Torralba et al. (2007), Torralba (2003)) to improve object detectors. Through the gist feature, the probability of a position containing objects in an image is enlarged, if at the position objects are more possibly present. And the position with less possibility of presenting objects is suppressed.

Depth is also useful information for object detection. The consistency of the depth of objects can be used as a good indicator for object detection. However, only little work takes the depth into account. For example, in Gould et al. (2008), the depth produced by a laser range scanner is integrated with image features to improve the object-dependent detector.

Generally speaking, the selection of features and models are two key issues for category-specific object detection. Features must have enough distinctness for distinguishing different category objects. Models should integrate features to form category templates to represent object categories. Similarly, category-independent object detection is also concerned in these two key issues. But there are several significant differences. First, features for category-independent object detection should generalize across categories and keep the distinctness for distinguishing objects from background, while features for category-specific object detection do not need high generalizability. Second, the development of models does not aim at training category-specific templates, but measuring a certain generic object-likeness. In this section, we review state-of-the-art methods about category-independent object detection by considering these two issues.

Recently, more and more approaches detecting and localizing category-independent objects have been proposed. They can be roughly divided into four classes: salient object detection based on visual saliency (Feng et al. (2011), Liu et al. (2011b), Goferman et al. (2010), Cheng et al. (2011)), object-likeness measurement based on structure regions (Endres and Hoiem (2010), Collet et al. (2011), Carreira and Sminchisescu (2012), Levinshtein et al. (2010), Saenko et al. (2011)), object-likeness ranking based on sliding windows (Alexe et al. (2010), Rahtu et al. (2011), Zhang et al. (2011)) and figure/ground segmentation (Carreira and Sminchisescu (2012), Ren et al. (2006), Ion et al. (2011), Bagon et al. (2008)). Lastly, although the co-segmentation framework does not explicitly detect category-independent objects, it also identifies similar objects from a pair or group of images, regardless of object categories. Therefore here we also introduce state-of-the-art methods about co-segmentation.

3.1.1. Salient Object Detection

Salient object detection is based on the human visual attention mechanism by which humans can rapidly attend to the region with highest distinctiveness. The common way to detect salient objects is first to compute the saliency map from an image by different approaches, such as local contrast and global distribution. After that, diverse methods are employed to determine which regions with a certain saliency contain the salient objects.

Since visual attention theories were proposed and applied in computer vision by Itti et al. (1998), all subsequent methods follow one or several of the four basic principles of human visual attention summarized by Goferman et al. (2010), which are local low-level considerations, global considerations, visual organization rules and

high-level factors. The most frequently used principle is the local low-level considerations that usually takes the contrast between the local area and its surroundings into account, which can be abstractively defined as:

$$sal(x) = D(f_{A_x}, f_{S_x}) \tag{3.1}$$

where $D(\cdot, \cdot)$ denotes the distance of two features, f_{A_x} and f_{S_x}, extracted from a local region A_x and its surroundings S_x, which are both centered at point x. For example, Li et al. (2010d) computed local saliency based on approximated local conditional entropy by the lossy coding length of multivariate Gaussian data, which is defined as:

$$sal(c_x) = L_\epsilon(SC) - L_\epsilon(S) \tag{3.2}$$

where c_x is a patch of an image centered at point x, $S(c)$ denotes the surroundings of this patch, $L_\epsilon(S)$ is the lossy coding length of surroundings of patch c, and $L_\epsilon(SC)$ is the lossy coding length of the area combining patch c and its surroundings. Here SC and S are represented by features. In Liu et al. (2011b) another effective local consideration, namely the center-surround histogram, is defined by the color histogram contrast between a center region and its surroundings.

$$sal(x) = \sum w_{xx'} \chi^2(R(x'), R(x)) \tag{3.3}$$

where x' denotes the surroundings of region x, $w_{xx'}$ are weights, and $\chi^2(R(x'), R(x))$ is the Chi-square distance between $R(x')$ and $R(x)$ which denote the color histogram of region x' and x respectively.

The global considerations are often computed based on the feature distribution on the whole image. For example, Cheng et al. (2011) defined the global saliency as the global rarity of L*a*b color. A L*a*b color histogram is constructed on the whole image, then the pixels taking a more rare color are considered as more salient positions. The saliency is defined as:

$$sal(c_l) = \sum_{j=1}^{n} f_j D(c_l, c_j) \tag{3.4}$$

where c_l is the color value corresponding to the l-th bin in the histogram, f_j is the probability of assigning pixels to bin j according to their colors. Liu et al. (2011b) used a RGB color Gaussian mixture model (GMM) instead of the color histogram to computed the global saliency. By the GMM model, regions where the color of pixels have lower probabilities in the GMM are regarded as the salient regions.

For the third principle, Liu et al. (2011b) and Goferman et al. (2010) assumed that the center regions in an image will obtain the higher saliency priori.

The fourth one is related to the model of determining which regions are the most salient regions to contain objects. The simplest method is to use a threshold to determine the object regions. For example, in Achanta et al. (2008) and Li et al. (2010d) the salient object map is defined as:

$$P(x) = \begin{cases} 1 & \text{if } sal(x) \geqslant threshold, \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

where 1 means that the corresponding pixel x belongs to an object. A more complicated method based on a CRF model is employed for salient object detection in Liu et al. (2011b), where unary potentials are computed from three kinds of visual saliency.

Although visual saliency can be regarded as a good category-independent feature, it is not sufficient to use the saliency as sole feature to precisely detect all objects of interest. This is because the definitions of two problems, salient object detection and category-independent object detection, are different. The former one is to find the most distinctive object with the highest saliency, but in the latter one objects in 2D images to be detected may not have a high saliency and the background may have a high saliency. Therefore, salient object detection methods may fail to detect only one object in an image (Liu et al. (2011b), Goferman et al. (2010)) or detect part of the background as objects (Feng et al. (2011), Cheng et al. (2011)). In the proposed method in this thesis, a novel 3D saliency is developed as one kind of category-independent feature together with 2D saliency. Thus more robust saliency features can be obtained. Furthermore, we do not assume that the region with the highest saliency corresponds to objects, but use a Support Vector Machine (SVM) to learn the probability of corresponding to objects of interest for different saliency values.

3.1.2. Objectness Ranking

The objectness ranking method was first proposed by Alexe et al. (2010). It combines several local cues into a naive Bayesian model to sample the windows in an image with high probability of containing objects by the sliding window approach. Four local cues are used, which are multi-scale saliency, color contrast, edge density and superpixel straddling. At first a saliency map, defined as $I_s^{MS}(p)$ for each pixel p, is computed on each scale s. Then the saliency of a window w at scale s can be defined as:

$$MS(w, \theta_s^{MS}) = \sum_{\{p \in w | I_s^{MS}(p) \geqslant \theta_s^{MS}\}} I_s^{MS}(p) \times \frac{|\{p \in w | I_s^{MS}(p) \geqslant \theta_s^{MS}\}|}{|w|} \tag{3.6}$$

where $|w|$ denotes the number of pixels in this window and the parameter θ_s^{MS} needs to be learned. The windows with higher density of salient pixels and higher salient values for pixels are computed as having a higher saliency, which are considered as having a high probability of containing objects.

The color contrast is a local measure of the dissimilarity of a window to its immediate surrounding area, which is based on the assumption that the higher contrast a

window has, the more possibly the window contains a whole object. Thus this cue is defined as:

$$CC(w, \theta_{CC}) = \chi^2(h(w), h(Surr(w, \theta_{cc}))) \tag{3.7}$$

where $h(w)$ denotes the color histogram of window w, and $h(Surr(w, \theta_{cc}))$ is the color histogram of the surroundings of w. The area of $Surr(w, \theta_{cc})$ is controlled by the parameter θ_{cc} that needs to be learned.

The edge density measures the density of edges near the window borders. This cue is based on the assumption that more edges occur at the objects' boundary. Thus the higher edge density a window has, the more possibly the window contains a whole object. This cue is defined as:

$$ED(w, \theta_{ED}) = \frac{\sum_{p \in Inn(w, \theta_{ED})} I_{ED}(p)}{LenInn(w, \theta_{ED})} \tag{3.8}$$

where $Inn(w, \theta_{ED})$ denotes the inner ring obtained by shrinking the window w by a factor θ_{ED} in all directions, $I_{ED}(p)$ denotes the binary edgemap which is obtained by Canny detector, and $Len(\cdot)$ measures the perimeter.

Finally, the most useful cue in Alexe et al. (2010) is the superpixel straddling based on the assumption that all pixels in a superpixel belong to the same object (Russell et al. (2006)). For a window w, if there is at least one pixel of a superpixel s in the window and at the same time at least one pixel of s outside w, this superpixel straddles this window. Thus the superpixel straddlling can be used to estimate if a window contains an object:

$$SS(w, \theta^{SS}) = 1 - \sum_{S \in \mathbf{S}(\theta^{SS})} \frac{min(|s \backslash w|, |s \cap w|)}{|w|} \tag{3.9}$$

where $\mathbf{S}(\theta^{SS})$ is the set of superpixels with a parameter, segmentation scale denoted by θ^{SS}, which needs to be learned. The sum in this equation counts the ratios between a minimal number of pixels in superpixels in a window or outside a window and this window's area. Thus when most parts of an object are in a window, this sum trends to zero and consequently $SS(w, \theta^{SS})$ closes to one.

These four cues are then combined into a simple Bayesian model, where each cue is treated independently. Thus the probability that a window contains an object is obtained by this formulation:

$$p(obj|\Omega) = \frac{p(\Omega|obj)p(obj)}{p(\Omega)} = \frac{p(obj) \prod_{\omega \in \Omega} p(c|obj)}{\sum_{c \in \{obj, bg\}} p(c) \prod_{\omega \in \Omega} p(\omega|c)} \tag{3.10}$$

where Ω is the set of local cues.

The sampled windows by Alexe et al. (2010) can cover most objects in an image regardless of their categories. Thus these outputs can be used as location priors for greatly reducing the number of windows evaluated by class-specific object detectors.

It further is improved by Rahtu et al. (2011) by more elaborate local cues and Zhang et al. (2011) by multimodal and global cues.

Although the output windows of objectness ranking methods can cover most objects, too many windows need to be sampled. To cover 70% objects one thousand windows with the highest objectness measurement need to be sampled as reported in Rahtu et al. (2011). Even in simple scenes, the number of sampled windows must be 3-5 times of the number of objects to cover most objects as reported in Zhang et al. (2011). Thus these objectness ranking methods cannot be directly used for category-independent object detection. Similarly to these methods, our method also integrates several cues into one uniform framework. But we use the more elaborate CMH-CRF model to obtain the more accurate detection.

3.1.3. Structured Segmentation

The structured segmentation method is first to segment images or 3D data (e.g. point cloud or depth map) into oversegments by some unsupervised segmentation algorithms (e.g. Shi and Malik (2000), Felzenszwalb and Huttenlocher (2004)). Then regions combined by adjacent oversegments are measured for whether they contain whole objects. Unlike the salient object detection, these methods can detect more than one object regardless of their categories.

However, these methods also have to sample a large number of regions to cover most objects in an image. Furthermore, they rely heavily on the results of oversegments. Thus on the one hand the small objects that are smaller than the minimal patch size of oversegments cannot be detected. On the other hand, objects with the similar color or texture as the background which cannot be segmented precisely may be detected as part of the background. To improve the oversegments, they employed hierarchical segmentation (Endres and Hoiem (2010)), multi-scale segmentation (Carreira and Sminchisescu (2012)) or utilized multimodal data (Collet et al. (2011), Saenko et al. (2011)). Although each oversegment may not exactly belong to only one object, it is still a kind of efficient mid-level representation of images. Based on oversegments, the computational cost is largely decreased with respect to that of directly using pixels. Furthermore, many mid-level features can be efficiently computed through oversegments. Similar to these structure region based methods, the proposed CMH-CRF model also takes the multimodal oversegments as basic nodes. But the details of how to utilize oversegments are significantly different.

3.1.4. Figure/ground Segmentation

As first introduced by Ren et al. (2006), the figure/ground segmentation in computer vision is defined as assigning two different labels (foreground and background) to different regions in an image. All regions with the same label have the largest consistency in color, texture and so on, while having the largest difference from other regions assigned another label. Recently, Ion et al. (2011) proposed a novel

framework for figure/ground segmentation, namely segmentation by composition, which composites regions with a certain consistency as foreground segmentation. After that, many researchers improved this framework and derived better segmentation performances as reported in Carreira and Sminchisescu (2012), Bagon et al. (2008). However, since the figure/ground segmentation only takes two labels, it is difficult to distinguish object instances when they overlap. Therefore each connected region of segmented results will be considered as one object, no matter how many objects this region does contain. Although the method proposed in Carreira and Sminchisescu (2012) takes the multi figure/ground segmentation hypothese and ranks them to detect more than one object, it still needs to sample a large set of different figure/ground segmentation hypothesis to find the best segments. Unlike figure/ground segmentation, our method in this thesis considers three labels (the object, the background and the boundary), by which the object instances can be easily distinguished even when they overlap.

3.1.5. Co-segmentation

In this thesis the proposed method for category-independent object detection simultaneously segments 2D image and corresponding 3D point clouds, which is related to co-segmentation methods. Current co-segmentation methods are all based on 2D images, and segment similar or same objects from a pair (Rother et al. (2006), Vicente et al. (2011), Vicente et al. (2010)) or group (Mukherjee et al. (2011), Glasner et al. (2011)) of images in an unsupervised way.

The common model of co-segmentation is defined as a Markov Random Field model with global constraints for which an energy function needs to be optimized (Vicente et al. (2010)):

$$E(x) = \sum_{p} w_p x_p + \sum_{(p,q)} w_{pq}|x_p - x_q| + \lambda E^{global}(h_1, h_2) \tag{3.11}$$

where the first two terms are the usual MRF terms for both images, w_p is the unary weight for each pixel and w_{pq} is the pairwise weight. The additional global constraint expressed by the third term encodes a similarity measure between the foreground histograms of both images, and λ is the weight for that term. Thus minimizing this energy function will find the most similar foreground objects in both images since their histograms reach the minimal distance.

Since Rother et al. (2006) first proposed the co-segmentation problem, a number of subsequent work extended and improved it from two aspects, model and optimization technique. Glasner et al. (2011) converted the model to a Quadratic Semi-Assignment Problem and used a Linear Programming for optimization. Although co-segmentation can successfully segment 'something similar' in a given set of images, it cannot guarantee that they are objects of interest, and some objects that are different in different images cannot be found. Vicente et al. (2011) added two more new aspects, objectness measurement and similarity learning, to current

co-segmentation framework. This yielded a new object co-segmentation, which increases the likelihood of the segmented results being objects. Although it is a further step carrying co-segmentation to category-independent object detection, it still has the limitations that the segmented results are binary and adjacent object instances cannot be distinguished. Different to current methods, we model the co-segmentation of the 2D image and the 3D point cloud as a new CRF, through which the object-likeness measurement can be learned from training data. Furthermore, three kinds of labels in the proposed method can be used to easily distinguish adjacent object instances.

3.2. Novel Category Detection and Discovery

The category-independent object detection tells us where objects are in a scene. The object recognition indicates objects belong to which categories that were learned before. The novel category detection and discovery is to identify and describe unknown objects, which can be treated as a synthesis of several relative problems, such as object description, object cluster and category organization, as mentioned in 1.4. Here we extensively review the related work. For object description, we employ and extend the object attributes to describe objects, therefore in the first part we only review state-of-the-art work about this. There is a lot of work on hierarchical object and category organizations, which will be introduced in the second part. Novelty detection is a kind of clustering technique which aims at finding novel patterns from data. We will review the state-of-the-art on this in the third part.

3.2.1. Object Attributes

Object attributes mean the external tokens of objects reflected in the human mind, which can be obtained by human or artificial sensors and used to describe and distinguish objects. Here the distinctions between object attributes and image features need to be clarified. Actually, they are different in several aspects. First, image features, such as color histograms, histograms of oriented gradients (Dalal and Triggs (2005)), local texture descriptors (Varma and Zisserman (2005)) and scale invariant feature transformation (Lowe (2004)), are computed directly from the original data of image pixels, while object attributes are obtained based on these image features. Second, their semantic meanings are different. Object attributes have certain kinds of semantic meanings while image features do not. For instance, when one talks about the 'metal' attribute, people can imagine certain objects, such as cars and knives which are made of metal. However, when people are given a color histogram, they usually do not know what kind of color it represents. Third, the generalizability across categories is different. An object attribute can be used to describe other categories than those in the training dataset, since the training of attributes is not category-specific (Farhadi et al. (2009)). However, image features are difficult

to generalize to other categories. At last, object attributes are more identical to human cognition since humans always describe an object by its attributes.

There are various attributes, such as color (e.g. 'blue'), texture (e.g. 'metal'), parts (e.g. 'has beak') and function (e.g. 'can fly'). Research on extracting and using attributes has recently received much attention (Farhadi et al. (2009), Lampert et al. (2009b), Farhadi et al. (2010), Ferrari and Zisserman (2008)). In Ferrari and Zisserman (2008), how to localize simple color and texture attributes has been learned. Farhadi et al. proposed an algorithm to learn a more broad set of complex attributes (Farhadi et al. (2009)) and subsequently extended it to localize objects (Farhadi et al. (2010)). In Lampert et al. (2009b) an algorithm was proposed to use attribute-based representations to recognize new categories of animals. This application is similar to a part of our work while the main difference is that there is no hierarchical structure of this model. More recently, further applications based on object attributes have been presented (Kulkarni et al. (2011), Douze et al. (2011), Liu et al. (2011a), Li et al. (2010a)). In Kulkarni et al. (2011), relevant sentences about images can be generated by detecting objects, adjectives and spatial relationships, which are described by attributes. The excellent results in image retrieval (Douze et al. (2011)) and human action recognition (Liu et al. (2011a)) are also demonstrated based on attributes. Rich semantic level image information based on object attributes is employed in Li et al. (2010a) to tackle higher level visual recognition problems.

Object attributes can be divided into semantic attributes and non-semantic attributes (Farhadi et al. (2009)). Semantic attributes have concrete semantic meanings which can be described clearly by simple language, while non-semantic attributes means those object characteristics that cannot be stated clearly by (at least simple) language.

The algorithm presented in Farhadi et al. (2009) will be extended in this thesis. Here we briefly state how it works. An object is located by a bounding box before extracting its attributes. The box is first divided into six parts with two rows and three columns (see fig. 3.1 as an example). Then the base features from which object attribute classifiers are learned are extracted from the six parts plus the whole box. There are four types of base features: texture descriptors extracted with a texton filter bank, HOG spatial pyramids, edges extracted by the Canny edge detector and color descriptors from the histogram of color. At last, a 9751 dimensional feature is formed for a bounded object.

Based on these base features, two kinds of attributes are learned. The first is semantic attributes. Four main types of semantic attributes - color, shape, part and material & texture - are used. However, it is difficult to learn an attribute classifier with which an occurrence attribute is associated. For instance, the 'metal' attribute is always present when training the 'wheel' attribute through buses, trains and cars samples. When a confused attribute classifier is used to identify the 'wheel', objects with the 'metallic' but without the 'wheel' may be regarded as objects with the 'wheel'. On the other hand, a wooden 'wheel' may not be recognized as a 'wheel'

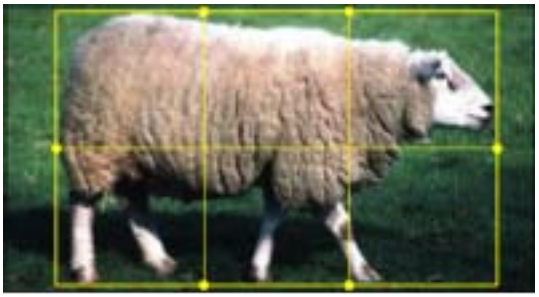


Figure 3.1.: An example of a bounded box on an object for extracting base features.

since it does not have the 'metallic' attribute. To overcome this problem and obtain more accurate classifiers, a feature selection criterion was proposed in Farhadi et al. (2009). For each class in the training set, samples with the 'wheel' attribute are positive examples and those without the 'wheel' attribute are negative examples. In each class, the features distinguishing positive and negative samples well are selected by using an L1-regularized logistic regression (Ng (2004)). Those selected features for all classes are then combined. A linear SVM is used to train the attribute classifier based on the combination of features.

Besides semantic attributes, to describe an object accurately requires non-semantic attributes due to the limitation of semantic attributes (Farhadi et al. (2009)). Only non-semantic attributes that take the difference among categories into account are learned in Farhadi et al. (2009). The algorithm is listed in alg. 3.1. This kind of non-semantic attributes are learned from multi-categories, therefore they can be generalized across categories. Since in line 3 the selected categories are divided into two sides, only the attributes that describe the difference among categories can be obtained.

Algorithm 3.1: Learn non-semantic object attributes.

1 Input: Base features of all object samples of all categories in training set.
for *each non-semantic object attribute* **do**
2 Randomly select 2-10 categories; Divide them into two halves with equal number of categories; Randomly select a subset of base features; Train a classifier using linear SVM;
3 Select a certain number of classifiers as inter-class non-semantic attributes which can classify the two halves best.

3.2.2. Hierarchical Object Models

Since the hierarchy is an important characteristic of human cognition, many research projects naturally utilize it for higher efficiency. The work based on object hierarchies can be roughly divided into three groups. The most popular work using hierarchies is based on the hierarchical representation of object parts (Larlus

et al. (2010), Fidler and Leonardis (2007), Epshtein and Uliman (2005), Ahuja and Todorovic (2007), Ommer and Buhmann (2010), Todorovic and Ahuja (2008)). These works decompose an object into several parts. These parts are organized as a tree-like structure. Parts in each level can be assembled into a whole object. Larger parts locate at upper levels, while the number of parts there is smaller. Small parts locate at lower levels in larger numbers. In (Larlus et al. (2010), Ommer and Buhmann (2010)), the spatial relationship between parts is also considered. In (Fidler and Leonardis (2007), Ahuja and Todorovic (2007), Todorovic and Ahuja (2008)), parts can be generalized to multi-categories. These methods obtain a better performance than traditional methods of object recognition, however, they cannot be used to discover novel categories and organize category hierarchies since they are all category-specific methods.

The second group uses hierarchies for scene detection and representation (Sudderth et al. (2008), Fei-Fei and Perona (2005)). In these methods, objects and regions are used to represent a scene in a hierarchical manner. Different scenes may share similar objects and regions which can be drawn from the same 'theme' or 'topic' and which can be looked at as intermediate representations of the scene. The procedure of training these intermediate representations is separated from the procedure of training a concrete scene type. Thus training a concrete scene is not closely related to original image features and a scene classifier can be easily generalized to other scenes. However, with these methods it is still difficult to discover and correctly categorize new scene types since these methods need samples to train new scene classifiers.

Methods in these two groups are concerned with concrete categories of objects or scenes and do not aim at organizing the relationship among large numbers of categories. Methods in the third group use hierarchies to represent the relationship among different categories of objects or scenes (Marszalek and Schmid (2008), Marszalek and Schmid (2007), Zweig and Weinshall (2007), Kapoor et al. (2009), Li et al. (2010b)). Hierarchies are organized as tree-like structures in which each node corresponds to one concrete category (a leaf node) or one superordinate of several concrete categories (a middle level node). The hierarchy can be built manually (Zweig and Weinshall (2007)), or from existing semantic networks (Marszalek and Schmid (2007)), or directly from image data (Marszalek and Schmid (2008), Kapoor et al. (2009), Li et al. (2010b)). The first two methods of building hierarchies are obviously not flexible enough to represent dynamic hierarchies. In Marszalek and Schmid (2008) and Kapoor et al. (2009), methods can build different hierarchies for different training image sets and can also be extended to find novel categories, however, the summarization of image features for each node in the hierarchies is not mentioned. In Li et al. (2010b), each node in the built hierarchy has a semantic tag which provides a certain semantic meaning. However, this method does not take into account how to change this tag when the hierarchy is changed as novel categories emerge.

3.2.3. Novelty Detection

Given a mixture of familiar and unfamiliar objects in a scene, unfamiliar objects should be identified before discovering novel categories to which they belong. Usually, these objects are detected and located by regions or bounding boxes. Although the problem of distinguishing unknown regions from all regions has not directly been addressed in the recognition literature (Lee and Grauman (2012)), we can borrow some ideas from novelty detection that has been researched widely in the machine learning community.

In the field of machine learning, there is much work that aims at novelty detection, such as Smola et al. (2009), Vieira Neto and Nehmzow (2007), Hoffmann (2007), Blanchard et al. (2010), Weinshall et al. (2008), Wu and Ye (2009). The common way of novelty detection usually assumes the training data obtained from the nominal classes and trains a distribution to represent these normal data. Thus a new sample that locates at the region with densities lower than a certain threshold will be treated as novel pattern (Schölkopf et al. (2000), Markou and Singh (2003a), Markou and Singh (2003b)). The recent achievements of novelty detection change this traditional manner by adding abnormal data at the training stage, or employ a soft threshold technique to obtain an improved performance. In Blanchard et al. (2010) and Wu and Ye (2009), the abnormal data is added to the training set which yields a smaller sphere and larger margin for the learned model. Smola et al. proposed to use a reference measure to be given in the form of a sample from an alternative distribution instead of a fixed threshold (Smola et al. (2009)).

A method of novelty detection that is suitable for our hierarchical category model is presented in Weinshall et al. (2008), where the conflicting predictions of a sample in the general level and specific level indicate the "incongruent" pattern. It defined two label hierarchies: the part-membership hierarchy where a concept requires a conjunction of parts, and the class-membership hierarchy where a concept is defined as the disjunction of more specific concepts. For different hierarchies, an observation resulting in these predictions is a novel pattern, if predictions satisfy the following conditions:

$$Q_a^g(X) \gg Q_a(X) \qquad \text{or} \qquad Q_a(X) \gg Q_a^s(X) \tag{3.12}$$

where $Q_a(X)$ denotes a probabilistic model of class a, $Q_a^s(X)$ is a probabilistic model of class a which is based on the probability of concepts in a more specific level than a, and $Q_a^g(X)$ represents a probabilistic model of class a which is based on the probability of concepts in a more general level than a. Thus the terms in eq. (3.12) give the novel pattern detection for the part-membership hierarchy (left) and class-membership hierarchy (right).

Since our category model in this thesis is also a hierarchy, this method is naturally employed for identifying novel categories. However, this method cannot be directly used for novel object category discovery, since it cannot tell us whether two novel patterns belong to the same category or different categories at all. And it is

inevitable that more than one novel category needs to be detected and recognized in our work. Therefore in this thesis we use the novelty detection to identify novel objects and employ our hierarchical category model to distinguish different novel categories.

3.3. Summary

In this chapter, we introduce diverse state-of-the-art methods related to the work presented in this thesis from different aspects. For some methods which are directly employed or extended in the presented work, we give a detailed introduction, including concepts and formulations. Furthermore, the limitations of these methods are also analyzed, explaining why some methods are not suitable for our work and some methods need to be extended for the presented work.

Chapter 4

Category-Independent Features based on Multimodal Data

4.1. Introduction

Feature is an important issue for category-specific object detection. Distinctive and representative features can simplify the complexity of models and achieve a better performance. For instance, some excellent features, such as SIFT (Lowe (2004)), SURF (Bay et al. (2006)) and MSER (Matas et al. (2004)) can yield good results even by a naive template matching method. To the problem of category-independent object detection, feature is also the key issue. Unlike those features used in category-specific object detection, category-independent object detection needs more generalizable features to describe generic objects. Thus the category-dependent features, such as aforementioned SIFT, SURF and MSER, cannot be directly used in this problem.

Color contrast, boundaries and edges are most often used category-independent features. Color contrast reflects the difference between an object and its surroundings in an image. The colors of an object usually have a certain consistency, and they are different from other objects and background. Thus its color contrast would be a good measure for generic object detection. However, when the appearance of the object is similar to others or the background is cluttered, the color contrast will lead to false results. And the boundaries and edges have similar shortages. They can efficiently detect simple objects in clear background. However, when the scene is complex, it is difficult to get complete boundaries for objects and the edges will be influenced by background. Consequently, it is necessary to develop more robust category-independent features for category-independent object detection.

Recently, visual saliency and superpixels are used to detect generic objects (e.g. Alexe et al. (2010)). They are computed independently to object categories and represent objects' common characters. The saliency highlights the foreground objects that are distinct to background. The superpixel straddling (Alexe et al. (2010),

Rahtu et al. (2011)) stands for concentrated foreground objects against spread background objects.

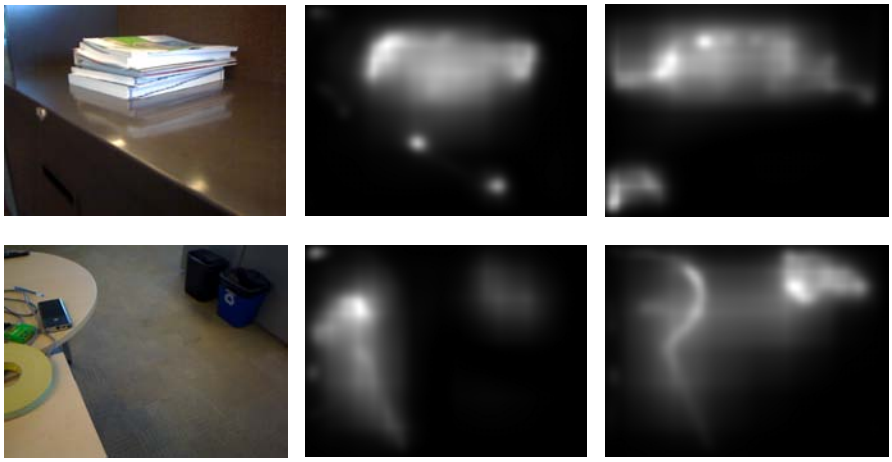


Figure 4.1.: 2D and 3D saliency maps. From left to right, the columns are the original images, 2D saliency maps, 3D saliency maps.

However, current visual saliency and superpixels are computed from only the color and brightness information of 2D images, which may lead to inaccurate results. Some foreground objects may have low saliency, while some background areas with surrounding noises may have high saliency. For example, as shown in Fig. 4.1, since the two dustbins in the second row have similar color with the background, their saliency is relative low. On the contrary, the two highlighted areas caused by specular reflection on the desk in the first row have high saliency.

Why the superpixel can be used as category-independent features is based on the assumption that all pixels in a superpixel belong to the same object (Russell et al. (2006)). Unfortunately, it is not always true in many cases. For example, in the top row of Fig. 4.2, the box marked by the red rectangle is segmented into a same region with a part of the floor in 2D oversegments, while in 3D oversegments it can be correctly organized. A similar situation can be found in the second row of the figure. This implies that simultaneously using 2D and 3D oversegments can yield good performance than using only one of them.

In general, there are two ways to incorporate 2D and 3D data for object detection. The first is to combine 2D and 3D information at the stage of extracting features. For example, 2D and 3D saliency maps can be integrated into one saliency map, and an oversegment method can combine colors and depth information to obtain RGB+D oversegments. Thus, the more accurate saliency maps and oversegments could be achieved. However, the interaction between 2D and 3D data is difficult to be modeled by this way, because the multimodal information is integrated into one saliency map and one oversegmentation, and the model cannot encode the interaction from the separated 2D features and 3D features. In our future work, we will investigate how to model the interaction from the features integrated color and depth information.

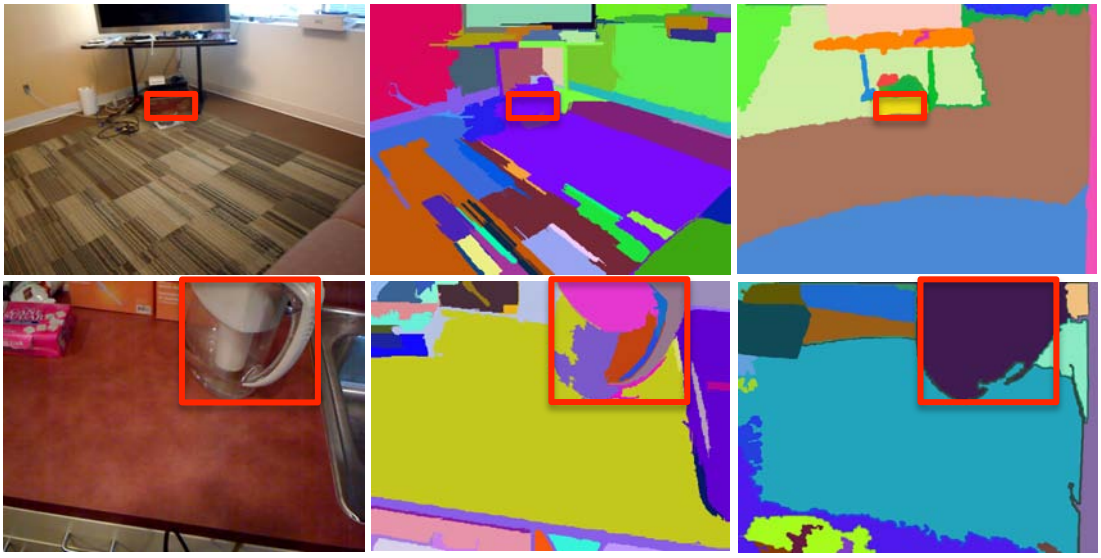


Figure 4.2.: Complementariness between 2D and 3D oversegments. From left to right, the columns are the original images, 2D oversegments, and 3D oversegments. The objects marked by red boxes have better segments in 3D oversegments.

In this thesis, we only concern the second way of incorporating 2D and 3D data, which obtains different features separately from different modalities and models the interaction of different modalities by integrating features of different modalities into one framework through some specific cross-modal terms (e.g. the cross-modal potentials introduced in the next chapter).

In this chapter, we develop a set of novel category-independent object features. These features are computed based on 2D and 3D saliency and oversegments. Extending 2D saliency and oversegments to 3D saliency and oversegments is introduced first. Then how to compute category-independent object features from them is presented. There are four kinds of features used in a novel cross-modal higher order conditional random field (CMH-CRF) model. This model will be introduced in the next chapter. In addition to the traditional features used in CRF model (i.e. the unary feature and pairwise feature), our model needs clique features and cross-modal features. All of these features have 2D and 3D versions, which are computed from 2D and 3D data respectively. Because clique features and cross-modal features are computed from several oversegments, it is difficult to directly evaluate their distinctiveness. Therefore, the experiments only show the distinctiveness of unary features and pairwise features, and the efficiency of clique features and cross-modal features are evaluated in the next chapter.

4.2. Saliency

The 2D saliency is computed by the algorithm proposed in Li et al. (2010d), and it is based on the lossy coding length of multivariate Gaussian data (Ma et al. (2007)). Given a set of vectors $\mathbf{w} = \{w_1, w_2, \cdots, w_M\} \in \mathcal{R}^{N \times M}$, a lossy coding scheme $L(\cdot)$ maps $\mathbf{w}$ to a sequence of binary bits $\widetilde{\mathbf{w}} = \{\widetilde{w}_1, \widetilde{w}_2, \cdots, \widetilde{w}_M\}$, from which the original vectors can be recovered up to an allowable distortion $E[||w_i - \widetilde{w}_i||^2] \leqslant \varepsilon^2$. If the data is i.i.d. sampled from a multivariate Gaussian distribution, the length of the encoded sequence is denoted by:

$$L_\varepsilon(\mathbf{w}) \doteq \frac{M+N}{2} \log_2 \det(I + \frac{N}{M\varepsilon^2} \bar{\mathbf{w}} \bar{\mathbf{w}}^T) + \frac{N}{2} \log_2(1 + \frac{\mu^t \mu}{\varepsilon^2}) \tag{4.1}$$

where $\mu = \frac{1}{M} \sum_{i=1}^{M} w_i$ and $\bar{\mathbf{w}} = [w_1 - \mu, w_2 - \mu, \cdots, w_M - \mu]$.

The saliency is then defined as the uncertainty of center regions with respect to their surroundings. A patch of image I and its surroundings are denoted by $c \in I$ and $S(c) = [s_1, s_2, \cdots, s_M]$, respectively. Let $SC(c)$ denote the union of c and $S(c)$, i.e. $SC(c) = S(c) \bigcup c$. Thus the saliency can be formulated as:

$$sal(c, S(c)) = L_\varepsilon(SC) - L_\varepsilon(S) \tag{4.2}$$

where $L_\varepsilon(S)$ is the lossy coding length of surroundings of patch c, and $L_\varepsilon(SC)$ is the lossy coding length of the area combining patch c and its surroundings. Here SC and S are represented by features. This method is insensitive to the features and the authors used the pixel values as features to obtain state-of-the-art results. The only parameter is the distortion ε which is tested in the experiments.

To extend this method to deal with 3D data, we use 3D information as features to represent SC and S. Saliency can be defined as the local contrast between a region and its surroundings. Intuitively, two kinds of 3D information, the normal direction and depth of each point, can be used as features to compute 3D saliency. When using normals as features, the object with sudden normal direction changes with respect to its surrounds will have high 3D saliency. For example, books on the table in Fig. 4.3have different normal directions from the desktop. When using depth as features, the object with different depth with respect to its surroundings will have high 3D saliency. For example, dustbins have different depth from the wall in Fig. 4.3. Finally, the 3D saliency map is obtained by combining the normal and depth saliency maps:

$$sal^{3d}(I) = (sal^{norm}(I) + sal^{depth}(I))/2 \tag{4.3}$$

In the experiments, we will show the promising performance of the 3D saliency. And the detection by combing 2D and 3D saliency is more robust than that of only using unimodal saliency.

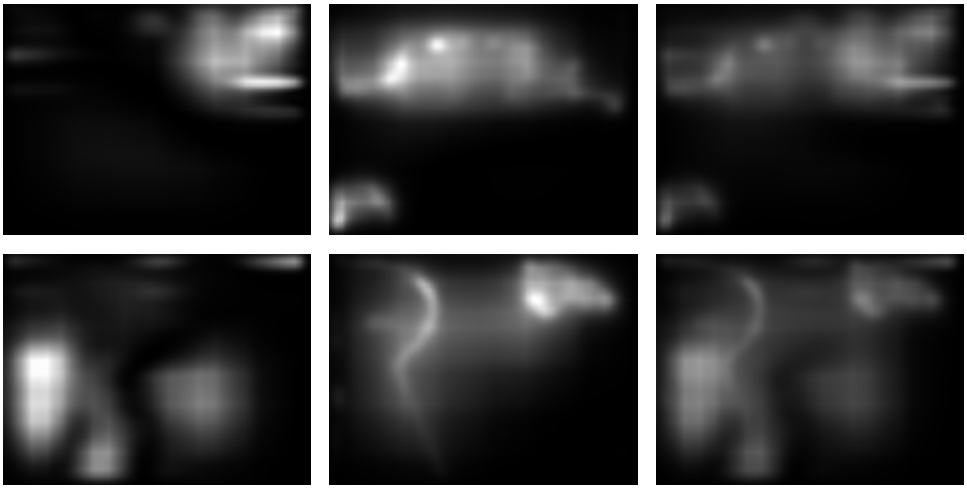
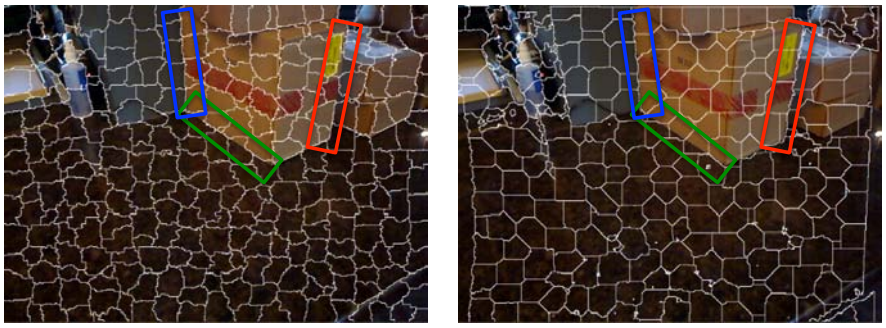


Figure 4.3.: Examples of 3D saliency corresponding to the original images shown in Fig. 4.1. From the left to the right, the saliency maps in the first and second column are computed from normals and depth, respectively. The last column are the final saliency maps.

4.3. Oversegmentation

That all pixels in a superpixel belonging to the same object (Russell et al. (2006)) is a basic assumption for many object detection methods (e.g. Alexe et al. (2010), Levinshtein et al. (2010)). Unfortunately, it is not always true in many cases as shown in Fig. 4.2. Therefore the 2D and 3D oversegments are computed for making using of their complementary effect. Since oversegments in two modalities are computed from different low level features, results of oversegments aligning to object boundaries are different for different modalities at the same position. For example, as shown in Fig. 4.4, oversegments in the red rectangle in (a) do not align the box boundaries well since two boxes have the same color and texture, but those in (b) align the boundaries well since in 3D point cloud data the point cloud of two boxes are separate and can be easily distinguished. However, as indicated by the green rectangle, oversegments in (a) align the object boundaries well while the alignments in (b) are not good, since the box and the tabletop have different color and texture, but have closed 3D point positions. The good alignments in both modalities are indicated by the blue rectangles since here the color, texture and 3D point position are obviously distinguishable.

Here the algorithm in Veksler et al. (2010) is employed to compute 2D oversegments and extended to compute 3D oversegments. This algorithm treats the segmenting task as a label assigned procedure. Given a set of pixels $\mathcal{P}$, a finite set of labels $\mathcal{L}$ and a neighbor system $\mathcal{N}$ which takes 8-connected grids, the segmentation is formulated as a problem of minimizing an energy function:



(a) (b)

Figure 4.4.: Difference of oversegments aligning to object boundaries for different modalities. The oversegments in (a) and (b) are computed from the 2D image and 3D point cloud data, respectively. The rectangles with the same color in two pictures indicate that the oversegments at the same position align the object boundaries differently.

$$E(f) = \sum_{p \in \mathcal{P}} D_p(f_p) + \sum_{\{p,q\} \in \mathcal{N}} w_{pq} \cdot V_{pq}(f_p, f_q) \tag{4.4}$$

where f denotes the label assignments of all pixels, f_p is the label assigned to p, $D_p(l)$ is the unary constraint which takes 1 if f_p is just equal to l or ∞ otherwise, $V_{pq}(f_p, f_q)$ denotes the smooth constraint which takes the Potts model $V_{pq}(f_p, f_q) = \min(1, |f_p - f_q|)$, and w_{pq} encourages discontinuities to coincide with intensity edges by taking the form:

$$w_{pq} = \exp(-\frac{(||I_p - I_q||^2)}{2\sigma^2 \cdot dist(p,q)}) \tag{4.5}$$

where I_p is the intensity of pixel p, $dist(p,q)$ denotes the Euclidean distance between p and q, and σ determines the penalty of discontinuity between p and q and can be referred as "camera noise" (Boykov and Funka-Lea (2006)).

The segments obtained by this algorithm have more regular size, which can align object boundaries better and lead to more regular connections as shown in Fig. 4.4. To extend this algorithm to compute 3D oversegments, normal directions and depth are used as features, instead of the the intensity of pixels as being used in eq. 4.5 to segment an image. Thus the w_{pq} for 3D oversegments is defined as:

$$w_{pq} = \exp(-\frac{(||NM_p - NM_q||^2 + ||DP_p - DP_p||^2)}{2\sigma^2 \cdot dist(p,q)}) \tag{4.6}$$

where NM_p and DP_p denote the normals and depth of point p. Another modification is that the neighbor system $\mathcal{N}$ consists of 16 nearest points in 3D space. Some results of 3D oversegments are illustrated in experimental section where we also compare

the boundary recall of using both 2D and 3D oversegments with that of using only 2D or 3D oversegments.

Note that in Fig. 4.4 the oversegment boundaries do not align well to picture edges. That is because the depth of points do not align exactly to image pixels due to the limitation of hardware (i.e. the Kinect camera). In Fig. 4.7, there are also some depth maps that do not align image edges well.

4.4. Unary Features

After obtaining 2D and 3D saliency and oversegments, the unary features can be computed for unary potentials in both modalities. Given a 2D segment s_i^I and 2D saliency map I_s for an image I, the 2D unary feature for this segment is defined as:

$$F_i^{u_I} = \frac{\sum_{j \in s_i^I} I_{s,j}/\sigma_j^2}{|s_i^I|} \tag{4.7}$$

where $|s_i^I|$ denotes the number of pixels in the segment s_i^I, $I_{s,j}$ is the saliency of the j-th pixel in s_i^I, and σ_j denotes the weight for the j-th pixel which is defined as $\sigma_j = d_j/3$ where the d_j is the Euclidean distance between pixel j and the mass of segment s_i^I. The 3D unary feature has the similar definition given a 3D segment s_m^T and 3D saliency map T_s:

$$F_m^{u_T} = \frac{\sum_{n \in s_m^T} T_{s,n}/\sigma_n^2}{|s_m^T|} \tag{4.8}$$

Note that in eq. (4.8) σ_n is computed from the Euclidean distance between point n and the mass of segment s_m^T in 3D space. In our experiments, we show the distinctiveness of the unary features.

After obtaining unary features for each segment in both 2D and 3D data, the probability of each variable assigned different labels can be computed. We choose a part of training data and train a classifier by multi-class SVM. Through this classifier, the probability can be obtained.

4.5. Pairwise Features

We use object boundaries to evaluate how likely two adjacent segments taking the same label. For the 2D image, the object boundaries are explicitly computed by the algorithm global probability of boundary (gpb) proposed in Maire et al. (2008). For the 3D point cloud data, since there is no an efficient method to explicitly extract object boundaries, we implicitly take their statistics into account when computing 3D pairwise features.

At first we explain how to formulate the 2D pairwise features. Give a gpb map B^I and two adjacent segments $\{s_i^I, s_j^I\} \in \mathbf{S}_I$, three kinds of features are computed

first for each point x in the edge $E_{i,j}$ between two segments, then integrate them into one feature whose average is the final 2D pairwise feature.

$$F_{i,j}^{p_I} = \frac{1}{|E_{i,j}|} \sum_{x \in E_{i,j}} (t_x \cdot (1 + \cos\theta_x)/d_x) \tag{4.9}$$

where d_x denotes the distance of a point x to the nearest boundary in gpb map B^I, t_x is the value of the nearest boundary in B^I, and θ_x is radian difference between the tangent to the edge point x and the tangent to the nearest boundary, as illustrated in Fig. 4.5. By this equation, the closer distance, higher gpb value and smaller radian difference result higher values for 2D pairwise feature which means there is higher probability of boundary between the two segments. Note that (4.9) also normalizes the 2D pairwise feature to the range $[0, 1]$.

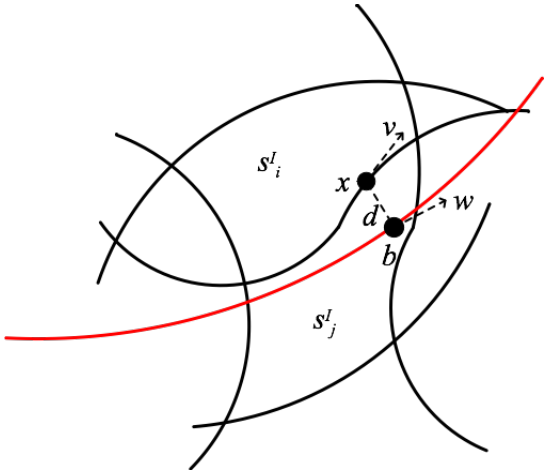


Figure 4.5.: 2D pairwise feature. The black curve between two segments is the edge with one point x marked. The red curve is the object boundary computed by gpb. b denotes the nearest object boundary to x. d is the distance. v and w are the tangent of x and b respectively.

Given two adjacent 3D segments $\{s_m^T, s_n^T\} \in \mathbf{S}_T$, three kinds of features are considered. The first is the Chi-square distance of point normal histogram between two adjacent segments. We transform a point normal from cartesian coordinates to spherical coordinates by following equations:

$$r = \sqrt{n_x^2 + n_y^2 + n_z^2}, \qquad \varphi = \arctan n_y/n_x, \qquad \theta = \arccos n_z/r \tag{4.10}$$

where $\varphi \in [0, 2\pi]$ and $\theta \in [0, \pi]$. Since r is always 1 for all point normals, the histogram of normals in one segment is constructed by only φ and θ as:

$$H_m^{norm}(i) = \frac{\sum |p_j|}{|s_m^T|} \qquad (p_j \in s_m^T, \{\varphi_j, \theta_j\} \in bin(i)) \tag{4.11}$$

where the histogram has 72 bins by dividing φ into 12 equal portions and dividing θ into 6 equal protions. Then we define the Chi-square distance of normal histogram of two adjacent segments as:

$$d_{m,n}^{norm} = \frac{1}{2}\sum_{i=1}^{72}\frac{(H_m^{norm}(i) - H_n^{norm}(i))^2}{H_m^{norm}(i) + H_n^{norm}(i)} \tag{4.12}$$

The second feature is computed from the difference of point cloud density between two segments, which is defined as:

$$\begin{aligned} d_{m,n}^{dens} &= \exp(-\frac{D_{m,n}}{D_m + D_n}) \\ D_{m,n} = \frac{|s_m^T|}{V_{box(s_m^T)}} \qquad & D_{m,n} = \frac{|s_{m,n}^T|}{V_{box(s_m^T + s_n^T)}} \end{aligned} \tag{4.13}$$

where D_m and D_n denote the point density of the segment s_m^T and s_n^T, $D_{m,n}$ denotes the joint point density of the joint segment $s_m^T + s_n^T$, and $V_{box}(s_m^T)$ is the volume of the minimal box containing all points in the segment s_m^T.

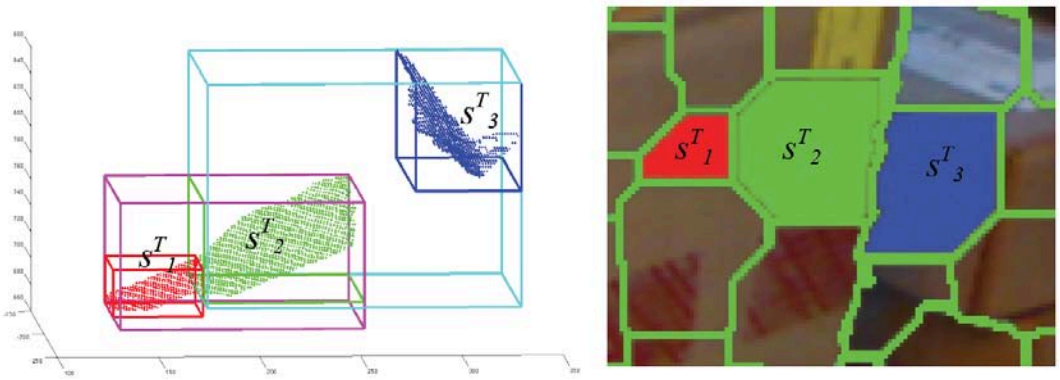


Figure 4.6.: Feature of the difference of point density. The right picture shows three adjacent segments whose point cloud in 3D space are shown in the left picture.

This feature is motivated by the following observation. If two adjacent segments are in the same object, their point density and their joint point density are more similar than that of two segments not in the same object. As shown in Fig. 4.6, considering three segments s_1^T, s_2^T and s_3^T. The former two segments are in the same object and the last one is in another object. It is obvious that s_1^T and s_2^T locate more closed and hence the different between $D_1 + D_2$ and $D_{1,2}$ is also smaller than that computed from s_2^T and s_3^T .

In general, the depth of points along with the edge of two segments changes gradually if two segments locate in the same object (e.g. s_1^T and s_2^T), otherwise the depth of points along with the edge changes suddenly (e.g. s_2^T and s_3^T). Motivated

by this intuition, we design the third feature that is computed from the difference of depth changes of points along with the edge of two adjacent segments. Given two adjacent segments s_m^T and s_n^T, the point along with the edge can be found more easily from 2D image as shown in Fig. 4.6. We define those points belonging to edge if the distance of the point to the edge of two segments is less than five pixels in 2D image. Then the third feature can be defined as:

$$\begin{aligned} D_{m_i} &= |\frac{\sum_{m_j \in e_m^T} d_{m_i,m_j}}{|e_m^T|} - \frac{\sum_{n_k \in e_n^T} d_{m_i,n_k}}{|e_n^T|}| \\ d_{m,n}^{depth} &= \frac{\sum_{m_i \in e_m^T} D_{m_i} + \sum_{n_i \in e_n^T} D_{n_i}}{|e_m^T| + |e_n^T|} \end{aligned} \tag{4.14}$$

where d_{m_i,m_j} is the 3D Euclidean distance of edge point m_i and m_j belonging to edge points in the s_m^T side, $|e_m^T|$ denotes the number of edge points. By this equation, the larger value of $d_{m,n}^{depth}$ means the higher probability of two segments locating in two objects.

Finally, we normalize three features to the range [0,1] by $(d_{m,n}^{feat} - \min(d^{feat}))/(\max(d^{feat}) - \min(d^{feat}))$ and compute the average of three features as the 3D pairwise feature.

$$F_{m,n}^{p_T} = (\widetilde{d}_{m,n}^{norm} + \widetilde{d}_{m,n}^{dens} + \widetilde{d}_{m,n}^{depth})/3 \tag{4.15}$$

where $\widetilde{d}_{m,n}^{feat}$ denotes the normalized value of $d_{m,n}^{feat}$. In our experiments, the distinctiveness of the 3D pairwise features will be evaluated.

4.6. Clique Features and Cross-modal Features

The clique potential is defined on an entire clique, so the clique feature is also computed from this clique. In the robust higher order CRF model (Kohli et al. (2009)), the higher order potential penalizes that the variables associated with segments in a clique do not have the same label. Therefore the clique feature will have higher value if all segments in a clique do not in the same object, otherwise the value is lower. A natural solution of clique feature is to count the boundaries in a clique, since the more boundaries there are, the more possible the clique straddles more than one objects. In this thesis, the 2D and 3D higher order features are defined as the mean values of all pairwise features between the segment center at a clique and all its adjacent segments:

$$\begin{aligned} F_{c_i}^{h_I} &= \frac{1}{|c|} \sum_{j \in c_i \backslash i} F_{i,j}^{p_I} \\ F_{c_m}^{h_T} &= \frac{1}{|c|} \sum_{n \in c_m \backslash m} F_{m,n}^{p_T} \end{aligned} \tag{4.16}$$

where $c_i \backslash i$ means all segments excluding s_i^I in clique c_i.

In the cross-modal higher order CRF model introduced in next chapter, the cross-modal potential also penalizes that the variables associated with segments in a cross-modal clique do not have the same label with the variable corresponding to this clique. Thus, the cross-modal potential is also defined as the mean values of all pairwise features in the cross-modal clique:

$$\begin{aligned} F^{h'_I}(c'_I, x_m^T) &= \frac{1}{|c'_I|} \sum_{i,j \in c'_I} F_{i,j}^{p_I} \\ F^{h'_T}(c'_T, x_i^I) &= \frac{1}{|c'_T|} \sum_{m,n \in c'_T} F_{m,n}^{p_T} \end{aligned} \tag{4.17}$$

where c'_I and c'_T are the cross-modal cliques corresponding to nodes x_m^T and x_i^I respectively, and $|c'_I|$ and $|c'_T|$ are the number of segments in the cross-modal cliques.

4.7. Experiments

In this section we evaluate the distinctiveness of the proposed features. The unary features and pairwise features are tested by using multi-class SVM. Experimental results show the promising distinctiveness. For the higher order and cross-modal features, since it is difficult to evaluate their distinctiveness, we do not explicitly test them but compare the performance of using or not using them in the evaluation of the whole cross-modal higher order CRF model in next chapter.

All experiments are based on a public RGB+D dataset (Lai et al. (2011)). The data in the RGB+D dataset is obtained by a Kinect style 3D camera. Some examples are shown in Fig. 4.7. The RGB and depth values are synchronized and aligned with 640×480 resolution. There are four classes of indoor scenes and 300 common household objects in it. Since the data is originally recorded in video sequences, we extract 200 images with significant difference in our experiments. Randomly selected 100 images are used as the training set and the rest is used for testing.

A set of parameters for computing the saliency and the oversegments needs to be set. In our experiments, considering the computational cost, we do not compute the saliency on the original images and point cloud data but on several down-sampled images and point cloud data to construct the multi-scale saliency. Particularly, three scales, $\{0.0625, 0.125, 0.25\}$, are used for both 2D and 3D saliency. For computing saliency, the distortion parameter ε need to be set. For computing oversegments, we need to set the parameters the allowed maximal patch size, and λ which controls the smoothness of boundaries.

First the results of saliency and oversegments under different parameters are tested to find the best parameters. For the parameter ϵ in (4.2), we evaluate different values to find the best parameters through which the 2D and 3D saliency maps yield the highest precisions and recall rates by considering pixels/points as objects if their

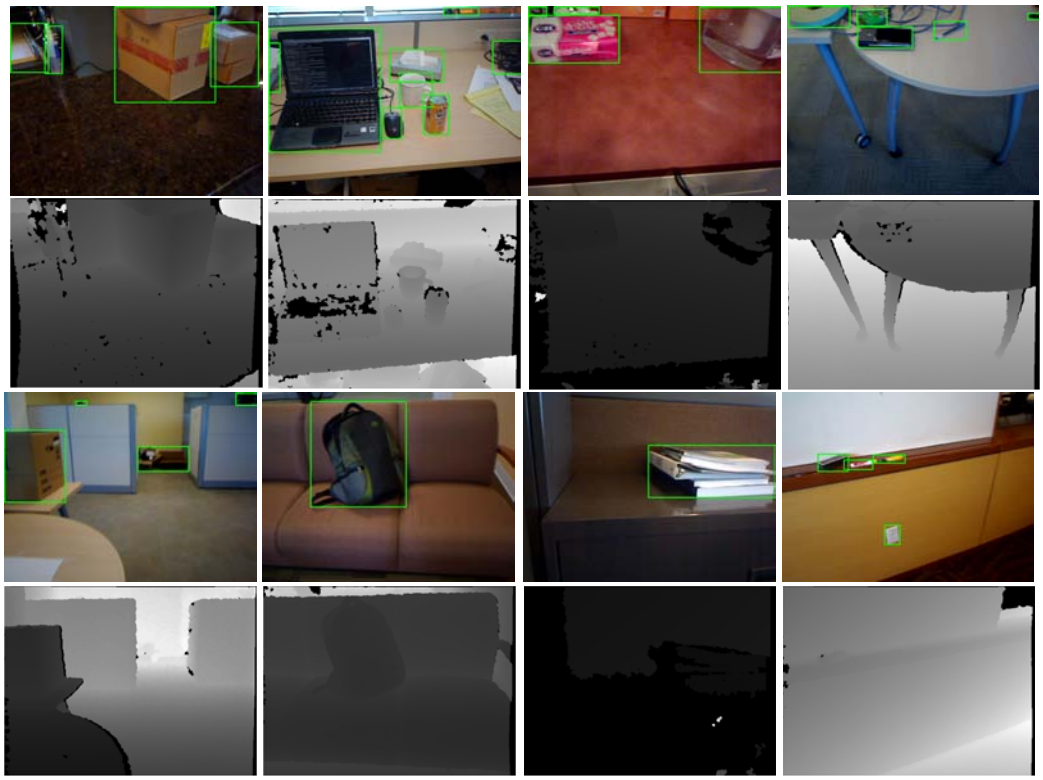


Figure 4.7.: Some examples in the RGB-D dataset used in our experiments. The first and third rows show some RGB images. The second and fourth rows show the corresponding depth maps. The bounding boxes indicate the annotated ground truth.

saliency is larger than a threshold $\tau \geqslant 0.6$. The parameter ε is variable from 0.05 to 0.8 with a interval 0.05. The precision and the recall rate are defined as

$$p = \frac{|A_{s\geqslant 0.6} \bigcap A_g|}{|A_{s\geqslant 0.6}|}, \qquad r = \frac{|A_{s\geqslant 0.6} \bigcap A_g|}{|A_g|} \tag{4.18}$$

where $A_{s\geqslant 0.6}$ is the region where the saliency of pixels or points is lager than 0.6, and A_g denotes ground truth regions.

We randomly select 50 images and corresponding 3D point cloud data from the training set to test the parameters. As shown in Fig. 4.8, the 2D saliency precision increases with the increment of the parameter ε. However, the recall is decreased when the value of ε increases. For the trade-off of precision and recall, therefore, we choose 0.5 as the value of ε for 2D saliency, by which the precision and recall are 0.72 and 0.8 respectively. Similarly, as shown in Fig. 4.9, ε is set to 0.8 for 3D saliency by which the precision and recall are 0.70 and 0.73 respectively. Furthermore, we combine the 2D and 3D saliency into one saliency map, and test if the combined saliency has higher precisions and recalls, which are illustrated in Fig.4.10. From this figure, it is obvious that the precision and recall of combined saliency maps are 1 and 3 percentage above that of single 2D and 3D saliency maps. Finally, we also show some 2D and 3D saliency maps in Fig. 4.11, from which the 2D and 3D saliency maps can be complementary for each other.

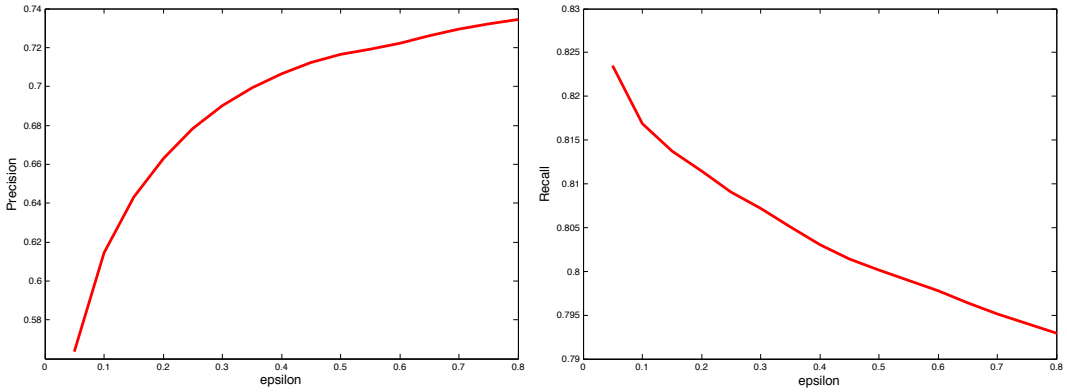


Figure 4.8.: 2D saliency precision and recall curves for different parameters.

For testing 2D and 3D oversegment parameters, we vary the patch size and λ from 10 to 80 and 10 to 100 with the same interval 10. The precision of oversegments is computed from the area of ground truth dividing by the area of oversegments overlapped with the region of ground truth. Fig. 4.12 shows the precision changes with different parameters. When the patch size is set to small values, the precisions is relative high. However, the number of oversegments is too large to fast infer the whole object detection model. Therefore, the patch sizes for 2D and 3D oversegments are both set to 20, which means the maximal oversegment not larger than 20×20. λ's for 2D and 3D oversegments are set to 40 and 10 respectively

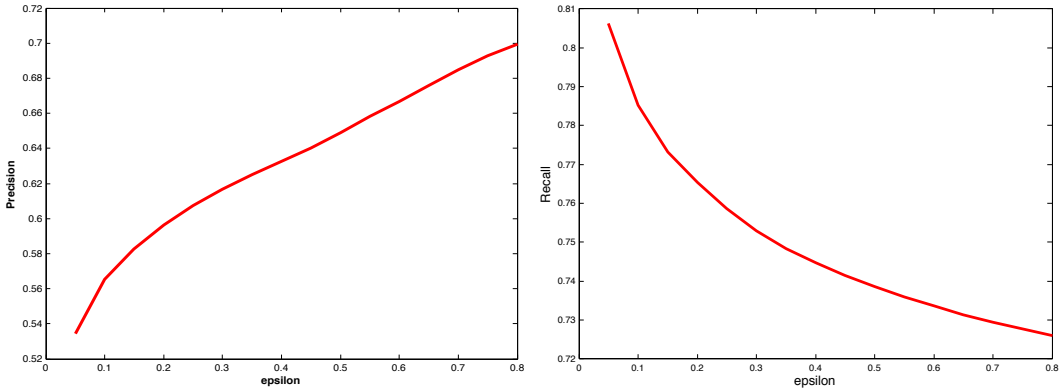


Figure 4.9.: 3D saliency precision and recall curves for different parameters.

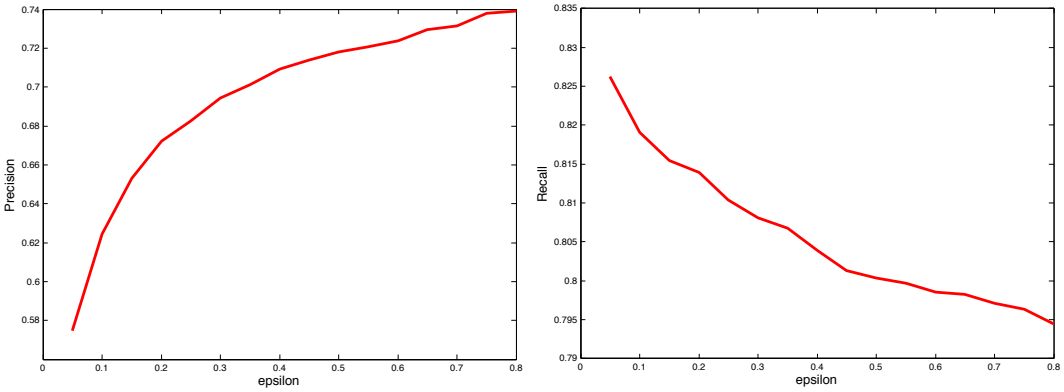


Figure 4.10.: Saliency precision and recall curves of combined saliency maps given different parameters.

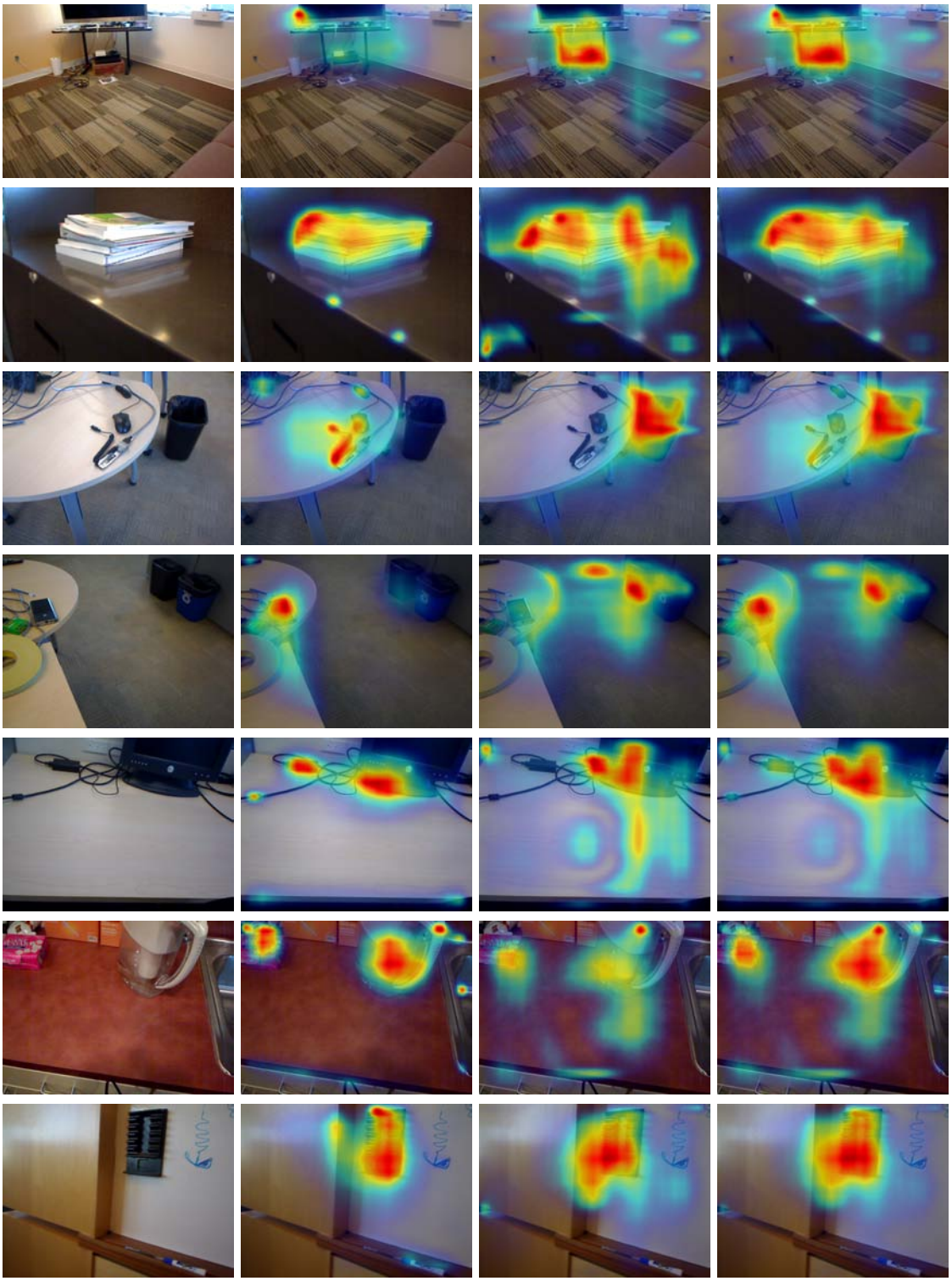


Figure 4.11.: Some examples of 2D and 3D saliency maps. From the left to right, the pictures are original images, 2D saliency maps, 3D saliency maps and combined saliency maps.

according to the precisions. As we explained in section 4.3, the 2D and 3D oversegments are complementary each other. Here an experiment is designed to prove it. After obtaining 2D and 3D oversegments for an image and its corresponding point cloud, we overlap them and obtain a new oversegment set with smaller oversegments. Since some oversegments cannot align to object boundaries well in one modality but can align well in another modality, theoretically, the overlapped oversegments can achieve better alignments than any unimodal oversegments. Actually, it is proved in our experiments as shown in Fig. 4.13, where the precision is improved averagely 20% for any parameters. Some 2D and 3D oversegments are also illustrated here in Fig. 4.14.

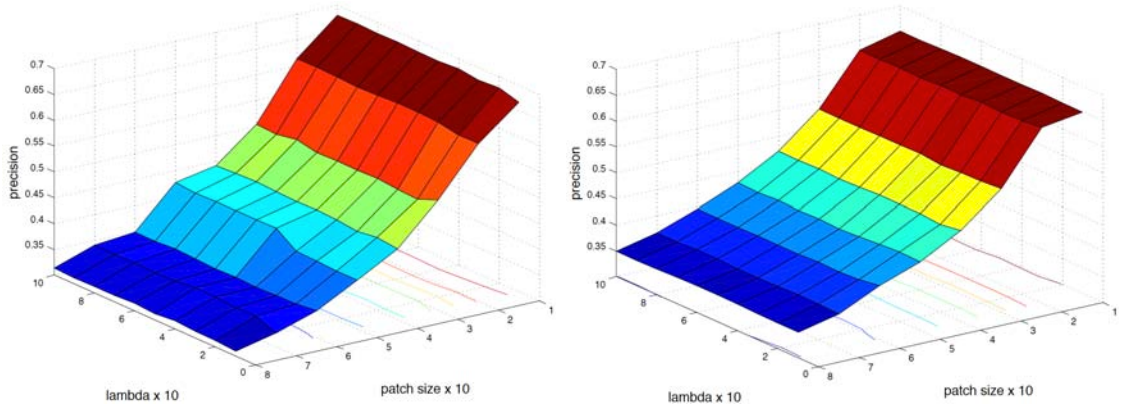


Figure 4.12.: Precision for 2D and 3D oversegments with different parameters. The left and right pictures show 2D and 3D oversegment precision repectively.

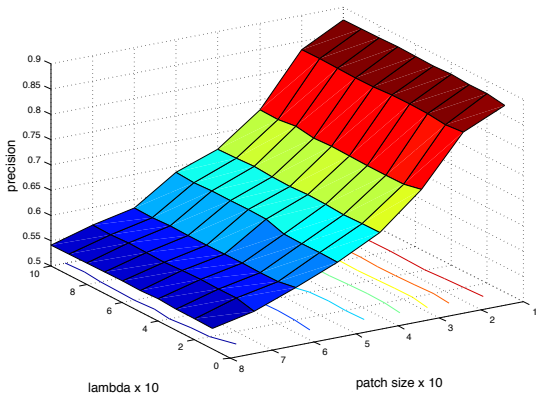


Figure 4.13.: Precision for overlapped oversegments with different parameters.

According to aforementioned experiments, we use the parameters that lead to best results for the following experiments. Now we can employ the multi-class SVM to evaluate the distinctiveness of unary and pairwise features. Randomly selected

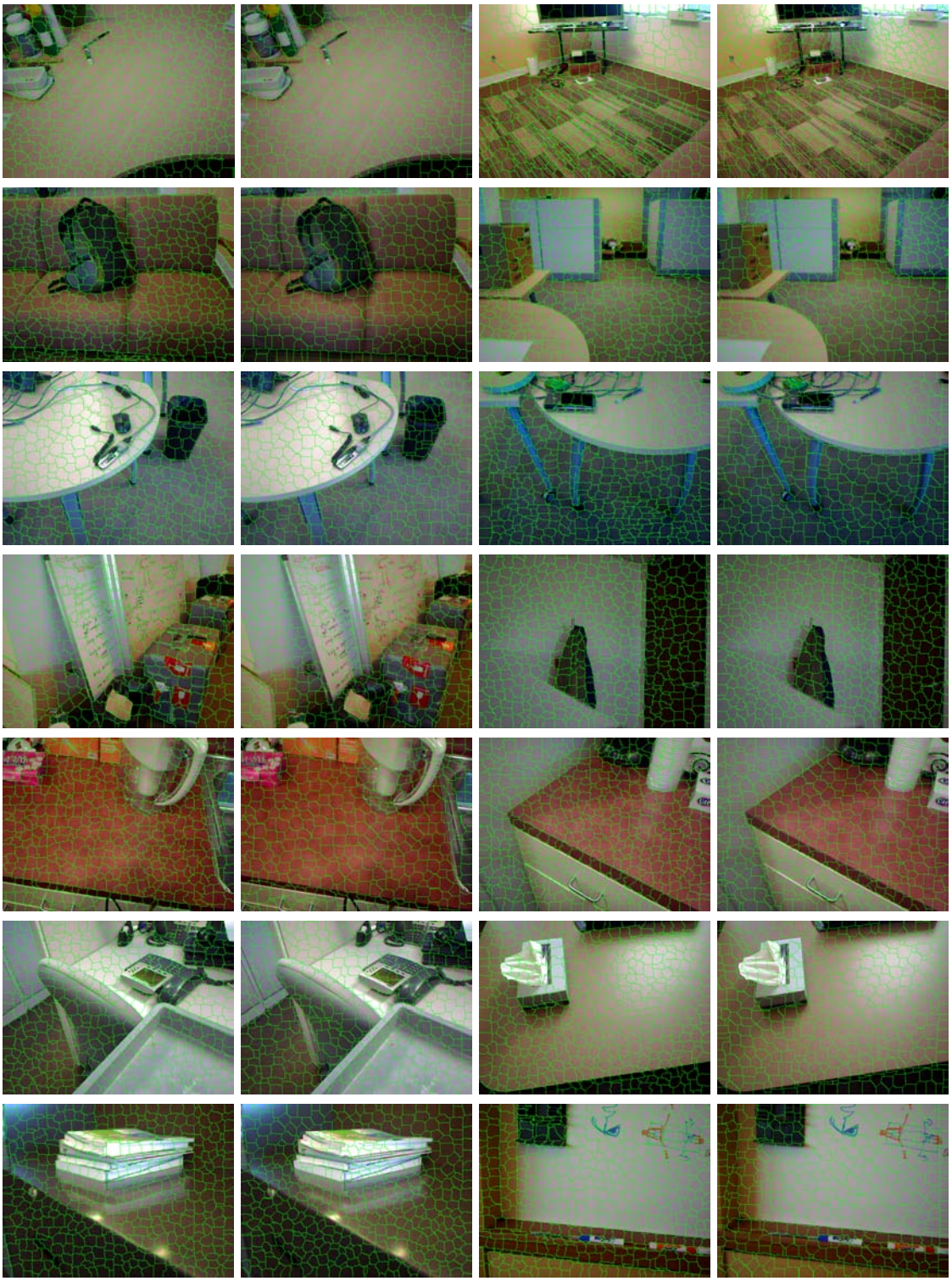


Figure 4.14.: Some examples of 2D and 3D oversegments. For each pair of pictures, the left one is the 2D oversegments and the right one is the 3D oversegments.

Table 4.1.: Precision of three-class SVM classifier for classifying unary features.

	Group1	Group2	Group3	Group4	Group5	Mean
O	0.84	0.83	0.86	0.84	0.85	0.845
B	0.82	0.81	0.82	0.80	0.81	0.812
G	0.85	0.84	0.85	0.83	0.82	0.836

Table 4.2.: Precision of six-class SVM classifier for classifying pairwise features.

	Group1	Group2	Group3	Group4	Group5	Mean
O-O	0.79	0.81	0.78	0.80	0.82	0.80
O-B	0.74	0.72	0.76	0.75	0.75	0.744
O-G	0.82	0.83	0.85	0.84	0.83	0.834
B-B	0.76	0.74	0.79	0.77	0.75	0.762
B-G	0.73	0.72	0.74	0.76	0.73	0.736
G-G	0.83	0.80	0.82	0.82	0.84	0.822

100 images and corresponding 3D point data are used for this evaluation. We obtain 490 average oversegments for each 2D image and 470 average oversegments for each 3D point cloud, which are divided into three classes corresponding to three kinds of labels. The unary features for all oversegments are computed first. Then half of unary features are randomly selected to train a three-class SVM classifier and the rest is used for evaluation. This experiment has been executed five times to obtain more convinced results. The results for both 2D and 3D unary features are listed in Table 4.1, where 'O', 'B' and 'G' refer to the label 'object', 'boundary' and 'background' respectively.

In the 100 training images and 3D point cloud, we averagely obtain 1220 and 1270 pairs of oversegments for 2D and 3D data respectively. For the pairwise feature, there are six classes of paired oversegments: 'object' to 'object' (O-O), 'object' to 'boundary' (O-B), 'object' to 'background' (O-G), 'boundary' to 'boundary' (B-B), 'boundary' to 'background' (B-G) and 'background' to 'background' (G-G). We randomly choose half of paired features to train a six-classes SVM classifier and the rest is used for evaluation. The experiments are also executed five times. The results are listed in Table 4.2. In this table, the precisions of classifying 'O-B', 'B-B' and 'B-G' are relatively low, that is because the boundaries between 'object' and 'boundary', 'boundary' and 'boundary', and 'boundary' and 'background' are more confusion. However, when using the CRF models, which take the spatial consistency into account, these confusion will be overcome and the results are better than the performance obtained by SVM.

As shown in these two tables, the best precisions of classifiers of unary and pairwise features are up to 86% and 85% respectively. The average precisions of them are

larger than 81.2% and 73.6%. From this, we can draw the conclusion that our unary and pairwise features have sufficient distinctiveness and are efficient for the proposed CMH-CRF model, which will be proved by experiments in the next chapter.

4.8. Conclusion

In this chapter we developed category-independent features used for the category-independent object detection model in the next chapter. To overcome the limitation of 2D saliency and superpixel, we extended algorithms to compute 3D saliency and oversegments. Based on 2D and 3D saliency and oversegments and boundary information, a set of novel features, including unary, pairwise, clique and cross-modal features, are developed. In the experiments, we first test the parameters for 2D and 3D saliency and oversegments. Then given the best parameters, we evaluated the distinctiveness of our unary and pairwise features, and proved that the new unary and pairwise features are efficient.

Chapter 5

Cross-Modal Co-segment for Category-Independent Object Detection

5.1. Introduction

To discover novel object categories from unexplored environments, the first thing is to detect and localize interesting objects in background. As a result, it is possible to recognize known objects and learn unknown objects. To detect and localize objects without recognizing them in scenes is an important ability of human beings. When doing this, the 3D information is indispensable. However, most of current category-independent object detection methods only consider the 2D images (e.g. Alexe et al. (2010), Rahtu et al. (2011), Endres and Hoiem (2010), Feng et al. (2011), Liu et al. (2011b)). Although some methods (e.g. Zhang et al. (2011), Collet et al. (2011)) utilize 2D and 3D data, they have not taken full advantage of the spatial consistency between 2D and 3D data. Furthermore, most of these methods are based on a procedure that ranks a lot of regions sampling from an image by measuring their object-likeness. They can achieve high object covering rates but the accuracy is very low, and consequently cannot be directly used for object localization.

In this chapter, we focus on detecting category-independent objects based on both 2D and 3D data. As convenient devices that can simultaneously obtain RGB+D (red, green, blue, + depth) data emerge, such as Microsoft Kinect, they provide a chance to computer vision system with efficiently integrating 3D and 2D data and improving object detection. Given a RGB+D image, a point belonging to an object in 2D space must have a corresponding point in 3D space belonging the same object. Furthermore, the 3D space points near to the corresponding point also have high probabilities to belong to the same object. It means that there must be spatial consistency between 2D and 3D data, and it motivates a novel method that treats the category-independent object detection as a cross-modal co-segmentation problem.

The co-segmentation is first proposed by (Rother et al. (2006)), which is defined as segmenting the same or similar regions in a pair or group of images by using the global constraint of all segmented images. After this, more and more methods are presented to improve its efficiency and performance (e.g. Vicente et al. (2011), Mukherjee et al. (2011), Vicente et al. (2010), Glasner et al. (2011)). The 2D image and corresponding 3D point cloud data also have some global constraints, so we believe that to simultaneously segment 2D images and 3D point data can improve the category-independent object detection. However, current co-segmentation methods are only based on 2D images, and do not explicitly detect object-likeness regions but unsupervised segment the same or similar regions among images. Therefore they cannot be directly applied to category-independent object detection.

We propose a novel co-segmentation method that overcomes these limitations and obtains promising category-dependent object detection in both 2D and 3D space. Fig. 5.1 shows the basic idea of the proposed method. This method is a kind of novel conditional random fields, namely Cross-Modal Higher-order Conditional Random Field (CMH-CRF) model, which uses the 2D superpixels and 3D supervoxels as basic nodes for decreasing computational costs. The CMH-CRF model takes the 2D and 3D potentials, and the cross-modal potentials in a uniform model. The 2D saliency, superpixels, and boundaries, and 3D saliency and supervoxels are computed first as features. Then the unary, pairwise and higher-order clique potentials are carried out. Next we compute the cross-modal higher-order potentials. After inferencing the CMH-CRF model, the 2D and 3D labeled results are simultaneously obtained. The pixel-wise results can then be produced by combining 2D and 3D results at pixel-level. The final results that bound object instances by boxes can be carried out by a simple algorithm. Unlike the general figure/ground segmentation methods that use 'object' and 'background' labels (e.g. Carreira and Sminchisescu (2012), Ion et al. (2011), Ren et al. (2006)), one more label 'boundary' is used in our method. Benefited from using three labels, object instances can be easily distinguished from segmenting results. We evaluate the proposed method in a public RGB+D dataset (Lai et al. (2011)). The experimental results show that the proposed method outperforms state-of-the-art category-independent object detection methods.

5.2. Overview of the Approach

The overview of the proposed CMH-CRF model is shown in Fig. 5.1. Given multi-modal data (i.e. 2D images and 3D point clouds), the 2D saliency and oversegments are computed by the methods proposed in Li et al. (2010d) and Veksler et al. (2010) respectively. Then we extend them to compute 3D saliency and oversegments. For 2D images, we also employ global probability of boundary (gpb) algorithm (Maire et al. (2008)) to compute object boundaries which are used to compute the 2D pairwise potentials. The 2D and 3D saliency can complement each other. In one modal data some regions may obtain incorrect saliency, while in the other modal data their

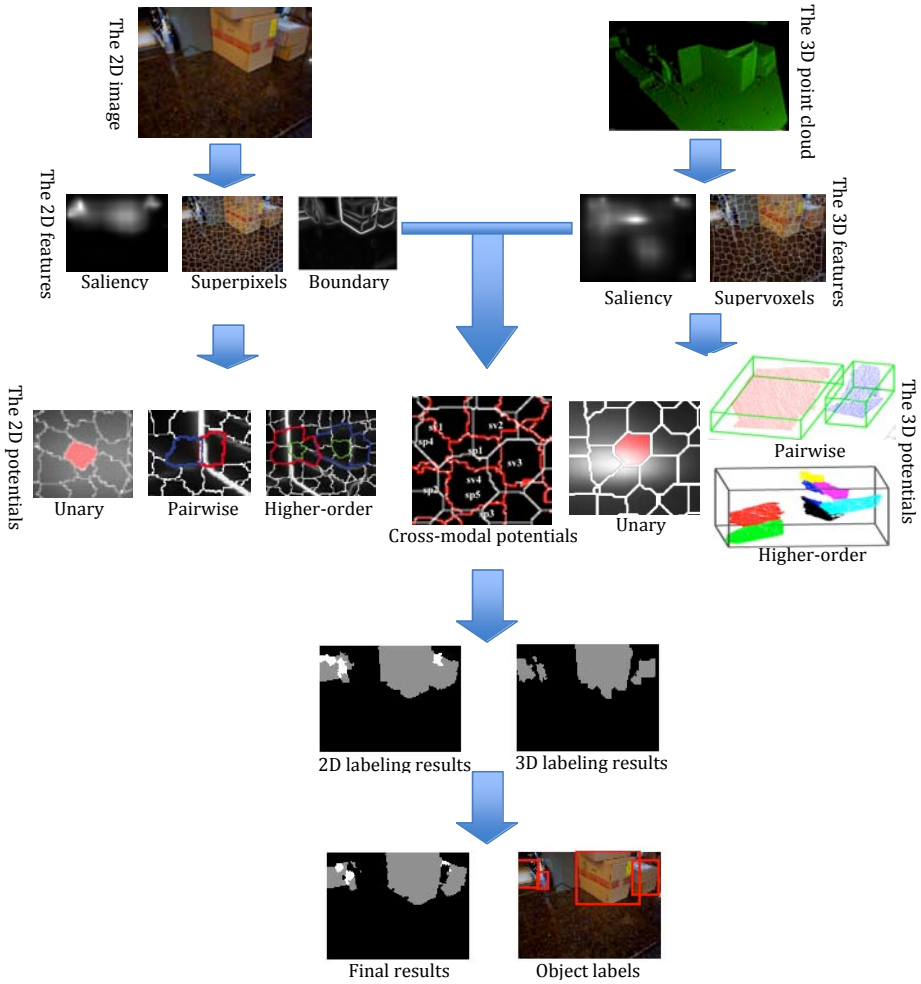


Figure 5.1.: Overview of the proposed method. From the first row to the third row, there are the original image and point cloud, their features and their potentials used in the proposed CMH-CRF model. In the fourth row, 2D and 3D labeling results are shown, where the black regions are backgrounds, the gray regions are objects, and the white regions are boundaries. By combining 2D and 3D results at pixel level, the final results are obtained and shown in the fifth row. Since three kinds of labels are used, object instances can be easily distinguished as single object instances.

saliency may be correct, which will enhance the robustness of saliency. The same complemental effect can also be found in 2D and 3D oversegments. Therefore, the proposed CMH-CRF model takes multimodal features to improve the robustness of detection.

The oversegments of two modalities are used as the basic node of CMH-CRF model, based on which we compute the unary, pairwise, higher-order and cross-modal potentials. The unary potentials in two modalities are computed from the averages of saliency of oversegments by support vector machine (SVM). The pairwise potentials in two modalities are computed by considering the boundaries between a pair of oversegments. Then for each oversegment, we construct a clique by all of its adjacent oversegment, and the higher-order potentials are computed from the boundaries among them. Given an oversegment in one modality, the oversegments in another modality overlapped by this oversegment are considered that have high probabilities of holding the same label. Thus the cross-modal potentials are computed by the boundaries among these overlapped oversegments. Then we use the swap and expansion moves algorithm to simultenously infer labels for all oversegments in two modalities. At last, the final label results are obtained by combining the labeled maps of two modalities at pixel level.

The CMH-CRF model defines three kinds of labels for oversegments. The object, boundary and background label refer to the locations inside an object, straddling object boundaries and outside an object respectively. Thus, the object instances can be accurately separated from the label results. Considering that a segment may not align the object boundary well, the concrete definition of three kinds of labels is as follows. If more than 80% pixels in an oversegment are within an object or background, the variable associated with this oversegment is regarded as taking the 'object' or 'background' label. If an oversegment straddles an object and the background or two objects (i.e. the pixels locate in one object or background is less than 80% of the whole oversegment), the variable associated with this oversegment takes the 'boundary' label.

5.3. Formulation of Single Modality

Given an image I and a corresponding point cloud T, we represent them by two sets of oversegments $\mathbf{S}_I = \{s_1^I, s_2^I, \cdots, s_{N_I}^I\}$ and $\mathbf{S}_T = \{s_1^T, s_2^T, \cdots, s_{N_T}^T\}$, where N_I and N_T are the number of oversegments of the image I and the point cloud T respectively. We first in this section formulate the higher-order conditional random field in 2D image case, and then extend it to cross-modal case by integrating 3D and cross-modal terms in the next section.

Given an image I, consider a discrete random field $\mathbf{X}_I$ defined over a set of vertices $\mathcal{V}_I = \{1, 2, \cdots, N_I\}$ with a neighbourhood system ε_I, where each vertex $i \in \mathcal{V}_I$ corresponds to an oversegment $s_i^I \in \mathbf{S}_I$. Each random variable $X_i^I \in \mathbf{X}_I$ associated with a vertex $i \in \mathcal{V}_I$ will take a value from the label set $\mathcal{L}_I = \{l_1, l_2, \cdots, l_k\}$, where in this study $k = 3$ for three kinds of labels. The neighborhood system ε_I consists of

variables directly connected by edges in the random field. A set of random variables $\mathbf{X}_c^I$ which are conditionally dependent on each other forms a clique c. Each random variable will be assigned a label and the configuration of labels is denoted by $\mathbf{x}_I$ which takes values from the set $\mathbf{L} = \mathcal{L}_I^N$.

Based on these notations, the CRFs framewrok (Lafferty et al. (2001)) models the probability of a labeling configuration $\mathbf{x}_I = \{x_i^I\}$ as a Gibbs distribution given a set of oversegments $\mathbf{S}_I$ and can be written as $Pr(\mathbf{x}_I|\mathbf{S}_I) = \frac{1}{Z}\exp(-E(\mathbf{x}_I|\mathbf{S}_I))$, where Z is the partition function and $E(\mathbf{x}_I|\mathbf{S}_I)$ is the Gibbs energy which is defined in the RH-CRFs framework (Kohli et al. (2009)) as:

$$E(\mathbf{x}_I) = \sum_{i\in\mathcal{V}_I} \psi_i^I(x_i^I) + \sum_{(i,j)\in\varepsilon_I} \psi_{ij}^I(x_i^I, x_j^I) + \sum_{c\in\mathcal{C}_I} \psi_c^I(\mathbf{x}_c^I) \tag{5.1}$$

Here $\mathcal{V}_I$ is the set of all oversegments and ε_I is the set of all edges, where each edge connects two vertex $i, j \in \mathcal{V}_I$. A clique c is defined over a segment i and all adjacent segments connecting to it. Fig. 5.2 shows the graphical model of the CMH-CRF model in unimodal case. As shown by red edges, nodes in this model do not take the regular connections of 4 or 8 neighbors that are usually used in pixel-level image segmentation.

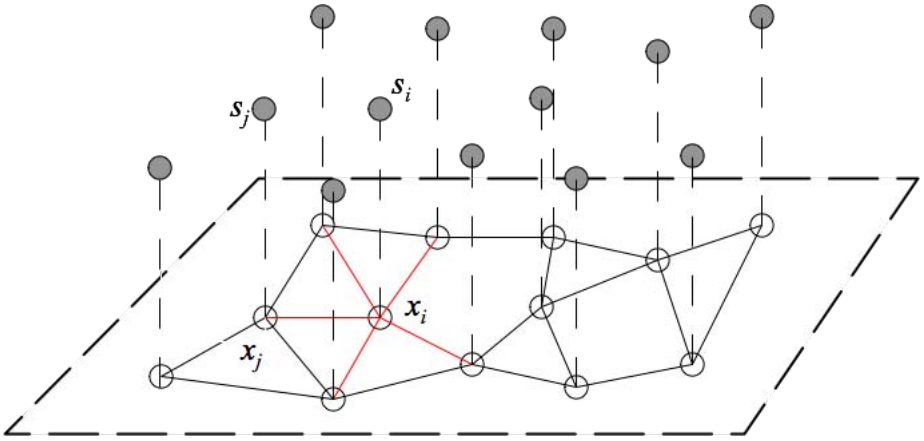


Figure 5.2.: Undirected graphical representation in a unimodal case. The shadow circles are the observed segments, while the corresponding white circles are the random variables indicating their labels. x_i^I and x_j^I are neighbors connected by an edge. An example of the clique c is represented by all vertices connected by all red edges.

The unary potential ψ_i^I in (5.1) is defined as the negative log of the likelihood of a label being assigned to segment s_i^I. In category-specific object detection methods, the unary potential is usually computed from color, texture, location and shape priors as shown in Kohli et al. (2009) and Shotton et al. (2009). However, when we aim at category-independent object detection, these category-specific features cannot be generalized well to represent category-independent objects. Therefore we

employ the visual saliency as category-independent features from which the unary potential of this model is computed. For each segment, its unary potential is defined as:

$$\psi_i^I(x_i^I) = -\log(p(x_i^I|F_i^{u_I})) \tag{5.2}$$

where the probability $p(x_i^I|F_i^{u_I})$ is computed by multi-class support vector machine (Chang and Lin (2011)) and $F_i^{u_I}$ is the saliency feature of the i-th segment in an image, computed as the weight sum of saliency within this segment.

The pairwise potential $\psi_{ij}^I(x_i^I, x_j^I)$ is also computed from category-independent features, instead of simple edge features based on the difference in colors, which has been widely used in category-specific image segmentation and object detection. In this study we employ boundaries as edge features, from which the pairwise potential is defined as the Potts model:

$$\psi_{ij}^I(x_i^I, x_j^I) = \begin{cases} 0 & \text{if } x_i^I = x_j^I, \\ \theta_p^I + \theta_v^I \exp(-\theta_\beta^I ||F_{i,j}^{pI}||^2) & \text{otherwise,} \end{cases} \tag{5.3}$$

where $F_{i,j}^{pI}$ is the pairwise feature between segments i and j, which is computed from the value of gpb of pixels along with the edge between two segments. The model parameters θ_p^I, θ_v^I and θ_β^I are learned from training data.

Kohli et al. (2009) proved that the higher-order CRF model can obtain better performance than that of the pairwise CRF model by adding an extra higher-order potential, and provided an efficient algorithm to solve it. In this study, we also use the robust P^n potential which is defined as:

$$\psi_c^I(\mathbf{x}_c^I) = \begin{cases} N_i(\mathbf{x}_c^I)\frac{1}{Q}\gamma_{max} & \text{if } N_i(\mathbf{x}_c^I) \leqslant Q, \\ \gamma_{max} & \text{otherwise,} \end{cases} \tag{5.4}$$

where $N_i(\mathbf{x}_c^I)$ denotes the number of nodes which have different labels to the dominant label in the clique c, and can be calculated by $N_i(\mathbf{x}_c^I) = \min_k(|c| - n_k(\mathbf{x}_c^I))$. Here $|c|$ is the number of all nodes in clique c and $n_k(\mathbf{x}_c^I)$ is the number of nodes taking label l_k. The truncation parameter denoted by Q is used to control the rigidity of this higher order potential. The γ_{max} is defined as:

$$\gamma_{max} = |c|^{\theta_\alpha^I}(\theta_{hp}^I + \theta_{hv}^I \exp(-\theta_{h\beta}^I ||F_c^{hI})||^2)) \tag{5.5}$$

where F_c^{hI} is the clique feature which is computed from object boundaries on all segments belonging to clique c. The parameters θ_α^I, θ_{hp}^I, θ_{hv}^I, and $\theta_{h\beta}^I$ are also learned from training data.

Here we emphasize why higher order potentials are necessary for labeling with oversegments. The pairwise potential which takes Potts model is a kind of hard constraint since it encourages two variables taking the same label and makes the CRF model favor smooth object boundaries. Although this smoothness potential sometimes over-smoothes object boundaries, the labeling results are still promising since the boundaries occupy a very small portion in the whole image. However,

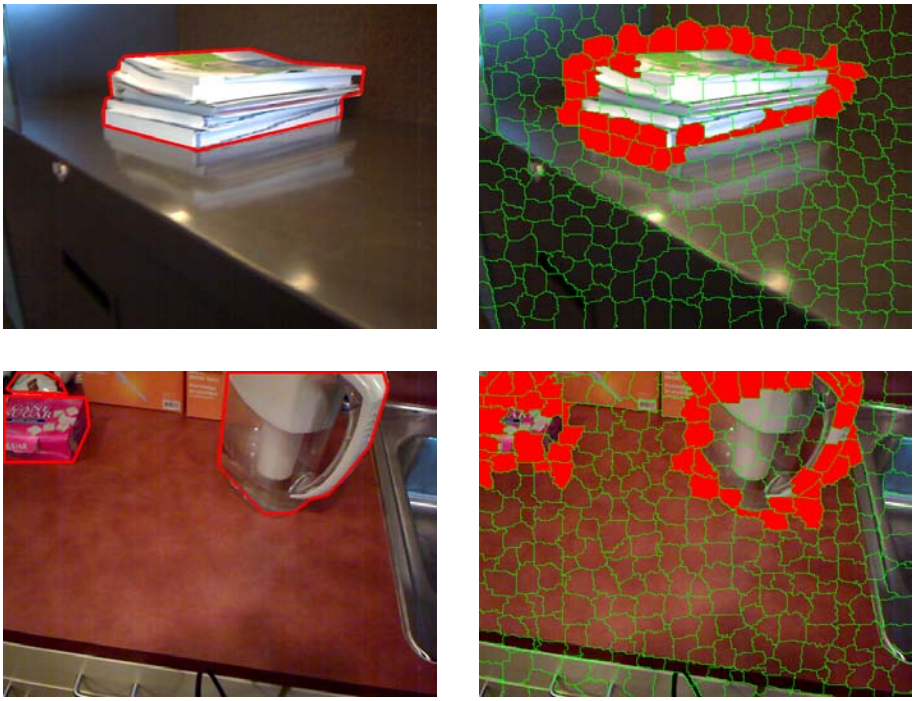


Figure 5.3.: Different proportion of two adjacent variables taking different labels or the same label. The red color means object boundaries. The green curves show the boundaries of oversegments

Table 5.1.: Different proportion of two adjacent variables taking different labels or the same label.

		Boundary nodes	All nodes	Ratio
Fig. 5.3a	Pixels	912	307200	0.0030
	Segments	33	307	0.0675
Fig. 5.3b	Pixels	1324	307200	0.0043
	Segments	48	324	0.0807

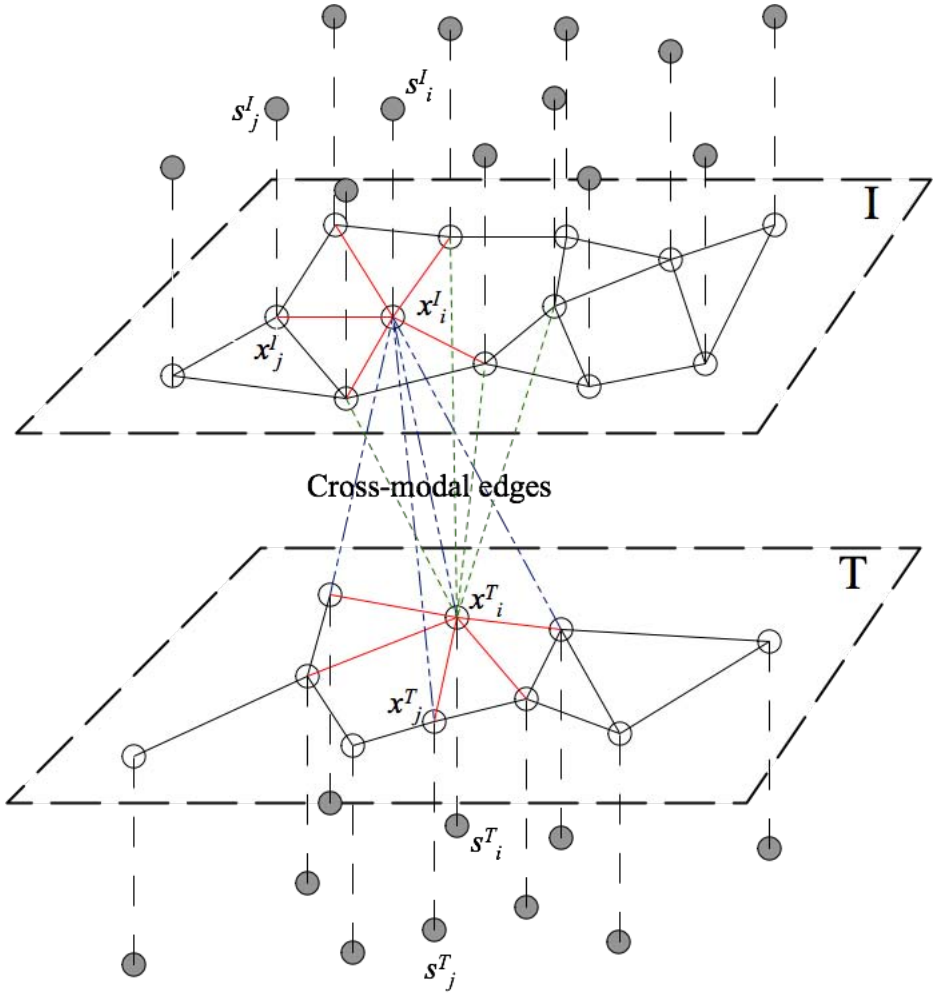


Figure 5.4.: Undirected graphical representation in a cross modality case. Two layers are in this model where the 'I' layer corresponds to image data and the 'T' layer means the 3D point cloud data layer. One node in a layer connects to other nodes that are not only in the same layer but also in another one. The blue and green dash lines are the edges between two layers which mean the interaction of related nodes across modalities. See text for the details about this model.

when using a set of oversegments to represent an image, over-smoothness decreases the accuracy of labeling results, since the portion of oversegments located at object boundaries with respect to all oversegments is much larger than that at pixel level. As illustrated in Fig. 5.3 and Table 5.1 where two images are both in 640×480 resolution, the portion of boundaries at pixel level is only 0.3%-0.4%, while the portion of boundaries at oversegment level is twenty times more than it. Unlike pairwise potentials, robust higher order potentials take a soft constraint and allow a part of nodes in a clique assigned different labels from the dominant label. Thus, using the robust higher order potential can preserve boundaries better as shown in Kohli et al. (2009). Therefore in this study the CMH-CRF model employs higher-order potentials to obtain better capabilities to keep object boundaries, which not only lead to higher precision of labeling results, but also improve the results being separated into single object instances.

5.4. Cross-Modal Higher-order CRF Model

Now we extend the oversegments based robust higher order CRFs to the cross-modal robust higher order CRF model. Kohli et al. (2009) only handled the single modal data, i.e. 2D images. In this study 2D images and 3D point cloud are labeled within a uniform CRF model, so potentials for different modalities will be integrated into (5.1). Furthermore, we also consider the global spatial consistency between an image and its corresponding 3D point clouds, therefore a set of novel cross-modal higher order potentials are designed to be also combined into (5.1). Finally the proposed CMH-CRF model is defined as:

$$\begin{aligned} E(\mathbf{x}_I, \mathbf{x}_T) = & \sum_{i \in \mathcal{V}_I} \psi_i^I(x_i^I) + \sum_{(i,j) \in \varepsilon_I} \psi_{ij}^I(x_i^I, x_j^I) + \sum_{c_I \in \mathcal{C}_I} \psi_{c_I}^I(\mathbf{x}_{c_I}^I) + \\ & \sum_{m \in \mathcal{V}_T} \psi_m^T(x_m^T) + \sum_{(m,n) \in \varepsilon_T} \psi_{mn}^T(x_m^T, x_n^T) + \sum_{c_T \in \mathcal{C}_T} \psi_{c_T}^T(\mathbf{x}_{c_T}^T) + \\ & \sum_{x_i^I \in \mathcal{V}_I, c_T' \in \mathcal{C}_T'} \psi_c^R(x_i^I, \mathbf{x}_{c_T'}^T) + \sum_{x_m^T \in \mathcal{V}_T, c_I' \in \mathcal{C}_I'} \psi_c^R(x_m^T, \mathbf{x}_{c_I'}^I) \end{aligned} \tag{5.6}$$

Here the first three items at the right hand side have been stated in (5.1). The following three items are the unary, pairwise and higher order potentials, which are computed from the 3D point cloud. The last two items are the cross-modal potentials which are motivated by the cross-modal spatial consistency of objects. Fig. 5.4 shows the graphical representation corresponding to this CMH-CRF model. There are two layers in the model, one for image data and another for 3D point cloud data. They are both represented by oversegments and therefore the connections between variables are not regularized. The potentials for 3D point cloud data are defined on the 'T' layer and have the similar formulation with those for image data. The 3D

unary potential is also the negative log probability of a label being assigned to the oversegment s_m^T:

$$\psi_m^T(x_m^T) = -\log(p(x_m^I|F_m^{u_T})) \tag{5.7}$$

where $F_m^{u_T}$ is the 3D saliency feature of the segment s_m^T and the probability $p(x_m^I|F_m^{u_T})$ is also computed by SVM. The pairwise potential of 3D modality is defined as:

$$\psi_{mn}^T(x_m^T, x_n^T) = \begin{cases} 0 & \text{if } x_m^T = x_n^T, \\ \theta_p^T + \theta_v^T \exp(-\theta_\beta^T ||F_{m,n}^{pT}||^2) & \text{otherwise,} \end{cases} \tag{5.8}$$

where the $F_{m,n}^{pT}$ is the pairwise feature between two 3D segments, s_m^T and s_n^T, which is a novel set of 3D boundary features developed in last chapter. The parameters θ_p^T, θ_v^T and θ_β^T are learned from training data like the parameters in (5.3). The 3D higher order potential is defined as:

$$\psi_{c_T}^T(\mathbf{x}_{c_T}^T) = \begin{cases} N_m(\mathbf{x}_{c_T}^T)\frac{1}{Q}\gamma_{max}^T & \text{if } N_m(\mathbf{x}_{c_T}^T) \leqslant Q, \\ \gamma_{max}^T & \text{otherwise,} \end{cases} \tag{5.9}$$

where $N_m(\cdot)$ and Q have the same meaning with those in (5.4). The γ_{max}^T is also defined similarly to (5.5):

$$\gamma_{max}^T = |c_T|^{\theta_\alpha^T}(\theta_{hp}^T + \theta_{hv}^T \exp(-\theta_{h\beta}^T ||F_{c_T}^{hT}||^2)) \tag{5.10}$$

where $F_{c_T}^{hT}$ is the 3D clique feature computed from 3D object boundaries on all segments belonging to clique c_T and the parameters $\theta^T I_\alpha$, θ_{hp}^T, θ_{hv}^T, and $\theta_{h\beta}^T$ are learned from training data.

The 3D potentials have the similar definition with 2D potentials. Their difference is that they are computed from different modal data. Besides combining the potentials of each modality, the interaction between two modalities is also taken into account in the proposed model, i.e. the cross-modal potentials. We use oversegments to represent an image and its corresponding 3D point cloud. Although they are synchronized well, 2D oversegments do not put into one-to-one correspondence with 3D oversegments. It means that a segment in the image overlaps to one or more segments in the 3D point cloud data if we overlay the segments of two modalities. Furthermore, since oversegments in two modalities are computed from different low level features, the results of the oversegments aligning to object boundaries are different for different modalities at the same position. For example, as shown in Fig. 4.4, the oversegments in the red rectangle in (a) do not align the box boundary well since two boxes have the same color and texture, but those in (b) align the boundary well since in 3D point cloud data the point cloud of two boxes are separated and can be easily distinguished. However, as indicated by the green rectangle, oversegments in (a) align the object boundary well while the alignment in (b) is not good, since the box and the tabletop have different color and texture, but have closed 3D point positions. The good alignments in both modalities are

indicated by the blue rectangles since here the color, texture and 3D point position are obviously distinguishable.

Therefore our goal is to find the appropriate configuration of labels for all segments in both two modalities, which leads to the best results that the segments with the same label can align the object boundaries better. We employ the cross-modal potential to achieve this goal. If a variable associated with the segment i in one modality is assigned a label l_k, the cross-modal potential encourages that variables associated with the segments in another modality that are overlapped by i take the same label. Thus the cross-modal potential can keep the spatial consistency across modalities. As shown in Fig. 5.5, when the variable associated with segment 's_j^I' is assigned the 'object' label, the variables associated with the overlapped segments 's_{n1}^T', 's_{n2}^T' and 's_{n3}^T' also trend to be assigned the 'object' label by this potential. More importantly, when a variable associated with a segment locating at object boundary, such as the segment 's_i^I', is assigned a 'boundary' label, it will enhance the probabilities of variables associated with 's_{m1}^T' to 's_{m4}^T' being also assigned 'boundary' label. Since the segment 's_i^I' does not align the object boundary well, it will result to inaccurate object detection. But the segments 's_{m1}^T' to 's_{m4}^T' in 3D data that align the object boundary well can improve the object detection if they are also assigned the 'boundary' label.

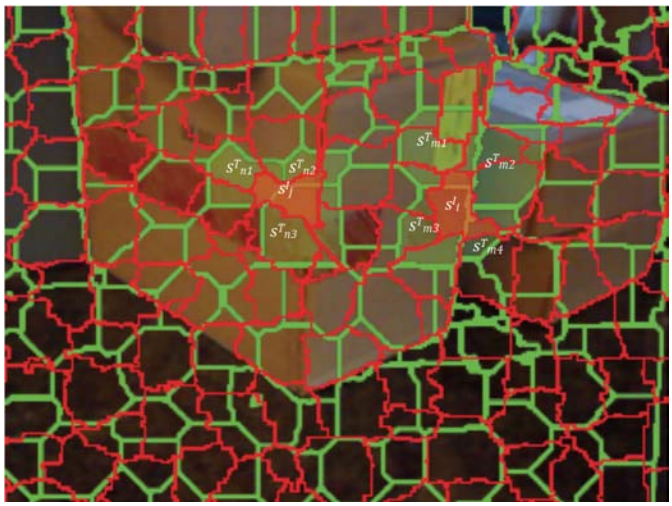


Figure 5.5.: Overlaying 2D and 3D oversegments. The red and green curves denote 2D and 3D oversegments respectively. Two examples of segments (marked by translucent red patches) and the corresponding overlapped segments (marked by translucent green patches) in another modality are shown.

Two cross-modal potentials respectively correspond to two modalities are defined as:

$$\psi_c^{R_1}(x_i^I, \mathbf{x}_{c'_T}^T) = \begin{cases} N_i'(x_i^I, \mathbf{x}_{c'_T}^T)\frac{1}{Q'}\gamma_{max}^{R_1} & \text{if } N_i' \leqslant Q', \\ \gamma_{max}^{R_1} & \text{otherwise,} \end{cases} \tag{5.11}$$

$$\psi_c^{R_2}(x_m^T, \mathbf{x}_{c'_I}^I) = \begin{cases} N_i'(x_m^T, \mathbf{x}_{c'_I}^I)\frac{1}{Q'}\gamma_{max}^{R_2} & \text{if } N_i' \leqslant Q', \\ \gamma_{max}^{R_2} & \text{otherwise,} \end{cases} \tag{5.12}$$

These equations are similar to (5.4) and (5.9) where Q' is also the truncation parameter, but the definitions of $N_i'(\cdot,\cdot)$ and $\gamma_{max}^{R_*}$ are different. $N_i'(x_i^I, \mathbf{x}_{c'_T}^T) = |c'_T| - n_{l_{x_i}}(c'_T)$ and $N_i'(x_m^T, \mathbf{x}_{c'_I}^I) = |c'_I| - n_{l_{x_m}}(c'_I)$ denote the number of nodes in $\mathbf{x}_{c'_T}^T$ and $\mathbf{x}_{c'_I}^I$ which have different labels from x_i^I and x_m^T, respectively. Here the $\gamma_{max}^{R_1}$ and $\gamma_{max}^{R_2}$ are defined as:

$$\gamma_{max}^{R_1} = |c'_T|^{\theta_\alpha^{T'}}(\theta_{hp}^{T'} + \theta_{hv}^{T'}\exp(-\theta_{h\beta}^{T'}||F_{c'_T}^{hT'}||^2)) \tag{5.13}$$

$$\gamma_{max}^{R_2} = |c'_I|^{\theta_\alpha^{I'}}(\theta_{hp}^{I'} + \theta_{hv}^{I'}\exp(-\theta_{h\beta}^{I'}||F_{c'_I}^{hI'}||^2)) \tag{5.14}$$

where parameters are learned from training set, $F_{c'_T}^{hT'}$ and $F_{c'_I}^{hI'}$ are cross-modal clique features computed from cliques c'_T and c'_I. Here the cliques are different from those used in unimodal higher order potentials. For each variable in one modality, there is a cross-modal clique in another modality corresponding to this variable. The cross-modal clique of a node in one modality consists of the nodes in another modality. Meanwhile the oversegments associated with these nodes are overlapped each other. As illustrated in Fig. 5.4, we give two examples of the cross-modal clique. The cross-modal clique c'_T corresponding to the node x_i^I in the 'I' layer forms by the nodes in the 'T' layer connected with x_i^I by the blue dash lines. Another cross-modal clique c'_I consists of the nodes connected with the node x_m^T by the green dash lines. Fig. 5.5 shows the cross-modal clique more intuitively. The cross-modal clique for the variable x_i^I associated with the segment s_i^I in the 2D image consists of four variables associated with 3D segments s_{m1}^T to s_{m4}^T.

5.5. Model Inference and Parameters Learning

Inference for CRF model is to find the configuration of all variables that minimizes the energy function. In general, exact computation for minimizing the energy function is intractable since this problem is NP-hard. Therefore, some algorithms are devised for approximate energy minimization and they can be divided into two categories (Kohli et al. (2009)): message passing algorithms (e.g. Ihler et al. (2006)) and move making algorithms (e.g. Boykov and Funka-Lea (2006)). The message passing algorithms are not suitable for inferring energy functions defined over large cliques since their computational complexity increases exponentially along with the increment of the size of the largest clique. To deal with large size cliques, Kohli et al. (2009) employed the optimal swap and expansion moves for energy function

containing higher order potentials. Our higher order potentials and cross-modal potentials take the general form:

$$\psi_c(\mathbf{x}_c) = \begin{cases} \gamma_k & \text{if } N_i(\mathbf{x}_c) \leqslant Q, \\ \gamma_{max} & \text{otherwise,} \end{cases} \tag{5.15}$$

which can also be reformulated as:

$$\psi_c(\mathbf{x}_c) = \min\{\min_{k \in \mathcal{L}}((|c| - n_k(\mathbf{x}_c))\theta_k + \gamma_k), \gamma_{max}\} \tag{5.16}$$

where $|c|$ is the number of variables in clique c, and potential function parameters γ_k, θ_k, γ_{max} satisfy the constraints which are $\theta_k = \frac{\gamma_{max} - \gamma_k}{Q}$ and $\gamma_k \leqslant \gamma_{max}$ for $\forall k \in \mathcal{L}$. Here $n_k(\mathbf{x}_c)$ has different meanings for unimodal higher order potentials and cross-modal potentials. For the former one, it denotes the number of variables in unimodal clique c which take the label k. For the later one, it denotes the number of variables in cross-modal clique c which take the label of the variable in another modality corresponding to c.

Kohli et al. (2009) proved that (5.16) can be transformed to submodular quadratic pseudo-boolean functions which can be minimized by graph cuts by adding only two auxiliary variables. Therefore in this thesis we also employ the algorithm proposed in Kohli et al. (2009) (Readers can refer to it for details).

There is a set of parameters in our CMH-CRF model that need to be learned from training data. A simple method to set these parameters is to cross-validate every combination of all parameters. However, due to the large number of parameters there is a very high dimensional parameter space to exhaustively search. This is obviously infeasible. Thus we employ a heuristic method, piecewise training (Sutton and McCallum (2005b)), which has been successfully used in Kohli et al. (2009) and Shotton et al. (2009). The particular training procedure consists of three steps.

1. Optimal parameters of two unimodal potentials are learned first. The pairwise potential parameters are learned by using unary and pairwise potentials. Next the parameters of higher order potentials are learned by using unary and higher order potentials. Then the ratios between unary, pairwise and higher order potentials are trained.

2. With all fixed unimodal terms, cross-modal potentials are trained separately by combining unimodal terms.

3. The ratios for cross-modal potentials are finally trained.

The final trained parameters and ratios for the RGB-D dataset will be listed in the experimental section.

5.6. Combination of Labeled Results at Pixel Level

After inferring the CMH-CRF model, two labeled maps corresponding to two different modalities are obtained. The detected results rely on the alignments of oversegments to object boundaries. Therefore, the best precision of results is no more than the precision of alignments of oversegments. As explained in section 5.4, combining 2D and 3D oversegments can enhance their alignments to object boundaries, so we can combine two labeled maps at pixel level to improve the precision of object detection.

For each paired RGB+D data, a pixel in an image and a point in a point cloud are one-to-one correspondence, but oversegments are not. Therefore, given one pixel/point p, we assume that oversegments s_m^I and s_n^T both contain p_i in 2D and 3D modal data respectively. And variables $x_m^I \in \mathbf{x}^I$ and $x_n^T \in \mathbf{x}^T$ may have different labels. Thus one position p may correspond to two labels. Therefore to combine 2D and 3D labeled maps, the label for each position p should be correctly chosen. Probabilities assigning labels to x_m^I and x_n^T are different. A position p can choose one of their labels with higher probability. Given a label configuration for all variables, the probability of one variable x_i being assigned a label is only dependent on its neighbors and the cliques with which x_i is associated. The probability can be computed as:

$$p(x_i|\mathbf{x}) = p(x_i)p(x_i|\mathbf{x}_{c_i})p(x_i|\mathbf{x}_{c_i'})\Pi_{j\in\mathcal{N}(i)}p(x_i|x_j) \tag{5.17}$$

where $\mathcal{N}(i)$ is the neighbor system of the variable x_i, c_i and c_i' denotes the unimodal and cross-modal cliques for x_i. The probability is inversely proportional to the value contributing to the energy function of the whole CMH-CRF model by assigning a label to x_i, which can be computed by:

$$E_i = \psi_i(x_i) + \sum_{j\in\mathcal{N}(i)} \psi_{i,j}(x_i, x_j) + \psi_{c_i}(\mathbf{x}_{c_i}) + \psi_{c_i'}(\mathbf{x}_{c_i'}) \tag{5.18}$$

where the terms denote unary, pairwise, higher-order and cross-modal potentials, respectively. Since each position p belongs to both oversegments in two modalities, p corresponds two values contributing to energy function. Thus, the label to which the variable are assigned with smaller energy value is chosen as p's label. The final labeled map is obtained by choosing appropriate labels for all positions.

5.7. Identification of Object Instances

In a complex scene, objects may be massed up together and occluded each other. If using only the 'object' and 'background', it is difficult to divide the objects into different regions where each region only contains one single object instance. For example, as the original images shown in the second and fourth row in Fig. 5.6. To overcome this problem, in this thesis the proposed model takes one more extra label

'boundary'. Thus patches between two objects or an object and the background are labeled as 'boundary' (e.g. the white patches in the label maps in Fig. 5.6.) and the object instances can be easily separated by alg. 5.1.

Algorithm 5.1: The simple algorithm for identifying all object instances labeled by three kinds of labels.

```
1  Input: The labeled map B;
2  Output: The set of object instances, o = {o_1, o_2, ··· , o_N};
3  Erode the labeled map B as B';
4  Compute the connected components for regions labeled as 'object' and
   'boundary', denoted by c = {c_1, c_2, ··· , c_M};
5  for each component c_i ∈ c do
6      Dilate c_i with the same parameters used for eroding the labeled map;
7      Ignore c_i if its area is less than a threshold T;
8      if c_i only contains one kind of label then
9          c_i corresponds to one object instance;
10     else
11         Compute the components for regions labeled as 'object' in c_i, denoted
           by c' = {c'_1, c'_2, ··· , c'_M'};
12         for each component c'_j do
13             Ignore c'_i if its area is less than the threshold T;
14             Obtain the closer halves of all regions labeled as 'boundary' and
               connected with c'_i;
15             Combine them with c'_i to forming an object instance.
```

In the line 3, a labeled map is eroded to obtain those independent components who connect other components by small patches. For example, as shown in the left picture on the first row in Fig. 5.6, two adjacent objects can be separated if their labeled regions are eroded slightly. A component dilated in line 6 is to recover it original region for accurate object detection. In line 7 and 13, a threshold T is set to filter those too small regions. In our experiments, this threshold is set to 225, which means the objects that can be detected by the proposed method are not smaller than 15×15 patch size. In line 14, we combine an independent region labeled as 'object' and some halves of patches connected to this region as a single object. This is because in the proposed method the label 'boundary' corresponds to those patches which straddle objects and background, or multiple objects, and therefore a part of 'boundary' patch may be a part of object. By this simple algorithm, the resulted labeled map can be efficiently separated into different single object instances. Some examples are shown in Fig. 5.6, where the black, gray and white patches correspond to the 'background', the 'object' and the 'boundary' labels respectively.

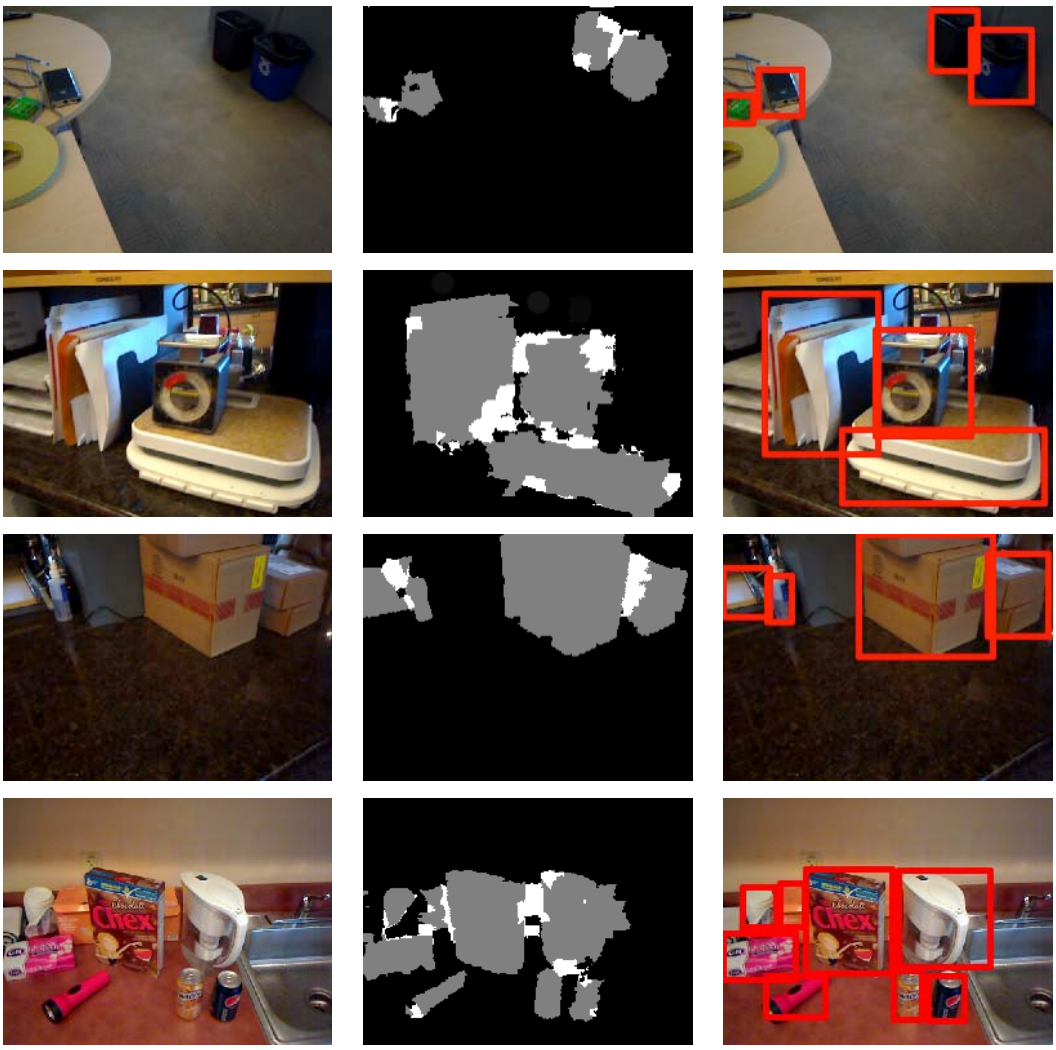


Figure 5.6.: Examples of conveniently identifying object instances by utilizing 3 labels. The object instances connecting together can be separated by the boundary labels and can be easily detected as different instances. From left to right columns, those pictures are original images, label maps and object instances bounded by boxes.

5.8. Experiments

In this section we evaluate the performance of category-independent object detection for the proposed CMH-CRF model. All experiments are based on a public RGB+D dataset (Lai et al. (2011)) which has been introduced in last chapter. We extract 500 images with significant difference from four scenes in our experiments. And randomly chosen100 samples from only two scenes are used as training set, to avoid all categories occur in the training stage, which make the experiments more objective for category-independent object detection. The rest 400 samples are used for testing.

The set of parameters for the CMH-CRF model is mentioned in last two sections. According to the method of learning parameters, these parameters for the RGB+D dataset are $\theta_p^I = 0.125$, $\theta_v^I = 16.0$, $\theta_\beta^I = 16.0$, $\theta_\alpha^I = 0.5$, $\theta_{hp}^I = 0.25$, $\theta_{hv}^I = 4.0$, $\theta_{h\beta}^I = 1.0$, $\theta_p^T = 0.1$, $\theta_v^T = 12.0$, $\theta_\beta^T = 9.5$, $\theta_\alpha^T = 1.0$, $\theta_{hp}^T = 0.1$, $\theta_{hv}^T = 8.0$, $\theta_{h\beta}^T = 16.0$, $\theta_\alpha^{I'} = 2.0$, $\theta_{hp}^{I'} = 0.1$, $\theta_{hv}^{I'} = 3.5$, $\theta_{h\beta}^{I'} = 2.0$, $\theta_\alpha^{T'} = 1.5$, $\theta_{hp}^{T'} = 0.3$, $\theta_{hv}^{T'} = 5.0$, $\theta_{h\beta}^{T'} = 4.0$.

The proposed CMH-CRF model is evaluated in four different settings. The first two only use 2D images or 3D point clouds which can be considered as normal single modal higher order CRF models (referred as Config I and Config II, respectively). The third uses both 2D and 3D data but without cross-modal potentials (referred as Config III), in which the final results are combined at pixel-level. The last uses all terms in our CMH-CRF model (referred as Config IV). The experiments are executed five times. Then their average precisions and recalls of each configuration are compared.

We compare these configurations at two different levels. The first is the object level where we use a box to bound each detected object and then compute the overlapped area to determine if this box really contains an object. To decide if a bounding box contains an object, we use the strict PASCAL-overall criterion which considers a box containing an object when the area of their overlapped region is more than 50% of their union area (Everingham et al. (2010)). The precision and the recall are determined by:

$$precision = \frac{N_C^B}{N_A^B} \qquad recall = \frac{N_C^B}{N_G^B} \tag{5.19}$$

where N_C^B denotes the number of detected bounding boxes containing objects, N_A^B is the number of all detected bounding boxes, and N_G^B denotes the number of all ground truth bounding boxes. The results are shown in Fig. 5.7 and Fig. 5.8. From these two figures, it can be seen that the cross-modal configuration achieves the highest values in both precision and recall. We also list their averages in the Table. 5.2. The average precision and recall of Config IV are improved about 10% to 15% with respect to the unimodal category-independent object detection. By comparison with the multimodal detection without cross-modal potentials, the best averages are also about 5% higher than it, which shows that the cross-modal higher order CRF model indeed improve the performance of category-independent object detection. Note that there are some results that have 0 or 1 precisions and recalls in

Table 5.2.: Average accuracy of object recognition

	Config I	Config II	Config III	Config IV
Object-level Precision	0.6047	0.6856	0.7014	0.7526
Object-level Recall	0.6393	0.6895	0.7289	0.7727
Pixel-level Precision	0.6316	0.6548	0.7309	0.7606
Pixel-level Recall	0.6553	0.6734	0.7694	0.7984

these two figures. In our experiments, there are some images and point clouds where only one objects need to be detected, such as the four column in Fig. 5.11 and the third object in Fig. 5.15. Thus if this object is correctly detected, the precision and recall of this image are 1. While this object is not correctly detected, the precision and recall of this image is 0. As shown in Fig. 5.9 and Fig. 5.10, these cases are not found in the evaluation at pixel level, since it is impossible for our method to output detected pixels that exactly match all annotated pixels.

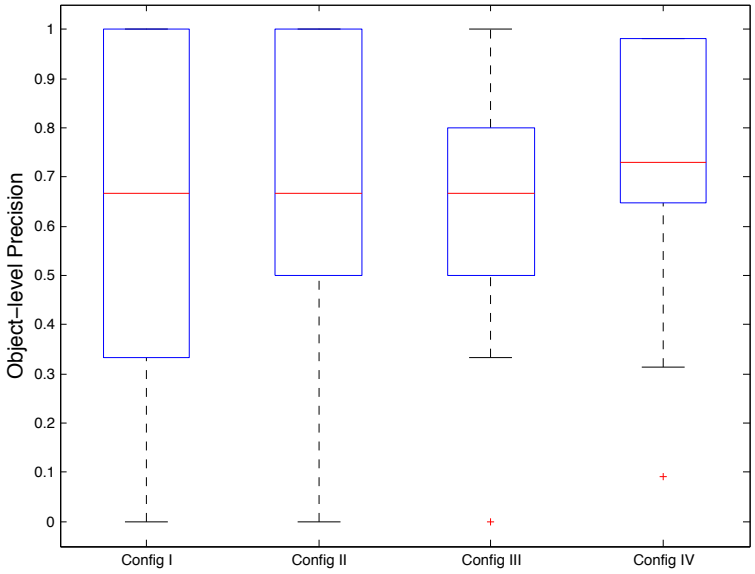


Figure 5.7.: Precision of object level detection results. From left to right, the 'Config I' to 'Config IV' are corresponding to four method configurations mentioned in text.

The second level is the pixel/point level. Our method can detect objects based on oversegments, therefore the precision and recall of detected regions can be computed as:

$$precision = \frac{N_C^P}{N_A^P} \qquad recall = \frac{N_C^P}{N_G^P} \tag{5.20}$$

where N_C^P is the number of pixels belonging to both detected regions and ground truth regions, N_A^P is the number of pixels belonging to all detected regions, and

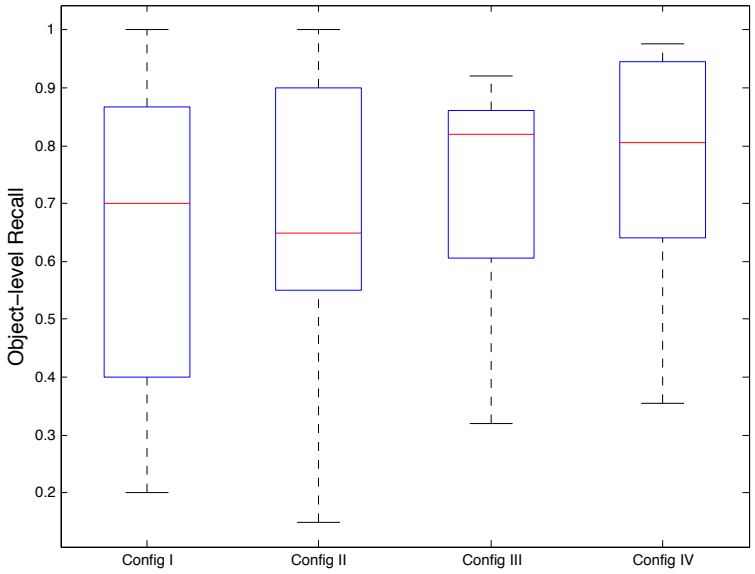


Figure 5.8.: Recall of object level detection results. From left to right, the 'Config I' to 'Config IV' are corresponding to four method configurations mentioned in text.

N_G^P is the number of pixels belonging to all ground truth regions. The results are shown in Fig. 5.9 and Fig. 5.10. These two figures also show that the cross-modal category-independent object detection achieves the best performance with 11%-14% improvements. Some results obtained from different configurations are illustrated in Fig. 5.11.

Moreover, three state-of-the-art methods that correspond to different classes of category-independent detection techniques are used for performance comparison. These three methods are the 'Global Contrast based Salient Region Detection' (GC, Cheng et al. (2011)), the 'Object Ranking based on Multimodal Cues' (OR, Zhang et al. (2011)) and the 'Constrained Parametric Min-Cuts' (CPMC, Carreira and Sminchisescu (2012)). Methods 'CPMC' and 'OR' do not directly give all objects' locations, but build a large set of region pool, ranking the objectness scores for all regions in this pool. Therefore, we evaluate their performance by computing the precisions and recalls given different number of regions drawn from their region pools. Actually, we draw the regions according to their scores. For example, when given a number 100, 100 regions are drawn from the pool with the top 100 scores. The maximal number is 200. Since our method and the 'GC' method directly give object positions, we can directly compute their precisions and recall rates. The first comparison is at object level. For convenient comparison, their precisions and recall rates are illustrated in two uniform figures, as shown in Fig. 5.12 and Fig. 5.13. For the recall curves, our method achieve comparable results with respect to the 'CPMC' method, which obtain similar object recall rate by sampling 200 regions. Therefore the proposed CMH-CRF model can detect category-independent

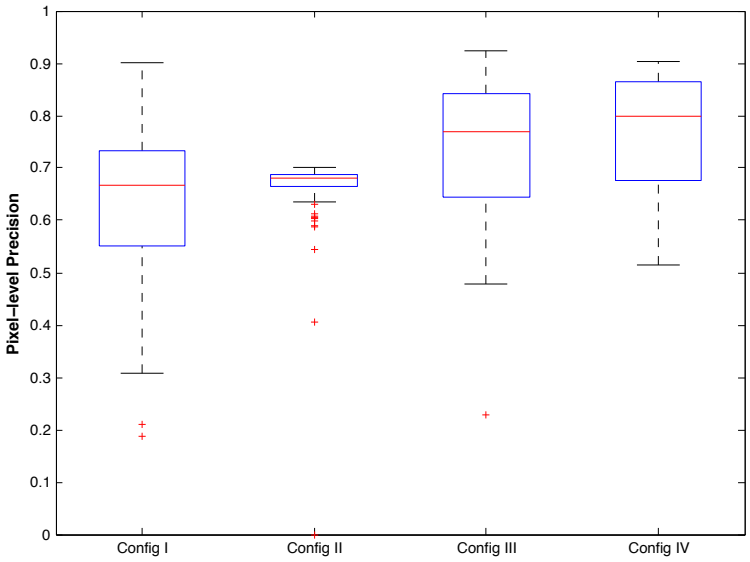


Figure 5.9.: Precision of pixel level detection results. From left to right, the 'Config I' to 'Config IV' are corresponding to four method configurations mentioned in text.

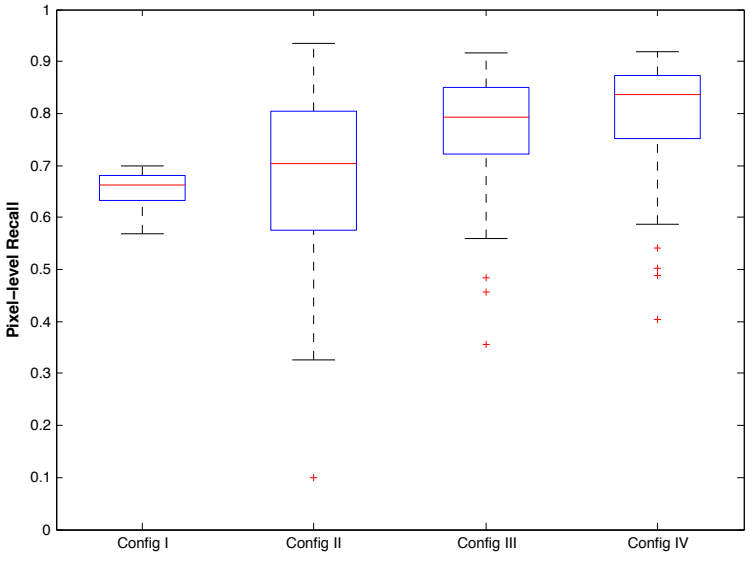


Figure 5.10.: Recall of pixel level detection results. From left to right, the 'Config I' to 'Config IV' are corresponding to four method configurations mentioned in text.

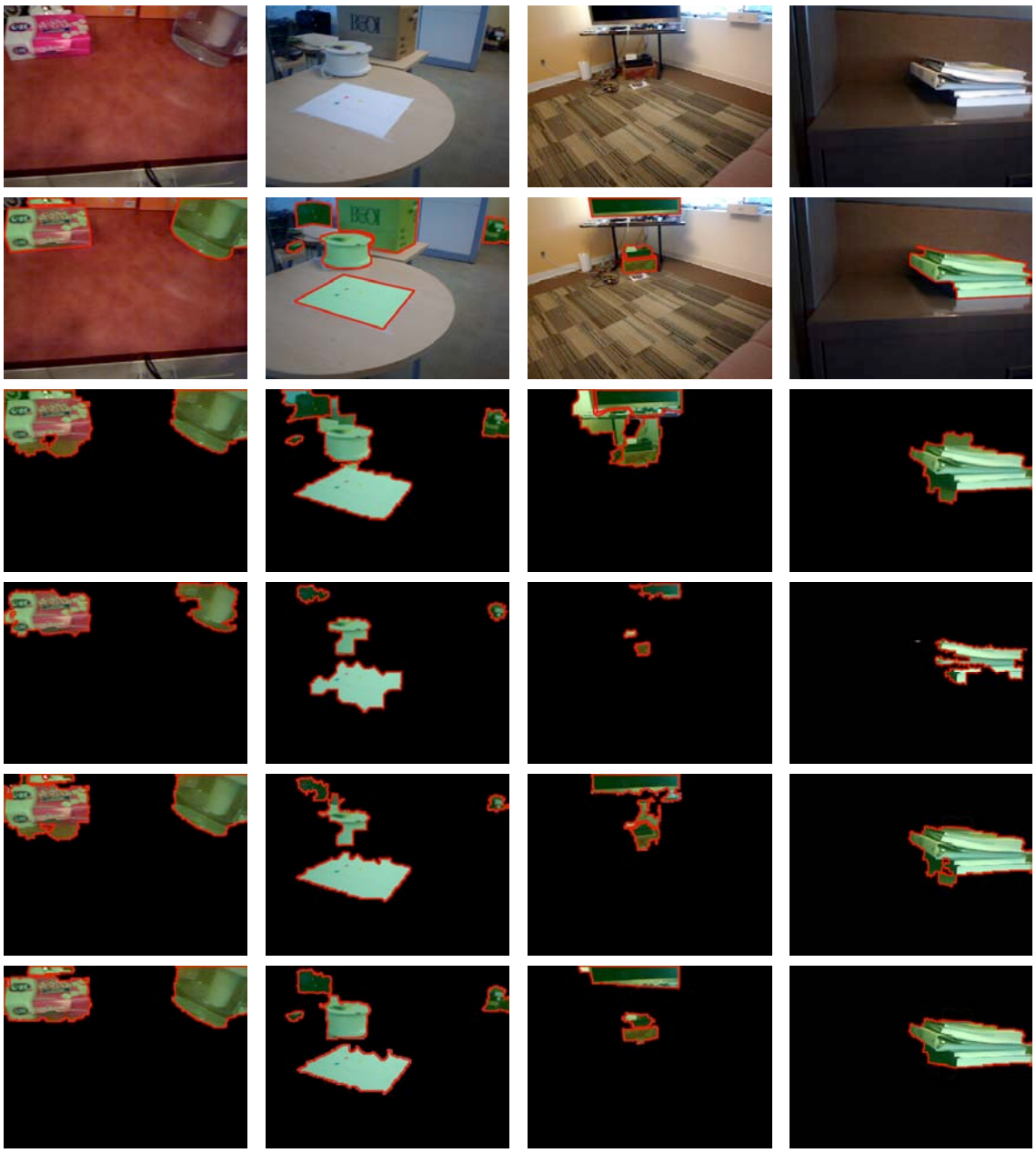


Figure 5.11.: Samples of results obtained from different configurations. From top to bottom, the rows correspond to original images, ground truth, 2D configure, 3D configure, multimodal configure without cross-modal potentials and cross-modal configure.

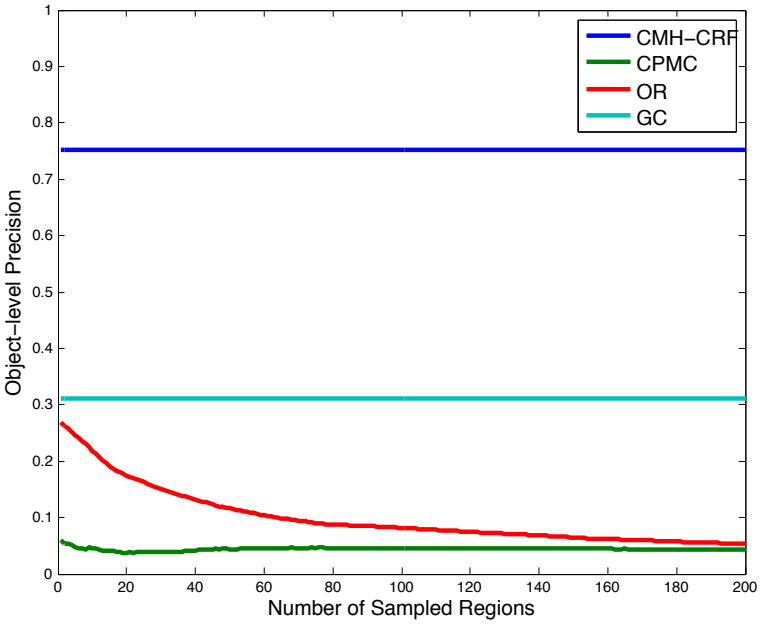


Figure 5.12.: Comparison of precision of object level detection results between three state-of-the-art methods and the proposed CMH-CRF model.

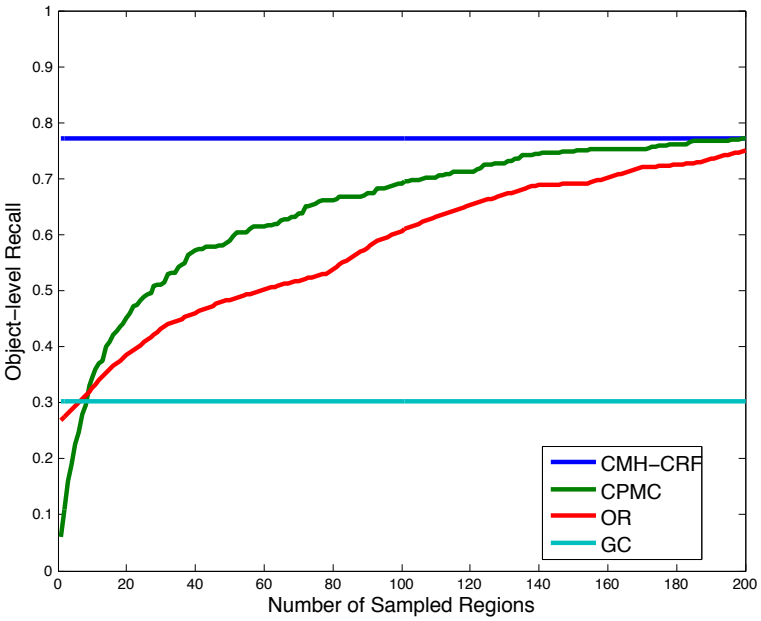


Figure 5.13.: Comparison of recall of object level detection results between three state-of-the-art methods and the proposed CMH-CRF model.

object in a more efficient manner, which does not need to sample a large number of regions but directly localizes objects. The efficiency of directly localizing objects can also be observed from the precision curves, where the proposed method carries out acceptable precision (78%) of detecting objects. In this evaluation, the 'CPMC' and 'OR' methods both have too low precisions to be suitable for novel object category discovery, because they propose too many unexpected regions which do not correctly contain objects. As shown in this figure, no less than 70% regions sampled by the 'CPMC' and 'OR' methods are associated with background. The salient detection method gains around 30% recall rate and precision, so it cannot efficiently detect category-independent object yet.

Except the 'OR' method, all other methods detect objects at pixel level. Therefore we can further compare their performance at pixel level. The results are shown in Fig. 5.14. We follow the criterion used in Carreira and Sminchisescu (2012) to evaluate the performance at pixel level. For the 'CPMC' method, the covering scores are computed by:

$$C(S, S'(r)) = \frac{1}{N} \sum_{R \in S} |R| \star \max_{R' \in S'(r)} O(R, R') \tag{5.21}$$

where S and S' denote the set of ground truth segments and the set of proposed segments respectively, N denotes the total number of pixels of all annotated objects in one image, $|R|$ is the number of pixels in the ground truth segment R, and O is the overlap measure between two regions which is defined as:

$$O(S, G) = \frac{|S \cap G|}{|S \cup G|} \tag{5.22}$$

which is also used for measuring the performance of the results carried out by the CMH-CRF model and the 'GC' method. Fig. 5.14 shows the similar results to that in object-level comparison. The 'CMH-CRF' method has the comparable scores with respect to the 'CPCM' method when it samples 200 regions. The overlap score is more than 77% which implies the CMH-CRF model can be practically used for category-independent object detection. The salient object detection method only obtain 30% overlap scores and there is still large gaps for accurately detecting category-independent objects. Furthermore, as shown in Fig. 5.15 and Fig. 5.16, the regions detected by the 'GC' method are difficult to identify multiple object instances since it aims at detecting single foreground object per image.

Finally, some examples of detected regions by different methods are illustrated in following figures (Fig. 5.15 and 5.16). For each group of pictures, from left to right and top to bottom, they are original images, ground truth, results of 'GC', results of 'CMH-CRF', results of 'CPMC', and results of 'OR'. Each detected object instances are marked by different solid curves and green transparent covers. The decimal in each object indicates overlapping or covering score computed by eq. (5.21) and eq. (5.22). For the method 'CPMC' and 'GC', we sample the region from the sampled region pool with best measuring scores and also give the ranking

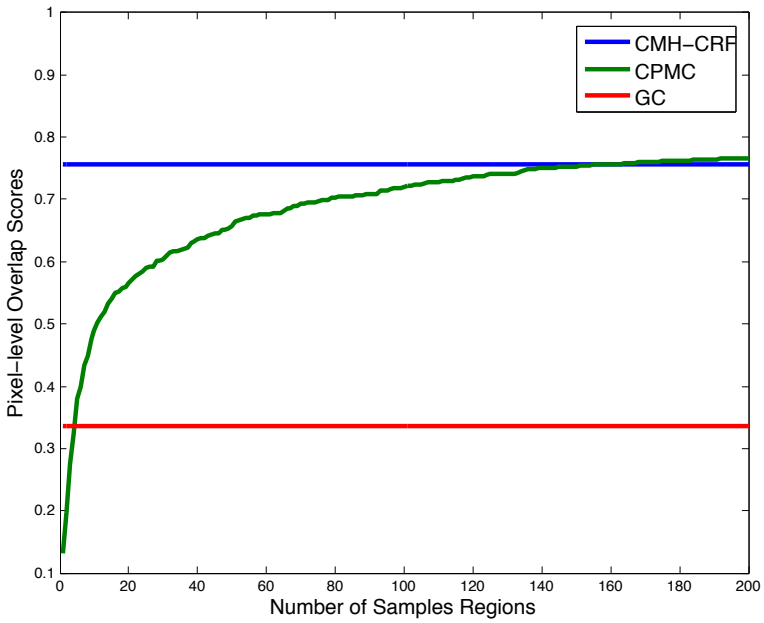


Figure 5.14.: Comparison of overlap scores at the pixel-level for detection results between two state-of-the-art methods and the proposed CMH-CRF model.

place by the integer number before their scores. The 'CPMC' method gives some better results than that of the proposed method, however, as shown in figures, most of these best regions locate out of the place of top 20 in the region pool. This implies that to get accurate detection of objects, more inaccurate regions may be sampled first, which is inevitable to lead to many wrong results of object category discovery since those regions without containing objects may be treated as novel categories. On the contrary, the proposed method can detect and localize object without sampling useless regions and consequently can improve the performance of novel object category discovery.

5.9. Conclusion

In this chapter we proposed a new cross-modal co-segmentation framework for category-independent object detection. For integrating 2D and 3D data, we developed the CMH-CRF model based on the robust higher order conditional random field model. This novel model takes the global spatial consistency in both 2D and 3D space into account, which leads to better detection results. By comparison with state-of-the-art methods, the extensively experiments show the novel model has the better performance. Based on the accurate detection results, we can develop the novel category discovery framework for our next goal.

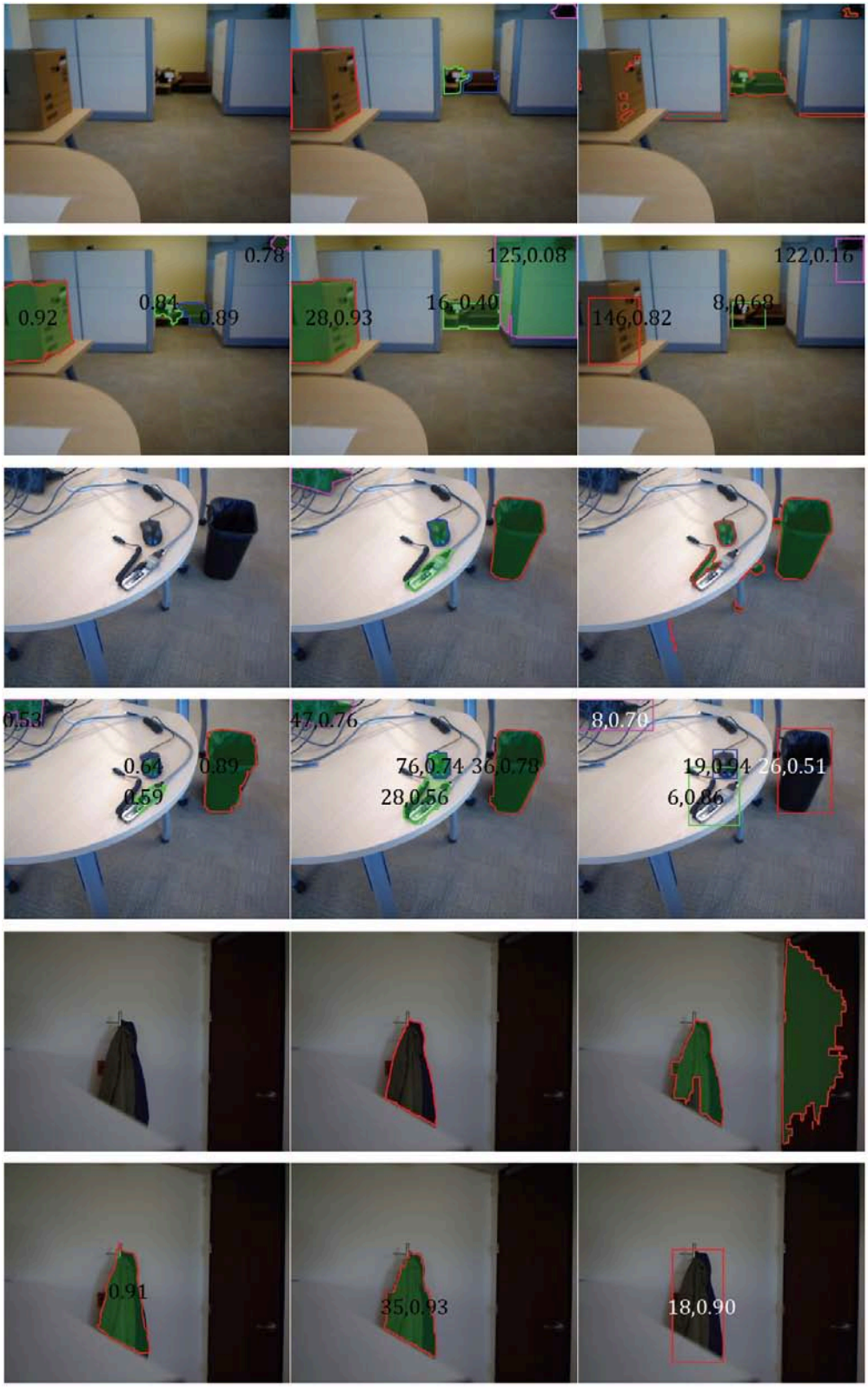


Figure 5.15.: Samples of comparison results among the proposed method and state-of-the-art methods. For each group of pictures, they are the original image, ground truth, results of 'GC', results of 'CMH-CRF', results of 'CPMC', and results of 'OR'.

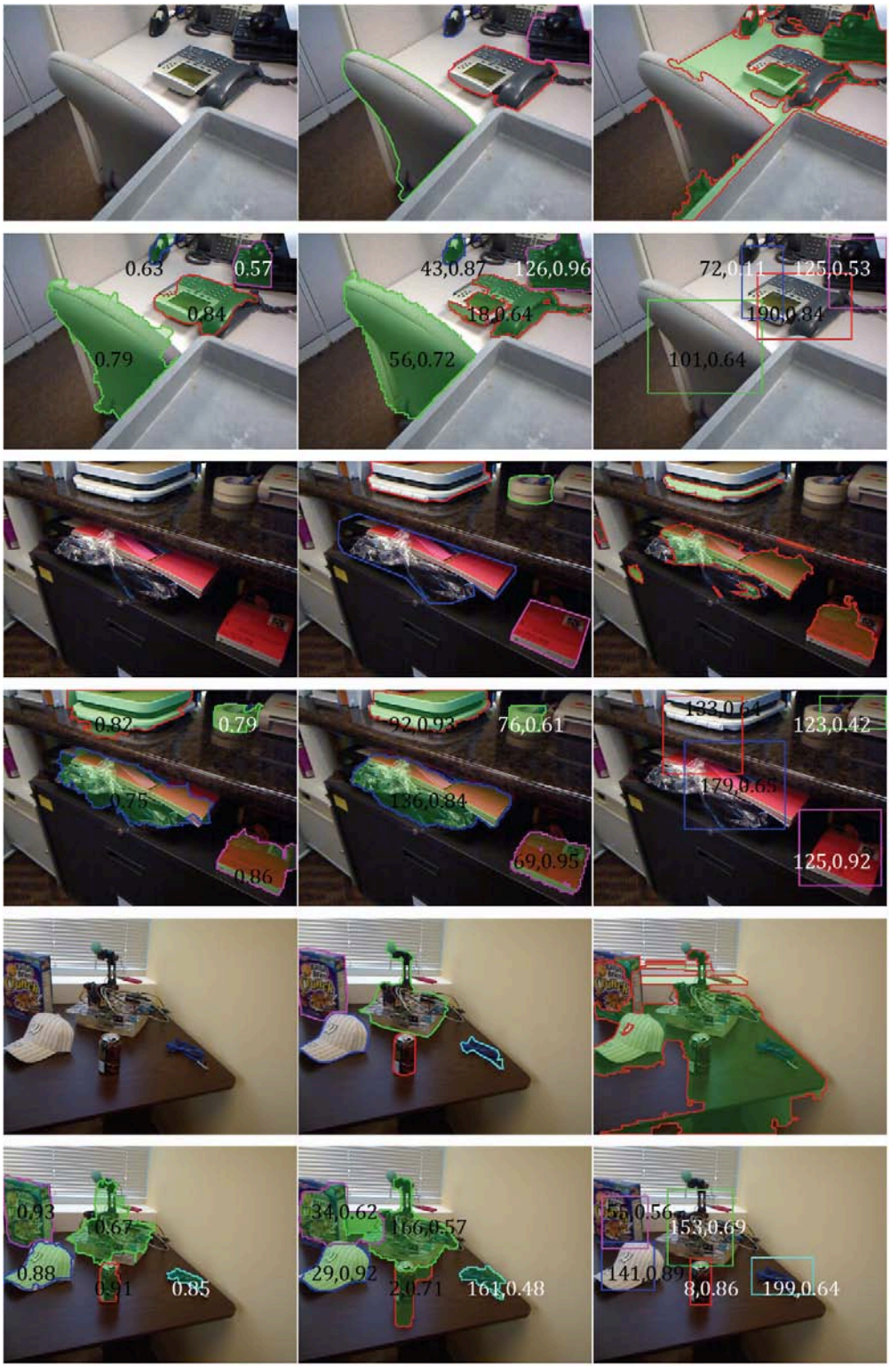


Figure 5.16.: More complicated samples of comparison results among the proposed method and state-of-the-art methods. For each group of pictures, they are the original image, ground truth, results of 'GC', results of 'CMH-CRF', results of 'CPMC', and results of 'OR'.

Chapter 6

Extended Object Attributes

6.1. Introduction

Humans always use attributes, such as color, shape, material and parts, to describe objects. For example, to describe a dog one may say this dog is white, has four feet and one tail with short dog hair. By using attributes, different objects can be distinctively described. To simulate this ability of human cognition, Farhadi et al. (2009) and Lampert et al. (2009b) originally proposed the extraction of visual attributes from images in the computer vision community. Recently, the object attributes have been improved from different aspects, such as using relative attributes to reveal what degree of an attribute an object has (Parikh and Grauman (2011)), and jointly learning object attributes and descriptions (Mahajan et al. (2011)).

Object attributes may possess semantic meanings, which are useful to transfer knowledge from known categories to unknown categories, since different categories may share the same attributes. However, using semantic attributes in a computer vision system is not enough to describe objects accurately. On the one hand, some similar categories are difficult to be distinguished by simple semantic attributes. For example, as shown in Fig. 6.1 some dogs and wolves look very similar and consequently are difficult to be distinguished even by human. It is necessary to use many complex words to describe their differences. But in the computer vision system, it is still a tough task to extract such complex semantic attributes. On the other hand, the semantic attributes have to be supervised trained from annotated data. To train thousands of attributes is infeasible due to limited manually annotated training data. Thus non-semantic attributes which can be trained unsupervised are necessary.

In Farhadi et al. (2009), a kind of non-semantic attribute is trained, which takes the difference among different categories into account. However, there is also large difference among some object instances within one category. Hence to describe objects more accurately, the intra-class non-semantic attributes are also required.

Figure 6.1.: Dog (left) and wolf (right) are difficultly distinguished by simple semantic description since they have similar appearance.

Most object attributes presented in current literature are extracted from 2D visual appearance. However, some attributes such as shapes are appropriate to be extracted from 3D data. For example, if extracting the 'sphere' attribute from 2D images, it may be confusion with the 'circle' attribute. On the other hand, 'sphere' and 'circle' shapes are discriminative in 3D space. As the convenient devices such as Microsoft Kinect camera emerged, it is possible to obtain color information and depth information simultaneously. Thus the object attributes can be extracted from multimodal features which enhance the representability of attributes.

In this chapter, we develop novel object attributes which is extended from two aspects based on the work in Farhadi et al. (2009). At first intra-class non-semantic attributes are developed to represent the difference among objects within the same category. Then multimodal attributes are implemented by adding 3D features into the original set of 2D visual features. The proposed object attributes are evaluated on two public datasets. They are a 2D image dataset used in Farhadi et al. (2009), and a RGB+D dataset (Lai et al. (2011)) used in our previous chapters. By comparison with the original object attributes in Farhadi et al. (2009), our experiments show that the novel object attributes are more robust and can recognize object categories more accurately.

6.2. Object Attributes

We have introduced the algorithm of extracting object attributes (Farhadi et al. (2009)) in section 3.2.1. Here it is briefly described to help readers recalling.

For each object in the training dataset, a feature vector with 9571 dimensions is first extracted as 2D base features. Since categories share attributes, an attribute classifier is trained across multiple categories. Since different attributes belonging to one object may influence each other, the training of attribute classifier is done by selecting positive and negative samples which can decorrelate attribute predictions.

For example, to train the 'wheel' attribute, the positive samples come from such as cars and buses which at least have one wheel, and the negative samples are collected from cars and buses without any wheel. Thus the 'wheel' attribute will not be effected by the 'metal' attribute. An L1-regularized logistic regression is used to train the classifier since it can produce a set of sparse representation of base features.

The training of the non-semantic attributes has two main steps.

1. Randomly select several categories and split them into two sides, each of which contains equal 1 to 5 categories.

2. Randomly select a subset of features for these categories and use a linear SVM to train a non-semantic attribute.

Through this training strategy, the generalizable non-semantic attributes across categories can be obtained, but it cannot describe the difference within one category. Thus we refer to this kind of non-semantic attributes as inter-class non-semantic attributes (Inter-NSA). In the next sections, why and how to extend this non-semantic attributes to intra-class non-semantic attributes (Intra-NSA) will be introduced first, and then the multimodal attributes will be presented.

Figure 6.2.: Large differences among object samples within one category.

6.3. Intra-class Non-semantic Attributes

Fig. 6.2 illustrates some examples to show how large the intra-class differences are. In the top row, the horns of the left and the middle cows have different shapes and colors. And the larger difference is in the right cow which has no horn. Furthermore,

the torso colors are also different. In the bottom row, two sheeps also have different appearances. Therefore, it is necessary to use intra-class non-semantic attributes.

The new algorithm of learning intra-class non-semantic attributes is listed in alg. 6.1. In line 5, samples in the same category are divided into two halves. In one half the samples are regarded as holding a non-semantic attribute but in the other half the samples do not have this non-semantic attribute. Obviously, the cases that samples in one category have and not have a non-semantic attribute can be represented by this algorithm, which cannot be obtained by the algorithm in alg. 3.1 used in Farhadi et al. (2009). In line 6, a feature selecting strategy is also employed to choose best features, which has been used to train semantic attributes in Farhadi et al. (2009). In line 7 those selected features of all selected categories are combined to train a non-semantic attribute classifier, which endows our classifiers with generalizability across categories.

Algorithm 6.1: Learning intra-class non-semantic object attributes.

```
1 Input: Base features of all object samples of all categories in training set.
2 for each non-semantic object attribute do
3     Randomly select 2-10 categories;
4     for each category do
5         Divide object samples into two halves, where each half contains 20%
          to 80% samples.
6         Use a feature selecting strategy to select a best sub-set of base
          features that can best distinguish the two halves;
7     Combine all selected features for all categories, train a classifier using
      linear SVM;
8 Select a certain number of classifiers as intra-class non-semantic attributes
  which can best classify object samples in two halves of each category.
```

6.4. Multimodal Attributes

6.4.1. Multimodal Base Features

The semantic and non-semantic attributes mentioned above are extracted from only 2D base features. In this section, the unimodal attributes are extended to multimodal attributes by using multimodal base features. Thus the first thing is to construct the multimodal base features. In this thesis the multi-modality refers to 2D and 3D modalities as used in last two chapters. Generally, the 3D data is represented by point clouds which are textureless data and only suitable for describing 3D shape characteristics. Two kinds of invariant 3D shape descriptors, the 3D shape context (3DSC, Kortgen et al. (2003)) and the fast point feature histogram (FPFH, Rusu et al. (2008b), Rusu et al. (2008a)), are employed to construct multimodal

base features. The original features, normals and curvatures are also used to form base features.

The 3DSC is extended from 2D shape context descriptor originally proposed by Belongie et al. (2002). The 3DSC constructs a multi-scale multi-sector sphere containing the 3D surface of an object, and counts the numbers of points for all bins to form a histogram. Thus this histogram models the distribution over relative positions of points as a robust and compact, yet discriminative descriptor. Particularly, a sphere is divided into 6 shells with logarithmically increasing radius:

$$r_i = \frac{1}{s}\log_a(a^s\frac{i}{s}) \tag{6.1}$$

where the i-th radius r_i depends on the number of shells s and log-base a. In this thesis, a is set to 2 which results in that each shell has the same volume. The longitude and latitude of the sphere are divided into 12 and 6 parts with the same angle interval 30^o respectively. Finally, the 3DSC descriptor of each point is a histogram with 432 bins.

To robustly describe an object shape, the descriptor should be invariant. The 3DSC is natural scale invariant since the size of sphere is relevant to the size of the entire object. The 3DSC is proposed for dealing with the matching problem of well shaped 3D object models, which cannot be guaranteed in this work since a point cloud of an object is segmented by the CMH-CRF model and consequently must involves some noise points. Therefore, by assuming that the point cloud of a 3D object obeys a multivariable Gaussian distribution and computing its mean μ and variance Σ, those points locating in the region $\mu \pm 3\Sigma^T\Sigma$ are considered as an object's points and contained into the sphere. The 3DSC naturally holds the translation invariance since it uses the relative positions among points to construct the histogram. Unnormalized 3DSC is not rotation invariant since the sphere is divided into multi-sectors. Following the same strategy as used in Kortgen et al. (2003), the Principal Axes Transform (PAT, Alpert et al. (1990)) method is also employed to perform rotation invariant normalization.

'Bag of words' style features are constructed for using the 3DSC descriptor. Some objects are chosen as training data and the 3DSC descriptors of all points for each object is computed. Then they are clustered by the K-means algorithm to form 256 centers. The 3DSC descriptors of all points of an object are quantized to the nearest one of 256 K-means centers. Thus for each object, its 3D point cloud can form a vector with 256 dimensions through the 3DSC descriptor.

An FPFH consists of 81 bins, forming a descriptor with 81 dimensions. Unlike the 3DSC that utilizes relative positions of points in an object, the FPFH models the relative normal directions of points to form an invariant descriptor. The FPFH has been proved to efficiently represent object shape and to be used for 3D object recognition (Rusu et al. (2008b)). In this work, similarly to utilize the 3DSC, 256 cluster centers are also obtained by the K-means algorithm. For a point of an object, its FPFH is computed and subsequently quantized to the nearest one of

256 K-means centers. Thus one more 256 dimensional vector feature is obtained to describe object's 3D shapes.

The point normals are quantized into a 72-bin histogram by dividing the zenith angle into 12 parts and the azimuth angle into 6 parts with equal 30^o interval. The curvatures are quantized into a 128-bin histogram by averagely dividing the difference between the maximal curvature and the minimal one.

An 3D object is first divided into 8 boxes with the same size. For each box, the four kinds of aforementioned features are computed. Then for the whole object, we also compute these four kinds of features. They are all concatenated to form a 6408 dimensional feature which is referred to the 3D base features. Finally, there is a 16159 dimensional base feature for each object by stacking 2D and 3D base features together.

6.4.2. Multimodal Attributes

Based on the multimodal base features, some more accurate and complex semantic attributes can be learned. For example, jointly leaning the material, part and 3D shape attributes can yield more accurate attributes such as "cup's handle" and "door's handle". Usually, a cup's handle and a door's handle are made of different material and have different shapes. But from the 2D perspective their shapes look very similar as shown in Fig. 6.3. That is because when projecting the cylinder and cuboid into 2D image, their projections are both rectangles and cannot be identified from only 2D data. Thus if only using 2D base features they would be difficultly classified. But combing 2D and 3D base features, these two attributes can learn the material difference from 2D base features and the shape difference from 3D base features, and they will be more distinctive than those of training from unimodal base features.

Furthermore, some shape attributes can be more efficiently trained by the 3D base features, such as 'sphere' and '3D concave'. Therefore some 3D semantic attributes are also trained from only 3D base features.

The learning strategy is similar as that of learning 2D semantic attributes. By choosing some objects possessing an attribute as positive samples and some objects without this attribute as negative samples, the L1-regularized logistic regression is employed to train the attribute classifier. Since the RGB+D dataset used in this work only contains indoor objects, only those multimodal attributes possessed by indoor objects are learned, such as 'cup handle' and 'door handle'. The details of all semantic attributes are listed in appendix A, including 2D, 3D and multimodal attributes.

We also train multimodal Inter-NSAs and Intra-NSAs. Because the multimodal base features are enhanced by the 3D features, the more discriminative non-semantic attributes can be obtained. Unlike 2D shape features whose discrimination is declined when 2D projections of 3D objects have the same shape, 3D shape features can

Figure 6.3.: Door's handle (left) and cup's handle (right).

always keep the difference between different shapes and therefore can describe object shapes more accurately. Thus the multimodal non-semantic attributes achieve better performance for classifiers, which will be proved in the next experiment section.

6.5. Experiments

To evaluate the efficiency of the Intra-NSAs and the multimodal attributes, two set of experiments are executed on a 2D image dataset and a public RGB+D dataset. For the 2D image dataset, we evaluate the performance of the Intra-NSAs. There are 32 classes in the image set where each class consists of hundreds of samples. For these 32 classes, 64 semantic attributes are trained. Furthermore, 1000 Inter-NSAs and 1000 Intra-NSAs are clustered. The distinctiveness of single attribute classifier is not evaluated, since in this work only the ability of describing objects of attributes is concerned. These semantic and non-semantic attributes are used as features to describe an object, and their performance of classifying object categories is evaluated. The training of semantic attributes is the same to that used in Farhadi et al. (2009), which is not explained here (readers can refer to Farhadi et al. (2009) for details). To train the non-semantic attributes, a half of 32 classes are randomly chosen as the training set, and the rest is used for test. Note that the test samples are described by the predicted semantic and non-semantic attributes. The predicted attributes have two sets: one consists of 64 semantic attributes and 1000 Inter-NSAs (referred to as 'Set I'), and the other set comprises 64 semantics and 1000 Intra-NSAs (referred to as 'Set II'). Then an object can be described by two vectors which both have 1064 dimensions. The value of each entity in the vector takes 1 or 0, which means whether the object has the corresponding attribute. The two vectors are treated as features and one-vs-all SVM is employed as to train classifiers. After executing five group experiments, the average accuracy of 73.6% by using the 'Set I' and 74.9% by using the 'Set II' are obtained. As shown in Fig. 6.4, the

experimental results prove that the Intra-NSAs have more discriminative ability for classifying objects.

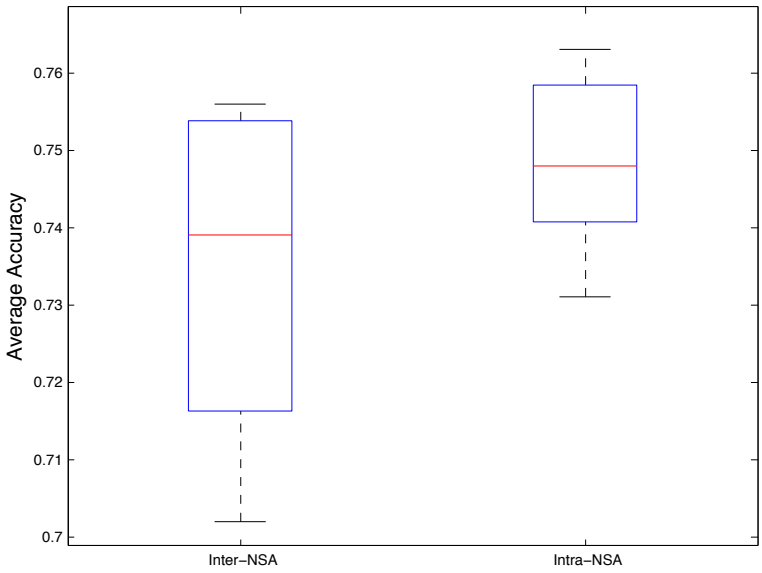


Figure 6.4.: Comparison of accuracy between Inter-NSAs and Intra-NSAs.

Since the RGB+D dataset is produced in the indoor environment, 32 semantic attributes that are possessed by indoor objects are trained, which are listed in appendix A. From this dataset, 20 categories are annotated out for evaluation, as listed in appendix B. Among these attributes, 16 attributes are the same with some 2D attributes, but trained from multimodal base features. There are 10 3D semantic shape attributes trained from only 3D base features. At last, 6 multimodal semantic attributes are trained from multimodal base features, which do not have alternative 2D ones. For comparison, 2D base features are used to train the last 16 3D and multimodal attributes by choosing the same training samples, though 2D base feature cannot accurately represent them.

These attributes are evaluated from several aspects. The first 16 attributes are compared with their corresponding 2D ones. The area under ROC curve of their classifiers are shown in Fig. 6.5. As observed from this figure, all of the 16 attributes have higher or comparable accuracy with respect to their corresponding 2D attributes. And there is average 4% improvement.

For the last 16 semantic attributes, we also compute their area under ROC curve of classifiers trained from 3D (the former 10 attributes in Fig. 6.6) and multimodal (the latter 6 attributes in Fig. 6.6) base feature and 2D base feature respectively. The results are shown in Fig. 6.6. The improvements here are much larger than the first 16 attributes. All of them have better distinctiveness than those corresponding attributes trained from only 2D base feature. The largest improvements are obtained by the “plane” attribute and the “desktop” attribute, which means that these two attributes are only suitable for being represented by 3D and multimodal

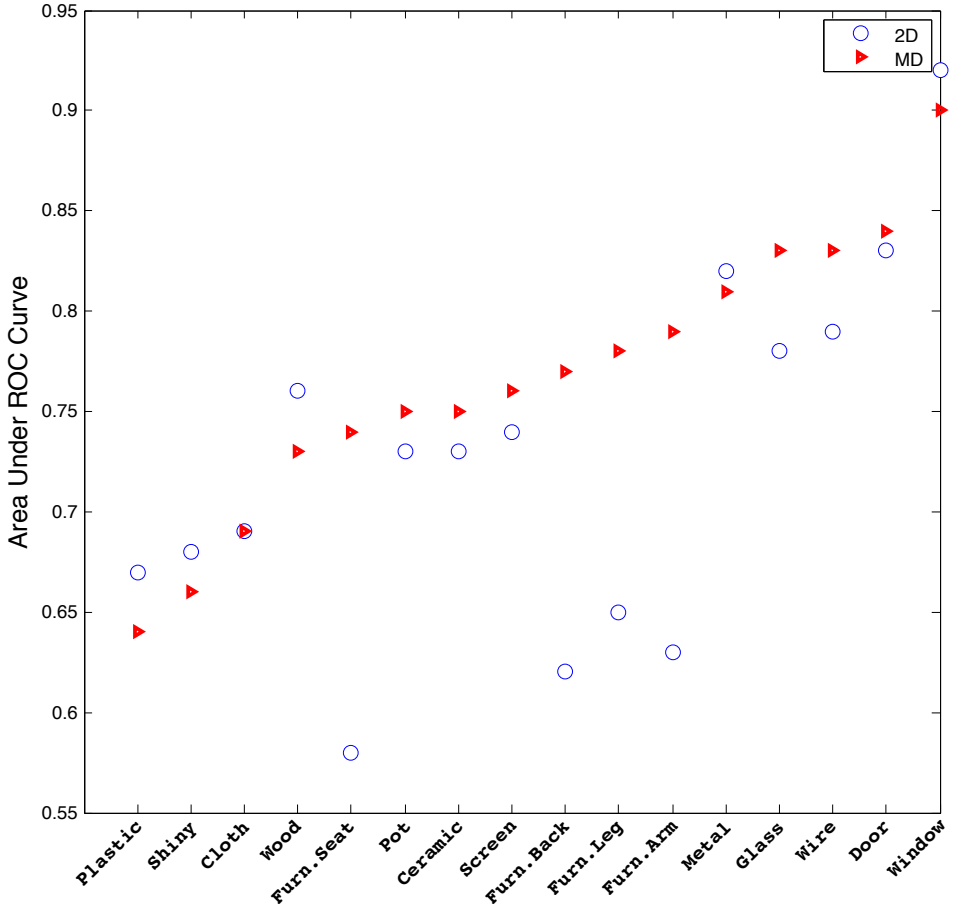


Figure 6.5.: Comparison of accuracy for 16 attributes trained by 2D base feature (blue circle) and multimodal base feature (red dot).

base features, respectively. The average improvements by using 3D and multimodal attributes are 19% and 18%, respectively.

At last these multimodal attributes' ability of describing objects is evaluated. Similar to the first group of experiment, these multimodal semantic and non-semantic attributes are used as features, which are trained from randomly selecting samples belonging to only 10 categories, to describe objects. And then the one-vs-all SVM is employed to learn classifiers. Their accuracy is shown in Fig. 6.7. As it can be seen from this figure, the classifiers trained by the attribute 'Set II' have average 2.3% higher accuracy than those classifiers trained by the attribute 'Set I'. This experiment also validates that the Intra-NSAs have better distinctiveness than the Inter-NSAs.

6.6. Conclusion

In this chapter, we presented the novel multimodal attributes and Intra-NSAs. By integrating 3D features into the original base features, the new set of base features can be used to efficiently train 3D semantic attributes and multimodal attributes, by which an object can be described more comprehensively. The Intra-NSAs take the intra-class difference into account and consequently describe objects more accurately. By using the novel multimodal attributes and Intra-NSAs, the better object description that simultaneously have excellent generalizability and discrimination is obtained. Furthermore, additional modal features can be continued integrating into the set of base features to obtain more powerful hybrid attributes. For example, by adding sound features we can classify objects when their visual features are similar; by adding mechanical features other important attributes such as "density" can be trained. In the further work, we will investigate how to combine different features extracted from more other modalities to improve the descriptive ability of attributes.

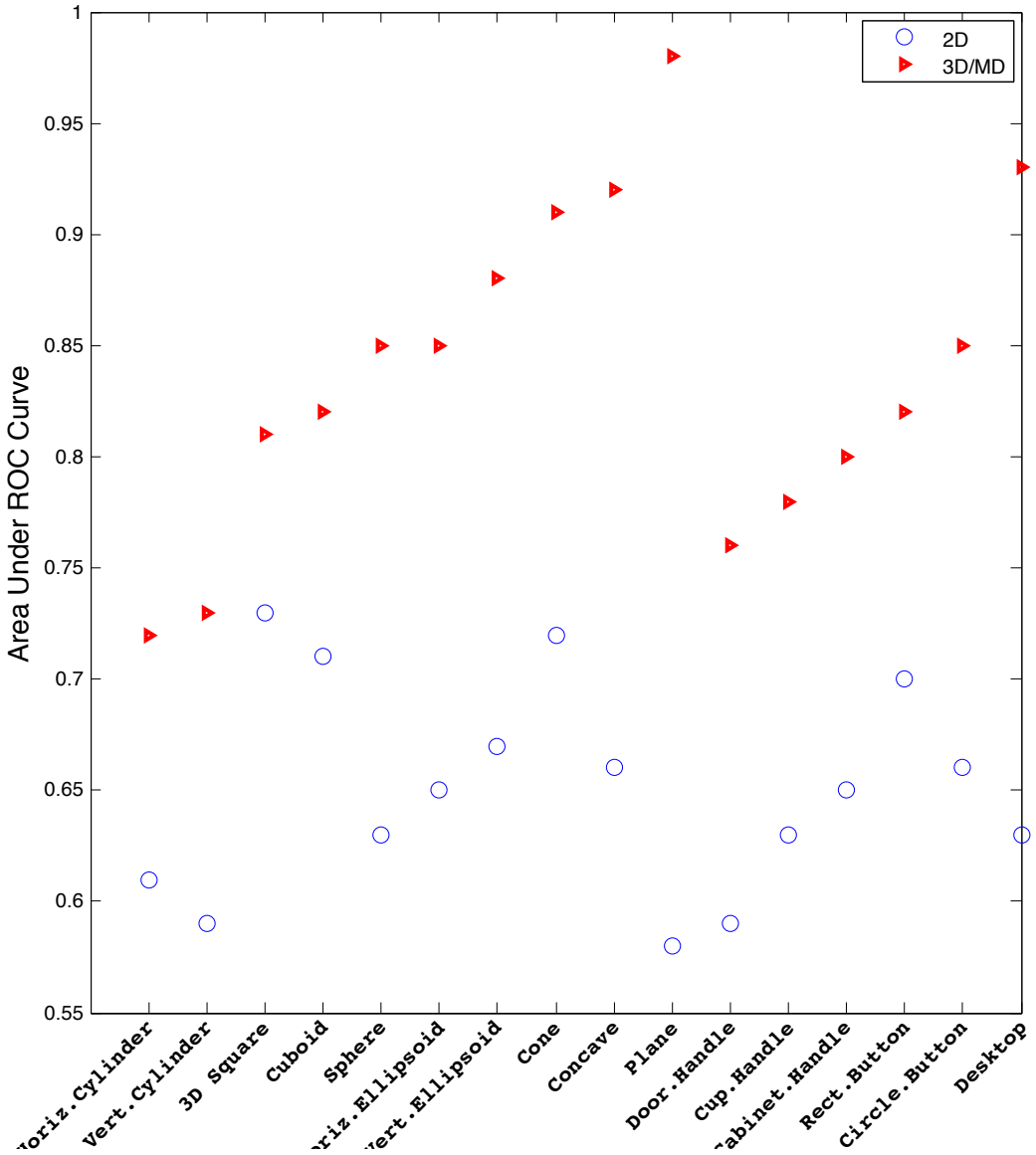


Figure 6.6.: Comparison of accuracy for 10 3D attributes (first ten in this figure) and 6 multimodal attributes (last six in this figure) trained by 2D base feature (blue circle) and 3D/multimodal base feature (red dot).

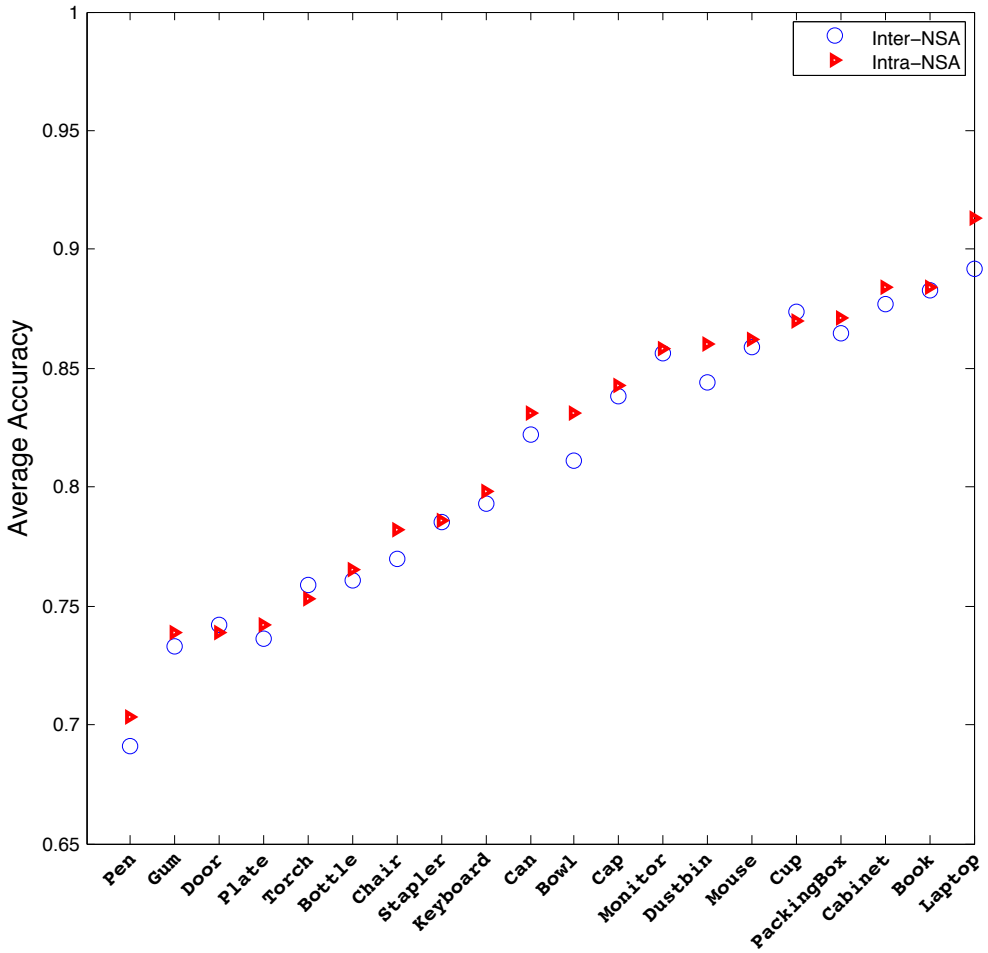


Figure 6.7.: Comparison of accuracy between Inter-NSAs and Intra-NSAs for 20 indoor categories extracted from RGB-D dataset.

Chapter 7

Supervised Hierarchical Latent Dirichlet Allocation

7.1. Introduction

Different categories share some attributes, no matter whether their relationships are near or far. For example, cat, dog, cow and sheep have four legs, bird and airplane have wings. Thus it is possible to cluster different categories, forming a compact category representation. This multi-category representation can be used for improving the object recognition and novel category identification.

One category may share different attributes with different categories. For example, bird can share attributes with cat since both have eyes and can see; at the same time, bird can share attributes with airplane since they have wings and can fly. The shared attributes among different categories can be treated as topics, from which the multi-category representation can achieve several advantages. First, organizing categories according to their shared topics is helpful for deriving the characteristics of novel categories. For example when a novel category shares some attributes with several categories which can fly, this novel category could fly too. Second, one category may have several topics and share with different categories by different topics. This makes the multi-category representation more flexible to be modeled in complicated environments. Third, topics can be treated as a latent layer, by which the category models built from the distributions over attributes can be more efficiently learned and inferred.

The topic models (e.g. Hofmann (2001), Blei et al. (2003)) are natural methods to model such a multi-category representation. Many literature employed the topic models to cluster categories by finding the shared topics, such as Fritz and Schiele (2008), Russell et al. (2006). These models organize the multi-category representation as a flat structure. However, evidences from education (Callanan (1985)), psychology (Murphy and Lassaline (1997), Gosselin and Schyns (2001)) and neurophysiology (Kiani et al. (2007), Kriegeskorte et al. (2008)) have been discovered

that the hierarchy is the natural structure of multi-category representation in human mind. Therefore many researchers proposed diverse methods to organize categories as hierarchical structures, such as Marszalek and Schmid (2008), Zweig and Weinshall (2007), Kapoor et al. (2009), Marszalek and Schmid (2007). However, these methods do not take the shared topics among categories into account (do not explicitly model topics).

Obviously, integrating hierarchies and topics into one uniform model can yield a better multi-category representative model. Therefore Sivic et al. (2008) employed the hierarchical latent Dirichlet allocation (hLDA, Blei et al. (2010)) to build a tree-like hierarchical structure to organize multiple categories. Compared to the flat topic models, the hLDA has several advantages. First, it has the hierarchical, tree-like structure. Second, the structure of the hLDA model can change dynamically as more samples are input. The hLDA, however, is an unsupervised model that cannot guarantee that one path in the tree-like hierarchy mainly corresponds to one category of data. Furthermore, its performance of predicting unseen data is not as good as supervised methods. The supervised latent Dirichlet allocation (sLDA, Hannah et al. (2011)) was extended from the latent Dirichlet allocation (LDA, Blei et al. (2003)) for predicting the response of documents by inferring its topic structure using a fitted model (see Fig. 7.1(a) for its graphical model). The experimental results in Hannah et al. (2011) showed that there was a large improvement of prediction by being compared with the LDA model. However, the sLDA is a flat structure model that cannot meet the requirement of a hierarchical representation. Therefore, in this chapter the hLDA is extended to the supervised hLDA that can integrate the advantages of the hLDA and the sLDA to build a more accurate hierarchical category model.

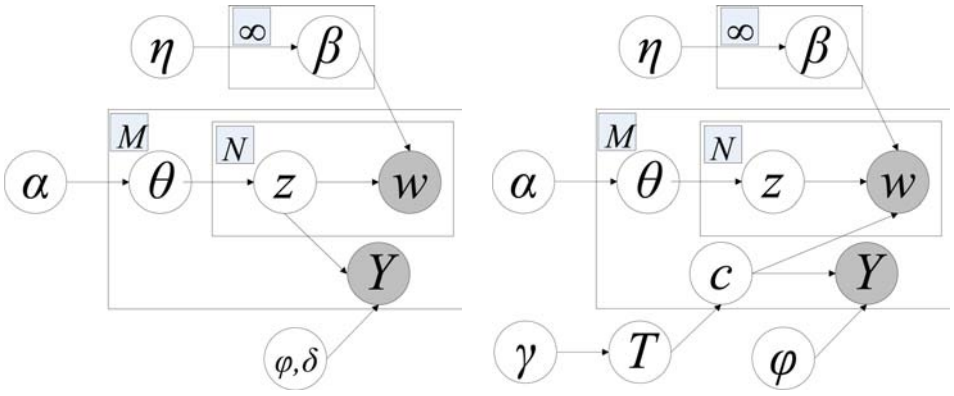


Figure 7.1.: Graphical models of sLDA (a) and supervised hLDA (b). The circles with shadow denote the observed variables.

The supervised hLDA also constructs a tree-like category structure, where each leaf node in the hierarchy exactly corresponds to a concrete category of objects.

Nodes at middle levels, which are constructed according to current object samples and their attributes, are the superordinates of those concrete categories. An example is shown in Fig. 7.2.

Since each topic is a well-defined distribution over attributes, the supervised hLDA can summarize the most distinctive and representative attributes for each category corresponding to each node in the hierarchy, and use these summarized attributes to represent categories and topics more efficiently.

The proposed supervised hLDA is evaluated on two public image datasets and one RGB+D dataset. The experiments test the accuracy of building appropriate category hierarchy. For the image datasets, we use 2D attributes as features to describe objects. For the RGB+D dataset, 2D, 3D and multimodal attributes are used to describe objects. Note the values of attributes for each object are not the ground truth attribute list, but are predicted by attribute classifiers. As it was proved in last chapter that the semantic attributes and Intra-NSAs can lead to better description for objects, they are also employed in this chapter.

7.2. Supervised Hierarchical Latent Dirichlet Allocation

Originally, the hLDA is described by the terms 'documents' and 'words' in Blei et al. (2010). Actually, when this model is applied to this study, the terms 'documents' and 'words' are equal to the terms 'objects' and 'attributes' respectively. Therefore we discuss the supervised hLDA by using all of these terms alternately.

Generally, a document contains several topics. In hLDA as introduced in section 2.3.2, therefore, each document is assigned a path from the root to a leaf node, which means that this document has L topics each of which is corresponding to one node in this path. It is worth noting that nodes located at the upper level correspond to more general topics and nodes located at the lower level correspond to more specific topics. Therefore, different documents that are assigned different paths may share the same superordinates if they have the same general topics but different specific topics.

When using attributes to describe objects, there is a similar situation to documents. Attributes of an object can also form several topics that can represent different aspects of an object with different generality. For instance, in fig. 7.2 the 'diningtable' and 'sofa' have similar furniture part attributes, so they are categorized into one superordinate at the third level. But they are different categories and have different category-specific attributes, so at the fourth level they are assigned to different leaf nodes. Similar examples can be found from other categories.

Generally, each node (i.e. topic) can sample any attributes from the attribute list, but the probabilities of sampling an attribute are different for different nodes. For those leaf nodes, the probability of attributes that can most specifically represent a concrete category will be highest, such as 'horn' and 'saddle' for 'cow' and 'horse'

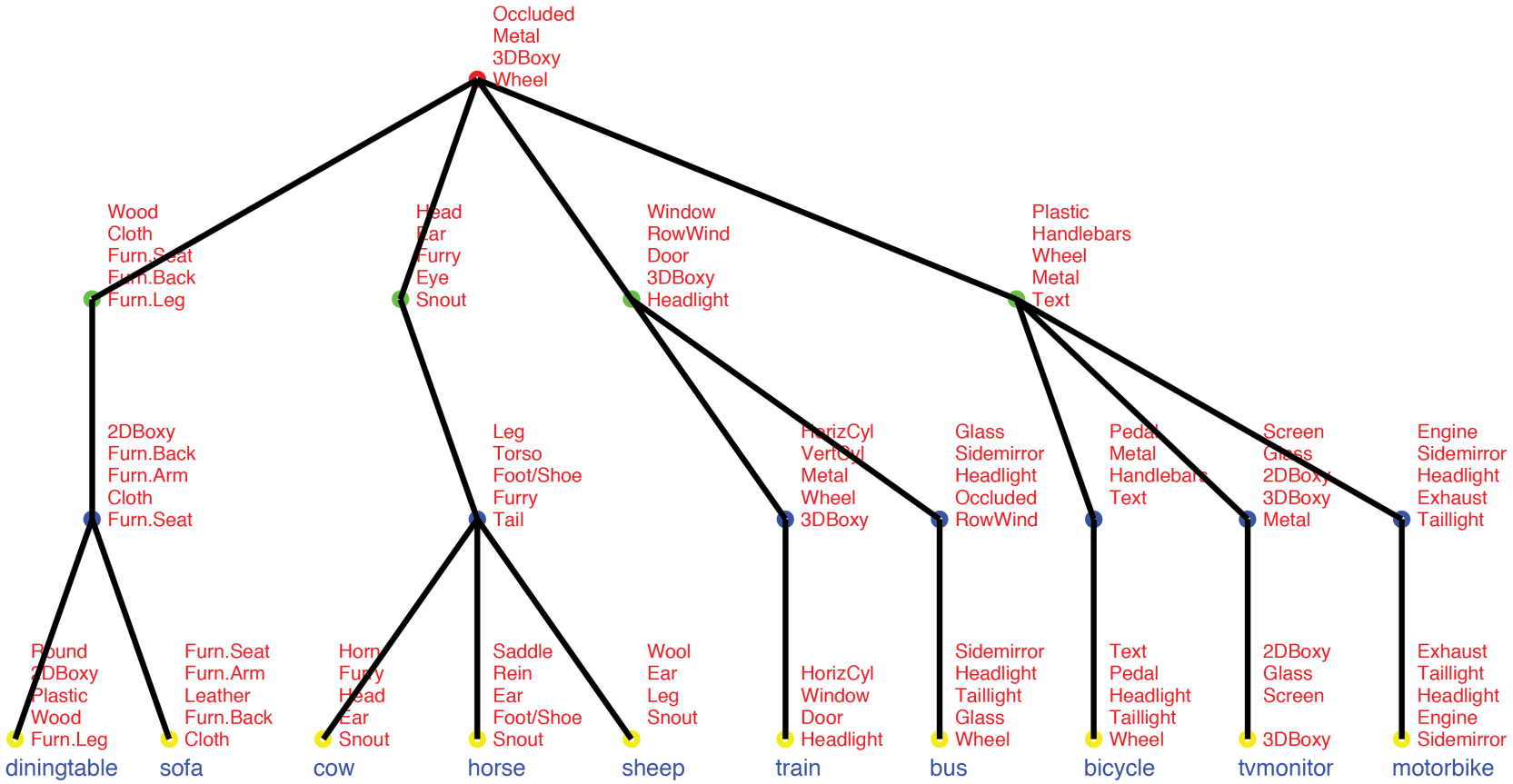


Figure 7.2.: An example of category hierarchy built by the supervised hLDA, where each leaf node exactly corresponds to only one category. The red words near each node is the most representative semantic attributes of the corresponding topic summarized by the supervised hLDA.

in Fig. 7.2, respectively. Nodes located at the middle level trend to sampling those attributes with high probability that represent the common characteristics of categories corresponding to child nodes of the current node, such as 'leg' and 'furry'. Note that a superordinate corresponding to nodes at upper levels does not simply aggregate its child categories, but forms a distribution of attributes to represent this superordinate more accurately. This hierarchical model differs from other models where the hierarchy is built by top-down segregated and/or bottom-up aggregated algorithms (e.g. Marszalek and Schmid (2008), Kapoor et al. (2009)), where categories located at the upper levels of the hierarchy are not an independent description but simply regarded as the aggregate of their child categories.

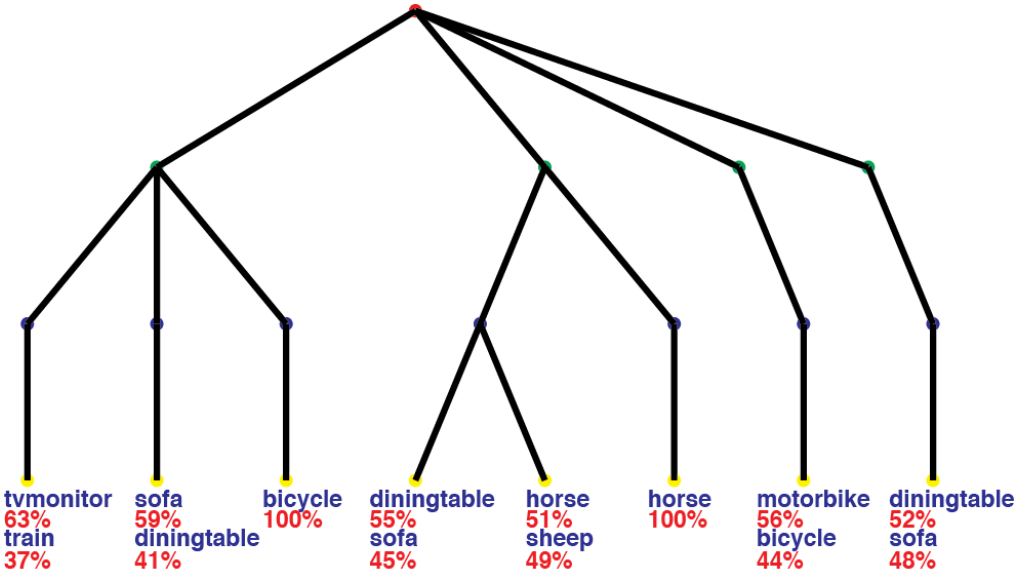


Figure 7.3.: The original hLDA cannot guarantee that each leaf node corresponds to only one concrete category due to its unsupervised manner. Some categories are assigned to more than one leaf nodes. The percentages indicate the ratios of each category in a leaf node.

As a notation of commonsense in the human mind, each leaf node in a category hierarchy corresponds to one concrete category. Moreover, because the samples belonging to each leaf node will be used to train classifiers for classifying new object samples, it is necessary to guarantee that each leaf node is only assigned object samples belonging to the same category. Because the hLDA is an unsupervised algorithm, however, it cannot guarantee that each path in a hierarchy corresponds exactly to one category (see Fig. 7.3 as an example). In our experiments, for each exact category, only 30-60% (the average is about 43%) of samples can be assigned to a correct leaf node. We therefore develop a supervised hLDA by extending the original hLDA whose graphical model is shown in Fig. 7.1(b). A category response is specified for each document in the training data. Thus the generative process of the supervised hLDA is:

1. For each node $k \in \mathbf{T}$ in the infinite tree, draw a topic $\beta_k \sim Dir(\eta)$,
2. For each document, $d \in \{1, 2, ..., D\}$,
   a) Draw $\mathbf{c}_d \sim nCRP(\gamma)$,
   b) Draw a distribution over levels in the tree, $\theta_d \mid \{m, \pi\} \sim GEM(m, \pi)$,
   c) For each word,
      i. Choose level $Z_{d,n} \mid \theta_d \sim Multi(\theta_d)$,
      ii. Choose word $W_{d,n} \mid \{z_{d,n}, \mathbf{c}_d, \beta\} \sim Multi(\beta_{\mathbf{c}_d}[z_{d,n}])$, which is parameterized by the topic in position $z_{d,n}$ on the path $\mathbf{c}_d$.
3. For each document, draw a response $Y_d \sim Multi(\varphi_d)$.

Thus the joint distribution becomes:

$$p(\mathbf{w}, \mathbf{z}, c, \mathbf{Y}, \theta, \beta | \alpha, \eta, T, \varphi) = \prod_{i=1}^{N} p(w_i | z_i, c, \beta) p(z_i | \theta) p(\theta | \alpha) p(\beta | \eta) p(c | T) p(\mathbf{Y} | c, \varphi) \tag{7.1}$$

When the model is training, a specified category response for objects can guarantee that each leaf node corresponds to *only one* category. A path from the root to one leaf node then corresponds to only one concrete category. But it is worth noting that one category may correspond to more than one leaf node. It is reasonable because there may be a situation where the difference of object samples is large enough for these object samples to be treated as two sub-categories. This characteristic is especially useful when distinguishing more than one novel category when they are detected in the same middle node by the algorithm introduced in the next chapter.

The limitation of a built category hierarchy is that its maximal level, L, is fixed since the supervised hLDA must be set a fixed level parameter. The large level parameter can lead to a more complicated hierarchical structure which can more accurately represent the relationship of all categories. However, it will also result in more isolated nodes without any branches and more time-cost. Therefore, the L is set to proportion in $log_2(C)$, where C is the number of categories.

7.3. Inference

Gibbs sampling is employed to perform posterior inference. There are three main steps in posterior inference: the sampling of level allocations, the sampling of path assignments and the sampling of category responses. The last two steps can be merged into one step. There are two situations when sampling the path. When the sampled path is a full path from the root to the leaf node, because the leaf node only corresponds to one category response, the path also corresponds to the same category response. When the sampled path ends at the middle level node, a new

branch will be generated and a new category response will be created. The leaf node generated in this new path will be fixed to correspond to a new category response. For both situations, assigning a path to a document also determines the category response of this document. Therefore, the last two steps can be executed in one step and only two sets of probability need to be computed.

1: When the current path is assigned, the level allocation variable $z_{d,n}$ for word n in document d needs to be sampled from its distribution given the current values of all other variables:

$$\begin{aligned} p(z_{d,n} \mid \mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta) \propto \\ p(z_{d,n} \mid \mathbf{z}_{d,-n}, m, \pi) p(w_{d,n} \mid \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}, \eta) \end{aligned} \tag{7.2}$$

where $\mathbf{z}_{-(d,n)}$ and $\mathbf{w}_{-(d,n)}$ denote the vectors of level allocations and observed words leaving out $\mathbf{z}_{(d,n)}$ and $\mathbf{w}_{(d,n)}$ respectively. And $\mathbf{z}_{d,-n}$ is the level allocations with excluding $\mathbf{z}_{d,n}$ in document d. This equation is the same as the step of sampling level allocations in hLDA. More details can be found in Blei et al. (2010).

2: When the level allocation variables are given, the path associated with each document conditioned on all other paths, all other category responses and the observed words need to be sampled:

$$\begin{aligned} p(\mathbf{c}_d \mid \mathbf{w}, \mathbf{c}_{-d}, \mathbf{Y}, \mathbf{z}, \eta, \gamma, \varphi) \propto \\ p(\mathbf{c}_d \mid \mathbf{c}_{-d}, \gamma) p(w_d \mid \mathbf{c}, \mathbf{w}_{-d}, \mathbf{z}, \eta) \\ p(\mathbf{Y}_d \mid \mathbf{c}, \mathbf{Y}_{-d}, \gamma, \varphi) \end{aligned} \tag{7.3}$$

The first two terms at the right hand of this equation are the same as hLDA. The method of how to compute them can be found in Blei et al. (2010). The third term is different from hLDA since the category response is added. It can be calculated as follows:

$$\begin{aligned} p(\mathbf{Y}_d \mid \mathbf{c}, \mathbf{Y}_{-d}, \gamma, \varphi) \propto \\ \frac{\Pi_y \Gamma(\#[\mathbf{Y}_d = y, \mathbf{c} = \mathbf{p}] + \varphi)}{\Gamma(\Sigma_y \#[\mathbf{Y}_d = y, \mathbf{c} = \mathbf{p}] + N_Y \varphi)} \\ \frac{\Pi_y \Gamma(\#[\mathbf{Y}_{-d} = y, \mathbf{c}_{-d} = \mathbf{p}] + N_Y \varphi)}{\Gamma(\Sigma_y \#[\mathbf{Y}_{-d} = y, \mathbf{c}_{-d} = \mathbf{p}] + \varphi)} \\ p(\mathbf{c}_d \mid \mathbf{c}_{-d}, \gamma) \end{aligned} \tag{7.4}$$

where N_Y is the number of responses at the current sampling iteration and $\mathbf{p}$ means the sampled path. $\#[\cdot]$ denotes the number of elements of an array that satisfies the conditions in the bracket. It is worth noting that the path must be drawn as a block, because its probability at each level depends on its probability at the previous level Blei et al. (2010). If the full path represented by a leaf node (from the root to one of the existing leaf nodes) is sampled, the category response is determined. Therefore, the probability will be one for the document sampling the correct response, or it will be zero for the document sampling the wrong one. If the sampled path end is a middle level node (represented by this middle-level node), the probability is calculated by sampling a new category response.

7.4. Summary of Representative Attributes

By using the supervised hLDA, the most representative attributes for each category and its superordinates can be summarized.

Two kinds of summaries of attributes are taken into account. One is the probability of a particular word at a particular node given the level assignment of words and the assignment of paths, with the assignment of responses not being taken into account. The other is the probability of a particular word at a particular node given the level assignment of words, the assignment of paths and responses. The first probability can be computed by the following equation:

$$p(w|\mathbf{z}, \mathbf{c}, \mathbf{w}, \eta) = \frac{\#[\mathbf{z} = l, \mathbf{c} = \mathbf{p}, \mathbf{w} = w] + \eta}{\#[\mathbf{z} = l, \mathbf{c} = \mathbf{p}] + V\eta} \tag{7.5}$$

This probability is roughly proportional to the number of times that word was generated by the topic at that node. The second probability can be computed by the following equation:

$$p(w|\mathbf{z}, \mathbf{c}, \mathbf{Y}, \mathbf{w}, \eta) = \frac{\#[\mathbf{z} = l, \mathbf{c} = \mathbf{p}, \mathbf{Y} = y, \mathbf{w} = w] + \eta}{\#[\mathbf{z} = l, \mathbf{c} = \mathbf{p}, \mathbf{Y} = y] + V\eta} \tag{7.6}$$

This probability is roughly proportional to the number of times that word was generated by the topic and specified category response at that node. Note that for a leaf node, the results of eq. (7.5) and eq. (7.6) are the same since a leaf node only has one category response. But the results are different for the middle node since it may correspond to more than one category response. As an example shown in fig. 7.2, the red words near nodes are the semantic attributes with the highest probability computed by eq. (7.5). The most representative attributes of the leaf node cow, horse and sheep are 'horn', 'saddle' and 'wool' respectively, which precisely corresponds to human knowledge of the categories cow, horse and sheep. The most representative attributes of their parent node are 'leg' and 'torso'. These attributes coincide with the human knowledge of four-foot animals that can be seen as the parent category of cow, horse and sheep. The detailed experiments and quantitative evaluation will be carried out in the next chapter.

7.5. Experiments

7.5.1. Dataset and Experimental Setup

Our experiments are based on two image datasets and one RGB+D dataset. The first one is used in Farhadi et al. (2009), called a-Pascal (a part of PASCAL VOC 08) and a-Yahoo (collected from the internet). The second one is the LabelMe database (Russell et al. (2008)). The third one is the RGB+D dataset published in Lai et al. (2011). They are widely used for evaluating the performance of object recognition, scene recognition, etc.. For the two image datasets, three experimental scenarios

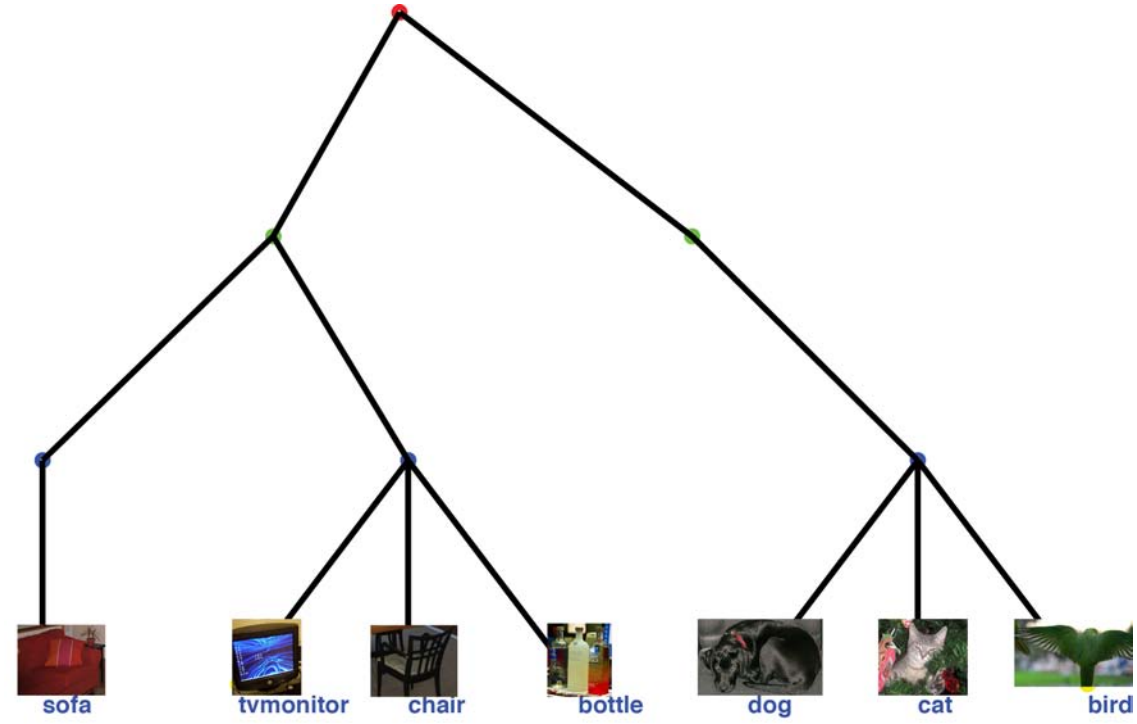
are designed. The first one uses 32 categories in a-Pascal and a-Yahoo (referred as S1). The last two scenarios are based on two kinds of scenes in LabelMe, the office (referred as S2) and the street (referred as S3). 25 and 28 categories which usually appear in offices and streets are used. For the RGB+D dataset, 20 categories used in indoor environments are extracted (referred as S4). All categories used in four scenarios are listed in Appendix B.

7.5.2. Evaluation of the Built Category Hierarchies

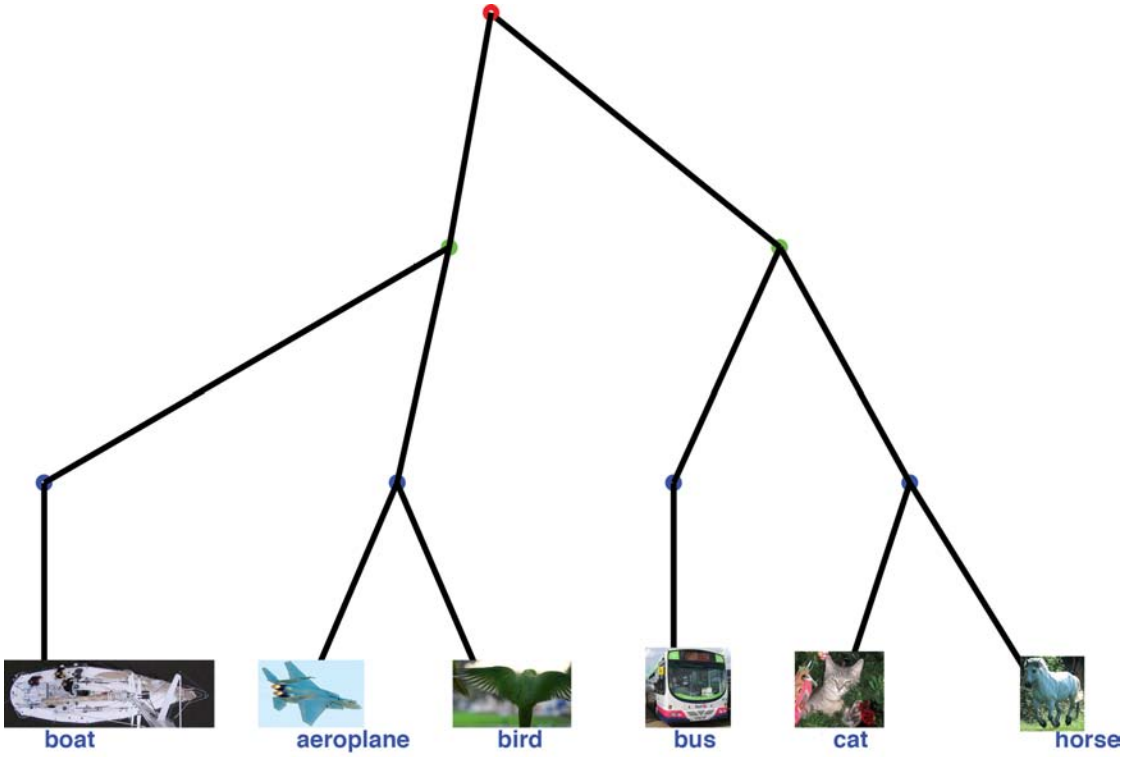
The precision in building correct category hierarchies by the proposed supervised hLDA are evaluated. Since one of purposes of this thesis is to explore novel categories in unexplored environments, the objects' locations and categories in such environments are not fixed. Thus we have to build the category hierarchy in a dynamic environment. For four scenarios, we simulate changes of circumstances by randomly selecting a part of categories and evaluate the precision of building correct category hierarchies based on these categories. Before extracting attributes to describe objects, they have to be bounded by boxes. In two image datasets, we use their ground truth bounding boxes. The 2D semantic attributes and Intra-NSAs are used as features to describe objects. In the RGB+D dataset, we use the CMH-CRF model introduced in section 5 to detect objects and obtain their bounding boxes. The 2D, 3D and multimodal semantic attributes and the multimodal Intra-NSAs are used to describe objects. In the first three scenarios, we choose 80 samples for each category to execute the following evaluation. In the fourth scenario, we choose 25 samples for each category to execute the experiments.

Under different circumstances, the category hierarchy is different since categories in different circumstances are different. In a certain circumstance some categories may belong to a superordinate, but in another circumstance they may belong to different superordinates, which is a reflection of the dynamic property of category hierarchies. An experimental result is shown in Fig. 7.4 (for the sake of clarity, a part of categories in this hierarchy is shown). Note the category 'bird' is categorized into different superordinates in two hierarchies. In the left one 'bird' is categorized into the superordinate shared by 'cat' and 'dog', since according to all of the categories used in this hierarchy it is most reasonable to group these three categories as 'animal'. In the right one, 'bird' belongs to the superordinate shared by 'aeroplane', since here all categories can be obviously divided into three groups of objects moving in the air, on land and on water, respectively.

Categories used to build hierarchies are changed randomly to simulate the changes in circumstances. In practice, 50 hierarchies are built for each experimental scenario. For each experiment, we randomly select 20, 15, 20 and 12 categories to build hierarchies for four corresponding scenarios respectively. The level of each hierarchy is set to 4. Other parameters of the supervised hLDA are set to $\eta = \{8.0, 6.0, 3.0, 3.0\}, \gamma = 0.5, m = 0.35, \pi = 100$, and $\varphi_d = 0.1$ for all categories. By using these parameters, the supervised hLDA can converge after several itera-



(a)



(b)

Figure 7.4.: Dynamically changed hierarchies according to different circumstances. Note that the 'bird' category has the same superordinate as the 'cat' and 'dog' categories in (a) and has the same superordinate with the 'aeroplane' category in (b).

Table 7.1.: The average correctness rate of category hierarchies

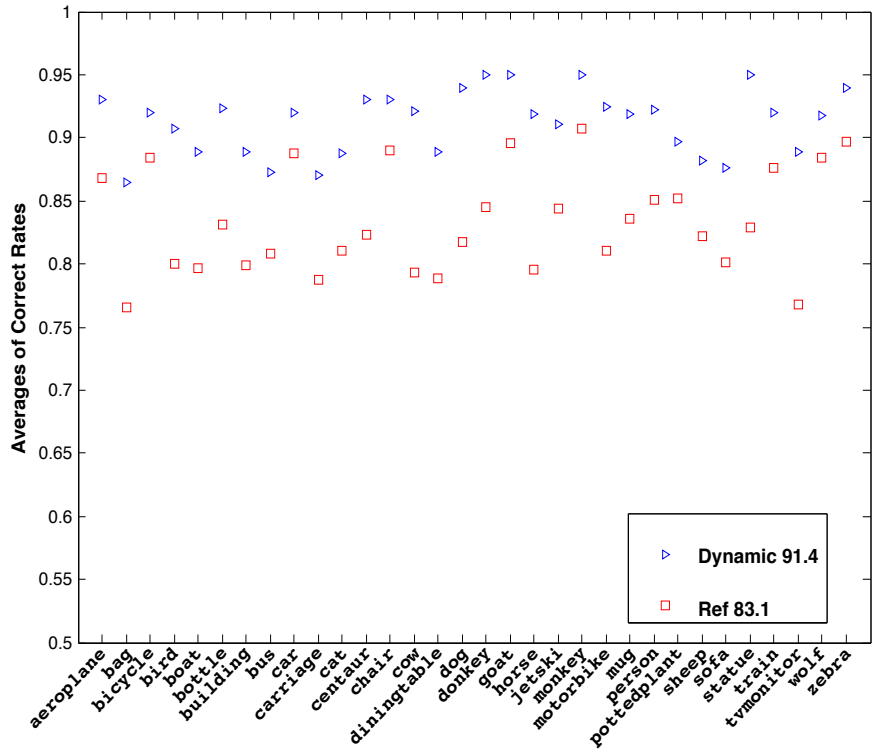
	S1	S2	S3	S4
Dynamic	91.4	90.3	92.1	85.2
Ref	83.1	79.1	78.5	72.5

tions (the average number of iterations is 15 in our experiment). Because there is no standard category hierarchy as a reference that can be used to evaluate the correctness of our built hierarchies, ten adults (five men and five women ranging from 20 to 50 years of age) are employed to determine the correctness for each category in a hierarchy. One state-of-the-art method in Marszalek and Schmid (2008) is used as a reference method for comparison. A softening parameter of this method is set to $\alpha = 0.5$ according to the suggestion in Marszalek and Schmid (2008). Fig. 7.5 shows the average correct rates for all categories in the four scenarios and Table 7.1 lists the average correctness rates of each scenario.

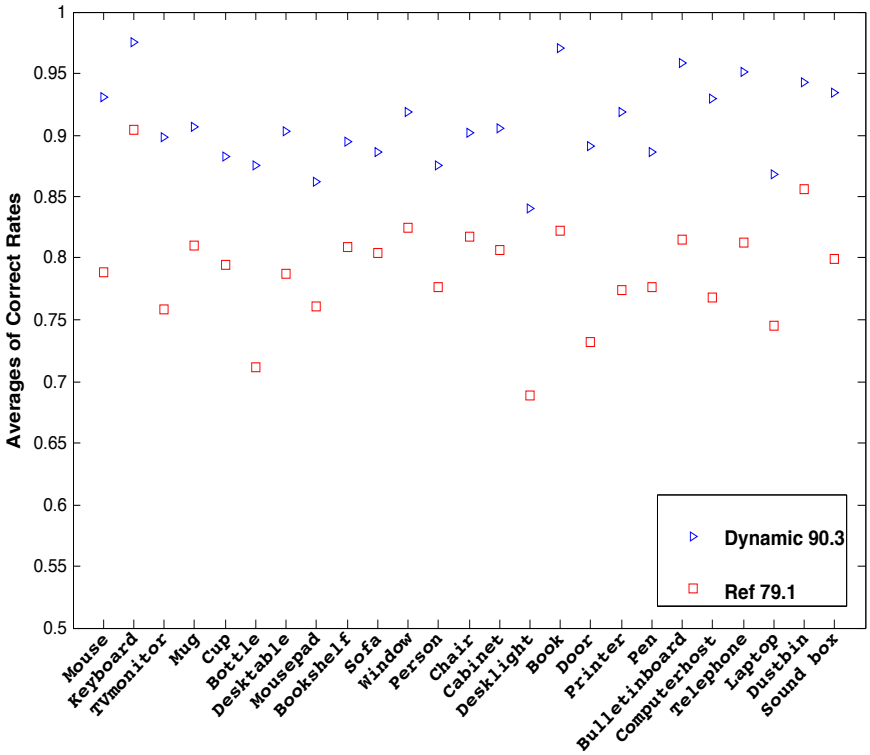
From these figures, it is obvious that the performance of the proposed framework to build category hierarchies is promising. Since the supervised hLDA uses the non-parametric Bayesian technology, it can obtain a more natural distribution for all topics than the reference method which uses the hierarchically normalized cuts to construct the hierarchy. The correct rates by the proposed method is higher than that of the reference method. The average of the correct rates is improved by 8-14% respectively. The averages of the first three groups are larger than 0.9. It is worth noting that there are several categories that have a very high accuracy. That is because these categories are ambiguous, so they can belong to different superordinates. For example, we can regard the 'monkey' category as animal, which is the superordinate for 'horse', 'sheep', etc, or regard it as a human-shape category which is the superordinate for 'person'. The accuracy of the fourth scenario is relative lower because the bounding boxes for objects used in this group experiments are not annotated manually but are detected by our CMH-CRF model, which accuracy is of course lower. But the result is still promising since it is higher than 85%.

7.6. Conclusion

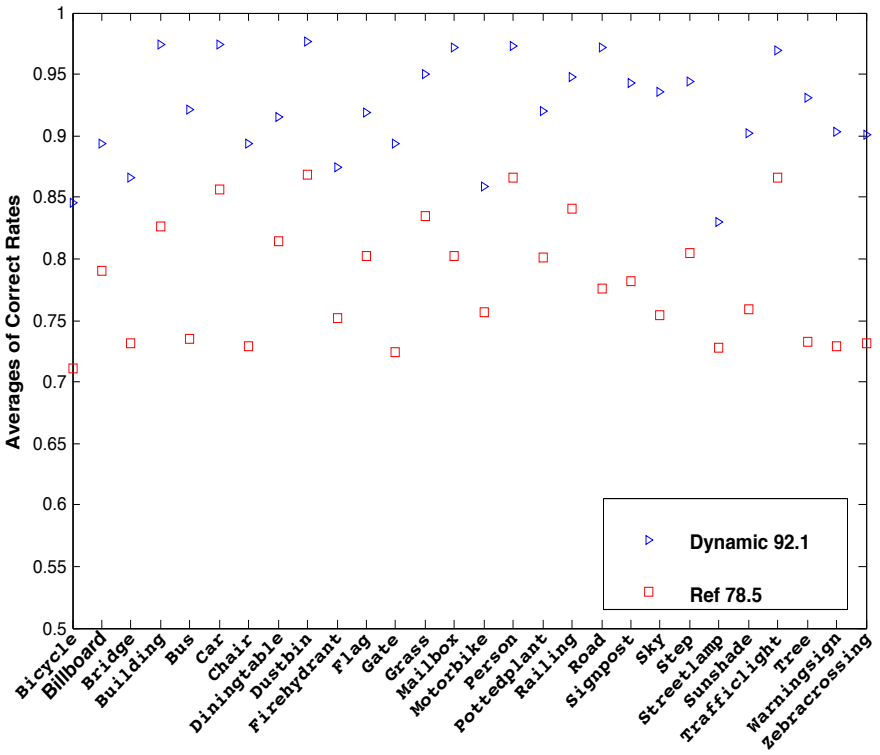
In this chapter, we proposed the novel supervised hLDA, which integrates the advantages holding by hierarchical topic models and supervised topic models into the uniform one. Through this supervised hLDA, the category hierarchies can be built more accurately. Furthermore, the built hierarchies have the adaptive context-aware structure which can change according to the changes of object categories in the environment. Based on it, we will conduct a novel dynamic hierarchical category model which can improve the object recognition for known categories and efficiently identify new objects and discover novel categories.


Averages of Correct Rates
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Dynamic 91.4
Ref 83.1
aeroplane
bag
bicycle
bird
boat
bottle
building
bus
car
carriage
cat
centaur
chair
cow
diningtable
dog
donkey
goat
horse
jetski
monkey
motorbike
mug
person
pottedplant
sheep
sofa
statue
train
tvmonitor
wolf
zebra

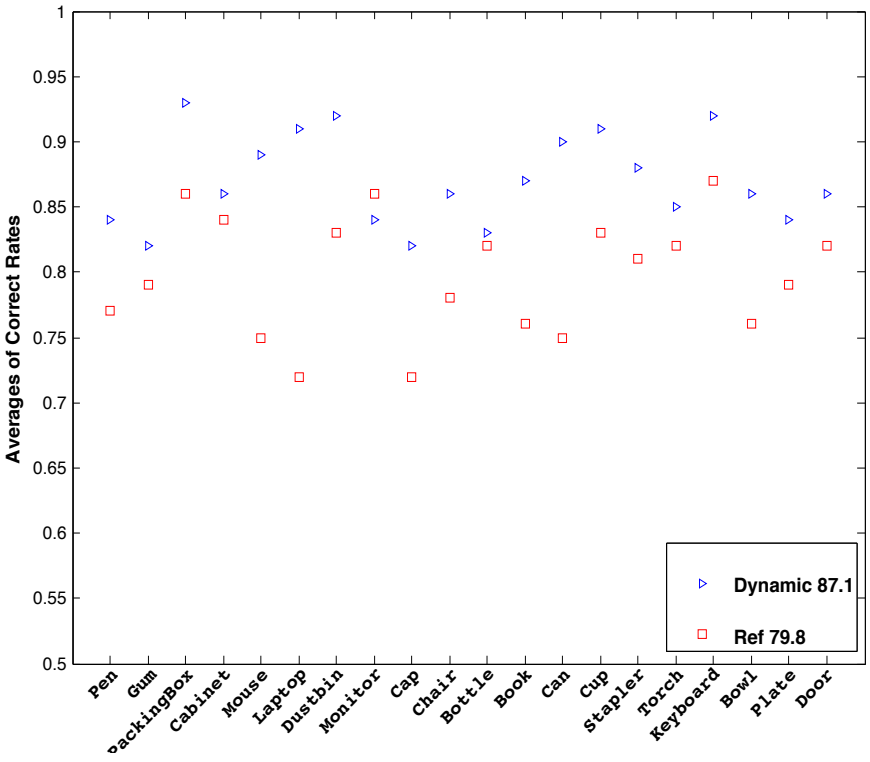
(a)


Averages of Correct Rates
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Dynamic 90.3
Ref 79.1
Mouse
Keyboard
TVmonitor
Mug
Cup
Bottle
Desktable
Mousepad
Bookshelf
Sofa
Window
Person
Chair
Cabinet
Desklight
Book
Door
Printer
Pen
Bulletinboard
Computerhost
Telephone
Laptop
Dustbin
Sound box

(b)



(c)



(d)

Figure 7.5.: Correctness rate of category hierarchies built from S1 (a), S2 (b), S3(c) and S4(d). 'Dynamic' and 'Reference' are the methods of using the proposed framework and the reference method to build hierarchies.

Chapter 8

Dynamic Category Hierarchies for Discovering Novel Category

8.1. Introduction

A category hierarchy is an inherent structure in the human mind. When people are 2-4 years old, they are taught to build category levels to help them to remember more and more objects (Callanan (1985)). Flexible hierarchies of categories of abstraction have become central to modern theories of categorization and recognition (Murphy and Lassaline (1997), Gosselin and Schyns (2001)). Researchers have also discovered neurophysiological evidence for category hierarchies (Kiani et al. (2007), Kriegeskorte et al. (2008)).

We argue that a basic characteristic of a category hierarchy is dynamics. There are two aspects of the dynamics of a category hierarchy. First, a category will be assigned to different superordinates in different cognitive environments. Since humans live in a dynamic world, the circumstances change dynamically. One object that belongs to one category at a certain moment may be categorized into another category in the next moment by human minds. For instance, a vase on a dining table is always looked at as a container for arranging flowers, while it will often be regarded as an artwork when people see it in a museum. Different cognitive goals also change a category hierarchy. For instance, when people intend to teach children knowledge about animals, a cat toy will be categorized as an animal. When people want to let children play with a toy, however, this cat toy is definitely used as a toy. This reveals that a built hierarchy should change its own structure according to different circumstances and purposes.

Second, when novel categories emerge, a category hierarchy can dynamically change its structures, branching off automatically at an appropriate node and generating new nodes to represent novel categories. This process can be found in human cognition for learning objects of a novel category. People will analyze properties of the new objects and find similar categories which have some of these properties.

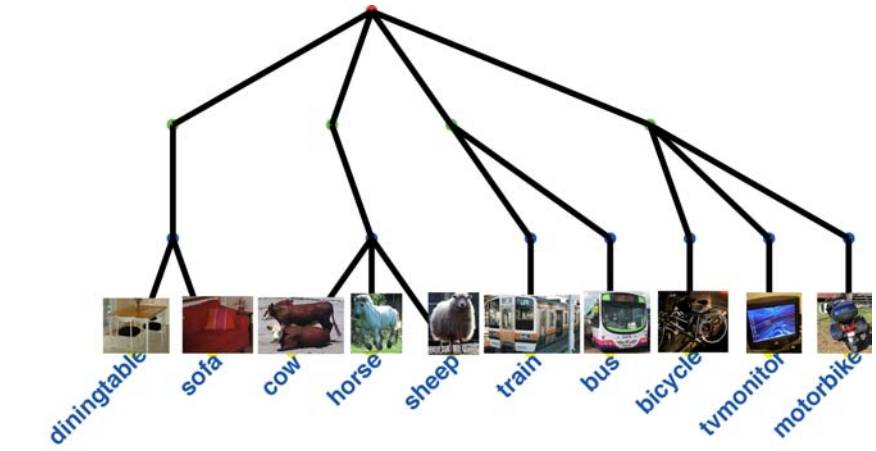
People can then categorize the new objects into a novel category that has the same superordinate as those similar categories.

In this chapter, motivated by human cognition of category hierarchies, we propose a novel framework for building a dynamic category hierarchy, which can simultaneously satisfy two aforementioned dynamic aspects. The proposed framework is based on object attributes (chapter 6) and the supervised hierarchical Latent Dirichlet Allocation (shLDA, chapter 7). The extended object attributes can describe objects more comprehensively, and their excellent generalizability across categories can help our framework to identify and describe the novel categories well. The shLDA can build a more accurate category hierarchy. By means of it, each leaf node in a hierarchy exactly corresponds to a concrete category of objects. Nodes at middle levels, which are constructed according to current object samples and their attributes, are the superordinates of those concrete categories. Since the hierarchy is totally built from those object samples without any assumptions, the structure of the built hierarchy can be changed, as the concerned object categories and attributes vary from different circumstances and for different goals (See Fig. 7.4 as an example). Because the supervised hLDA can help the existing hierarchy to generate new branches when novel categories appear, and object attributes can describe objects across categories, even though those categories have not been seen before, the proposed framework can meet the second aspect as well (See Fig. 8.1 as an example).

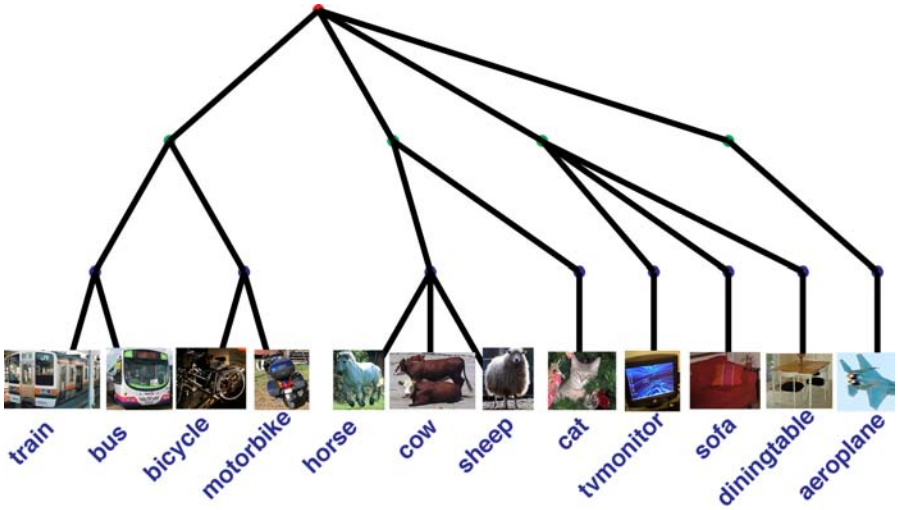
Moreover, our framework has an extra advantage that the built hierarchy can keep on enhancing its ability for novel category discovery and object recognition by incrementally learning from those incoming objects. By using topic models and object attributes, it can summarize those most distinctive and representative attributes for each category corresponding to each node in the hierarchy, and use these summarized attributes to recognize objects so as to improving the recognition efficiency (see Fig. 7.2 as an example). This is also similar to human cognition. As people access more and more objects, they can improve the recognition of known objects and the categorization of novel objects. For instance, it is obvious that adults have a better cognitive ability than children.

To summarize, our framework contributes several important features. First, a dynamic category hierarchy can be built by our framework. Second, based on the dynamic category hierarchy, novel categories can be effectively discovered and described. Third, the recognition of known objects can also be improved. Fourth, object attributes can be summarized for each category to describe an object more compactly.

The proposed framework is evaluated on three public datasets by extensive experiments. Based on the built category hierarchies, we test the accuracy of object recognition and prove that the performance of object recognition using hierarchical representation is significantly improved with respect to that of not using hierarchical structure. More importantly, experiments show that the proposed framework can efficiently discover different novel categories and describe them. The summarized representative and distinctive attributes are also evaluated.



(a)



(b)

Figure 8.1.: Automatically branching off at a corresponding level with each new category. Note that in (b) there are two new categories, 'cat' and 'aeroplane'. The 'cat' has the same super-superordinate with other animal categories since they have similar attributes. The 'aeroplane' here is treated as totally different from the other categories and consequently the hierarchy branches off a new path at root level. The results closely conform to the general hierarchy in the human mind.

8.2. Dynamic Category Hierarchies

Our framework can be outlined by a block diagram shown in Fig. 8.2. The attributes of the annotated objects in the training image dataset are computed based on the extracted base features. Given these attributes, a hierarchy representing the relationship among current categories can be built by using supervised hLDA. For each node in the built hierarchy a classifier can then be trained based on those object samples assigned to this node. The attributes of new object samples are also computed. Then node classifiers are used to determine if they belong to known concrete categories, or unseen categories. If new object samples are corresponding to unseen categories, our framework can indicate their superordinates and the hierarchy will branch off at appropriate nodes, then generate new paths to represent new categories. Therefore our framework feeds back predicted results to the built hierarchy, by which the hierarchy can change dynamically.

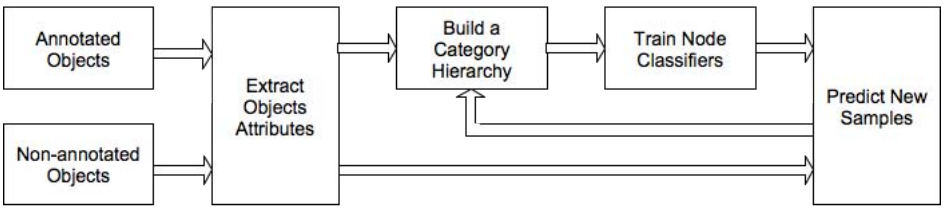


Figure 8.2.: The block diagram of our proposed framework for building dynamic category hierarchies.

8.2.1. Extraction of Object Attributes

We use the algorithm introduced in chapter 6 to extract object attributes. For 2D images, 64 ordinary and simple semantic attributes from several aspects such as color, shape, material & texture, and part are extracted. For RGB+D data, the 2D, 3D and multimodal semantic attributes with an amount of 80 are extracted. These semantic attributes, as listed in Appendix A, can be expressed by simple words such as red, cubic, metallic, feather. For different application circumstances, of course, some semantic attributes may mainly be considered while the others may be omitted. Some semantic attributes can be filtered to describe objects more efficiently. By using filtered attributes, category hierarchies will be built more accurately. In this chapter, however, we aim at evaluating the general performance of our framework and therefore use all semantic attributes without filtering.

We also train 1000 Intra-NSAs to enhance the object description. Finally, for each object there is a vector consisting of semantic and non-semantic features denoted by $\mathbf{a} = \{a_1, a_2, \cdots, a_N\}$ where N is the number of all attributes and is equal to 1064 or 1080 for pure 2D data or RGB-D data respectively. In such a vector, if an object

has any attributes, the values of corresponding items are 1. Otherwise, the values are 0.

8.2.2. Construction of Category Hierarchies

With the list attributes at hand, we can treat it as a vocabulary, and consider each object described by attributes as a document consisting of a subset of this vocabulary. The only difference is that in such a document each word in the vocabulary is sampled at most once. For all samples of labelled object, the initial category hierarchy, H^0, can be built by using the supervised hLDA. Based on a built category hierarchy, category labels of any new object samples can then be predicted. After predicting a certain number of new samples, these samples are combined with the samples which are previously used to build H^0 to build a new category hierarchy H^1. Thus a current category hierarchy H^t based on all known object samples at any time t can be obtained. Therefore, the category hierarchy can change its structure dynamically according to current object samples and their categories.

However, it is worth noting that the category hierarchy will not change immediately after each and every new sample categorization. That is because a category hierarchy is a relatively stable structure in human cognition. Although it may change when cognitive goals or circumstances change, it will not change when only a few new objects are met. Therefore, it is unnecessary to change the built category hierarchy for each new object sample. In practice, the built hierarchy will be updated only when the number of new object samples exceeds a threshold.

When a category hierarchy is built, the distribution of attributes associated with one node can also be determined. The probability of one node sampling an attribute is roughly proportional to the sum of occurance times of this attribute in all object samples belonging to this node. The higher the occurance frequence of an attribute is, the larger the number of object samples possessing this attribute is. The most representative attributes of one category must be those attributes that are most often held by the instances of this category. Therefore, the most representative attributes of each category can be extracted according to the distribution of attributes of each node. These most representative attributes can be used to form a more compact description of objects.

When a built category hierarchy is updated, the distributions of attributes associated with all nodes will be changed. One possible situation of an updated category hierarchy is that the structure does not change but the number of samples of nodes changes (all new samples belong to known categories). Another situation is that the structure has changed (some new samples belong to unknown categories and new nodes are added to represent these new categories). For the first situation, along with the input of new samples, the distribution of attributes of each node can describe the state of the corresponding categories more accurately. For the second situation, with the new nodes appended, the state of their superordinates will be

represented more precisely. Under both situations, therefore, the efficiency of object recognition and novel category discovery will be improved.

8.2.3. Training of Classifiers and Prediction of New Objects

After building a category hierarchy, it can be used to recognize objects and discover novel categories. Two methods are developed to predict the category of a new object sample. The first method directly uses supervised hLDA to compute the probabilities of sampling every path in the hierarchy, then the path corresponding to maximal probability is the label of this new object sample. If this path exists, the new object sample belongs to a known category. If this path does not exist, the new object sample belongs to an unknown category and the node that branches off this new path is the superordinate of this new category. For the second method we first train a classifier for each node. Starting from the root of the hierarchy, node classifiers are used to predict the label for a new object sample. If this new sample is classified as belonging to current node, then classifiers corresponding to child nodes of this node are also used to predict if this new sample belongs to these child nodes. Until this procedure is established as the new sample is classified as belonging to a leaf node, or cannot find any child nodes that this sample belongs to, the prediction of new sample is finished. If this sample is predicted to belong to a leaf node, it is an object of a known category. If this sample belongs to a node but not to any child nodes, it is an object of an unknown category and the hierarchy will branch off a new path at this node. Both methods are introduced in detail as follows.

Prediction by Supervised hLDA

To predict category responses for new inputs, $p(\mathbf{Y}_d \mid \mathbf{c}, \mathbf{Y}_{-d}, \gamma, \varphi)$, the probability of sampled responses based on given word level assignments and path assignments needs to be computed according to eq. (7.4). Here we need to consider all possible paths, including existing paths (from the root to leaf nodes) and new paths that will branch off at middle level nodes. Therefore, the response is obtained by the maximal probability of the following equation:

$$y = \underset{\mathbf{c}\in\mathcal{C}}{arg\max}\, p(\mathbf{Y}_d \mid \mathbf{c}, \mathbf{Y}_{-d}, \gamma, \varphi) \tag{8.1}$$

where $\mathcal{C}$ denotes all possible paths in the hierarchy. At this time, the correct response is unknown. The probabilities for all possible responses need to be computed, and the predicted result is the response with the maximal probability. Before the level allocations for all words are sampled for computing the probability of sampling a category response, the initial path assignment for the document needs to be determined. However, the correct response for this document is unknown. Therefore all existing responses may be sampled and all middle level nodes may also branch off a new path to generate a new response. Thus, for obtaining the level allocation of all words, all possible initial path assignments are considered. For each level

allocation, the maximal probability of sampling responses is obtained. The final decision is the respond corresponding to the global maximal probability by comparing with all these maximal probabilities. For an object belonging to one existing category, the path tends to sample an existing full path represented by a leaf node and consequently samples an existing category response. For an object belonging to an unknown category, the path tends to sample a path represented by a middle level node and branches off a new full path, adding a new category response.

After predicting a certain number of new inputs, the selection of responses and the allocation of words will be added into the model for the following prediction. Consequently, the structure of the model will change when a new category is input and the distinctive attributes of existing categories will be enhanced when a known category is input.

However, the accuracy of predicting new object samples is relative low. In our experiments, the accuracy ranges from 30%-60% for known categories and 40%-60% for unknown categories. Therefore it is necessary to develop a better method to predict new object samples.

Prediction by SVM

To improve the accuracy of predicting new object samples, we employ the SVM to train classifiers for all nodes in a built category hierarchy. The reason for choosing the SVM is because it can achieve high accuracy of classification and has convenient pre-built library (i.e. LibSVM in Chang and Lin (2011)).

Assuming a current category hierarchy is H^t. The ith node located at the lth level is denoted as D_i^l. Its parent node and the set of child nodes are denoted as $PD^{D_i^l}$ and $\mathbf{CD}^{D_i^l} = \{CD_1^{D_i^l}, \cdots, CD_m^{D_i^l}\}$ respectively. At first, the classifier for the root D_1^0 in H^t is trained. Because the root D_1^0 contains all object samples used for building H^t which can be regarded as positive samples, no negative samples are provided. Thus the traditional two-class SVM cannot be used. The one-class SVM introduced in Chang and Lin (2011) is used to train the classifier for the root, denoted as $S_{D_1^0}$. For other nodes, a one-class SVM or a two-class SVM can be used to train classifiers. If using a one-class SVM to train a classifier for a node D_i^l, only the object samples belonging to this node are considered as positive samples and other object samples are not taken into account. According to Chang and Lin (2011), a one-class SVM constructs a hyper-sphere in the feature space enclosing the image of all object samples belonging to this node. If using a two-class SVM to train a classifier for a node D_i^l, the object samples belonging to this node are considered as positive samples and others as negative samples. We are denoting the classifier of node D_i^l as $S_{D_i^l}$. Similarly, classifiers of the parent node and the children node set of D_i^l are denoted as $S_{PD^{D_i^l}}$ and $\mathbf{S}_{\mathbf{CD}^{D_i^l}} = \{S_{CD_1^{D_i^l}}, \cdots, S_{CD_m^{D_i^l}}\}$. In practice, we use LibSVM to train these classifiers. After obtaining classifiers for all nodes in a current category hierarchy, a recursive function that is listed in alg. 8.1 is used to predict the category label of a new object sample.

Algorithm 8.1: Recursive Function for Predicting a new object sample.

Input:
The attribute vector of a new object sample, A^{new};
Current node D_i^l, the classifier and category label of the current node, $S_{D_i^l}$ and $L_{D_i^l}$;
The parent node and child node set of the current node, $PD^{D_i^l}$, $\mathbf{CD}^{D_i^l}$;
The associated classifier for the child node, $\mathbf{S}_{\mathbf{CD}^{D_i^l}}$;
Output:
The category label of the new object sample, B^{new};
The node of the corresponding superordinate of the predicted result, PD^{new}
if D_i^l *is a leaf node* **then**
    if A^{new} *is classified as belonging to this node by* $S_{D_i^l}$ **then**
        Return: $B^{new} = L_{D_i^l}$, $PD^{new} = PD^{D_i^l}$;
    else
        Return: $B^{new} = L_{new}$, $PD^{new} = PD^{D_i^l}$;
else
    if A^{new} *is classified as belonging to this node by* $S_{D_i^l}$ **then**
        for *each* m *in* $\boldsymbol{CD}^{D_i^l}$ **do**
            Select $CD_m^{D_i^l}$ as current node;
            Find associated classifier $S_{CD_m^{D_i^l}}$ for $CD_m^{D_i^l}$;
            Find category label for $CD_m^{D_i^l}$;
            Use D_i^l as parent node for $CD_m^{D_i^l}$;
            Find child node set and associated classifier set for $CD_m^{D_i^l}$;
            Invoke this function recursively;
    else
        Return: $B^{new} = L_{new}$, $PD^{new} = PD^{D_i^l}$;

By using this algorithm, a category label can be predicted. Note that if the new object sample belongs to a known category, the algorithm will return this known category label and its superordinate, and if the new object sample belongs to an unknown category, the algorithm will return 'new' as its category label but will also indicate its superordinate which means the hierarchy will branch off at the node corresponding to its superordinate. In our experiments, the average accuracy of recognizing known objects ranges from 74% to 86% by using this method, which is a considerably greater improvement than the accuracy resulting by directly using supervised hLDA. Similarly, the average accuracy of the novel category discovery ranges from 69% to 84%.

8.2.4. Determination of Categories for New Objects

When more than one object sample belonging to novel categories is detected , these samples may belong to different novel categories within one superordinate. However, previous steps can only give a 'new' label to all of them but cannot determine how many novel categories there are, nor further distinguish which sample belongs to which novel category. This is the problem of unsupervised clustering object categories and many methods can be employed to solve this problem (Tuytelaars et al. (2010)), including latent variable methods, and spectral clustering methods. The method to clustering object categories used in the proposed framework is a kind of latent variable method.

As mentioned in section 7.2, the supervised hLDA has an important characteristic that it will create more than one leaf node for one category if the differences among objects are large enough. After novel object detection, all new objects are assigned to one new leaf node. If these objects belong to different categories, there must be enough difference. Thus the generative procedure in section 7.2 will automatically cluster these new object samples into different new leaf nodes, and consequently change the structure of the built category hierarchy to form a new hierarchy.

Based on a category hierarchy, the proposed framework has at least one advantage with respect to the determination of categories for new object samples. Since only object samples belonging to similar novel categories will be assigned to the same new node, the number of novel categories which need to be distinguished at the same time is decreased. Thus the performance of distinguishing categories will be improved.

8.3. Experiments

The experimental setup is the same to that used in chapter 7. Based on the built category hierarchies in chapter 7, four groups of experiments based on four scenarios are used to evaluate the performance of: 1) object recognition for known categories,

Table 8.1.: Average accuracy of object recognition

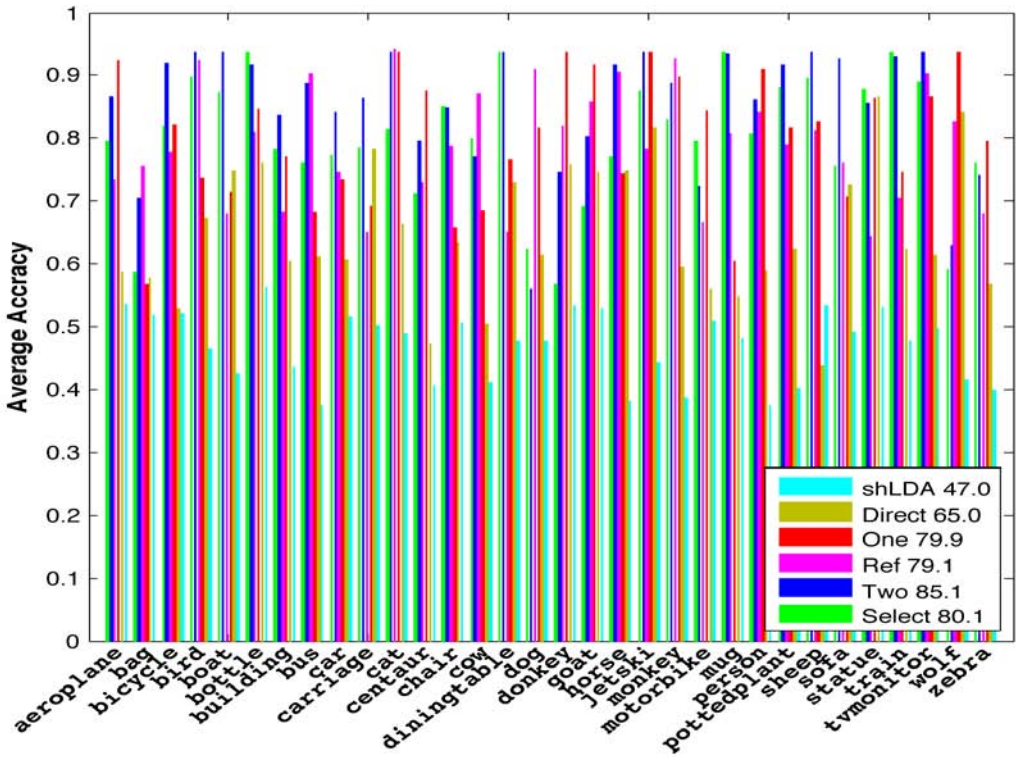
	S1	S2	S3	S4
One-SVM	79.9	77.0	79.0	69.9
Two-SVM	85.1	86.1	85.9	80.1
Select	80.1	85.9	84.7	78.6
Direct	65.0	71.8	74.1	63.1
shLDA	47.0	40.2	47.5	38.9
Ref	79.1	76.4	75.2	70.3

2) detecting novel categories, 3) and distinguishing these novel categories, respectively. At last, the effectiveness of the summarization of attributes for each category is evaluated.

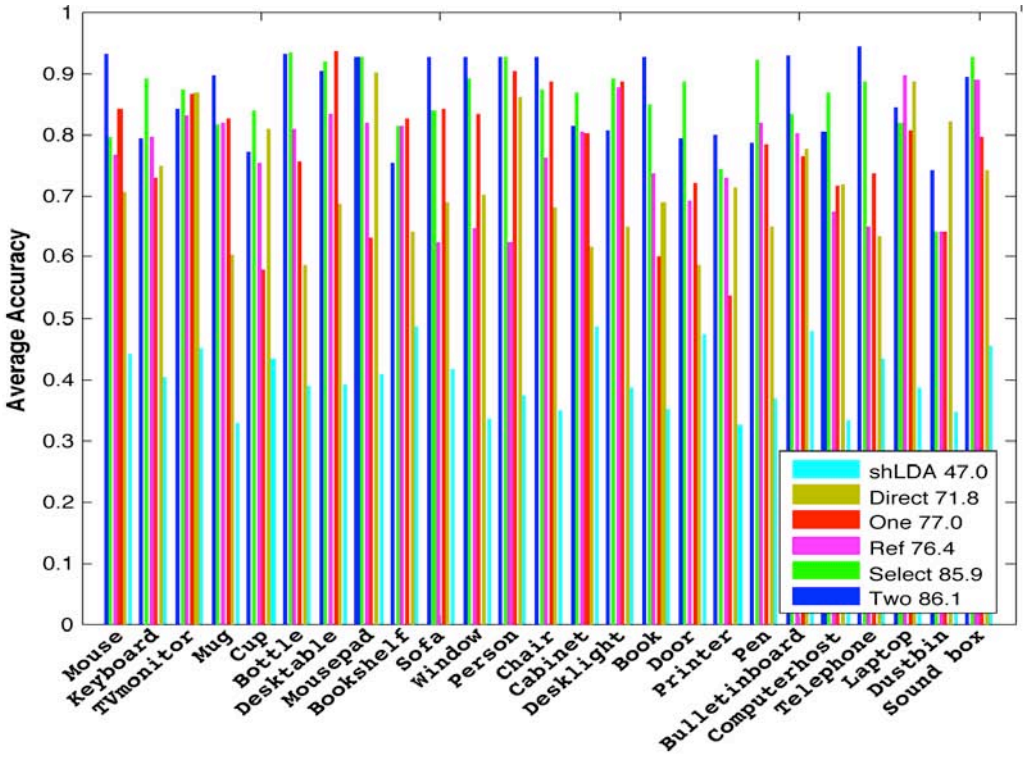
8.3.1. Evaluation of Object Recognition

The accuracy of object recognition can be improved by using a hierarchical structure as shown in Marszalek and Schmid (2008), Marszalek and Schmid (2007), Zweig and Weinshall (2007), Kapoor et al. (2009). In this section we show this improvement by using the proposed category hierarchies. By comparison with a state-of-the-art method (Marszalek and Schmid (2008)), that also builds the category hierarchy to recognize objects, experimental results show that the proposed method achieves better performance. Based on the category hierarchies built in the last chapter, the algorithms introduced in section 8.2.3 are used to predict object categories for the remaining 40 samples of each category. The average accuracy is computed for every category from 50 hierarchies built in the last chapter. If not using category hierarchies, classifiers are directly trained by an SVM for each category. 40 samples, the same as those used to build category hierarchies in the last experiments, are used as positive samples. 500 samples randomly selected from other categories are used as negative samples. The performance of object recognition is tested by using 6 different experimental setups, which are shown in Fig. 8.3. And Table 8.1 lists the average accuracy of object recognition for all categories in each scenario. 'Direct' means that the methods utilize an SVM directly to train classifiers for all categories without considering category hierarchies. 'One'/'Two' mean that the methods use a one/two-class SVM to train classifiers for categories based on category hierarchies. 'Select' means that the methods use selected attributes and a two-class SVM to train classifiers for categories based on category hierarchies. 'shLDA' means that the methods use the supervised hLDA model directly to predict labels of test samples. 'Ref' means a reference method was used in these experiments. The mean of average accuracy for each method is also listed in the legends of the figures.

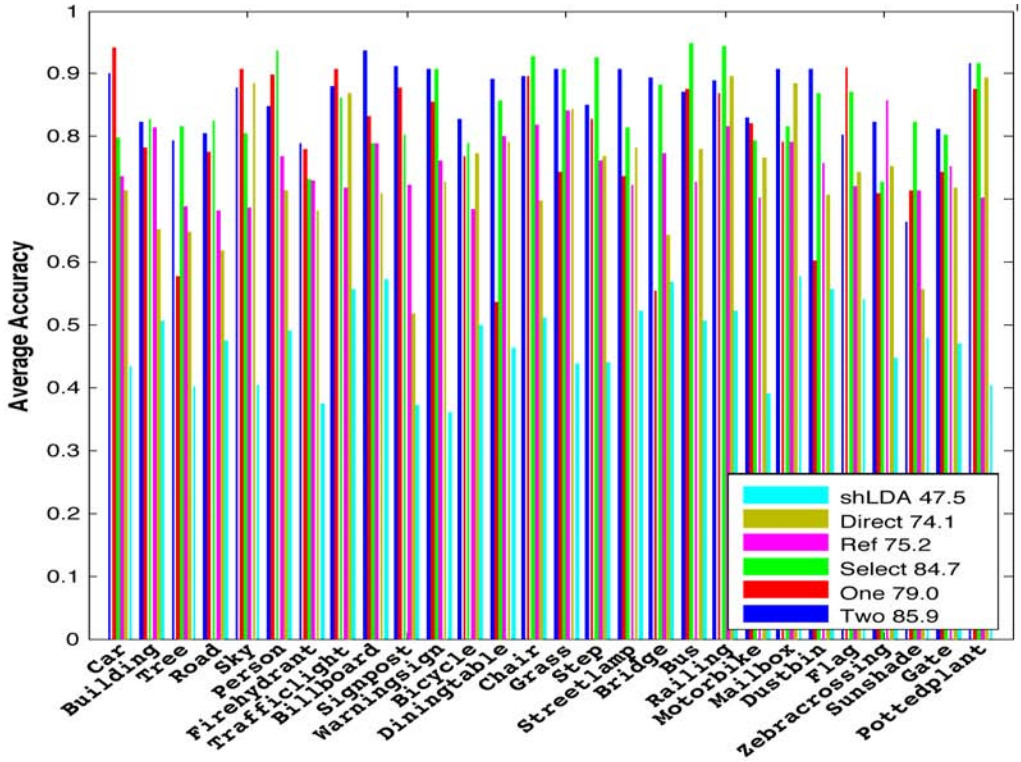
When the goal is to recognize object categories, using a two-class SVM is better than using a one-class SVM. This is because the best results are obtained by the method 'Two-SVM' as shown in Fig. 8.3. When using selected attributes, the


Average Accracy
shLDA 47.0
Direct 65.0
One 79.9
Ref 79.1
Two 85.1
Select 80.1
aeroplane
bag
bicycle
bird
boat
bottle
building
bus
car
carriage
cat
centaur
chair
cow
diningtable
dog
donkey
goat
horse
jetski
monkey
motorbike
mug
person
pottedplant
sheep
sofa
statue
train
tvmonitor
wolf
zebra

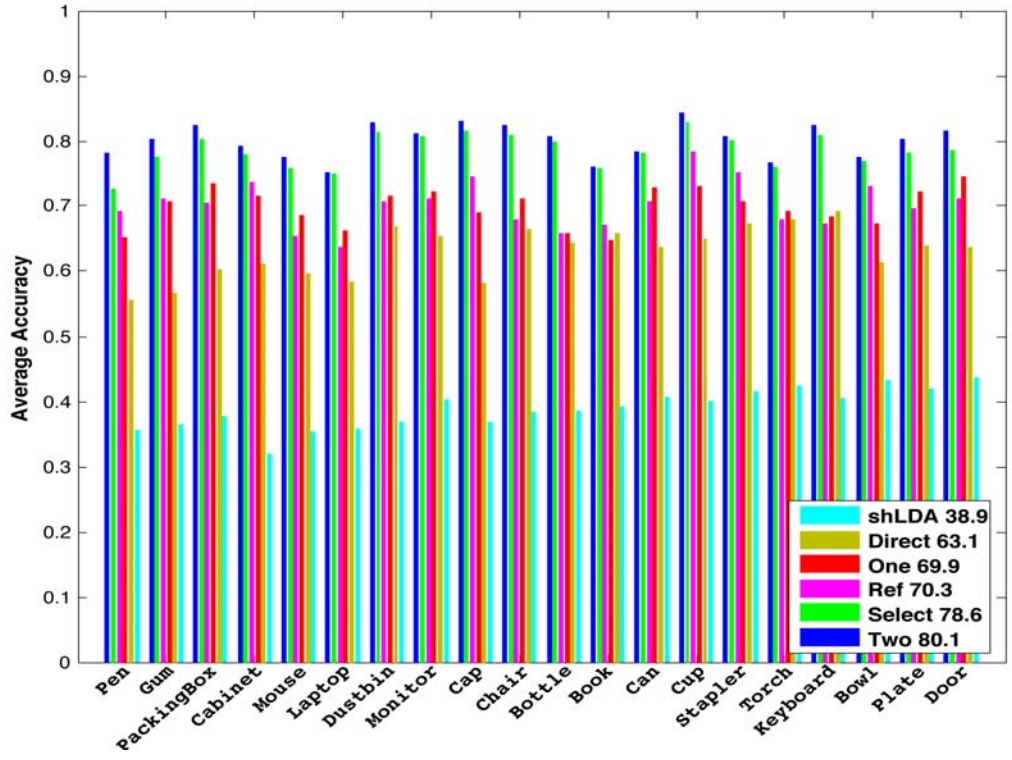
(a)


Average Accuracy
shLDA 47.0
Direct 71.8
One 77.0
Ref 76.4
Select 85.9
Two 86.1
Mouse
Keyboard
TVmonitor
Mug
Cup
Bottle
Desktable
Mousepad
Bookshelf
Sofa
Window
Person
Chair
Cabinet
Desklight
Book
Door
Printer
Pen
Bulletinboard
Computerhost
Telephone
Laptop
Dustbin
Sound box

(b)



(c)



(d)

Figure 8.3.: Accuracy of object recognition for categories in S1 (a), S2 (b), S3(c) and S4(d). We use ten methods to test the performance. The index of each method is illustrated in the left sub figure.

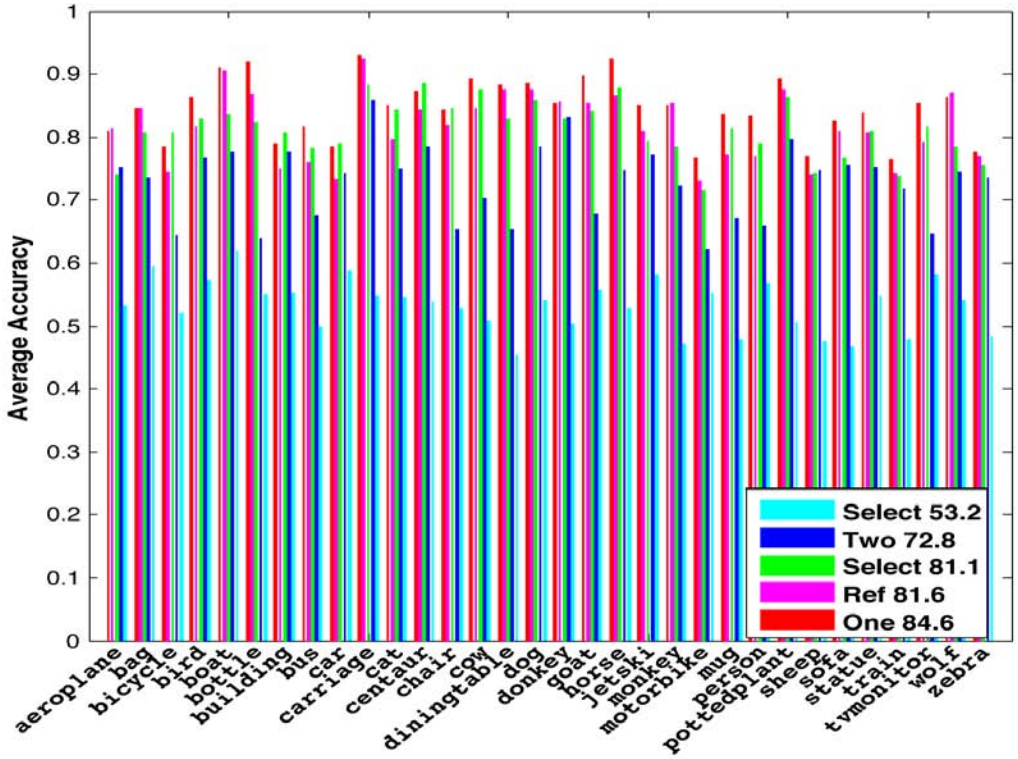
accuracy is closed to best results, which means that a promising performance of object recognition can be obtained by using lower dimensions but more representative attributes to describe objects. It is also worth noting that the performance of methods tagged by 'Direct' is worse than the performance of methods tagged by 'Two-SVM' and 'Select', which means that using category hierarchies can improve the object recognition. However, when directly using supervised hLDA to predict object labels, the worst results are obtained, because this topic model is a kind of generative model and is not suited for tasks like supervised object recognition. According to the experimental results shown here, therefore, we suggest that using a two-class SVM to train object classifiers based on (selected) attributes consisting of non-semantic attributes and semantic attributes and built category hierarchies can achieve an improved performance for object recognition. For the reference method, a two-class SVM is used to train classifiers for each node. The performance of the reference method is worse than that of the proposed method, since the better category hierarchies can be obtained by the proposed framework.

8.3.2. Evaluation of Novel Category Detection

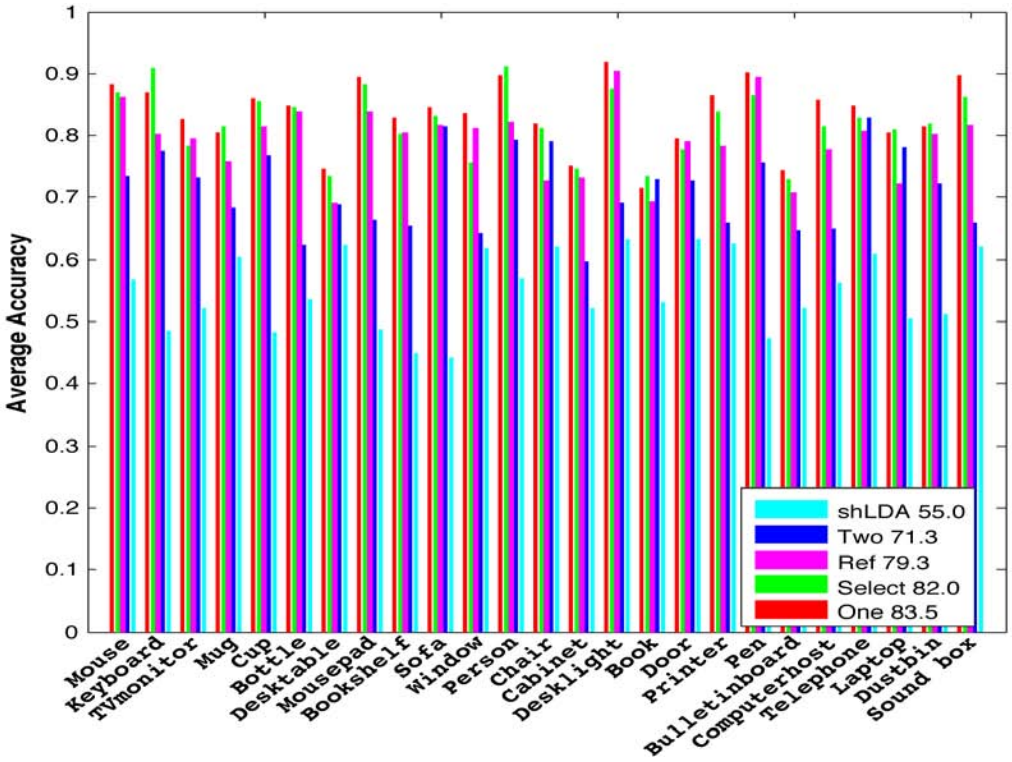
Under dynamic circumstances, it is an important ability for human cognition to identify a new object as belonging to a novel category and distinguish different novel categories for new objects. This is also one of the most important goals of the proposed framework. In this subsection novel category detection will be evaluated. The performance of distinguishing novel categories will then be evaluated in the next subsection.

Based on category hierarchies built in last chapter, four experimental setups which are listed in the legends of the result figures, are used to test for the proposed framework. Moreover, a state-of-the-art method in Smola et al. (2009) for novelty detection is used for comparison with the proposed framework. For the proposed framework, only when novel object samples are detected as belonging to a novel category and a node representing this category is branched off at a correct superordinate, the result is correct. Because there is no standard reference category hierarchy, the method used to evaluate the correctness is the same as the one in last chapter. In this reference method, all categories used to build a hierarchy are regarded as known categories, while the remaining categories are treated as a novelty to be detected. Object samples in each of novel categories are input to the reference method and the accuracy of correct detection is computed. Because the reference method does not utilize a hierarchy, the location of a novel category in the hierarchy is not concerned.

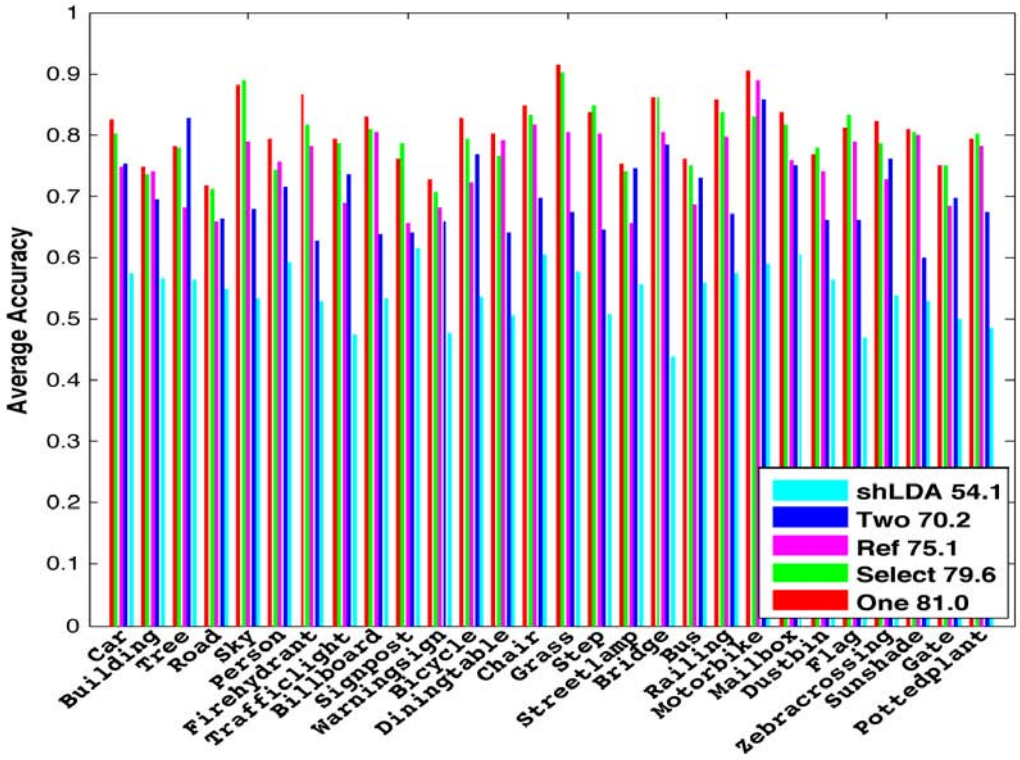
The average accuracy for each category in four scenarios is shown in Fig. 8.4. Note that the method 'Select' is trained by a one-class SVM in this group. Table 8.2 lists the average accuracy of novel category discovery in each scenario. The performance of the reference method is lower than these of the methods marked by 'One-SVM' and 'Select'. The best results for novel category discovery are obtained with the method 'One-SVM'. That is different from the results in the last subsection where


Average Accuracy
Select 53.2
Two 72.8
Select 81.1
Ref 81.6
One 84.6
aeroplane
bag
bicycle
bird
boat
bottle
building
bus
car
carriage
cat
centaur
chair
cow
diningtable
dog
donkey
goat
horse
jetski
monkey
motorbike
mug
person
pottedplant
sheep
sofa
statue
train
tvmonitor
wolf
zebra

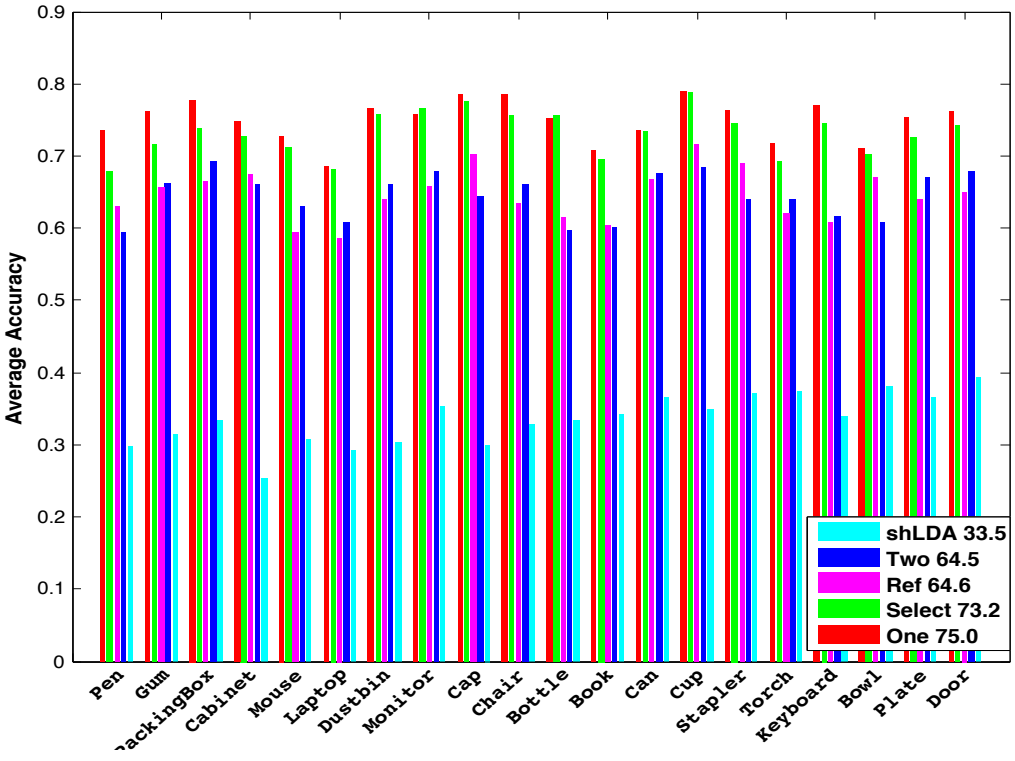
(a)


Average Accuracy
shLDA 55.0
Two 71.3
Ref 79.3
Select 82.0
One 83.5
Mouse
Keyboard
TVmonitor
Mug
Cup
Bottle
Desktable
Mousepad
Bookshelf
Sofa
Window
Person
Chair
Cabinet
Desklight
Book
Door
Printer
Pen
Bulletinboard
Computerhost
Telephone
Laptop
Dustbin
Sound box

(b)



(c)



(d)

Figure 8.4.: Accuracy of novel category discovery in S1 (a), S2 (b), S3(c) and S4(d). We use eight methods to test the performance. The marks are explained in detail in the text.

Table 8.2.: Average accuracy of novel category discovery

	S1	S2	S3	S4
One-SVM	84.6	83.5	81.0	75.0
Two-SVM	72.8	71.3	70.2	64.5
Select	81.1	82.0	79.6	73.2
shLDA	53.2	55	54.1	33.5
Ref	81.6	79.3	75.1	64.6

the best results are obtained by the method 'Two-SVM'. The goal of experiments in this group is to detect novel categories whose object samples are not included in any existing category. As mentioned in section 8.2.3, to accurately detect a novel category, object samples belonging to this novel category must be classified as belonging to a superordinate, but not belonging to any existing child node of this superordinate. That means in feature space the boundary of the superordinate classifier must contain these object samples, while the boundaries of its children classifiers need not contain these object samples. Since a one-class SVM constructs a hyper-sphere boundary in feature space, the more similar samples can be contained in this hyper-sphere. On the other hand, a two class SVM constructs two hyper-plane boundaries in feature space for two categories where both hyper-planes are as close as possible to one class but as far as possible from the other class, and therefore there is a margin between two hyper-planes in which object samples cannot be correctly classified. However, these object samples belonging to novel categories are most likely located at this margin. Therefore the method 'One-SVM' is better than the method 'Two-SVM' for novel category detection.

As experiments also show that the performance of a two-class SVM is better than that of a one-class SVM when the goal is to recognize object samples belonging to known categories. Therefore we suggest that, for a category hierarchy, the classifiers associated with the root and middle level nodes are trained by a one class SVM and the classifiers associated with leaf nodes are trained by a two class SVM. In this way, we can simultaneously obtain a better performance for both known object recognition and novel category discovery.

For more intuitively understanding novel category discovery, there is an example shown in Fig. 8.1. The initial hierarchy (Fig. 8.1(a)) has ten categories. It is changed to a new hierarchy (Fig. 8.1(b)) as two novel categories are added. The 'cat' category shares the same superordinate with 'horse', 'cow' and 'sheep' since all of them are animal categories. However, the path corresponding to 'aeroplane' is branched off at the root since this category is totally different from the other categories in this hierarchy. This result obviously conforms to human cognition.

Table 8.3.: Average accuracy of distinguishing novel categories

	S1	S2	S3	S4
Dynamic	68.7	72.5	69.9	64.2
Ref	62.1	60.8	56.8	54.0

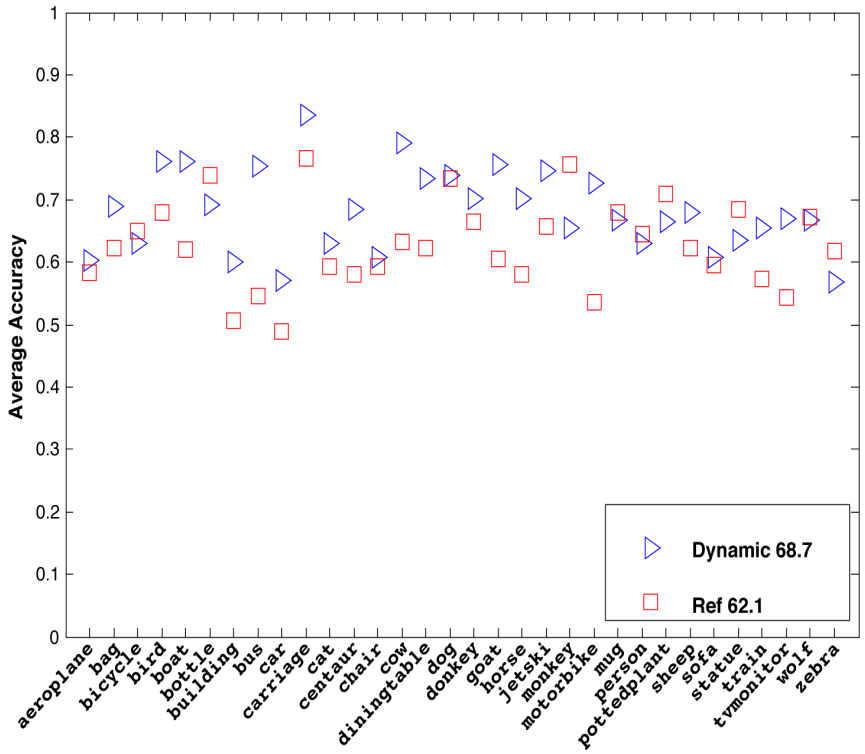
8.3.3. Evaluation of Distinguishing Novel Categories

The performance of clustering novel categories is also compared with different configuration and a state-of-the-art method. This state-of-the-art spectral clustering method, global kernel k-means (Tzortzis and Likas (2009)) is employed as a reference method for performance comparison. Since in the last group of experiments it has been proved that methods marked by 'One-SVM' obtain the best performance of detecting novel categories, only these methods are used to evaluate the performance of clustering novel categories on the proposed framework.

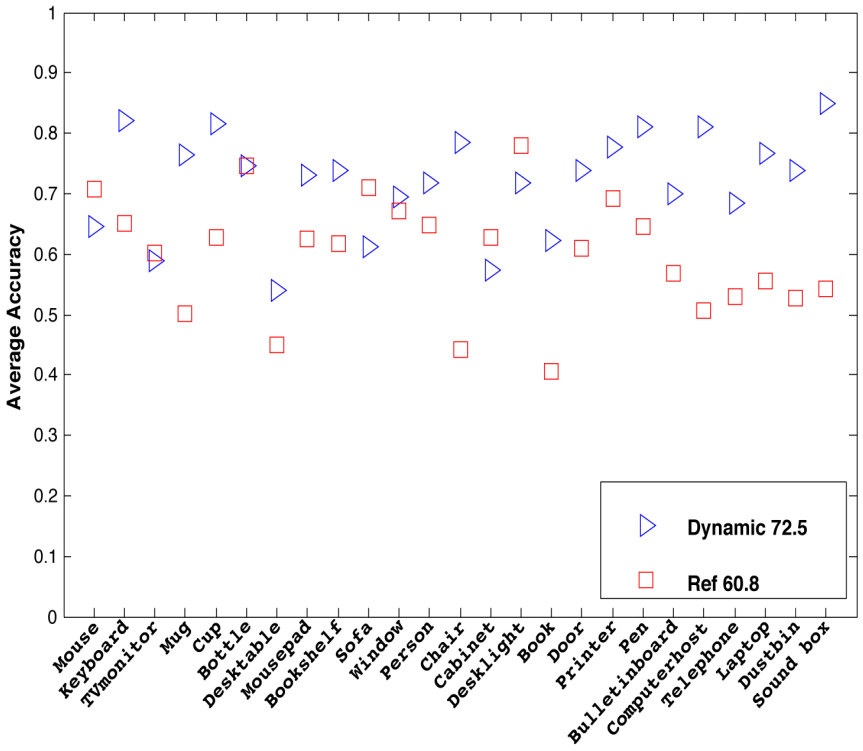
Note that here the term ''novel object' means an object sample belonging to a novel category. Based on the detection results from the last group of experiments, objects detected as belonging to learned categories are assigned corresponding labels of known categories, while novel objects detected as belonging to novel categories are assigned the same new category label. Thus a new temporal hierarchy, $H^{t'}$, can be obtained. In the proposed framework, the generative procedure mentioned in section 7.2 can be used to build a new hierarchy, H^{t+1}, from all training samples and detected results. If objects belonging to different novel categories but being assigned to the same new leaf node in $H^{t'}$ can be divided into different new leaf nodes in H^{t+1}, the results are regarded as correct. For the reference method, it does not need to indicate the cluster number because of its global property. Therefore this method can be directly applied to all leaf nodes. The accuracy can be directly computed by counting the number of objects correctly distinguished. The average accuracy for each category is shown in Fig. 8.5 and the average accuracy for each scenario is listed in Table 8.3. Since this evaluation is based on the correct result of novel category detection, the accuracy is lower than the results in the last group of experiments. However, it is still obvious that the performance of the proposed method is better than that of the reference method. The average accuracy is improved by 10-16%.

8.3.4. Evaluation of Object Attribute Summarization

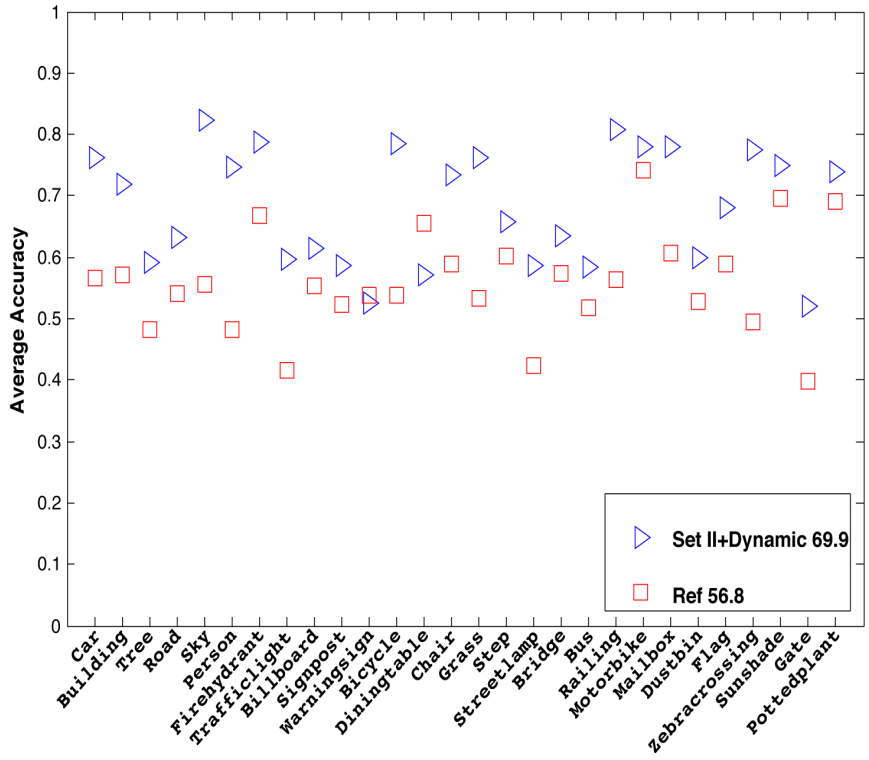
Humans can gradually extract the most representative and distinctive features for a category and quickly recognize objects in this category. The proposed framework can also extract the most representative and distinctive attributes for all categories. Each topic in the hierarchy built by the supervised hLDA corresponds to a distribution of attributes. Through this distribution, it is possible to summarize those attributes that can represent a category best. Thus we only need to use these representative attributes rather than all attributes to describe and recognize objects. According to the method introduced in section 7.2, the probabilities of all attributes


Average Accuracy
Dynamic 68.7
Ref 62.1
aeroplane
bag
bicycle
bird
boat
bottle
building
bus
car
carriage
cat
centaur
chair
cow
diningtable
dog
donkey
goat
horse
jetski
monkey
motorbike
mug
person
pottedplant
sheep
sofa
statue
train
tvmonitor
wolf
zebra

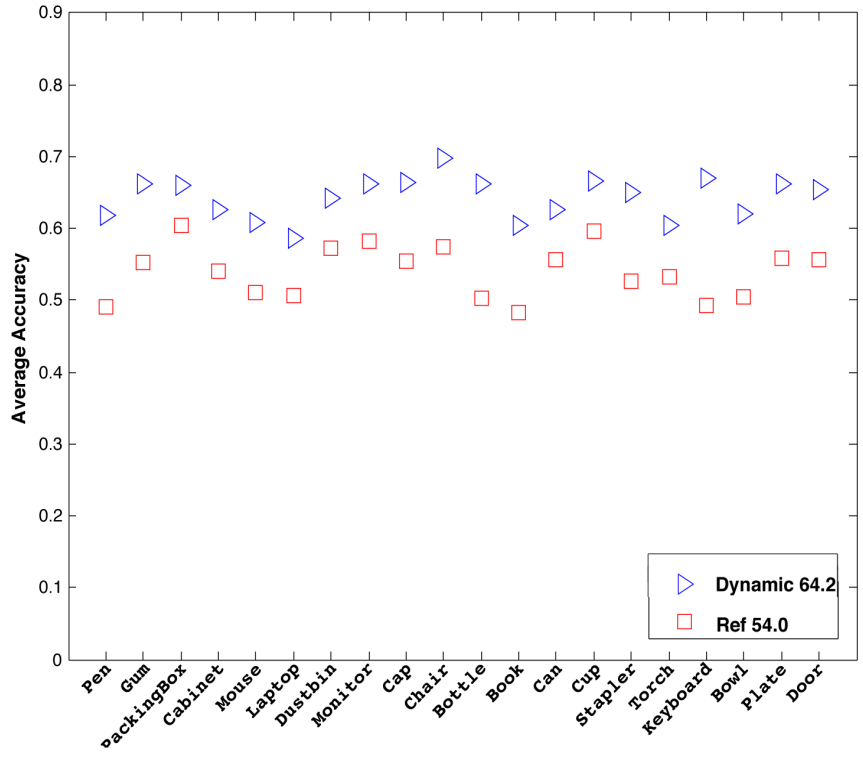
(a)


Average Accuracy
Dynamic 72.5
Ref 60.8
Mouse
Keyboard
TVmonitor
Mug
Cup
Bottle
Desktable
Mousepad
Bookshelf
Sofa
Window
Person
Chair
Cabinet
Desklight
Book
Door
Printer
Pen
Bulletinboard
Computerhost
Telephone
Laptop
Dustbin
Sound box

(b)



(c)



(d)

Figure 8.5.: Accuracy of distinguishing novel categories in S1 (a), S2 (b), S3(c) and S4(d).

in the distribution associated with one node are computed and sorted in descending order. The upper part attributes whose sum of probabilities is larger than 80% is selected as the summarized attributes.

In the first two evaluations of this section, category classifiers only using selected attributes are also trained to evaluate the performance of the object attribute summarization. The experimental results are shown in Fig. 8.3 and Fig. 8.4. The difference between the average accuracy of object recognition of two methods ('Two-SVM' and 'Select') ranges from 0.3% to 5%. The difference of the average accuracy of novel category discovery between two methods ('One-SVM' and 'Selected') ranges from 1.4% to 3.5%. It is obvious that the average accuracy of using selected attributes is close to that of using all attributes. For some categories, the performance of using selected attributes is even better than that of using all attributes. Therefore we can save time and computational cost and improve the efficiency for our framework by using selected attributes. These results prove that extracting attributes in this manner is promising for both object recognition and novel category discovery.

Furthermore, we can validate the effectiveness of summarized attributes from the perspective of human cognition since here semantic attributes are used. As shown in Fig. 7.2, the red words close to each node have the highest probabilities compared with other semantic attributes. Using these attributes to represent corresponding categories obviously conforms to human cognition. For instance, attributes 'horn', 'saddle' and 'wool' are of course the most representative attributes for categories 'cow', 'horse' and 'sheep' respectively.

8.4. Conclusion

We proposed a framework of building dynamic category hierarchies in which objects are described by attributes and hierarchies are constructed by topic models. Through our framework, category hierarchies that are in better keeping with a current circumstance can be obtained and can dynamically change their structure when this circumstance changes. Thus a high accuracy of object recognition can be obtained. More important, if objects belong to novel categories which have never been learned previously, they can be detected and clustered accurately, and corresponding novel categories can be added precisely to current hierarchies. These conclusions have been proved through our extensive experiments.

Because of the advantages of the proposed framework, it can be applied to, for example, unknown environment exploration for robots. It is inevitable for robots to meet unknown objects that have never been learned. Robots therefore can only classify these unknown objects according to their experiences. By using traditional object recognition technologies, robots will categorize unknown objects into those known categories most similar to them, which will lead to failures and influence the following work. Through our framework, however, robots will infer that unknown objects belong to novel categories and relate novel categories to known categories by

built category hierarchies. This is helpful for the consecutive reasoning work, such as functions and characteristics of these novel categories.

Moreover, because objects are described by attributes, the proposed framework can easily integrate multi modal information, such as auditory and tactile attributes. For objects in certain special categories, they can produce a unique sound or have a unique surface. We can extract these unique features as attributes and add them to the attribute list to improve the distinctiveness of these categories.

Chapter 9

Summary and Conclusion

In this chapter, we first summarize the main contributions of this thesis and draw several conclusions about it. Then some limitations of these proposed methods are figured out, through which an outlook on future research directions can be motivated in the following section.

9.1. Thesis Summary and Conclusion

9.1.1. Summary

The work of this thesis is try to implement a key cognitive ability inspired by human visual cognition in artificial intelligent systems. This key ability is discovering novel categories in an unexplored environment, and simultaneously connecting these novel categories with known categories by building an appropriate relationships among them. It is extremely useful for the next tasks, for example, reasoning functions of these novel categories, and subsequently replacing known categories for the following work. According to human cognitive procedure, we divide such a problem into two successive sub-tasks.

Before recognizing and identifying objects, we first need to known objects' positions in a scene. Thus at first objects should be detected and localized. But in an unexplored environment there is no prior about objects' number and categories. Therefore the first sub-task to be solved is category-independent object detection. After given object positions, the second task which are recognizing known objects, identifying novel categories and building category relationships can then be solved. To solve these two tasks, multimodal data, i.e. the 2D and 3D information, are employed, because human visual system always simultaneously utilizes them for visual cognition.

Recognizing learned categories, identifying and discovering novel categories need to use object features to represent objects. For the particular problem in this thesis, object features should have excellent generalizability across categories since there is

no prior information about object categories. For the first sub-task, a set of category-independent object features is developed. They only consists of those basic features, which make an object to be a stand-alone instance regardless of its category, such as multimodal saliency, multimodal oversegments, and boundaries. These category-independent features represent the characteristic of an object against background and other objects. Our extensive experiments show their satisfactory performance.

For the second task, it also needs to consider the descriptive ability of features besides their generalizability. Human beings always describe objects by their attributes. Furthermore, objects belonging to different categories can share the same attributes. Therefore object attributes are employed as object features for object recognition and novel category discovery. To utilize multimodal data, a set of novel 3D and multimodal object attributes, and novel intra-class non-semantic attributes are developed. These new attributes can describe object more comprehensively. Their performance is also promising by experimental evaluation.

Motivated by the observation that 2D and 3D spatial positions of an object keep consistency, a novel multimodal co-segmentation framework is developed based on state-of-the-art higher order CRF model, called cross-model higher order conditional random field (CMH-CRF) model. This model simultaneously labels patches (i.e. oversegments) in 2D and 3D space. By integrating cross-modal potentials, the overlapped patches in different modalities are constrained to obtain consistent labels. At the same time, the model takes three kinds of labels which extremely simplifies the difficulty and improves the accuracy of identifying all single object instances from a clutter scene. Through extensive experiments, the CMH-CRF model shows its satisfactory performance that the accuracy is high enough for the practical application.

To efficiently organizing object categories, we adopt the hierarchical structure which is inspired by human cognition to represent the relationship among object categories. By developing the supervised hierarchical latent Dirichlet allocation (hLDA), a precise category hierarchy can be constructed. Based on this hierarchy, we further developed a dynamic category hierarchy framework. Because the dynamic category hierarchy takes the novel categories into account, it can efficiently improve the object recognition, discover novel categories and simultaneously refine the built hierarchy. Extensive experiments are carried out, which prove the efficiency of the proposed supervised hLDA, as well as the dynamic category hierarchy framework.

9.1.2. Conclusion

Based on aforementioned summaries of this thesis, a conclusion can be drawn that multimodal data indeed improve object detection and novel category discovery. We list the details about this conclusion as follows:

- We found 2D and 3D saliency can complement each other, as well as 2D and 3D oversegments. Therefore the more accurate saliency and oversegments can be obtained from multimodal data.

- An objects have spatial consistency in 2D and 3D space. We found that the co-segmentation based on cross-modal spatial consistency can efficiently detect and localize objects, regardless of their categories.

- Generally, humans describe an object through multimodal attributes. In this thesis, we also proved that utilizing multimodal attributes can describe objects more accurately and comprehensively.

- Based on supervised hLDA and object attributes, a dynamic category hierarchy can be efficiently built. We found such a dynamic hierarchy coincides with human cognition better and can obtain better object recognition and discover novel object categories than a static category hierarchy.

9.2. Limitations and Future Work

9.2.1. Limitations

Although two sub-tasks in this thesis can be solved efficiently by the proposed methods, there is a shortage that these methods are not integrated into a uniform framework. We first need to use the CMH-CRF model to detect and localize objects. Then the dynamic category hierarchy framework can separately recognize object and discover novel categories. Thus the spatial context among objects in a single scene is not utilized in the proposed methods. Furthermore, the results of the second sub-task cannot be fed back to the first sub-task to refine the results of object detection.

The proposed methods in this thesis are based on multimodal information, but only two visual modalities are utilized. Other modal information which may also improve the performance is not integrated into these methods. For example, the audio information can help object localization and the tactile information can help object recognition.

With respect to the particular technical details, there are also several limitations to be eliminated. When detecting and localizing objects, it requires a lot of computational resource to extract category-independent object features. Training multimodal object attributes is also a time-consuming task for extracting multimodal base features and training SVM classifiers. For the supervised hLDA, the number of levels of the hierarchy has to be set to a fixed constant. On one hand, the large level number can lead to a more complicated hierarchical structure. Although a complicated hierarchy can more accurately represent the relationship of all categories, it may also result in more isolated nodes without any branches and more time-cost. On the other hand, a small level number may result the hierarchy more flat and cannot represent category relationship precisely.

9.2.2. Discussion on Future Research Directions

The first part of further work is to overcome those limitations mentioned above. At beginning, those methods of solving two sub-tasks should be integrated into a uniform framework to improve their performance. By using the context information of detected objects, the relationship among categories can be built more precisely. For example, some categories, such as monitor, mouse and keyboard occur together, which can be categorized into one superordinate. The recognition results by the dynamic hierarchical framework can improve the results of object detection. For example, when one object is recognized as belonging to one known category, the prior information of this category can then be utilized.

Furthermore, the utilization of depth information can be further explored. For example, a good extension would be to use depth information to measure object scales (e.g. Zhang et al. (2011) and Fritz et al. (2010)). Thus, the category-independent object detection could achieve more accurate results and the size of objects could be an efficient attributes.

Next, we will integrate more modalities into the proposed methods. Audio information can be used as a kind of good category-independent object features, which can be used to localize object positions. Tactile information can be integrated into the base feature to further extending object attributes, such as density, material and weight.

Besides these further improvements, another part of further work is to describe novel categories and reason their functions according to the relationship related to other known categories, which is also an important ability for humans. For example, the proposed methods may discover a novel category that has the same superordinate with other known categories, such as cup and glass. And this novel category shares some attributes with known categories, such as 'Vert.Cylinder' and 'Concave'. Thus we can reason it as a kind of container like a cup and use it to fill beverage.

After further improving the efficiency of the proposed methods, especially reducing their computational cost, we also plan to deploy the proposed methods to a particular artificial intelligent agent such as a service robot to improve robot' autonomous abilities.

Appendix A

Object Semantic Attributes

All attributes used in this thesis here are listed as follows. Note that the attributes with italic style word have been trained by multimodal base feature for comparison in chapter 6.

- **Shape**: 2D Boxy, Round, Triangle, Occluded.
- **Part**: Tail, Beak, Head, Ear, Snout, Nose, Mouth, Hair, Face, Eye, Torso, Hand, Arm, Leg, Foot/Shoe, Wing, Propeller, Jet engine, *Window*, Row Window, Wheel, *Door*, Headlight, Taillight, Side mirror, Exhaust, Pedal, Handlebars, Engine, Sail, Mast, Text, Label, *Furn. Leg*, *Furn. Back*, *Furn. Seat*, *Furn. Arm*, Horn, Rein, Saddle, Leaf, Flower, Stem/Trunk, *Pot*, *Screen*, *wire*.
- **Material & Texture**: Skin, *Metal*, *Plastic*, *Wood*, *Cloth*, Furry, *Glass*, Feather, Wool, Clear, *Shiny*, Vegetation, Leather, Grass, *Ceramic*.
- **3D**: Sphere, Vert.Cylinder, Horiz.Cylinder, Cone, Vert.Ellipsoid, Horiz.Ellipsoid, 3D Square, 3D Cuboid, Plane, Concave.
- **Multimodal Attributes**: Cup.Handle, Door.Handle, Cabinet.Handle, Rect.Button, Circle.Button, Desktop.

Appendix B

Object Categories

All object categories used for experiments in chapter 6, 7 and 8 are listed as follows:

- **a-Pascal & a-Yahoo**: Aeroplane, Bicycle, Bird, Boat, Bottle, Bus, Car, Cat, Chair, Cow, Diningtable, Dog, Horse, Motorbike, Person, Pottedplant, Sheep, Sofa, Train, TVmonitor, Donkey, Monkey, Goat, Wolf, Jetski, Zebra, Centaur, Mug, Statue, Building, Bag, Carriage.
- **LabelMe Office scene**: Mouse, Keyboard, TVmonitor, Mug, Cup, Bottle, Desktable, Mouse pad, Book shelf, Sofa, Window, Person, Chair, Cabinet, Desklight, Book, Door, Printer, Pen, Bulletin board, Computerhost, Telephone, Laptop, Dustbin, Sound box.
- **LabelMe Street scene**: Car, Building, Tree, Road, Sky, Person, Fire-hydrant, Traffic light, Billboard, Signpost, Warning sign, Bicycle, Diningtable, Chair, Grass, Step, Street lamp, Bridge, Bus, Railing, Motorbike, Mailbox, Dustbin, Flag, Zebra crossing, Sunshade, Gate, Pottedplant.
- **RGB-D Dataset**: Pen, Gum, Packing Box, Cabinet, Mouse, Laptop, Dustbin, Monitor, Cup, Chair, Bottle, Book, Can, Cap, Stapler, Torch, Keyboard, Bowl, Plate, Door.

Bibliography

Achanta, R., Estrada, F., Wils, P., and Süsstrunk, S. (2008). Salient region detection and segmentation. *Computer Vision Systems*, 5008/2008:66–75.

Ahuja, N. and Todorovic, S. (2007). Learning the taxonomy and models of categories present in arbitrary images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

Alexe, B., Deselaers, T., and Ferrari, V. (2010). What is an object? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 73–80.

Alpert, N., Bradshaw, J., Kennedy, D., and Correia, J. (1990). The principal axes transformation–a method for image registration. *Journal of Nuclear Medicine*, 31(10):1717.

Arora, H., Loeff, N., Forsyth, D., and Ahuja, N. (2007). Unsupervised segmentation of objects using efficient learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7.

Bagon, S., Boiman, O., and Irani, M. (2008). What is a good image segment? a unified approach to segment extraction. In *Europe Conference on Computer Vision (ECCV)*, pages 30–44.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *Europe Conference on Computer Vision (ECCV)*, pages 404–417.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

Bengio, Y., Delalleau, O., Roux, N., Paiement, J., Vincent, P., and Ouimet, M. (2004). Learning eigenfunctions links spectral embedding and kernel pca. *Neural Computation*, 16(10):2197–2219.

Binder, J., Koller, D., Russell, S., and Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2):213–244.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. springer New York.

Blanchard, G., Lee, G., and Scott, C. (2010). Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009.

Blei, D., Griffiths, T., and Jordan, M. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):7:1–7:30.

Blei, D. and Lafferty, J. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, (1):17–35.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Bodesheim, P. (2011). Spectral clustering of rois for object discovery. In *Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pages 450–455.

Boykov, Y. and Funka-Lea, G. (2006). Graph cuts and efficient n-d image segmentation. *International Journal of Computer Vision*, 70(2):109–131.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239.

Bradley, J. and Guestrin, C. (2010). Learning tree conditional random fields. In *International Conference on Machine Learning (ICML)*.

Buntine, W. and Jakulin, A. (2006). Discrete component analysis. In Saunders, C., Grobelnik, M., Gunn, S., and Shawe-Taylor, J., editors, *Subspace, Latent Structure and Feature Selection*, pages 1–33. Springer, Berlin.

Callanan, M. (1985). How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child Development*, pages 508–523.

Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

Carreira, J. and Sminchisescu, C. (2012). Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transaction on Pattern Analysis and Machine Intelligence*.

Chang, C. and Lin, C. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Chang, J. and Blei, D. (2009). Relational topic models for document networks. In *Conference on AI and Statistics (AISTATS)*.

Cheng, M., Zhang, G., Mitra, N., Huang, X., and Hu, S. (2011). Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 409–416.

Cho, M., Shin, Y., and Lee, K. (2010). Unsupervised detection and segmentation of identical objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1617–1624.

Chu, Y., Ye, X., Qian, J., Zhang, Y., and Zhang, S. (2007). Adaptive foreground and shadow segmentation using hidden conditional random fields. *Journal of Zhejiang University-Science A*, 8(4):586–592.

Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J., and Yang, C. (2001). Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):64–78.

Collet, A., Srinivasay, S., and Hebert, M. (2011). Structure discovery in multi-modal data: A region-based approach. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5695–5702.

Cozman, F. (2000). Generalizing variable elimination in bayesian networks. In *Workshop on Probabilistic Reasoning in Artificial Intelligence*, pages 27–32.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893.

Dechter, R. (1999). Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, 113(1):41–85.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Do, T. and Artières, T. (2005). Conditional random field for tracking user behavior based on his eye's movements. In *Advances in Neural Information Processing System (NIPS) Workshop*, page 19.

Douze, M., Ramisa, A., and Schmid, C. (2011). Combining attributes and fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 745–752.

Dueck, D. and Frey, B. (2007). Non-metric affinity propagation for unsupervised image categorization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

Duvenaud, D., Marlin, B., and Murphy, K. (2011). Multiscale conditional random fields for semi-supervised labeling and classification. In *Canadian Conference on Computer and Robot Vision (CRV)*, pages 371–378.

Endres, I. and Hoiem, D. (2010). Category independent object proposals. In *Europe Conference on Computer Vision (ECCV)*, pages 575–588.

Epshtein, B. and Uliman, S. (2005). Feature hierarchies for object classification. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 220–227.

Everingham, M., Gool, V., Williams, C., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338.

Farhadi, A., Endres, I., and Hoiem, D. (2010). Attribute-centric recognition for cross-category generalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2352–2359.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing objects by their attributes. In *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785.

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Felzenszwalb, P. and Huttenlocher, D. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181.

Feng, J., Wei, Y., Tao, L., Zhang, C., and Sun, J. (2011). Salient object detection by composition. In *IEEE International Conference on Computer Vision (ICCV)*.

Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. (2005). Learning object categories from google's image search. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1816–1823.

Ferrari, V. and Zisserman, A. (2008). Learning visual attributes. In *Europe Conference on Computer Vision (ECCV)*.

Fidler, S. and Leonardis, A. (2007). Towards scalable representations of object categories: Learning a hierarchy of parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Frey, B. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972.

Fritz, M., Saenko, K., and Darrell, T. (2010). Size matters: Metric visual search constraints from monocular metadata. In *Advances in Neural Information Processing System (NIPS)*, volume 23.

Fritz, M. and Schiele, B. (2008). Decomposition, discovery and detection of visual categories using topic models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Fruehwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

Gallager, R. (1962). Low-density parity-check codes. *IRE Transactions on Information Theory*, 8(1):21–28.

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

Ghosh, S., Ungureanu, A., Sudderth, E., and Blei, D. (2011). Spatial distance dependent chinese restaurant processes for image segmentation. In *Advances in Neural Information Processing System (NIPS)*, pages 1585–1592.

Glasner, D., Vitaladevuni, S., and Basri, R. (2011). Contour-based joint clustering of multiple segmentations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2385–2392.

Goferman, S., Zelnik-Manor, L., and Tal, A. (2010). Context-aware saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2376–2383.

Gosselin, F. and Schyns, P. (2001). Why do we slip to the basic level? computational constraints and their implementation. *Psychological Review*, 108(4):735.

Gould, S., Baumstarck, P., Quigley, M., Ng, A., Koller, D., et al. (2008). Integrating visual and range data for robotic object detection. In *Europe Conference on Computer Vision (ECCV)*.

Grauman, K. and Darrell, T. (2006). Unsupervised learning of categories from sets of partially matching image features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19–25.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

Hannah, L., Blei, D., and Powell, W. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 1:1–33.

Hartigan, J. (1983). *Bayes Theory*. Springer.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., and Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help? In *ACM Conference on Information and Knowledge Management*, pages 957–966.

He, X., Zemel, R., and Ray, D. (2006). Learning and incorporating top-down cues in image segmentation. *Europe Conference on Computer Vision (ECCV)*, pages 338–351.

Herbst, E., Henry, P., Ren, X., and Fox, D. (2011). Toward object discovery and modeling via 3-d scene comparison. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2623–2629.

Hoffmann, H. (2007). Kernel pca for novelty detection. *Pattern Recognition*, 40(3):863–874.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Huang, Q., Han, M., Wu, B., and Ioffe, S. (2011). A hierarchical conditional random field model for labeling and segmenting images of street scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1953–1960.

Ihler, A., Fisher, J., and Willsky, A. (2006). Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, 6(1):905.

Ion, A., Carreira, J., and Sminchisescu, C. (2011). Image segmentation by figure-ground composition into maximal cliques. In *IEEE International Conference on Computer Vision (ICCV)*.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jo, Y. and Oh, A. (2011). Aspect and sentiment unification model for online review analysis. In *ACM International Conference on Web Search and Data Mining*, pages 815–824.

Jordan, M. and Weiss, Y. (2002). Graphical models: Probabilistic inference. In Arbib, M., editor, *The Handbook of Brain Theory and Neural Networks,*. MIT Press, Cambridge.

Kapoor, A., Urtasun, R., and Darrell, T. (2009). Probabilistic kernel combination for hierarchical object categorization. Technical report, EECS Tech. Rep. 2009–2016.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296–4309.

Kim, G., Faloutsos, C., and Hebert, M. (2008). Unsupervised modeling of object categories using link analysis techniques. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

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

Kohli, P. and Torr, P. (2007). Dynamic graph cuts for efficient inference in markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2079–2088.

Kortgen, M., Park, G., Novotni, M., and Klein, R. (2003). 3d shape matching with 3d shape contexts. In *Central European Seminar on Computer Graphics*, volume 3, page 5.

Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A., and Berg, T. (2011). Baby talk: Understanding and generating simple image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1608.

Kumar, S. and Hebert, M. (2003a). Discriminative fields for modeling spatial dependencies in natural images. In *Advances in Neural Information Processing System (NIPS)*.

Kumar, S. and Hebert, M. (2003b). Discriminative random fields: A discriminative framework for contextual interaction in classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157.

Ladicky, L., Russell, C., Kohli, P., and Torr, P. (2009). Associative hierarchical crfs for object class image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 739–746.

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

Lai, K., Bo, L., Ren, X., and Fox, D. (2011). A large-scale hierarchical multi-view rgb-d object dataset. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1817–1824.

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

Lampert, C., Nickisch, H., and Harmeling, S. (2009b). Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference Computer Vision and Pattern Recognition (CVPR)*, pages 951–958.

Larlus, D., Verbeek, J., and Jurie, F. (2010). Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields. *International journal of computer vision*, 88(2):238–253.

Lee, Y. J. and Grauman, K. (2009). Foreground focus: Unsupervised learning from partially matching images. *International Journal of Computer Vision*, 85(2):143–166.

Lee, Y. J. and Grauman, K. (2012). Object-graphs for context-aware visual category discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):346–358.

Lehmann, A., Leibe, B., and Van Gool, L. (2009). Feature-centric efficient subwindow search. In *IEEE International Conference on Computer Vision (ICCV)*, pages 940–947.

Levinshtein, A., Sminchisescu, C., and Dickinson, S. (2010). Optimal contour closure by superpixel grouping. In *Europe Conference on Computer Vision (ECCV)*, pages 480–493.

Li, L., Su, H., Xing, E., and Fei-Fei, L. (2010a). Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Advances in Neural Information Processing System (NIPS)*.

Li, L., Wang, C., Lim, Y., Blei, D., and Fei-Fei, L. (2010b). Building and using a semantivisual image hierarchy. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3336–3343.

Li, P., Jiang, J., and Wang, Y. (2010c). Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Annual Meeting of the Association for Computational Linguistics*, pages 640–649.

Li, S. (2009). *Markov Random Field Modeling in Image Analysis*. Springer-Verlag New York Inc.

Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *International Conference on Machine Learning (ICML)*, pages 577–584.

Li, Y., Zhou, Y., Yan, J., Niu, Z., and Yang, J. (2010d). Visual saliency based on conditional entropy. In *Asian Conference on Computer Vision (ECCV)*, pages 246–257.

Li, Z., Wang, Y., Geers, G., Chen, J., Yang, J., and Laird, J. (2010e). Saliency based joint topic discovery for object categorization. In *IEEE International Conference on Image Processing (ICIP)*, pages 4581–4584.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *ACM Conference on Information and Knowledge Management*, pages 375–384.

Liu, D. and Chen, T. (2007a). A topic-motion model for unsupervised video object discovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Liu, D. and Chen, T. (2007b). Unsupervised image categorization and object localization using topic models and correspondences between images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–7.

Liu, J., Kuipers, B., and Savarese, S. (2011a). Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3344.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H. (2011b). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367.

Liu, Y., Carbonell, J., Gopalakrishnan, V., and Weigele, P. (2007). Protein quaternary fold recognition using conditional graphical models. In *International Joint Conference in Artificial Intelligence*.

Lou, Z., Ye, Y., and Liu, D. (2010). Unsupervised object category discovery via information bottleneck method. In *International Conference on Multimedia*, pages 863–866.

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

Lu, W., Ng, H., and Lee, W. (2009). Natural language generation with tree conditional random fields. In *Conference on Empirical Methods in Natural Language Processing*, pages 400–409.

Lu, Y. and Zhai, C. (2008). Opinion integration through semi-supervised topic modeling. In *International Conference on World Wide Web (WWW)*, pages 121–130.

Ma, Y., Derksen, H., Hong, W., and Wright, J. (2007). Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562.

Mahajan, D., Sellamanickam, S., and Nair, V. (2011). A joint learning framework for attribute models and object descriptions. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1227–1234.

Maire, M., Arbeláez, P., Fowlkes, C., and Malik, J. (2008). Using contours to detect and localize junctions in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Mao, Y. and Lebanon, G. (2007). Isotonic conditional random fields and local sentiment flow. In *Advances in Neural Information Processing System (NIPS)*, volume 19, page 961.

Markou, M. and Singh, S. (2003a). Novelty detection: A review–part 1: Statistical approaches. *Signal Processing*, 83(12):2481–2497.

Markou, M. and Singh, S. (2003b). Novelty detection: A review–part 2: Neural network based approaches. *Signal Processing*, 83(12):2499–2521.

Marszalek, M. and Schmid, C. (2007). Semantic hierarchies for visual object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7.

Marszalek, M. and Schmid, C. (2008). Constructing category hierarchies for visual recognition. *Europe Conference on Computer Vision (ECCV)*, pages 479–491.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767.

McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In *International Conference on World Wide Web (WWW)*, pages 171–180.

Minka, T. (2001). Expectation propagation for approximate bayesian inference. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 17, pages 362–369.

Modayil, J. and Kuipers, B. (2004). Bootstrap learning for object discovery. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 1, pages 742–747.

Morency, L., Quattoni, A., and Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Mukherjee, L., Singh, V., and Peng, J. (2011). Scale invariant cosegmentation for image groups. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1881–1888.

Murphy, G. and Lassaline, M. (1997). Hierarchical structure in concepts and the basic level of categorization. *Knowledge, Concepts, and Categories*, pages 93–131.

Murphy, K. and Paskin, M. (2002). Linear-time inference in hierarchical hmms. *Advances in Neural Information Processing System (NIPS)*, 2:833–840.

Murphy, K., Weiss, Y., and Jordan, M. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 467-475.

Murray, I. and Ghahramani, Z. (2004). Bayesian learning in undirected graphical models: Approximate mcmc algorithms. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 392–399.

Ng, A. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *International Conference on Machine Learning (ICML)*, page 78.

Ngo, H., Kim, H., Han, M., and Lee, Y. (2010). Semi-markov conditional random fields for accelerometer-based activity recognition. *Applied Intelligence*, 35(2):226–241.

Ommer, B. and Buhmann, J. (2010). Learning the compositional nature of visual object categories for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):501–516.

Parikh, D. and Grauman, K. (2011). Relative attributes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 503 – 510.

Pearl, J. (1985). *Bayesian Networks: A Model of Self-Activated: Memory for Evidential Reasoning*. Computer Science Department, University of California.

Pineda, G., Koga, H., and Watanabe, T. (2010). Object discovery by clustering correlated visual word sets. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 750–753.

Quattoni, A., Wang, S., Morency, L., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852.

Rahtu, E., Kannala, J., and Blaschko, M. (2011). Learning a category independent object detection cascade. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1052–1059.

Ren, X., Fowlkes, C., and Malik, J. (2006). Figure/ground assignment in natural images. In *Europe Conference on Computer Vision (ECCV)*, pages 614–627.

Reynolds, D. (2008). Gaussian mixture models. *Encyclopedia of Biometric Recognition*.

Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., and Schiele, B. (2010). What helps where and why: Semantic relatedness for knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 910–917.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 487–494.

Rother, C., Minka, T., Blake, A., and Kolmogorov, V. (2006). Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 993–1000.

Russell, B., Freeman, W., Efros, A., Sivic, J., and Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1605–1614.

Russell, B., Torralba, A., Murphy, K., and Freeman, W. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1):157–173.

Rusu, R. B., Marton, Z. C., Blodow, N., and Beetz, M. (2008a). Learning Informative Point Classes for the Acquisition of Object Model Maps. In *International Conference on Control, Automation, Robotics and Vision (ICARCV)*.

Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., and Beetz, M. (2008b). Towards 3D Point Cloud Based Object Maps for Household Environments. *Robotics and Autonomous Systems Journal (Special Issue on Semantic Knowledge)*.

Saenko, K., Karayev, S., Jia, Y., Shyr, A., Janoch, A., Long, J., Fritz, M., and Darrell, T. (2011). Practical 3-d object detection using category and instance-level appearance models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 793–800.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J. (2000). Support vector method for novelty detection. In *Advances in Neural Information Processing System (NIPS)*, pages 582–588.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Shimosaka, M., Mori, T., and Sato, T. (2007). Robust action recognition and segmentation with multi-task conditional random fields. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3780–3786.

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

Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W. (2005). Discovering objects and their location in images. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 370–377.

Sivic, J., Russell, B., Zisserman, A., Freeman, W., and Efros, A. (2008). Unsupervised discovery of visual object class hierarchies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Slonim, N. and Tishby, N. (1999). Agglomerative information bottleneck. In *Advances in Neural Information Processing System (NIPS)*, volume 12, pages 617–623.

Smola, A., Song, L., and Teo, C. (2009). Relative novelty detection. In *International Conference on Artificial Intelligence and Statistics*, volume 5, pages 536–543.

Stein, A. N., Stepleton, T., and Hebert, M. (2008). Towards unsupervised whole-object segmentation: Combining automated matting with boundary detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Sudderth, E. and Jordan, M. (2009). Shared segmentation of natural scenes using dependent pitman-yor processes. In *Advances in Neural Information Processing System (NIPS)*, pages 1585–1592.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2008). Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77(1-3):291–330.

Sung, Y., Boulis, C., Manning, C., and Jurafsky, D. (2007). Regularization, adaptation, and non-independent features improve hidden conditional random fields for phone classification. In *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 347–352.

Sutton, C. and McCallum, A. (2005a). Composition of conditional random fields for transfer learning. In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 748–754.

Sutton, C. and McCallum, A. (2005b). Piecewise training of undirected models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Sutton, C., McCallum, A., and Rohanimanesh, K. (2007). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *The Journal of Machine Learning Research*, 8:693–723.

Sutton, C., Sindelar, M., and McCallum, A. (2006). Reducing weight undertraining in structured discriminative learning. In *Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 89–95.

Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, pages 985–992.

Teh, Y. W. (2010). Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer.

Todorovic, S. and Ahuja, N. (2008). Unsupervised category modeling, recognition, and segmentation in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2158–2174.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191.

Torralba, A., Murphy, K., and Freeman, W. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):854–869.

Torralba, A., Murphy, K., and Freeman, W. (2010). Using the forest to see the trees: Exploiting context for visual object detection and localization. *Communications of the ACM*, 53(3):107–114.

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766.

Triebel, R., Shin, J., and Siegwart, R. (2010). Segmentation and unsupervised part-based discovery of repetitive objects. In *Proceedings of Robotics: Science and Systems (RSS)*.

Tuytelaars, T., Lampert, C., Blaschko, M., and Buntine, W. (2010). Unsupervised object discovery: A ccomparison. *International Journal of Computer Vision*, 88(2):284–302.

Tzortzis, G. and Likas, A. (2009). The global kernel-means algorithm for clustering in feature space. *IEEE Transactions on Neural Networks*, 20(7):1181–1194.

Varma, M. and Zisserman, A. (2005). A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1):61–81.

Vedaldi, A., Gulshan, V., Varma, M., and Zisserman, A. (2009). Multiple kernels for object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 606–613.

Veksler, O., Boykov, Y., and Mehrani, P. (2010). Superpixels and supervoxels in an energy optimization framework. In *Europe Conference on Computer Vision (ECCV)*, pages 211–224.

Vicente, S., Kolmogorov, V., and Rother, C. (2010). Cosegmentation revisited: Models and optimization. In *Europe Conference on Computer Vision (ECCV)*, pages 465–479.

Vicente, S., Rother, C., and Kolmogorov, V. (2011). Object cosegmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2217–2224.

Vieira Neto, H. and Nehmzow, U. (2007). Visual novelty detection with automatic scale selection. *Robotics and Autonomous Systems*, 55(9):693–701.

Wang, L. and Suter, D. (2007). Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Wang, S., Quattoni, A., Morency, L., Demirdjian, D., and Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1521–1527.

Wang, X. and Grimson, E. (2007). Spatial latent dirichlet allocation. In *Advances in Neural Information Processing System (NIPS)*, pages 1577–1584.

Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *IEEE International Conference on Data Mining (ICDM)*, pages 697–702.

Wang, Y. and Mori, G. (2009). Max-margin hidden conditional random fields for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 872–879.

Weinshall, D., Hermansky, H., Zweig, A., Luo, J., Jimison, H., Ohl, F., and Pavel, M. (2008). Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. In *Advances in Neural Information Processing System (NIPS)*, volume 8.

Welling, M. and Hinton, G. (2002). A new learning algorithm for mean field boltzmann machines. In *International Conference on Artificial Neural Networks (ICANN)*, pages 82–82.

Welling, M. and Sutton, C. (2005). Learning in markov random fields with contrastive free energies. *International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pages 389–396.

Winn, J. and Jojic, N. (2005). Locus: Learning object classes with unsupervised segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 756–763.

Winn, J. and Shotton, J. (2006). The layout consistent random field for recognizing and segmenting partially occluded objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 37–44.

Wu, M. and Ye, J. (2009). A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2088–2092.

Wu, T., Lian, C., and Hsu, J. (2007). Joint recognition of multiple concurrent activities using factorial conditional random fields. In *Conference on Artificial Intelligence (AAAI)*.

Zhang, J., Xiao, J., Zhang, J., Zhang, H., and Chen, S. (2011). Integrate multi-modal cues for category-independent object detection and localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 801–806.

Zhang, N. and Poole, D. (1996). Exploiting causal independence in bayesian network inference. *Journal of Artificial Intelligence Research*, 5:301–328.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. *Advances in Information Retrieval*, pages 338–349.

Zouhar, A., Baloch, S., Tsin, Y., Fang, T., and Fuchs, S. (2010). Layout consistent segmentation of 3-d meshes via conditional random fields and spatial ordering

constraints. In *Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 113–120.

Zweig, A. and Weinshall, D. (2007). Exploiting object hierarchy: Combining models from different category levels. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

Publications Rleated to This Thesis:

Journal

Shengyong Chen, Jianhua Zhang, Youfu Li and Jianwei Zhang, "A Hierarchical Modal Incorporating Segmented Regions and Pixel Descriptors for Video Background Subtraction", IEEE Transactions on Industrial Informations, vol. 8, no. 1, pp. 118-127, 2012.

Jianhua Zhang, Sheng Liu, Y. F. Li and Jianwei Zhang, "Target Contour Recovering for Tracking People in Complex Environments", Computational and Mathematical Methods in Medicine, vol. 2012 (2012), Article ID 506908, doi:10.1155/2012/506908

Conference

Jianhua Zhang, Jianwei Zhang, Shengyong Chen and Ying Hu， "Multimodal Mixed Conditional Random Field Model for Category-Independent Object Detection", Accepted by IEEE First International Conference on Cognitive Systems and Information Processing, 2012.

Jianhua Zhang, Jianwei Zhang, Shengyong Chen, Ying Hu and Haojun Guan, "Constructing Dynamic Category Hierarchies for Novel Visual Category Discovery", Accepted by IROS 2012

Jianhua Zhang, Junhao Xiao, Jianwei Zhang, Houxiang Zhang and Shengyong Chen, "Integrate Multi-Modal Cues for Category-Independent Object Detection and Localization", IROS 2011.

Junhao Xiao, Jianhua Zhang, Houxiang Zhang, Jianwei Zhang, and Hans Petter Hildre, "Fast Plane Detection for SLAM from Noisy Range Images in Both Structured and Unstructured Environments", IEEE International Conference on Mechatronics and Automation(ICMA), Beijing, China, Aug., pp 1768 - 1773, 2011.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den Unterschrift

_____