Towards Unified Object Analytics

JIAN DONG

(B.ENG.(Hons.), USTC)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Jian Dong

July. 2014

Dong Jian

20. Nov, 2014

Acknowledgements

This thesis is not possible without the support from so many people around me. I would like to express my sincere gratitude to all of them.

My first debt of gratitude must go to my advisor, Dr. Shuicheng Yan. It has been an honor to be his Ph.D. student. For the last four years, I have been deeply inspired by his vision and passion to research, his attention and curiosity to details, his dedication to the profession, his intense commitment to his work, and his humble and respectful personality. I have learned from him not only how to conduct high-quality research, but also how to be a man of humility, kindness and patience.

My sincere thanks also goes to Professor Alan Yuille of the University of California at Los Angeles for providing me with the opportunity of visiting his group. I was impressed by his enthusiasm and curiosity, and there I met many great researchers. I am very grateful to Dr. Zhongyang Huang for offering me the internship opportunities at Panasonic and leading me working on diverse exciting projects.

I would thank all the members at LV group. They have created a very pleasant atmosphere in which to conduct research and live my life. I enjoyed all the vivid discussions we had and had lots of fun being a member of this fantastic group. I am very grateful to my senior Qiang Chen and Jiashi Feng for helping me at the beginning of my PhD career. I also would like to give my gratitude to my lovely roommates, Wei Xia and Li Zhang for the time with laughter, mutual encouragement, and love.

Last but not the least, I would like to thank my parents for their love and support throughout my life.

Contents

Summary

Object recognition is one of the fundamental challenges in computer vision and robots. Because of its complexity, object recognition is usually decomposed into simplified tasks by the research community. Although different recognition tasks may seem diverse, they share the ultimate target (visual recognition) and can thus be regarded as the same problem from different views, such as at the whole-image level - object classification, at the sub-window level - object detection, and at the pixel level - object segmentation/parsing. Due to the intrinsic consistency, these tasks should be strongly correlated. Unfortunately, the current system usually treats them separately.

In this thesis, we propose to unify the approaches for visual recognition to effectively leverage the intrinsic consistency among different recognition tasks. By reconsidering current recognition techniques, we explore several ways to integrate approaches for different tasks, such as fusing the outputs from different tasks in a principled way or directly solving multiple tasks in a unified framework, to boost the state-of-the-art performance. In addition, we further extend the idea of "unified analytics" from general object recognition to the specific field of human analysis.

In summary, we develop effective unified frameworks for both general object recognition and specific field of human analysis by employing the intrinsic consistency among different recognition tasks in a principled way.

List of Tables

List of Figures

Chapter 1

Introduction

Visual recognition is one of the fundamental challenges in computer vision and robots. The core tasks for visual recognition, such as classification, detection, segmentation and pose estimation, have drawn much research attention due to the wide application in robotics, human-computer interaction, health care, and Web data mining. As these tasks essentially handle the same problem from different views, they should be strongly related. In this thesis, we aim to explore the intrinsic consistency among different tasks for visual recognition.

In artificial intelligence, visual recognition refers to the task of automatically understanding the world using visual information, aiming to mimic the fascinating perception abilities of humans. Owing to its potential application in various domains, visual recognition has gained extensive attention for over four decades. Significant efforts have been devoted to developing representation schemes and algorithms for recognizing generic objects in real-world images [18]. However, so far even devising vision systems that can match the cognitive abilities of children is still very challenging.

Because of its complexity, visual recognition is usually decomposed into simplified tasks by the research community. Among various tasks, several of them receive special attention for their wide application: (1) Object Classification which aims to predict the existence of certain objects in the images, (2) Object Detection which targets to predict and localize the objects in the images, (3) Object Segmentation

which tries to obtain the per-pixel object level indication masks for the images, and (4) Pose Estimation which desires to estimate the 2D/3D spatial configuration of the objects in the images. Previous works on visual recognition often solve each task separately, which ignore the strong correlation among different tasks. For example, many state-of-the-art image classification systems follow the popular local feature extraction-coding-pooling pipeline [18]. Each image is represented globally by a feature vector. Though such representation has demonstrated to be robust to occlusion and pose variance, it is sensitive to scale of the object. If the size of the concerned object is too small, the information from it is easy to be suppressed by clustered background. In contrast, the current de facto systems for object detection employ the sliding window approach. Assisted with the multi-scale strategy, this sliding window based system can effectively detect the object of small scale [44]. However, such approach only relies on the information inside the bounding box region and thus ignores the valuable background information, which may lead to inferior performance. Owing to the complementary properties of classification and detection, combining them properly should boost the performance of each other [86]. Similarity, segmentation and detection are highly related. The bounding boxes from the detection methods will significantly simplify the segmentation task while the results from segmentation can directly convert to the bounding boxes for detection.

This thesis focuses on exploring the intrinsic correlation among different recognition tasks to boost the final recognition performance. Instead of improving the existing models for a specific task, we believe that it is more important to look at the recognition in a bigger picture.

1.1 Background and Related works

This section presents a survey of literature for classical tasks in visual recognition, focusing on the intrinsic consistency between these tasks. After briefly reviewing the traditional works to handle each task separately, we introduce the recent advances in combining the techniques designed for different tasks.

1.1.1 Classical Tasks for Visual Recognition

Object Classification

Objects usually come with specific background. For example, airplanes often appear in the sky. Hence, the context information should be valuable to predict the existence of certain objects in the images. Many state-of-the-art image classification systems follow the popular local feature extraction-coding-pooling pipeline [42], which effectively utilizes both foreground and background information. Specifically, local features like HOG [27], SIFT [75] and LBP [78] are first extracted on the dense grids or sparse interest points. They are then encoded with a predefined visual dictionary by vector quantization (VQ), locally-constrained linear coding (LLC) [99] or Fisher kernel (FK) [49, 99]. Finally the encoded vectors are pooled together to form the image-level representation [67, 20]. Much research on image classification has been focused on improving this pipeline [42, 99, 49, 18]. However, this traditional framework assigns equal weight to each local feature. Thus, it is sensitive to scale of the object. It would risk suppressing information of the concerned object by clustered background if its size is too small.

Object Detection

Object detection [44], which is complementary to object classification, is another central problem in visual recognition [86, 58]. As an object can appear at any position and scale in the image, sliding window scanning has shown to be extremely effective, and consequently become the dominant paradigm for a long time [2]. By exhaustive search, the original complicated detection problem can be converted into much simpler binary classification problems. However, such approach makes use of only the image inside the bounding box and thus ignores the valuable context information, which may lead to inferior performance.

Sematic Segmentation

Sematic segmentation [14], which aims to provide more detailed information (pixel-level labeling) than classification and detection, is usually cast as an optimization problem under the MRF framework [65]. Traditional methods usually only rely on the low level information. More specifically, such methods utilize appearance features to construct unary term. Similar to object detection, such segmentation methods fail to capture the informative context information. More importantly, the shape and other top-down information are often discarded, which may heavily decrease performance [65].

Human Parsing

Unlike other classical tasks, there exist several inconsistent definitions for human parsing in literature. Some works [94, 101, 102] treat human parsing as a synonym of human pose estimation. In this thesis, we follow the convention of scene parsing [71, 89] and define human parsing as partitioning the human body into semantic regions. Though human parsing plays an important role in many human-centric applications [19], it has not been fully studied. Yamaguchi *et al.* [111] performed human pose estimation and attribute labeling sequentially for clothing parsing. However, such sequential approaches may fail to capture the correlations between human appearance and structure, leading to unsatisfactory results.

1.1.2 Contextualization and Unification for Visual Recognition

Due to the strong correlation between these classical tasks, some recent works began to investigate how to effectively combine the techniques designed for separate tasks [26, 107].

Contextualization

Contextualization refers to combine the results from separate tasks in the late fusion stage. The results of many tasks, such as object classification and detection, are

complementary and thus should be able to boost each other. Harzallah et al. [58] introduced the pioneering work for detection and classification contextualization. Though contextualization is intuitive, how to perform contextualization effectively and efficiently is still rarely studied. In this work, we expect to investigate how to fuse the results from the leading techniques to further improve the state-of-the-art pipeline [86, 107].

Unification

As discussed above, different tasks share the ultimate target and the techniques designed for each task should have the potential to corporate with each other. Unification here refers to combining many related techniques in a unified framework. Selective search based recognition [14] is a representative approach. This line of works first generate set of object hypotheses based on bottom-up segmentation methods and then convert the recognition problem into a classification problem. Though great success has been achieved in the past few years [14, 13], current works mainly focus on limited type of unification, such as unifying segmentation and classification. In this work, we aim to explore more kinds of unification under a novel framework to reveal the power of unification.

1.1.3 Datasets

General Object Recognition

Many datasets exist for general object recognition. Unfortunately, most of them, such as Oxford Flowers [77], Caltech 101 [41] and Caltech 256 [52], are single-label, object-centric and deficient in variance of pose and appearance, which makes these datasets insufficient to represent the visual world. In addition, these datasets are near saturation and not discriminative enough to distinguish different leading algorithms. Hence, we validate the proposed framework on the challenging PASCAL Visual Object Challenge (VOC) datasets [39, 36, 37], which provide a common evaluation platform for both object classification and detection. These datasets

are extremely challenging since the images are crawled from the real-world photo sharing website and the objects contained vary significantly in size, pose, view point and appearance. VOC 2007, 2010 and 2012 datasets are selected for experiments. These datasets contain 20 object classes and are divided into "train", "val" and "test" subsets. We conduct our experiments on the "trainval" and "test" splits. We follow the standard PASCAL protocol by employing Average Precision (AP) and Intersection over Union (IoU) as evaluation metric for object classification/detection and semantic segmentation, respectively.

Human Parsing

Our experiments are conducted on two datasets. The first one is the Fashionista (FS) dataset [111], which has 685 annotated samples with 56 different clothing labels. This dataset is originally designed for fine-grained clothing parsing. To adapt this dataset for our human parsing, we merge their labels according to our Parselet definition as in [32].. As there is no direct link between their annotation and our "coat" Parselet, we ignore the "coat" Parselet and merge all upper body clothing into the "upper clothes" Parselet. The second dataset, called Daily Photos (DP), contains 2500 high resolution images, which are crawled following the same strategy as the FS dataset [111]. In order to obtain quantitative evaluation results, we thoroughly annotate the semantic labels at pixel-level. Compared with FS, the DP dataset contains much more images and has consistent labels with Parselet definition for human parsing. we label the common 14 joint positions in the same manner as in [111].

1.2 Thesis Focus

Based on the above review, although different tasks for visual recognition seem diverse, they share the ultimate target (visual recognition) and can thus be regarded as the same problem from different views, i.e. at the whole-image level - object classification, at the sub-window level - object detection, and at the pixel level -

object segmentation. Due to the intrinsic consistency, these tasks should be strongly correlated. Unfortunately, the current system usually treats them separately, failing to utilize the information from different levels effectively. The detailed research gaps are summarized below:

- Information from different levels, such as whole-image level and sub-window level, are essentially complementary. As classical models usually focus on a specific level, the resulting system fails to capture the complementary information and thus cannot utilize such complementary information to distinguish ambiguous samples at a specific level, leading to inferior performance.

- Many new applications rely on the proper combination of various recognition techniques. Current leading approaches usually brutally decompose these new problems into well-defined tasks. For example, human parsing is decomposed into sequential human pose estimation and region labeling. However, these brute decompositions may ignore the intrinsic properties of each task and thus harm the overall performance.

Instead of improving the existing models for a specific level, we believe that it is more important to look at the recognition in a bigger picture. Thus, the main aim of this thesis is to explore the intrinsic correlation among different recognition tasks to boost the final recognition performance for each task. More specifically, we conduct research on the following aspects:

- Subcategory Aware Object Recognition. We explore the subcategory structures embedded in semantic categories, which are effective to link the outputs of different tasks. We then build a subcategory-aware recognition framework to boost category level object classification performance. Different from the existing monolithic model approaches, we aim to automatically leverage the embedded subcategory structure to assist the further category level recognition. Motivated by the observation of considerable intra-class diversities and inter-class ambiguities in many current object classification datasets, we explicitly split data into subcategories by ambiguity guided subcategory mining.

The resulting subcategories are seamlessly integrated into the state-of-the-art detection assisted classification framework [34, 30].

- Unified Object Detection and Semantic Segmentation. Object detection and semantic segmentation are two strongly correlated tasks, yet typically solved separately or sequentially with substantially different techniques. Motivated by the complementary effect observed from the typical failure cases of the two tasks, we propose a unified framework for joint object detection and semantic segmentation. By enforcing the consistency between final detection and segmentation results, the proposed unified framework can effectively leverage the advantages of the leading techniques for these two tasks [33].

- Human Parsing based on Parselets. Previous works often consider solving the problem of human pose estimation as the prerequisite of human parsing. We argue that these approaches cannot obtain optimal pixel level parsing due to the inconsistent targets between different tasks. To overcome this limitation, we directly address the problem of human parsing by using the novel Parselet representation as the building blocks of our parsing model. We then build a Deformable Mixture Parsing Model (DMPM) for human parsing to simultaneously handle the deformation and multi-modalities of Parselets. The DMPM thus directly solves the problem of human parsing by searching for the best graph configuration from a pool of Parselet hypotheses without intermediate tasks to guarantee the overall performance [32].

- Unified Human Parsing and Pose Estimation. Human parsing and human pose estimation, *i.e.* identifying the semantic regions and body joints respectively over the human body image, are intrinsically highly correlated. However, previous works generally solve these two problems separately or iteratively. In this thesis, we propose a unified framework for simultaneous human parsing and pose estimation based on semantic parts. By utilizing Parselets and Mixture of Joint-Group Templates as the representations for these semantic parts, we seamlessly formulate the human parsing and pose estimation prob-

lem jointly within a unified framework via a tailored And-Or graph. A novel Grid Layout Feature is then designed to effectively capture the spatial co-occurrence/occlusion information between/within the Parselets and MJGTs. Thus the mutually complementary nature of these two tasks can be harnessed to boost the performance of each other [31].

1.3 Thesis Overview

In Chapter 2, we propose a subcategory aware recognition approach to contextualize object detection and classification. Then in Chapter 3, we show how to perform object detection and semantic segmentation in a unified framework. In Chapter 4, we reconsider the human parsing problem and propose to use Parselets as the basic elements for human parsing. Finally, we demonstrate a unified framework for simultaneous human parsing and pose estimation in Chapter 5.

Chapter 2

Looking Inside Category: Subcategory-aware Object Recognition

In this chapter, we show how to automatically mine the embedded visual subcategory structure in semantic categories. The resulting subcategory information can help to link the outputs of current leading object classification and detection methods and improve the category level recognition performance.

2.1 Introduction

Visual categorization is a core problem in computer vision. Bag-of-Words (BoW) approaches to category level classification advanced significantly during the past few years [42, 67, 99, 49, 20]. This framework utilizes the local feature extraction, feature encoding and feature pooling pipeline to generate global image representations. Each object category is then represented by a monolithic model, such as a support vector machine classifier. However, the large intra-class diversities induced by pose, viewpoint and appearance variations [76] make it difficult to build an accurate monolithic model for each category, especially when there are many ambiguous samples. For example, the chair category in Figure 2.1 includes three

obvious subcategories, namely, sofa-like chairs, rigid-material chairs and common chairs. In feature space, these subcategories are essentially far away from each other. Furthermore, the ambiguous sofa-like chairs look more like sofas than common chairs. In this case, representing all chairs with a monolithic model will weaken the model separating capacity and cannot distinguish sofas from chairs. Hence, it is intuitively beneficial to model each subcategory independently. These considerable intra-class diversities and inter-class ambiguities are common in the challenging real world datasets [39, 109], which makes the subcategory awareness necessary.

To effectively employ the subcategory information for category level classification in a principled way, the first step is to mine the subcategory structure automatically. At first glance, clustering all training data of an object category based on intra-class similarity seems to be a natural strategy, since objects belonging to the same subcategory should intuitively have larger similarity in terms of appearance and shape. However, in the context of generic object classification, subcategories mined with only intra-class visual similarity cues are unnecessary to be optimal due to the ignorance of valuable inter-class information [25]. More specifically, if the samples are clustered by standard clustering methods, we are unable to utilize the valuable inter-class information to handle the ambiguous samples. Then all ambiguous samples, which often lie near the decision boundary, may be grouped together and preserve the original complicated decision boundary. On the contrary, with the assistance of inter-class information ambiguous samples can be grouped into proper subcategories, which leads to easier subproblems and further improves the overall performance. For instance, the chair category and other categories in Figure 2.1 have non-linear decision boundary. By noting the ambiguous chair sample distribution near the decision boundary, these chairs should be intuitively divided into separate subcategories. The proper split as indicated in Figure 2.1 will make all subcategories linearly separable from other categories, which is only achievable with the assistance of inter-class information. The above observation inspires us to propose an ambiguity guided subcategory mining approach to explore the intrinsic subcategory structure embedded in each category.

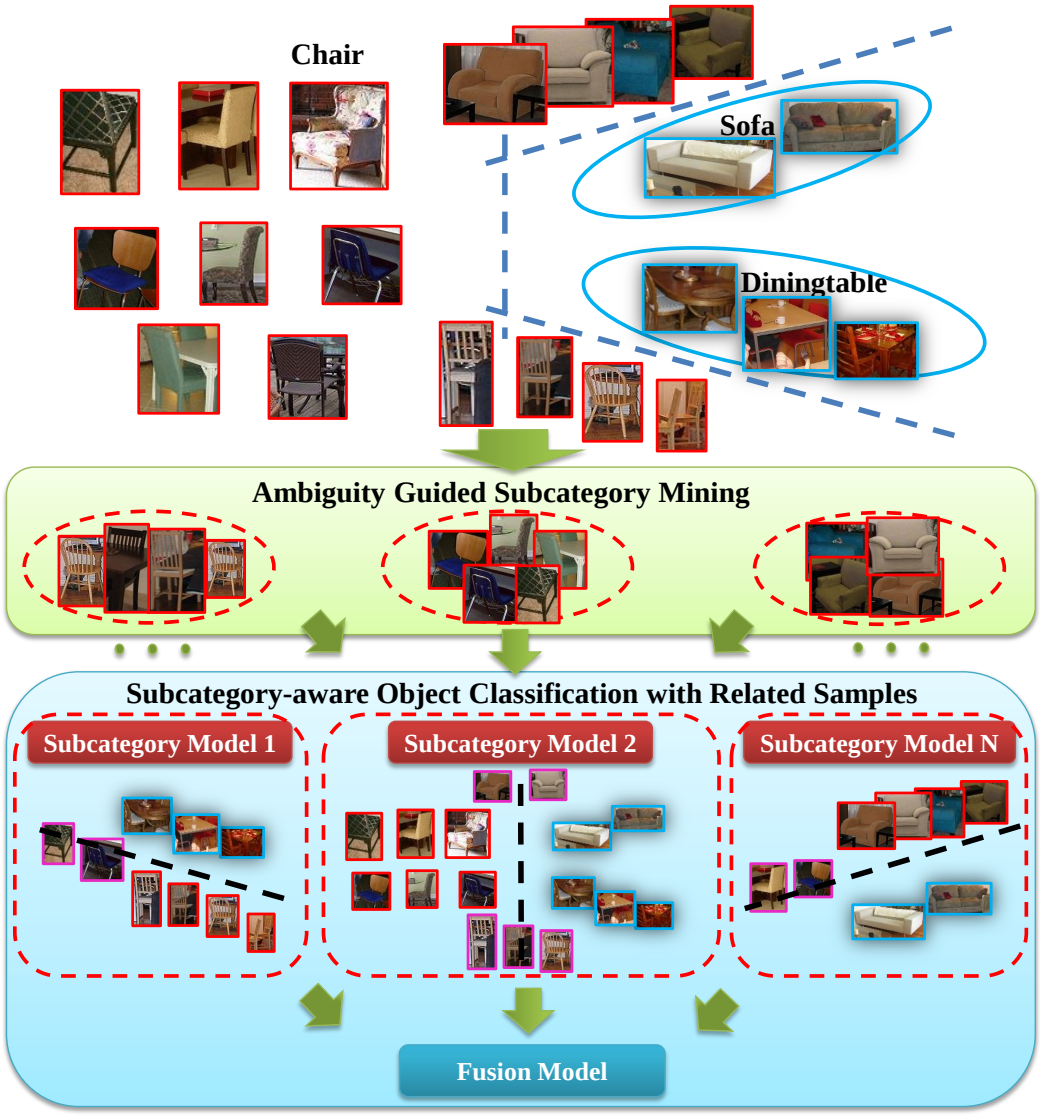


Figure 2.1: Overview of the proposed ambiguity guided subcategory mining and subcategory-aware object classification framework. For each category, training samples are automatically grouped into subcategories based on both intra-class similarity and inter-class ambiguity. An individual subcategory model is constructed for each detected subcategory. During training, the samples assigned to the target subcategory, the other subcategories belonging to the same category and other categories are treated as positive, related and negative samples, respectively. The final classification results are obtained by aggregating responses from all subcategory models.

With mined subcategories, designing an effective strategy to train subcategory classifiers tailored for category level classification is not trivial. A naive approach is assigning the samples for the mined subcategory as positive samples and samples

in the other categories as negative samples. However, such approach ignores the informative related samples (samples from other subcategories of the same category) and is unstable for some subcategories with small number of samples. Instead, we propose to employ the related samples under the "Universum" SVM framework [85], which can stabilize and regularize the subcategory classifier to further boost the category level performance.

Overall, with subcategory awareness we can boost category level classification by subcategory-aware object classification (SAOC). As indicated in Figure 2.1, we split data into subcategories by ambiguity guided subcategory mining and train an individual model for each subcategory. During subcategory classifier training, besides positive and negative samples we further leverage the related samples to regularize the subcategory classifier for better fitting the overall category level data distribution. Since the diversities in each subcategory and ambiguities between subcategories and other categories are reduced, more accurate shape-based [27, 44]/appearance-based [96, 70] detectors and foreground classification model [20] can be built, which fits nicely with the state-of-the-art detection assisted classification framework [58, 86]. The final classification results are generated by aggregating subcategory responses through subcategory-aware kernel regression.

The main contributions of this chapter are summarized as follows.

- We propose a novel ambiguity guided subcategory mining approach, which gracefully integrates the intra-class similarity and inter-class ambiguity for effective subcategory mining.

- We design an effective strategy to employ "related samples" under the "Universum" SVM framework. Such informative related samples will fine-tune the subcategory classifiers to be more suitable for category level classification.

- We provide a subcategory-aware object classification framework based on the detection assisted classification scheme [58, 86] to demonstrate how to effectively employ the subcategory information for visual recognition. Our ambiguity guided subcategory mining approach can be seamlessly integrated into

such framework. Utilizing mined subcategories can improve both detection and classification performance and allow more effective subcategory level interaction in the fusion model. The state-of-the-art classification results on the PASCAL VOC datasets verify the effectiveness of our new framework.

The rest of the chapter is organized as follows. Section 2.2 briefly reviews the related literature. Section 2.3 describes the overview of the proposed subcategory aware classification framework. Detailed explanation of subcategory mining and subcategory classification with related samples is presented in Section 2.4 and Section 2.5. Extensive experiments are conducted in Section 2.6. Section 2.7 concludes the chapter.

2.2 Related Work

Current leading detection assisted classification framework relies on the cooperation of many recognition techniques, such as classification, detection and even segmentation. A detailed review of all the fields is beyond the scope of this chapter, hence we only focus on the topics that are most related to the proposed framework.

Object Classification. Traditional works for image classification usually focused on improving the popular local feature extraction-coding-pooling pipeline [18]. Some recent works [86, 58, 70, 96, 82, 6, 92] have begun to investigate out of this pipeline. Harzallah et al. [58] introduced the pioneering work for detection and classification contextualization, the extension of which leads to the state-of-the-art results [86, 20, 82]. Segmentation results [14] have also been employed to boost the classification performance [96, 70]. However, all the above methods train a monolithic model for each category, and there are few works analyzing the data structure embedded in each category. In this chapter, we show that properly splitting the data into subcategories will boost the performance of the state-of-the-art pipeline. Another line of work design a large number of weakly trained classifiers and treat the output of these classifiers as image descriptor [6, 92]. Such weak classifiers are usually obtained from semantic annotation, such as visual concepts, and bear the

mid-level information to some extent. Unlike these methods, our work automatically discovery the structure embedded in each category without relying on manual annotation.

Object Detection. For object detection, mixture models are proposed and have become the standard approach [121, 44], as most semantic categories do not form coherent visual categories. Early works only investigate heuristics based on meta-data or manual labels such as bounding box aspect ratio [44], object scale [79], object viewpoint [55] and part labels [10] to group the positive samples into clusters. However, each of these methods has its own limitations and ignores other more general intra-class variations such as appearance and shape variance [76, 53]. Malisiewicz et al. [76] handled the intra-class variation by training a separate model for each positive instance, which inevitably reduces the generalization capacity of each model. Some recent works begin to investigate the visual subcategory structure embedded in each category [28, 53, 24, 121, 2, 29], which leads to considerable improvement in object detection performance. Gu et al. [53] grouped the samples into components based on the key point and mask annotations. Aghazadeh et al. [2] built a similarity graph based on intra-class information and utilized spectral clustering to split the data. In contrast to our method, these methods either require manual annotation or are fragile to outliers corresponding to highly occluded or strange samples. Furthermore, most of previous works focus on object detection and are not suitable for object classification. Finally, these methods discard the inter-class information during data grouping, which is critical for object classification.

Locally Adaptive Classifiers. When the data has a complex non-linear structure, locally adaptive classifiers are usually superior to the use of a single global classifier [93, 62, 25]. Kim and Kittler placed the local classifiers at the clusters obtained by the K-means clustering algorithm [62]. Instead of placing the classifiers based on the data distribution only, Dai et al. [25] proposed a responsibility mixture model that uses the uncertainty associated with the classification at each training sample. Using this model, the local classifiers are placed near the decision boundary where they are most effective. Hoai and Zisserman [59] learn sub-categories by in-

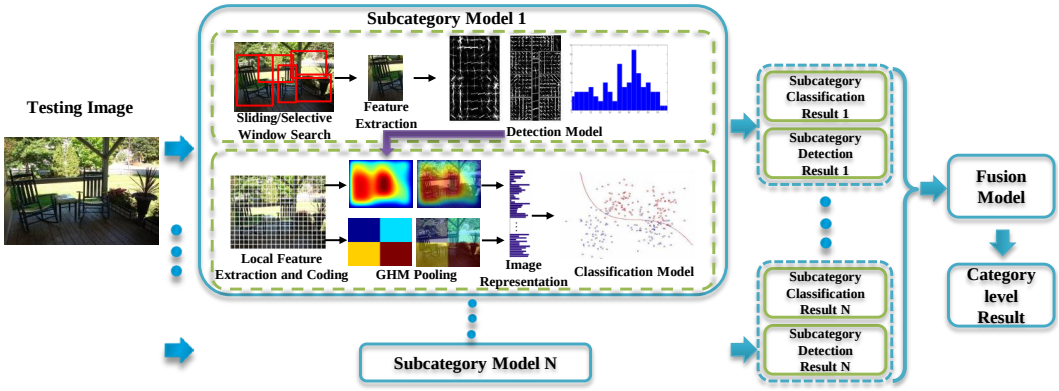


Figure 2.2: Diagrammatic flowchart of the proposed subcategory-aware object classification framework. Given a testing image, they are first processed by each learnt subcategory model including detection and classification models. Then the responses from all subcategory models are fed into the fusion model to generate the final category level classification results.

vestigating a weakly supervised approach using both positive and negative samples of the category. In this chapter, we borrow the idea of uncertainty piloted classification and propose an ambiguity guided subcategory mining approach under the graph shift [72] framework.

2.3 Subcategory-aware Object Classification

Our subcategory-aware object classification (SAOC) framework relies on the automatically mined subcategory information to boost the category level recognition. In this section we mainly demonstrate how to effectively utilize the mined subcategory information in current leading detection assisted classification scheme. Details on subcategory mining and strategy of training a subcategory classifier are shown in the sequent sections.

The diagrammatic flowchart of our SAOC framework is depicted in Figure 2.2. The whole framework consists of three main components - detection, classification and fusion models. We will first introduce each component of the framework and then emphasize how subcategory information fits into each step.

2.3.1 Classification Model

For classification, we follow the state-of-the-art Generalized Hierarchical Matching (GHM) pipeline [20] and train a classifier for each subcategory individually. GHM generalizes the Spatial Pyramid Matching by allowing image adaptive pooling instead of pre-defined grid-based pooling. Both the detection confidence map and saliency map have shown to be effective to guide the pooling process for certain datasets [20]. In this chapter, since we focus on the scenarios where background is usually cluttered and many of the concerned object classes may co-occur in a single image, detection confidence maps are employed as the side information for GHM. The details for classifier training are explained in Section 2.5.

2.3.2 Detection Model

Detection and classification are two strongly correlated and complementary tasks. Most leading classification systems employ the detection techniques to some extent. In our framework, the raw detection results are fed into the final fusion model as middle level features as well as provide the confidence map for the GHM pooling. Specifically, each subcategory is characterized by one shape-based sliding window detector [44, 118] and one appearance-based selective window detector [97, 96], respectively. The usage of two detectors is to guarantee both high precision and high recall on object detection since none of the detectors can achieve this alone and they complement each other.

2.3.3 Fusion Model

The fusion model mainly aims to: (1) boost the classification performance by complementary detection results, (2) utilize the context of all categories for reweighting, and (3) fuse the subcategory level results into final category level results. All of these are achieved by kernel regression. First, we construct a middle level representation for each training/testing image by concatenating classification scores and the leading two detection scores from each subcategory model. The final cat-

egory level classification results are then obtained by performing Gaussian kernel regression on this representation. Without sophisticated models and complicated postprocessing [37, 86], our subcategory-aware kernel regression is very efficient and still performs well experimentally.

2.3.4 Subcategory Awareness

Subcategory awareness, which benefits each model separately and then boosts the overall performance of the framework, plays a critical role in extending current detection assisted classification framework.

- The subcategory information can be used to initialize both detection and classification models to better handle the rich intra-class diversities in challenging datasets. Less diversity in each subcategory will lead to a simpler learning problem, which can be better characterized by current state-of-the-art models, such as the Deformable Part based Model (DPM) for detection and the foreground BoW models involved in GHM.

- The subcategory awareness will lead to more effective fusion models. First, subcategory awareness allows us to model the subcategory level interaction. For example, occluded chairs and sitting persons often occur together. The co-occurrence of occluded chairs and sitting persons then should boost the classification scores of each other. On the contrary, unoccluded chairs and pedestrians are independent, the co-occurrence of which should not improve the classification scores of either one. However, these two different cases cannot be differentiated in the category level. Only by subcategory awareness can such underlying correlation be captured effectively. Second, the subcategory awareness is able to reduce the false mutual boosting when samples from ambiguous categories are wrongly classified. More specially, diningtables often appear together with common chairs. Then the co-occurrence of diningtables and common chairs should lead to mutual boosting of classification scores. On the contrary, sofas and diningtables are usually independent and thus

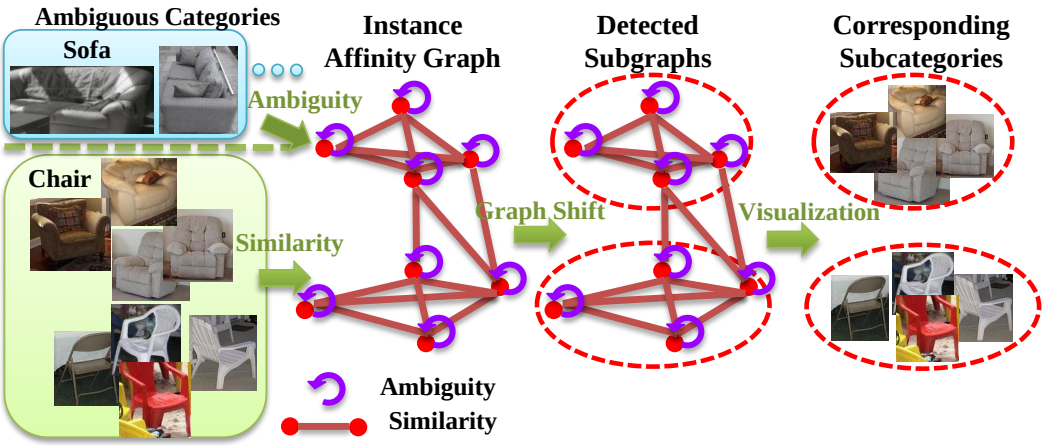


Figure 2.3: Ambiguity guided subcategory mining approach. First instance affinity graph is built by combining both intra-class similarity and inter-class ambiguity. Then dense subgraphs are detected within the affinity graph by performing graph shift. Each detected dense subgraph corresponds to a certain subcategory.

should not mutually boost the classification scores of each other. However, for category level interaction, if sofas are misclassified as chairs, the dinningtable scores may be boosted and thus lead to false alarms on diningtables. With subcategory awareness, the response of diningtable will not be boosted as there exists no mutual boosting between the sofa-like chairs and diningtables.

2.4 Ambiguity Guided Subcategory Mining

In this section, we will introduce how to find the subcategories by our ambiguity guided subcategory mining approach as illustrated in Figure 2.3. Before digging into details, we first summarize the notations used in this work. For a classification problem, a training set of M samples are given and represented by the matrix $X = [x_1, x_2, \ldots, x_M] \in \mathbb{R}^{d \times M}$. The class label of x_i is $c_i \in \{1, 2, \ldots, N_c\}$, where N_c is the number of classes. We also denote the number of samples belonging to the cth class by n_c, and the corresponding index set of samples by π_c.

2.4.1 Similarity Modeling

In this work, we define the appearance similarity as the Gaussian similarity between classification features ($\exp\{-||x_i - x_j||^2/\delta^2\}$), where δ^2 is the empirical variance of x. Though it is a common similarity metric for object classification, appearance similarity only is not enough for our SAOC framework, as in SAOC classification and detection are closely integrated. Subcategory mining only based on appearance similarity may lead to poor detectors, which in turn harms the overall performance. Hence detection and classification feature spaces ought to be taken into count simultaneously for similarity calculation.

The HOG based sliding window methods are the dominant approaches for object detection, which concatenate all the local gradients to form the window representation. These grid based HOG representations roughly capture object shapes and thus are sensitive to highly cluttered backgrounds and misalignments. Directly computing distance in concatenated HOG feature space often leads to poor results due to image misalignments [76]. To better measure the shape similarity between samples, we train a separate Exemplar-SVM detector[76, 56] for each positive sample. The misalignments can thus be partially handled by sliding the detector. The calibrated detection scores [76] are defined as the pair-wise shape similarity.

The final instance similarity is defined by fusing the appearance similarity and pair-wise shape similarity. More specifically, we denote the appearance similarity as $S(A)_{i,j}$ and the pair-wise shape similarity as $S(P)_{i,j}$. Both $S(A)$ and $S(P)$ are normalized to $[0, 1]$. The final instance similarity is defined as $S_{i,j} = S(A)_{i,j} \times S(P)_{i,j}$.

2.4.2 Ambiguity Modeling

As discussed above, inter-class information is crucial for object classification. Dai et al. [25] have shown that placing local classifiers near the decision boundary instead of based on the data distribution only leads to better performance. This is intuitive as even there are many subcategories spreading separately in the feature space, if

none of subcategories are close to samples of other categories, a single classifier may be enough to correctly classify all these subcategories. On the contrary, if some subcategories are near the decision boundary, separate classifiers should be trained for these ambiguous subcategories. Otherwise the ambiguous subcategories may decrease the classification performance of categories near the decision boundary.

As ambiguity is critical for object classification, subcategory mining should be guided by ambiguity instead of only relying on intra-class data distribution. Before introducing how to combine sample similarity and ambiguity into a unified framework, we need to first explicitly define the ambiguity measure. Here, we consider the L-nearest neighbours[1] of a particular sample x_i. If most of its neighbours share the same class label as x_i, the classification of x_i should be easy. Otherwise, x_i will be ambiguous and likely to be classified incorrectly. We thus define the ambiguity $A(x_i)$ of a training sample x_i as:

$$A(x_i) = \frac{\sum_{j \in N_i^L, j \notin \pi_{c_i}} S_{i,j}}{\sum_{j \in N_i^L} S_{i,j}}, \tag{2.1}$$

where N_i^L is the index set of the L-nearest neighbours of x_i. From the definition, a large $A(x_i)$ means that the neighbouring samples are likely to be of different classes, and hence the classification of x_i is more uncertain. On the contrary, a small $A(x_i)$ indicates that more neighbouring samples share the same class label of x_i. Note that computing the ambiguity relies on not only the intra-class information but also the inter-class formation. The ambiguity will be high for those training samples lying close to the decision boundary, and thus such samples should be more likely to form a separate subcategory.

2.4.3 Subcategory Mining by Graph Shift

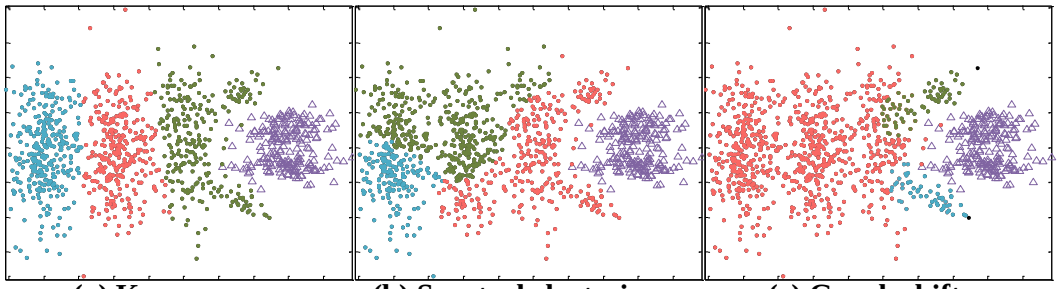
Intuitively, the subcategory mining algorithm is expected to satisfy the following three properties. (1) It should be compatible with graph representation. Many similarity metrics are defined based on pair-wise relation, such as our pair-wise shape

[1] In the experiments, we simply use $L = n_c/10$ for the cth class.

similarity. Hence, non-graph based algorithms, such as mean shift, k-means and [59], may not be suitable due to the lack of ability to directly utilize the pair-wise information. (2) It is able to utilize the informative inter-class ambiguities. Clustering methods based on only intra-class data distribution may fail to detect the ambiguous subcategories on the decision boundary and lead to subcategories imperfect for classification. Hence the expected algorithm should be able to adaptively cluster the data guided by ambiguity. (3) It should be robust to outliers. Some samples, such as highly occluded or strange images, may not belong to any subcategory. Methods insisting on partitioning all the input data into coherent groups without explicit outlier handling may fail to find the true subcategory structure.

The traditional partitioning methods, such as k-means and spectral clustering methods, are not expected to always work well for subcategory mining due to their insistence on partitioning all the input data and inability to integrate the inter-class information. Hence we need a more effective algorithm satisfying the above three properties. The graph shift algorithm [72], which is efficient and robust for graph mode seeking, appears to be particularly suitable for our subcategory mining problem as it directly works on graph, allows one to extract as many clusters as desired, and leaves the outlier points ungrouped. More importantly, the ambiguity can be seamlessly integrated into the graph shift framework. The graph shift algorithm shares the similar spirit with mean shift [23] algorithm and evolves through iterative expansion and shrink procedures. The main difference is that mean shift operates directly on the feature space, while graph shift operates on the affinity graph. The simulation results for comparing our ambiguity guided graph shift (AGS) with kmeans and spectral clustering are provided in Figure 2.4, from which we can see that our AGS can lead to subcategories more suitable for boosting classification.

Formally, we define an individual graph $G = (V, A)$ for each category. $V = \{v_1, \ldots, v_n\}$ is the vertex set, which represents the positive samples for the corresponding category. A is a symmetric matrix with non-negative elements. The diagonal elements of A represent the ambiguity of the samples while the non-diagonal



(a) Kmeans **(b) Spectral clustering** **(c) Graph shift**

Figure 2.4: The subcategory mining results on synthetic data from kmeans, spectral clustering and graph shift. Here, triangles ($\triangle$) and dots ($\cdot$) represent samples from two different categories, respectively. Dots are split into subcategories, and different colors represent different subcategories. Kmeans and spectral clustering cluster the dots relying on only intra-class information, which leads to non-linearly separable subcategories from triangles. However, by utilizing the inter-class information, all three subcategories mined by the ambiguity guided graph shift are linearly separable from triangles, which is desired for classification. For better viewing, please see original colour pdf file.

element measures the similarity between samples. The modes of a graph G are defined as local maximizers of graph density function $g(y) = y^T Ay, y \in \Delta^n$, where $\Delta^n = \{y \in R^n : y \geq 0 \text{ and } ||y||_1 = 1\}$. More specifically, in this chapter sample similarity and ambiguity are integrated and encoded as the edge weights of a graph, whose nodes represent the instances of the specific object category. Hence subcategories should correspond to those strongly connected subgraphs. All such strongly connected subgraphs correspond to large local maxima of $g(y)$ over simplex, which is an approximate measure of the average affinity score of these subgraphs.

Since the modes are local maximizers of $g(y)$, to find these modes, we need to solve following standard quadratic optimization problem (StQP) [8]:

$$\begin{aligned} &\text{maximize} \quad g(y) = y^T Ay \\ &\text{subject to} \quad\quad y \in \Delta^n. \end{aligned} \tag{2.2}$$

Replicator dynamics, which arises in evolutionary game theory, is the most popular method to find the local maxima of StQP (2.2). Given an initialization $y(0)$, corresponding local solution y^* of StQP (2.2) can be efficiently computed by the

discrete-time version of first-order replicator equation, which has the following form:

$$y_i(t+1) = y_i(t)\frac{(Ay(t))_i}{y(t)^T Ay(t)}, i = 1, \ldots, n. \tag{2.3}$$

It can be observed that the simplex Δ^n is invariant under these dynamics, which means that every trajectory starting in Δ^n will remain in Δ^n . Moreover, it has been proven in [104] that, when A is symmetric and with non-negative entries, the objective function $g(y) = y^T Ay$ strictly increases along any non-constant trajectory of Eqn. (2.3), and its asymptotically stable points are in one-to-one correspondence with strict local solutions of StQP (2.2). One of the main drawbacks of replicator dynamics is that it can only drop vertices and be easily trapped in any local maximum. The graph shift algorithm provides a complementary neighbourhood expansion procedure to expand the supporting vertices [72]. The replicator dynamics and the neighbourhood expansion procedure thus have complementary properties, the combination of which leads to better performance. In addition, as the diagonal elements may prevent the expansion to other vertices with no diagonal elements, vertices with large diagonal elements tends to form a local subgraph.

Like mean shift algorithm, the graph shift algorithm starts from each individual sample and evolves towards the mode of G. The samples reaching the same mode are grouped as a cluster. Each large cluster corresponds to one subcategory, while small clusters usually result from noises and/or outliers.

2.5 Subcategory Classification with Related Samples

With the subcategory mining results, the following step is to construct subcategory classifiers tailored for category level classification. One intuitive approach is to employ standard binary SVM while treating samples in the target subcategory as positive samples and samples in other categories as negative samples. However, this strategy may lead to sub-optimal results for the final category level classification due to several reasons. First, this hard separation of the whole training samples

may result in limited samples for some subcategories, which will lead to unstable classifiers. Second, this approach is unable to exploit other informative samples in the same category. As our main goal is to construct classifiers suitable for category level classification instead of for accurate subcategory classification, classifiers only relying on the samples in the target subcategory may decrease the final category level performance.

To overcome the difficulty mentioned above, we propose the concept of "related samples". For a target subcategory, related samples are defined as samples from other subcategories of the same category. Though unlabeled, the related samples should be informative for classification. A prominent example for utilizing unlabeled data is semi-supervised learning [17], where an additional set of unlabeled data are assumed to follow the same distribution as the training inputs. However, for our subcategory classification problem, related samples should have different distribution from either positive or negative samples. In other word, these related samples, which are considered potentially helpful for classification, should represent a third class. We note that the related samples can be viewed as a special form of "Universum" set as in [85]. Hence, we employ the "Universum" SVM framework [85, 105] for subcategory classification.

"Universum" SVM is an extension of the standard SVM by introducing the "Universum" set. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\, y_i \in \{-1, 1\}\}_{i=1}^n$ be the set of labeled examples and let $\mathcal{U} = \{(\mathbf{x}_j) \mid \mathbf{x}_j \in \mathbb{R}^p\}_{j=1}^m$ denote the set of related samples. $H_a[t]$ is the hinge loss ($H_a[t] = \max\{0, a - t\}$) and $I_\epsilon[t]$ is ϵ-insensitive loss ($I_\epsilon[t] = \max\{0, |t| - \epsilon\}$). Besides penalizing the wrongly classified samples in $\mathcal{D}$, we also bring the related examples close to the separating hyperplane by minimizing the ϵ-insensitive loss on the related samples in $\mathcal{U}$. For a linear discriminant functions $f_{w,b}(x) = (w \cdot x) + b$, the final object function is formulated as follows:

$$\min_{w,b} \frac{1}{2}||w||^2 + C_{\mathcal{D}} \sum_{i=1}^{n} H_1[y_i f_{w,b}(x)] + C_{\mathcal{U}} \sum_{j=1}^{m} I_\epsilon[f_{w,b}(x_j)]. \tag{2.4}$$

Noting that $I_\epsilon[t] = H_{-\epsilon}[t] + H_{-\epsilon}[-t]$, one can use the simple trick of adding the

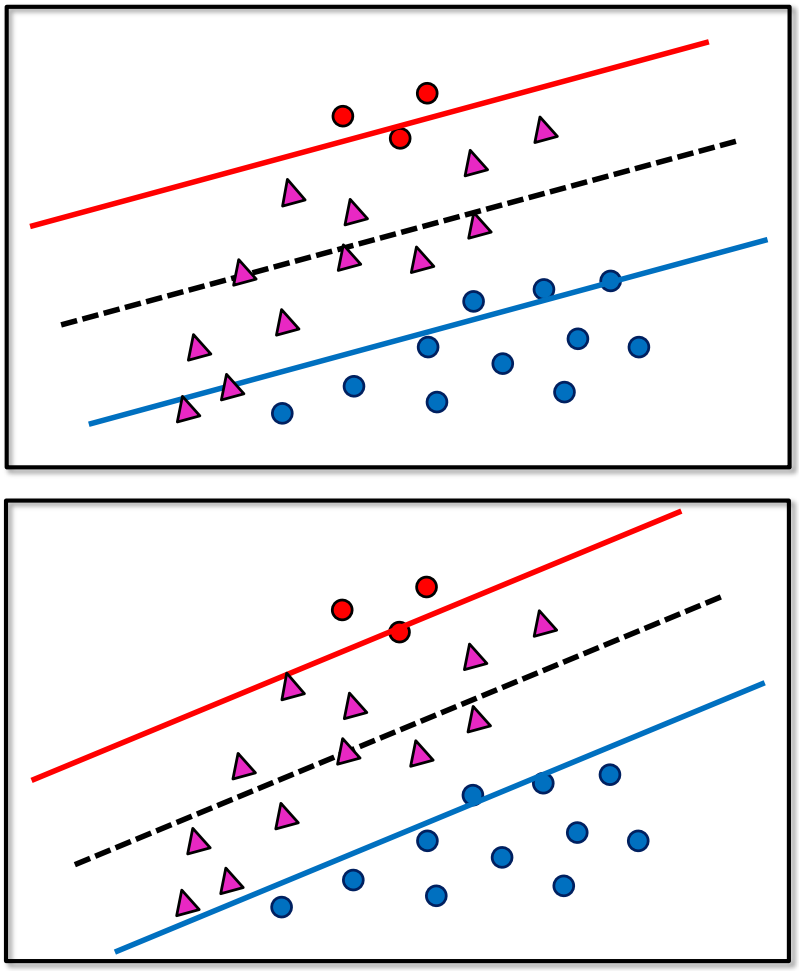


Figure 2.5: The influence of related samples for subcategory classifier. Red and blue circles represent labeled samples for positive and negative class, respectively. Pink triangles represent related samples. The upper figure shows the decision boundary (black dashed line) obtained only based on labeled data. The resulting subcategory classifier is not optimal for category level classification as some related samples will be classified as negative samples with high confidence. The lower figure shows the decision boundary (black dashed line) based on the proposed related sample augmented approach, which is more suitable for category level classification as no strong assertion is made about the labels of related samples. For better viewing, please see original colour pdf file.

"Universum" examples twice with opposite labels and obtain an SVM like formulation, which can be easily extended to the kernel form and solved with a standard SVM optimizer [16].

Thus, we treat the mined samples, related samples and samples from other categories as the positive samples, "Universum" set and negative samples, respectively

for the subcategory classification problem. As shown in Figure 2.5, the related samples will tune the classifier to better distinguish between the negative samples and the positive + related samples. The classifier obtained only based on labeled samples (upper figure) classifies the related samples as negative samples with hight confidence. On the contrary, besides correctly classifying the labeled samples with high confidence, the proposed related sample augmented approach (lower figure) will not make a strong assertion about the labels of related samples, which is beneficial for the final category level classification.

2.6 Experiments

In the following experiments, we first show our ambiguity guided subcategory mining results for the bus and chair categories in Section 2.6.1. We then extensively compare different subcategory mining methods and subcategory classifier training strategies using VOC 2007 "trainval/test" datasets (i.e. "trainval" set for training and "test" set for test) for proof of concept and ease of parameter tuning in Section 2.6.2 and 2.6.3. Finally, we evaluate the optimal configuration of our method on 2010 "trainval/test" datasets and compare with the state-of-the-art performance ever reported in Section 2.6.4.

2.6.1 Ambiguity Guided Subcategory Mining Results

It has been shown that models trained by "clean" subsets of images usually perform better than trained with all images [121]. The importance of "clean" training data suggests that it is critical to cluster training data into "clean" subsets and remove outliers simultaneously. Figure 2.6 displays our subcategory mining results for bus and chair categories. Each row on the left side shows one discovered subcategory while right side images are detected as outliers and left ungrouped.

For the bus category, the first 3 subcategories correspond to 3 different views of buses. This is mainly due to the discriminative pair-wise shape similarity for different views of buses, as the Exemplar-SVM works well for the categories with

Figure 2.6: Visualization of our ambiguity guided subcategory mining results for bus and chair category on VOC 2007. Each row on the left shows one mined subcategory. Images on the right are detected as outliers.

common rigid shapes. We note the shape and appearance of the last subcategory show much larger diversity than other subcategories. Though these images are not very similar to each other, the strong ambiguity with the person category still guides them to form a separate subcategory.

For chairs, there are no common rigid shapes as buses and the shapes of various chairs are very diverse, which leads to much noisier pair-wise shape similarity. Hence the subcategory mining results should be the combination effects of both appearance similarity and shape similarity, which can be observed from the discovered subcategories. Some subcategories may not have common shapes, but have similar local patterns. For example, chairs of the 2nd subcategory all have the stripe-like patterns. We note again the last detected subcategory looks like sofas. Besides being different from other chair subcategories, the ambiguity with sofa is also one of the main reasons that these images form a separate subcategory.

Table 2.1: Classification results (AP in %) comparison for different subcategory mining approaches on VOC 2007. For each category, the winner is shown in **bold** font.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
FV [49]	75.7	64.8	52.8	70.6	30.0	64.1	77.5	55.5	55.6	41.8	56.3	41.7	76.3	64.4	82.7	28.3	39.7	56.6	79.7	51.5	58.3
FVGHM [20]	76.7	74.7	53.8	72.1	40.4	71.7	83.6	66.5	52.5	57.5	62.8	51.1	81.4	71.5	86.5	36.4	55.3	60.6	80.6	57.8	64.7
FVGHM-CTX	78.5	80.0	54.9	71.9	55.4	75.1	87.1	67.2	58.4	60.3	60.0	47.3	83.0	76.3	90.5	44.9	59.6	63.2	83.5	68.9	68.3
FVGHM-CTX-spectral	81.2	82.1	56.7	73.5	56.2	76.5	88.5	67.8	58.0	60.1	61.7	48.1	85.1	77.8	90.7	45.5	60.6	64.4	84.3	69.2	69.4
FVGHM-CTX-GS	81.8	82.3	**58.5**	74.1	**56.5**	77.2	88.7	68.4	59.4	61.5	63.0	49.8	84.9	80.0	**91.3**	47.7	61.3	65.9	85.7	70.8	70.4
FVGHM-CTX-AGS	**82.2**	**83.0**	58.4	**76.1**	56.4	**77.5**	**88.8**	**69.1**	**62.2**	**61.8**	**64.2**	**51.3**	**85.4**	**80.2**	91.1	**48.1**	**61.7**	**67.7**	**86.3**	**70.9**	**71.1**

Table 2.2: Detection results (AP in %) comparison for different subcategory mining approaches on VOC 2007.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
E-SVM [76]	20.8	48.0	7.7	14.3	13.1	39.7	41.1	5.2	11.6	18.6	11.1	3.1	44.7	39.4	16.9	11.2	22.6	17.0	36.9	30.0	22.7
MC [53]	33.4	37.0	**15.0**	15.0	22.6	43.1	49.3	**32.8**	11.5	**35.8**	17.8	**16.3**	43.6	38.2	29.8	11.6	**33.3**	23.5	30.2	39.6	29.0
DPM [43]	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.3
DPM-spectral	32.9	60.3	9.6	15.9	29.2	52.6	58.1	21.6	21.1	24.6	26.1	10.8	58.2	48.1	37.6	11.9	21.5	35.3	48.6	43.1	33.4
DPM-GS	34.3	60.7	11.4	17.5	29.9	53.0	**58.9**	23.7	22.9	25.8	30.3	12.6	60.8	**49.2**	**42.6**	**13.3**	22.9	37.0	50.2	45.4	35.1
DPM-AGS	**34.7**	**61.4**	11.5	**18.6**	**30.0**	**53.8**	58.8	24.7	**24.7**	26.8	**31.4**	13.8	**61.4**	**49.2**	42.2	12.9	23.9	**38.5**	**50.8**	**45.5**	**35.7**

2.6.2 Subcategory Mining Method Comparison

We extensively evaluate the effectiveness of different subcategory mining approaches on the VOC 2007 dataset, as the ground-truth of its testing set is released. To allow direct comparison with other popular works [49, 18, 20], we only implement a simplified SAOC framework. More specifically, we choose the state-of-the-art FVGHM method [20] as the classification pipeline (dense SIFT feature [75] with FK coding [49] plus GHM pooling [67, 20]) and the customized DPM [43] as object detector. The only difference between customized DPM and the standard DPM is the model initialization step. Unlike standard DPM, which utilize the aspect ratio to cluster training samples into different groups, DPM-spectral, DPM-GS and DPM-AGS replace the aspect ratio based initialization with spectral clustering, graph shift and ambiguity guided graph shift mining results, respectively. For the standard DPM, we use the publicly available implementation with the default settings (8

parts) [43]. As detection assisted classification has become a standard approach for classification on PASCAL VOC. We augment FVGHM with detection context information as in [86] and utilize the resulting FVGHM-CTX as the starting point to evaluate different subcategory mining methods. Dense SIFT is extracted using multiple scales setting (spatial bins are set as 4, 6, 8, 10) with step 4. The size of Gaussian Mixture Model in FK is set to 256. For GHM [20], we construct the hierarchical structure with three-level clusters, each of which includes 1, 2, 4 nodes respectively. One-vs-All SVM is learnt for each category/subcategory. For our graph shift based approach, the subcategory number is determined by the expansion size (the number of selected nearest neighbors for the expansion stage [72]). In experiments the expansion size is decided by cross-validation, and the subcategory number is generally from 2 to 5. For fair comparison, We did not compare with the non-graph based approaches, such as k-means and [59], as they are difficult to directly utilize our pair-wise shape similarity. Spectral clustering, the representative graph based partition method, is chosen for comparison. We extensively evaluate spectral clustering with the cluster number from 2 to 5 and report the best results.

The detailed classification results are shown in Table 2.1. It can be concluded from the table that:

- Subcategory awareness does improve the performance of current detection assisted classification framework. Subcategory information provides an effective approach to decompose the original difficult problem into several easier sub-problems. Such simplified sub-problem can be better captured by current classification methods, which then improves the overall performance. Even with the naive spectral clustering for category mining, we can still boost the state-of-the-art classification performance;

- Our ambiguity guided graph shift approach is effective for subcategory mining. The resulting subcategories can obviously improve the classification performance; By adaptively grouping the samples into subcategories and rejecting the outliers, our ambiguity guided graph shift approach performs much better

than the spectral clustering.

- Ambiguity is informative for subcategories mining. The sample ambiguity implicitly provides information about other categories and enables the algorithm to focus on the samples near the decision boundary, which are more important to the classification problem. With the assistance of sample ambiguity, the graph shift algorithm can obtain better results for 17 out of 20 categories.

Figure 2.7 shows some exemplar results for the baseline method (FVGHM-CTX) and the proposed algorithm (FVGHM-CTX-AGS) from the VOC "test" set. It can be observed that the monolithic model (FVGHM-CTX) fails to recognize many samples due to the variance of pose, view point and appearance. On the contract, such samples can be successfully recognized by some subcategory classifiers. The less diversities in each subcategory will make the corresponding classifier more reliable and accurate. The final subcategory-aware classifier, which fuses the responses from all subcategory classifiers, can successfully recognize more samples than the baseline method.

As object detection is an inseparable component of our SAOC framework, we also show the intermediate detection results in Table 2.2. Besides standard DPM, we add two more baselines, which also use the multiple components/models for object detection [53, 76]. When compared with other leading techniques in subcategory based detection, our method obtains the best results for most categories, achieving superior performance on categories with rigid shape or high ambiguity. We note the MC [53], which requires manually labelling the pose of each image, performs quite well on articulated categories. The inferior performance of our ambiguity guided mining framework on articulated categories is mainly due to the limited discriminative ability of current similarity metric.

The number of subcategories: We have proposed the ambiguity guided graph shift for subcategory mining and verified its effectiveness. Here we evaluate the influence of the number of subcategories. Particularly, we select the bus, chair and horse category as representative.

Figure 2.7: Exemplar results for the baseline method (FVGHM-CTX) and FVGHM-CTX-AGS from the VOC 2007 "test" set. The classification results are compared by the confidence scores for each classifier. The blue and green bars represent the baseline classifier and the subcategory classifiers, respectively. The subcategory-aware classifier, which fuses the scores of all subcategory classifiers to obtain the final score, is represented by the red bar. For better viewing, please see original colour pdf file.

From Figure 2.8, the optimal number of subcategories depends on the characteristics of the specific category. We can summarize the observations for the different categories as follows:

- For small number of subcategories (K) the performance gradually increases with increasing K, but stabilizes around $K = 4$. As there are large variation for samples in each category due to pose, viewpoint and appearance variance,

Table 2.3: Classification results (AP in %) comparison for different subcategory classifier training strategies on VOC 2007. For each category, the winner is shown in **bold** font.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
FVGHM-CTX-ASM	82.2	83.0	**58.4**	76.1	56.4	77.5	**88.8**	69.1	62.2	61.8	64.2	51.3	85.4	80.2	**91.1**	48.1	61.7	67.7	86.3	70.9	71.1
FVGHM-CTX-ASM-2	76.9	79.0	55.9	72.7	53.6	76.4	88.4	67.8	60.6	59.9	61.3	50.4	79.9	74.5	87.8	44.8	59.2	62.2	83.7	69.7	68.2
FVGHM-CTX-ASM-RS	**82.6**	**85.3**	58.2	**78.5**	**57.7**	**79.2**	88.6	**70.4**	**63.8**	**64.1**	**65.4**	**53.7**	**86.1**	**80.6**	90.8	**48.9**	**63.4**	**69.7**	**87.8**	**71.6**	**72.3**

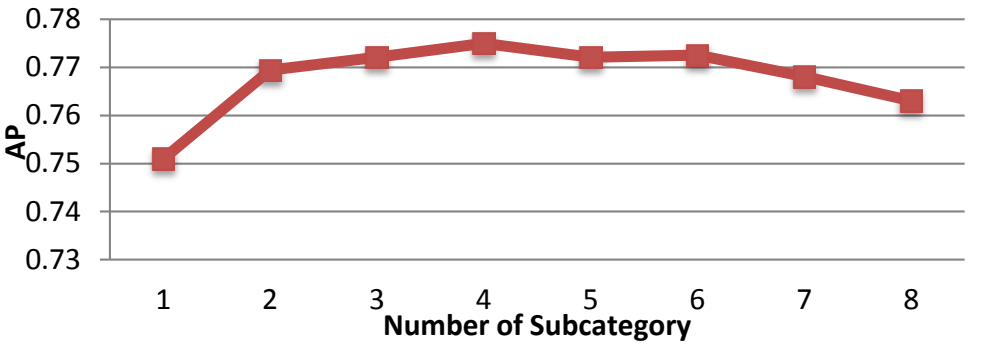
properly dividing them into subcategory will lead to easier sub-problems and thus improve the overall performance.

- Further increasing K may decrease the performance. One of the reasons for such decrease is the lack of data. Larger K will lead to fewer samples in each subcategory. Such small number of samples may be insufficient for training a reliable subcategory model and hurt the overall performance.

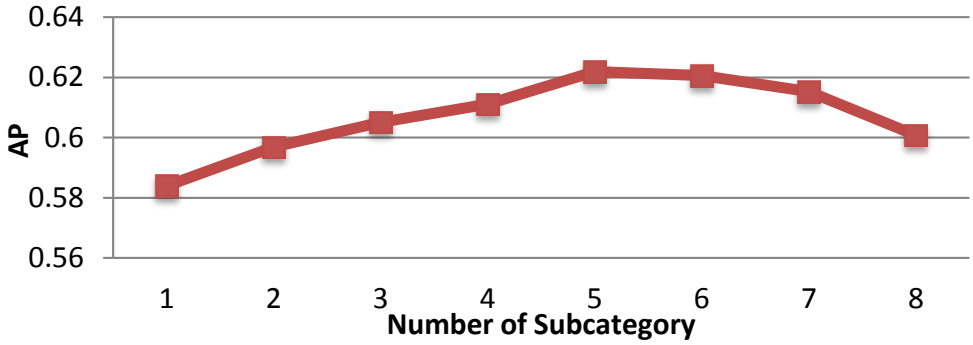
As the running time increases with K and $K = 5$ is large enough to get the optimal performance for most categories, we select the best K from 2 to 5 for the balance of accuracy and speed.

2.6.3 Subcategory Classifier Training Strategy Comparison

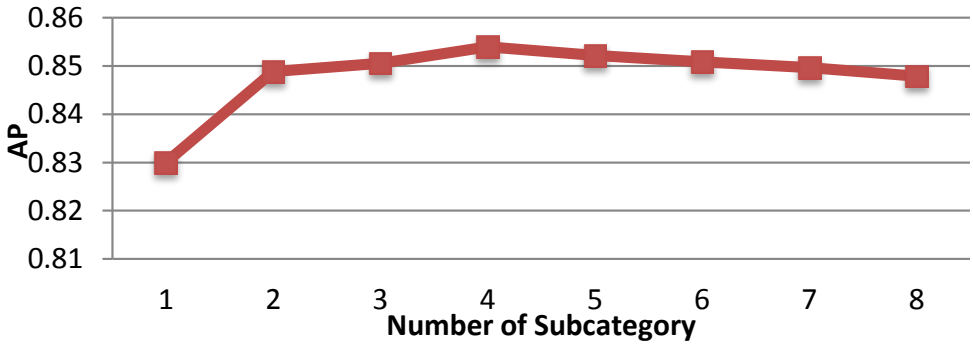
In this subsection, we evaluate different strategies for training subcategory classifiers. We compare the related samples augmented approach described in Section 2.5 with two baseline strategies. For the target subcategory, the first strategy assigns the samples in this subcategory as positive samples and the samples belonging to the other categories as the negative samples. This is the approach used in Subsection 2.6.2. The second strategy assigns the samples in this subcategory as positive samples and all other samples as the negative samples. The difference between two baseline strategies lies in how to handle the related samples. The first strategy simply abandons them while the other one assigns them as the negative samples. We use the same experiment setting as in Subsection 2.6.2 and the experimental results are shown in Table 2.3. The penalty parameters $C_{\mathcal{D}}$ and $C_{\mathcal{U}}$ in Eqn. 2.4 are decided



(a) bus



(b) chair



(c) horse

Figure 2.8: Variation in classification accuracy as a function of number of subcategories for three distinct categories on VOC 2007 dataset. The A.P. gradually increases with increasing number of subcategories and stabilizes beyond a point.

by cross-validation. From the Table 2.3, it can be observed that:

- The baseline strategy 2 (FVGHM-CTX-AGS-2: assign samples for the target subcategory as the positive samples and all other images as negative samples) leads to the worst results. This is intuitive as our concern is category level classification. However, because the samples from the same category are usu-

Table 2.4: Classification results from the proposed framework with comparison to other leading methods on VOC 2010.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
NLPR [37]	90.3	77.0	65.3	75.0	53.7	85.9	80.4	74.6	62.9	66.2	54.1	66.8	76.1	81.7	89.9	41.6	66.3	57.0	85.0	74.3	71.2
NEC [37]	93.3	72.9	69.9	77.2	47.9	85.6	79.7	79.4	61.7	56.6	61.1	71.1	76.7	79.3	86.8	38.1	63.9	55.8	87.5	72.9	70.9
ContextSVM [86]	93.1	78.9	73.2	77.1	54.3	85.3	80.7	78.9	64.5	68.4	64.1	70.3	81.3	83.9	91.5	48.9	72.6	58.2	87.8	76.6	74.5
GHM ObjHierarchy [20]	94.3	81.3	77.2	80.3	56.3	87.3	83.8	82.2	65.8	73.7	67.0	75.9	82.3	86.5	92.0	51.7	75.1	63.3	89.9	77.3	77.2
FVGHM-CTX-AGS	95.9	83.2	**79.0**	84.0	**57.5**	91.4	84.3	83.4	70.2	75.1	68.9	78.2	85.4	88.4	**92.8**	52.4	**78.5**	67.8	93.0	77.4	79.3
FVGHM-CTX-AGS-RS	**96.4**	**84.8**	78.3	**85.5**	57.0	**91.7**	**85.6**	**85.1**	**72.7**	**77.2**	**70.6**	**80.1**	**86.4**	**89.4**	92.0	**54.2**	78.0	**70.5**	**93.4**	**79.4**	**80.4**

ally more similar, this strategy will make subcategory classifier focus on the boundary between the target subcategory and the other subcategories of the same category instead of the boundary between the target subcategory and other categories. Hence, the final subcategory classier is not discriminative for the category level classification.

- Unlike the baseline strategy 1 (FVGHM-CTX-ASM), which abandons the informative related samples, the proposed related samples augmented approach (FVGHM-CTX-ASM-RS) effectively utilize them under the "Universum" SVM framework. These related samples are effectively exploited to tune the classifier for category level classification, especially for the subcategory with small number of samples, which leads to the best performance.

2.6.4 Comparison with the State-of-the-arts

In this section we compare the performance of the proposed SAOC framework with the reported state-of-the-art results on the VOC 2010 dataset. To obtain the state-of-the-art performance, we conduct the experiments with more complicated setting. For classification, we extract dense SIFT, HOG, color moment and LBP features in a multi-scale setting. All these features are encoded with VQ, LLC and FK [18] and then pooled by GHM. The pooling results are concatenated to form the final image representation. During SVM training, χ^2 and linear kernel is employed for VQ/LLC and FK, respectively. For object detection, we train one shape-based detector and

one appearance-based object detector for each object category. The augmented DPM [118, 86] employing both the HOG and LBP features is adopted as the shape-based model. For appearance-based approach [97, 96], we sample 4000 sub-windows of different sizes and scales, and perform the BoW based object detector on these sub-windows. The number of subcategories is also determined by cross-validation as mentioned above.

We compare with the best known VOC 2010 performance from several recent papers and the released results from the VOC 2010 challenge [37], which are all obtained through the combinations of multiple methods in order to obtain better performance. The comparison results are presented in Table 2.4, from which it can be observed that:

- Our proposed method outperforms the competing methods on all 20 object categories. We note that all the leading classification methods combine object classification and object detection to achieve higher accuracy. However, most of the previous methods simply fuse the outputs of a monolithic classification model and a monolithic detection at category level. This limitation prevents them from grasping the informative subcategory structure and the interaction among the subcategories. By properly employing the subcategory structure, we can improve the state-of-the-art performance by 2.1%.

- Related samples are informative for the category level classification. The proposed related samples enhanced approach can further boost the overall performance by 1.1%.

- Note that our methods can significantly improve the performance of rigid categories (bus, train) and ambiguous categories (sofa, chair). For the rigid categories, the proposed subcategory mining approach is able to split the data effectively, which leads to "clean" subcategories and boosts the performance. For ambiguous categories, our model can implicitly re-rank the results. The scores for subcategories without ambiguity are raised and scores for ambiguous subcategories are depressed (still larger than samples not belonging to the

corresponding category), which will also improve the AP.

When measured with object detection, we can achieve the performance of 37.1% compared to the state-of-the-art results of 36.8 % [37], which is obtained by much more complicated detection models than ours. As our framework focuses on classification, detailed detection results are omitted due to the space limitation.

2.7 Chapter Summary

In this chapter, we proposed an ambiguity guided subcategory mining and subcategory-aware object classification framework for object classification. We modeled the subcategory mining as a dense subgraph seeking problem. This general scheme allows us to gracefully embed intra-class similarity and inter-class ambiguity into a unified framework. The subcategories, which correspond to the dense subgraphs, can be effectively detected by the graph shift algorithm. Ambiguity guided subcategory mining results are then seamlessly integrated into the subcategory-aware detection assisted object classification framework. The usage of "relate samples" allows us to effectively tailor the subcategory classifiers for category level classification. Extensive experimental results on both PASCAL VOC 2007 and VOC2010 clearly demonstrated the proposed framework achieved the state-of-the-art performance.

Chapter 3

Towards Unified Object Detection and Semantic Segmentation

In this chapter, we show how to jointly solve object detection and semantic segmentation in a unified framework. By enforcing the consistency between final detection and segmentation results, our unified framework can effectively leverage the advantages of leading techniques for both tasks to improve the overall performance.

3.1 Introduction

Object detection and semantic segmentation are two core tasks of visual recognition [27, 44, 95, 7, 115, 13, 90, 108, 103, 106]. Object detection is often formulated as predicting a bounding box enclosing the object of interest [44] while semantic segmentation usually aims to assign a category label to each pixel from a pre-defined set [13]. Though strongly correlated, these two tasks have typically been approached as separate problems and handled using substantially different techniques.

Template based detection using sliding window scanning (*e.g.* HoG [27] and DPM [44]) has long been the dominant approach for object detection. Though good at finding the rough object positions, this approach usually fails to accurately lo-

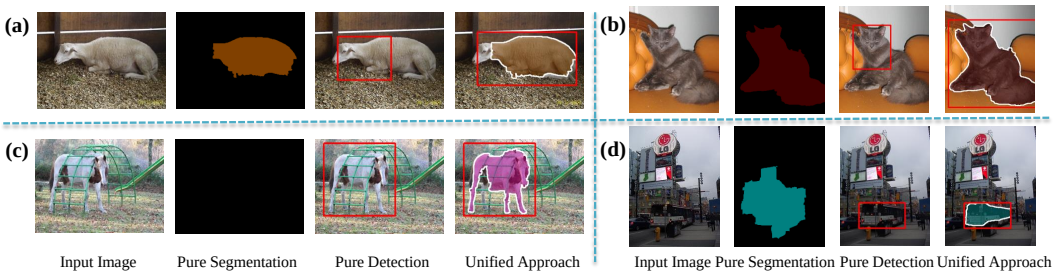


Figure 3.1: The inconsistency of failure cases for object detection and semantic segmentation. The images in the top row show the scenario where detection is imperfect due to pose variance while the semantic segmentation works fine. The images in the bottom row show the scenario where semantic segmentation is not accurate while detectors can easily locate the objects. Thus, the two tasks are able to benefit each other, and more satisfactory results can be achieved for both tasks using our unified framework.

calize the whole object via a tight bounding box. In fact, it has been found that the largest source of detection error is inaccurate bounding box localization ($0.1 \leq$ overlap < 0.5) [26, 60]. This may arise from the limited representation ability of template-based detectors for non-rigid objects. For example, the deformable part-based model (DPM) [44] detector works much better for localizing rigid cat heads than for more amorphous cat bodies [80]. As shown in Figure 3.1 (a) and (b), the DPM detector often locates the head region only, which leads to the localization error. On the other hand, owing to their homogeneous appearances, the whole objects (cat and sheep) can be easily segmented out by the leading semantic segmentation techniques [13]. If poor localizations can be corrected with the help of semantic segmentation techniques [13], the overall detection performance would be improved considerably from additional true positives and fewer false positives.

Hypotheses based semantic segmentation has achieved great success during the past few years, which works by directly generating a pool of segment hypotheses for further ranking [4, 13]. However, due to the lack of global shape models, these approaches may fail to recognize the hypotheses of objects with heterogeneous appearances in the cluttered background, especially when all the generated hypotheses have some artifacts. As shown in Figure 3.1 (c) and (d), the leading hypotheses based semantic segmentation approach [13] either fails to segment out the object

of interest or selects a much larger segment hypothesis. In contrast, if the target object has strong shape cues, the template-based detector [44] can easily locate the object and thus provide valuable information for semantic segmentation. Recently, a line of works, called detection-based segmentation, explored directly utilizing the detection results as top-down guidance and then performing segmentation within the given bounding boxes [11, 107]. However, such approaches usually have to make a hard decision about detection results at the early stage. Hence the error for detection, especially the localization error, will propagate to the segmentation results and could not be rectified. Intuitively it is beneficial to postpone making a hard decision till the last step of the pipeline [110].

Based on the above observations, we argue that object detection and semantic segmentation should be addressed jointly. Object detections should be consistent with some underlying segments to integrate local cues for better localization as shown in Figure 3.1 (a) and (b). Similarly, hypotheses based semantic segmentation should benefit from template-based object detectors to select better segment hypotheses as shown in Figure 3.1 (c) and (d). To this end, we propose a principled framework to unify current leading object detection and semantic segmentation techniques. By enforcing the consistency, our unified approach can benefit from the advantages of both techniques. In addition, some ambiguous object hypotheses may be difficult to classify from the information within the window/segment alone, but contextual information, such as local context around each object hypothesis and global image-level context, can help [69, 86, 22]. Hence, we further integrate contextual modeling into our framework. The major contributions of this chapter can be summarized as follows:

- We propose a principled framework for joint object detection and semantic segmentation. By enforcing the consistency between detection and segmentation results, our unified framework can effectively leverage the advantages of both techniques. Furthermore, both local and global context information are integrated into our unified framework to distinguish the ambiguous examples.

- With our unified framework, all information is accumulated at the final stage of the pipeline for decision making. Hence, it is avoided to make any hard decision at the early stage. The relative importance of different components is automatically learned for each category to guarantee the overall performance.

- Extensive experiments are conducted for both object detection and semantic segmentation tasks on the PASCAL VOC [39] datasets. The state-of-the-art performance of the proposed framework verifies its effectiveness, showing that performing object detection and semantic segmentation jointly is beneficial for both tasks.

3.2 Related Work

Recently, by noticing the limitation and complementarity of techniques for both tasks, some researchers have begun to investigate their correlations [64, 4, 12, 112]. The early work [64] simply employs the masks from detectors to initialize graph-cuts based segmentations. In [68, 107], more sophisticated models are proposed to refine the region within ground-truth bounding boxes. Rather than focusing on entire objects, Brox *et al.* employed Poselet detectors to predict masks for object parts [12]. Arbeláez *et al.* aggregated top-down information from detectors as activation features for bottom-up segments [4]. Conversely, segmentation techniques have also been explored to assist object detection in different ways. Dai *et al.* utilized segments extracted for each object detection hypothesis for better localization [26]. Fidler *et al.* [48] proposed to improve object detection based on semantic segmentation results [13]. The segments and detection windows are associated with several manually designed geometry features. Unfortunately, nearly all the above approaches utilize a sequential manner to fuse detection and segmentation techniques. Hence, the overall performance heavily relies on the correctness of the initial results as the errors in the early stage are difficult to rectify.

Probably the most similar approach to ours is [66], which also aims to perform joint object detection and semantic segmentation. Our framework is different in the

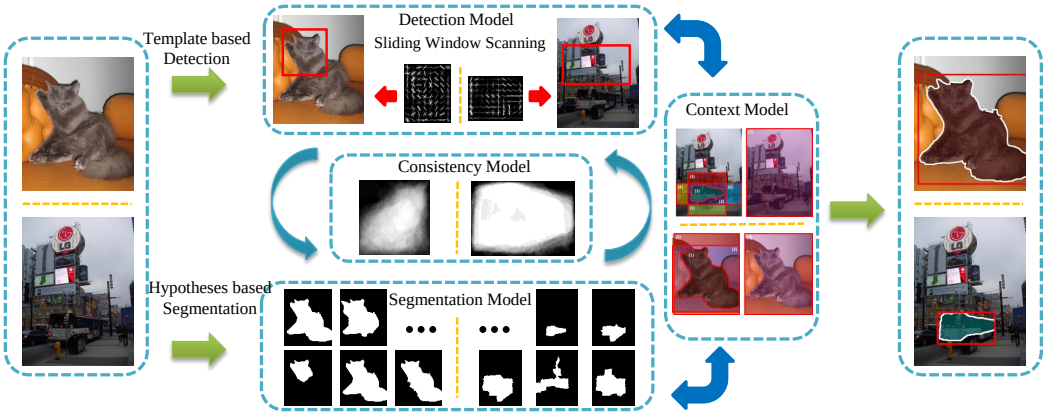


Figure 3.2: Overview of the proposed unified object detection and semantic segmentation framework. Give a testing image, our UDS framework performs template based detection using sliding window scanning and hypotheses based semantic segmentation jointly. The agreement of the predictions from these two approaches is ensured by the consistency model. Both local context around the object hypothesis and global image context are also seamlessly integrated into our framework. The final output is the bounding box position and the index of the selected segment hypothesis.

sense that we avoid making any hard decision at the early stage. All the information is aggregated at the final stage of the pipeline for decision making. On the contrary, [66] has to make initial decision about detection results. Hence, the initial detection errors, such as localization error, are difficult to rectify. Furthermore, unlike the CRF based model used in [66], we employ a hypotheses based approach for semantic segmentation. Hence, it is easier to ensure the shape consistency of top-down and bottom-up information in our framework.

3.3 Unified Object Detection and Semantic Segmentation

In this section, we introduce the details of the proposed unified object detection and semantic segmentation (UDS) framework. We start with an overview of the system and then detail each key component.

Figure 3.2 illustrates the pipeline of the proposed UDS framework. For the segmentation component, we employ the hypotheses based approach. Thus, with a

pool of generated segment hypotheses, the segmentation problem is converted into choosing the appropriate hypothesis. Given a testing image, we perform template based detection using sliding window scanning and hypotheses based semantic segmentation jointly. Successful detection and segmentation require the agreement of both detection and segmentation predictions, which is achieved by utilizing a consistency model. In addition, as context plays an important role in distinguishing ambiguous object hypotheses, we further design a context model to aggregate both local (around the target object) and global (image-level) context information. For different object categories, each of these four components may have a different level of importance, which is automatically decided during the learning process. The final output of our system is the bounding box position (p_0) and the selected segment index (id) for the target object.

Formally, the joint detection and segmentation is achieved via the maximization of the following score function:

$$\begin{aligned} S(I, z, id) = \lambda^{Dt} S^{Dt}(z|w^{Dt}, I) + \lambda^{Sg} S^{Sg}(id|w^{Sg}, I) \\ +\lambda^{Ct} S^{Ct}(z, id|w^{Ct}, I) + S^{Cs}(z, id|w^{Cs}), \end{aligned} \tag{3.1}$$

where w^{Dt}, w^{Sg}, w^{Ct} and w^{Cs} are the parameters for detection, segmentation, context and consistency component, respectively. λ^{Dt}, λ^{Sg}, λ^{Ct} are scalar weights for the corresponding components. z captures the information for the template based detector and id denotes the index of the selected segment. The details of each component are introduced in the following subsections. Based on the proposed unified approach, we avoid making any hard decision at the early stage. The final decision is delayed to the last step of the pipeline with all the integrated information, which implicitly relies on the learning mechanism to assess the relative importance of different components for each object category to guarantee the overall performance.

Finally, we want to emphasize that the proposed UDS framework provides a principled way to unify detection and segmentation techniques. We can directly employ the existing techniques or design new approaches for each component. Hence, it is

easy to tailor UDS for specific applications, such as simultaneous person detection and segmentation. In this chapter, we will focus on utilizing the UDS framework for general object detection and semantic segmentation to verify its effectiveness.

3.3.1 Template based Detection Component

For the detection component, we aim to utilize the template based approach [44, 28], as it is good at capturing the shape cue and thus complementary to the appearance based segmentation techniques [13, 110]. In addition, through the mixture model strategy [44], these approaches can easily encode sub-category level top-down information (subcategory specific soft shape mask in this work). In this chapter, we utilize the state-of-the-art deformable part-based model (DPM) [44]. Following [44], we define $z = \{c, p\}$, where $p = \{p_i\}_{i=0,\cdots,m}$. Here, c denotes the mixture component index. p_0 encodes the location and scale of the root bounding box in an image pyramid and $\{p_i\}_{i=1,\cdots,m}$ encodes the m part bounding boxes at the double resolution of the root. By concatenating the parameters for all mixtures as in [44], the score of a configuration can be written as

$$S^{Dt}(p, c|w^{Dt}, I) = \sum_{i=0}^{m} w_i^{Dt} \cdot \phi^{Dt}(I, p_i, c) + \sum_{i=1}^{m} w_{i,def}^{Dt} \cdot \phi^{Dt}(p_0, p_i, c), \tag{3.2}$$

where $\phi^{Dt}(I, p_i, c)$ and $\phi^{Dt}(p_0, p_i, c)$ are the HoG pyramid features and spring deformation features, respectively, as in [44]. As Eqn. (3.2) is linear in model parameters, it can be written compactly as:

$$S^{Dt}(p, c|w^{Dt}, I) = w^{Dt} \cdot \phi^{Dt}(I, p, c). \tag{3.3}$$

3.3.2 Hypotheses based Segmentation Component

Hypotheses based semantic segmentation has achieved great success during the past few years [14, 13, 110]. This line of approaches mainly consist of two stages. The first stage generates a pool of segment hypotheses. The second stage ranks the generated hypotheses based on category-dependent information. The top ranked

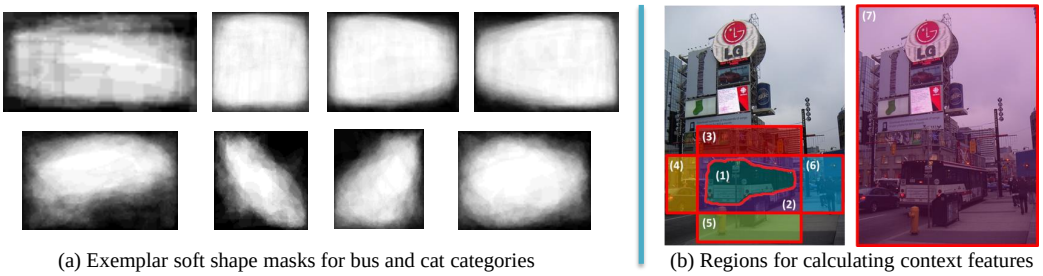
segments are returned as the final solution. Many efforts have been devoted to hypotheses generation through either a pure bottom-up approach [14, 95, 4] or a CRF based approach [110]. For the second stage, most approaches [14, 95, 110] simply employ the appearance based classification/regression for ranking. However, due to the limited discriminative ability of the appearance based ranking function, there exists a large gap between upper-bound accuracy of generated hypotheses (larger than 80%) and predicted accuracy of selected hypotheses (less than 50%) [14, 110]. As shown in Figure 3.1, due to the lack of global shape models, semantic segmentation relying on pure appearance based ranking may fail to find the appropriate hypotheses.

Based on the above observation, it may be expected that considerable improvement over the current segmentation performance can be achieved by means of simply selecting better hypothesis without generating more hypotheses. Hence, in this work we use standard methods for hypotheses generation and focus on selecting better segment hypotheses. To allow direct comparison, we utilize the publicly available code of the second order pooling (O_2P) approach [13] for hypotheses generation. For the feature representation $\phi^{Sg}(I, id)$ of the selected hypothesis id, a naive strategy is directly employing the second order pooling features as in [13]. However, training a latent model with high dimension features may be intractable. Hence, rather than keeping $\phi^{Sg}(I, id)$ as a high dimensional vector of raw second order pooling features, we represent $\phi^{Sg}(I, id)$ as the scores of pre-trained support vector regressors (SVR) [13]. Then, the score function of the segmentation component can be written as:

$$S^{Sg}(id|w^{Sg}, I) = w^{Sg} \cdot \phi^{Sg}(I, id). \tag{3.4}$$

3.3.3 Consistency Component

The consistency component mainly aims to enforce the consistency between detection and segmentation prediction and thus leverage the advantages of both approaches. Soft shape mask has demonstrated to be effective for many detection



(a) Exemplar soft shape masks for bus and cat categories

(b) Regions for calculating context features

Figure 3.3: (a) Examples of subcategory-specific soft shape masks for buses (top row) and cats (bottom row). (b) Illustration of regions defined for computing the context features. Based on the selected segment hypothesis and bounding box, we adaptively divide the image into 7 regions as described in Section 3.3.4.

guided techniques [112, 4, 15]. Hence, in this work, we measure the consistency between results of detection and segmentation approaches by calculating the correlations between their masks as shown below:

$$S^{Cs}(z, id|w^{Cs}) = \sum_{i=0}^{m} w_i^{Cs} \cdot m(p_i, id, c) = w^{Cs} \cdot \phi^{Cs}(p, id, c), \tag{3.5}$$

where $m(p_i, id, c)$ is the binary map $\{1, -1\}$ clipped from the segmentation hypothesis id by the localized bounding box p_i. Here, c in $m(p_i, id, c)$ is only used for padding 0 to make the equation with mixture models more compact, which is a common trick for the DPM approach [44].

Intuitively, the learned soft mask w^{Cs} from top-down detection techniques can be seen as a shape guidance for bottom-up segmentation techniques. Enforcing the correlation between masks from both approaches will guarantee the consistency of top-down and bottom-up information. In addition, the mixture model strategy is critical to cope with variance in the poses as well as the view points. To ensure obtaining a reliable shape mask for each mixture component, we employ a shape guided mixture initialization as introduced in Section 3.4.2. Some examples of such soft shape masks are visualized in Figure 3.3 (a).

3.3.4 Context Component

Both the local context around the target object [69] and the global image context [86, 4, 22] have shown to be effective for visual recognition. The local context directly models the interaction of the target object and the surrounding environment. For example, a horse is often occluded by a person riding on it. In contrast, the global context mainly captures the image level information and coexistence/exclusion relation between objects.

In order to leverage such informative context cues, we further enhance the framework with an adaptive context model. Specifically, given a bounding box p_0 and a segment id, we divide the image into 7 regions (segment region, surrounding region within p_0, 4 context boxes and the whole image) as shown in Figure 3.3 (b). The area of the context box is half of that of the bounding box p_0. Hence, the spatial extent of the local context will vary adaptively based on p_0. If a context box crosses the boundary of the image, we consider only the area within the image. Fisher Vector (FV) [49, 18] is employed as region feature representation, as it has demonstrated the state-of-the-art performance for both object classification and detection [20, 22]. Furthermore, the average pooling strategy for FV enables effective calculation by utilizing the integral graph. Thus, the raw context representation is the concatenation of FVs on the 7 regions mentioned above.

Similar to the segmentation component, the dimension of the raw context features is too high. Hence, we first train a separate classifier for each object category and then use the predicted scores as the final context features. Then, the context component can be written as:

$$S^{Ct}(z, id|w^{Ct}, I) = S^{Ct}(p_0, id|w^{Ct}, I) = w^{Ct} \cdot \phi^{Ct}(I, id, p_0), \tag{3.6}$$

where $\phi^{Ctx}(I, id, p_0)$ is the concatenation of predicted scores for all classifiers. In fact, our context model can be seen as a variant of the appearance based detection approach to some extent. We still call it "context model" as it can provide valuable and complementary context information to the other three components.

3.4 Inference and Learning

This section introduces inference and learning of the proposed UDS framework. We begin with the general inference and learning procedure and then describe the implementation details in practice.

3.4.1 Inference

Similar to DPM [44], we employ the sliding windows strategy for inference. For a fixed root bounding box position p_0 and mixture index c, inference in our model can be done by solving the following optimization problem:

$$\begin{aligned} S(p_0, c) = & \max_{p_1,\cdots,p_m,id} S(p, id, c) = \max_{id}[\lambda^{Dt} w_0^{Dt} \cdot \phi^{Dt}(I, p_0, c) \\ & + \lambda^{Sg} w^{Sg} \cdot \phi^{Sg}(I, id) + \lambda^{Ct} w^{Ct} \cdot \phi^{Ct}(I, id, p_0) + w_0^{Cs} \cdot m(p_0, id, c) \\ & + \max_{p_1,\cdots,p_m} \sum_{i=1}^{m} (\lambda^{Dt} w_i^{Dt} \cdot \phi^{Dt}(I, p_i, c) + \lambda^{Dt} w_{i,def}^{Dt} \cdot \phi^{Dt}(p_0, p_i, c) + w_i^{Cs} \cdot m(p_i, id, c))]. \end{aligned} \tag{3.7}$$

By defining

$$\begin{aligned} R_0(p_0, id, c) = & \lambda^{Dt} w_0^{Dt} \cdot \phi^{Dt}(I, p_0, c) + \lambda^{Sg} w^{Sg} \cdot \phi^{Sg}(I, id) \\ & + \lambda^{Ct} w^{Ct} \cdot \phi^{Ct}(I, id, p_0) + w_0^{Cs} \cdot m(p_0, id, c) \\ R_i(p_i, id, c) = & \lambda^{Dt} w_i^{Dt} \cdot \phi^{Dt}(I, p_i, c) + w_i^{Cs} \cdot m(p_i, id, c), \end{aligned}$$

the Eqn. (3.7) can be written compactly as:

$$S(p_0, c) = \max_{id}[R_0(p_0, id, c) + \max_{p_1,\cdots,p_m} \sum_{i=1}^{m} (R_i(p_i, id, c) + \lambda^{Dt} w_{i,def}^{Dt} \cdot \phi^{Dt}(p_0, p_i, c))]. \tag{3.8}$$

With fixed segment index id, this scoring function is similar to that of DPM and can thus be passed to an off-the-shelf DPM solver. Hence, the inference algorithm works as follows: First, we compute $R_0(p_0, id, c)$ for each root filter position p_0 and segment index id. Then, we prune the object hypotheses based on the score of R_0 without sacrificing the overall recall rate (validated on the validation set). For each

retained segment hypothesis, we further run the full model (3.7) locally with the dynamic programming approach as in [44]. Finally, we compute the maximum over the mixture components to obtain the final score of the object hypothesis.

3.4.2 Learning

By defining the output variable $y = \{p_0, id\}$ and latent variable $h = \{p_1, \cdots, p_m, c\}$, the scoring function (3.1) can be rewritten as

$$S(I, y, h) = w \cdot \Phi(I, y, h), \tag{3.9}$$

where w is the concatenation of all model parameters (w^{Dt}, w^{Sg}, w^{Ct} and w^{Cs}). $\Phi(I, y, h)$ is the concatenation of all four components features weighted by their weights (λ^{Dt}, λ^{Sg} and λ^{Ct}) with respect to the label y and latent variable h.

We note that Eqn. (3.9) is linear in the model parameter w, thus this model can be effectively learned based on the latent structure SVM framework [114, 51]:

$$\min_{w} \frac{1}{2}||w||^2 + C[\sum_{j=1}^{n} \max_{\hat{y},\hat{h}} (w \cdot \Phi(x_j, \hat{y}, \hat{h}) + \Delta(y_i, \hat{y}, \hat{h})) - \sum_{j=1}^{n} \max_{h} (w \cdot \Phi(x_i, y_i, h))], \tag{3.10}$$

where the loss function $\Delta(y_i, \hat{y}, \hat{h})$ is defined as the weighted sum of the Intersection over Union of the root filters and segment hypotheses (in current implementation, we simply use the average value of two IoUs).

The standard approach to solve the optimization problem (3.10) is the Concave-Convex Procedure (CCCP) [116, 114]. However, as the CCCP algorithm only guarantees to converge to a local minimum, we learn the model progressively to ensure a reasonable initialization. More specifically, we first train each component separately and jointly learn the overall model with Eqn (3.10).

For the object detection component, we follow the original training approach of DPM [44] except for the mixture initialization and part discovery. Aspect ratio based clustering is used in [44] for mixture initialization. However, such an approach may ignore the potential pose/view variance. Hence, we employ the idea of "sub-

category mining" [34, 3, 28] by utilizing the additional segmentation annotation to ensure a more reliable shape mask for each component. Specifically, we resize all the cropped segmentation masks to the same height and l_2 normalizes all the resized masks. Then, the similarity between two normalized masks a and b is defined as the maximal value of the convolution response map of a and b. Finally, the graph shift algorithm [72] is employed to discover the dense subgraphs, which correspond to the subcategories, as in [34]. The resulting subcategories are then used for mixture initialization. The original DPM approach [44] discovers the salient parts greedily by covering the high-energy region of the root HOG-template. Recently, [15] suggests that modifying this "saliency" measure by multiplying the HOG magnitude by the average segmentation mask for each component will lead to more semantic meaningful parts. Hence, we follow their approach by utilizing the modified 'saliency" measure for part discovery. For the consistency component, the pixel-wise mean of all segmentation masks for each component is utilized for initialization.

In the final joint learning stage, all model parameters (w^{Dt}, w^{Sg}, w^{Ct} and w^{Cs}) in Eqn. (3.10) are jointly optimized. Thus, the relative importance of each component will be automatically tuned for each category.

3.4.3 Implementation Details

As discussed in Section 3.3.3 and 3.3.4, we employ the predicted scores of the basic-level classifiers as features for both the segmentation ($\phi^{Sg}(I, id)$ in Eqn. (3.4)) and context ($\phi^{Ct}(I, id, p_0)$ in Eqn. (3.6)) components to improve the efficiency of the UDS framework. For the segmentation component, we follow the second-order pooling approach [82] by utilizing the public available implementation provided by the author. 150 top-ranked object hypotheses are generated with the CPMC method for each image [14]. The concatenation of scores from support vector regressors of all categories is employed as the segmentation component feature for each hypothesis. For the context component, the dense SIFT [75] and color moment are extracted as low-level features. Both features are projected to 64 dimensions using PCA and the size of Gaussian Mixture Model in FV [18] is set to 64. The concatenation of

Table 3.1: Proof-of-Concept experiments for object detection on VOC 2010 validation set.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
DPM	43.6	51.1	4.4	3.4	21.7	57.4	40.4	17.0	16.4	15.3	10.2	11.1	37.2	39.1	40.4	5.2	27.4	18.9	39.7	37.1	26.9
S-DPM	48.2	52.7	4.9	5.7	25.3	60.6	40.8	21.6	**16.6**	16.3	17.0	12.5	40.5	38.8	**41.3**	6.9	32.5	23.2	44.3	40.8	29.5
S-DPM+Sg	57.6	55.4	22.6	15.8	27.9	**64.3**	45.8	54.8	10.7	26.9	21.9	35.2	48.2	49.8	38.8	13.3	36.3	32.5	49.0	45.3	37.6
S-DPM+Sg+Ct	**59.2**	**56.7**	**22.8**	**16.4**	**28.9**	63.7	**46.6**	**56.2**	15.6	**29.1**	**25.1**	**36.9**	**49.5**	**50.7**	39.3	**14.4**	**38.2**	**36.1**	**49.2**	**46.2**	**39.0**

resulting FVs in all regions is then trained with the LibLinear library [40] in a similar manner with [27]. Finally, the confidence scores of classifiers for all categories are utilized as the context component features.

For the shape-guided DPM, the number of subcategories is automatically decided by the graph shift algorithm based on the expansion size, which is decided by cross-validation [72]. The resulting subcategory number for different object categories is generally from 4 to 8.

The weights λ^{Dt}, λ^{Sg} and λ^{Ct} in Eqn. (3.1) are set as 0.1, 0.2 and 0.2, respectively, based on cross-validation. In fact, the final accuracy is not very sensitive to the variation of these parameters, as our UDS framework can automatically learn w to adjust the relative weights of different components.

3.5 Experiments

In the following section, we extensively evaluate the proposed UDS framework on the challenging PASCAL Visual Object Challenge (VOC) datasets [39]. We first conduct multiple Proof-of-Concept experiments on the validation set to assess the relative importance of each individual component. Then, we evaluate the optimal configuration of the proposed framework on the test set to compare with the state-of-the-art performance ever reported for both object detection and semantic segmentation tasks.

Table 3.2: Proof-of-Concept experiments for semantic segmentation on VOC 2010 validation set.

Method	b/g	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	avg
O_2P	83.2	70.0	22.0	43.8	39.6	40.3	60.3	64.9	55.7	13.2	37.1	20.2	42.5	**37.3**	47.1	50.5	31.9	51.5	27.2	58.6	50.6	45.1
S-DPM+ Sg	82.5	74.2	20.5	45.0	42.7	38.4	65.1	66.9	55.8	16.1	37.3	23.3	41.3	34.7	49.6	49.5	34.1	54.6	33.4	63.7	53.5	46.8
S-DPM+Sg+Ct	**83.2**	**74.9**	**22.9**	**45.7**	**43.4**	**40.6**	**66.2**	**68.1**	**56.4**	**16.8**	**39.8**	**24.0**	**44.2**	36.3	**49.9**	**50.9**	**34.4**	**56.7**	**34.1**	**64.8**	**54.4**	**48.0**

3.5.1 Proof-of-Concept Experiments

In this subsection, we evaluate the relative importance of individual components in our framework on VOC 2012 "train/val" datasets (i.e. "train" set for training and "val" set for test) with the extra segmentation annotation from [57] for proof of concept and ease of parameter tuning.

Table 3.1 and 3.2 show the detailed object detection and semantic segmentation results, respectively. It can be concluded from the tables that:

- Shape-guided subcategory mining does improve the detection performance. By better capturing the pose/viewpoint variance and adaptively deciding the number of subcategories, shape-guided DPM (S-DPM) can provide more reliable shape masks for our UDS framework.

- Object detection and semantic segmentation techniques are complementary. Performing two tasks jointly will boost the performance of each other. As shown in Table 3.1, the joint approach (S-DPM+Sg) significantly outperforms the detection baseline (S-DPM) by 8.1%. In fact, the DPM based detector mainly captures the shape cues. Hence, it may locate rigid parts only and thus leads to localization error. On the contrary, the underlying segmentation component mainly relies on the appearance cues and thus can help to rectify the bounding box position, especially for the objects with homogeneous appearances. Table 3.2 demonstrates that the joint approach (S-DPM+Sg) also outperforms the segmentation baseline (O_2P). For objects in the cluttered background, shape based detectors can provide valuable information to assist in selecting better segment hypotheses. More examples to illustrate the

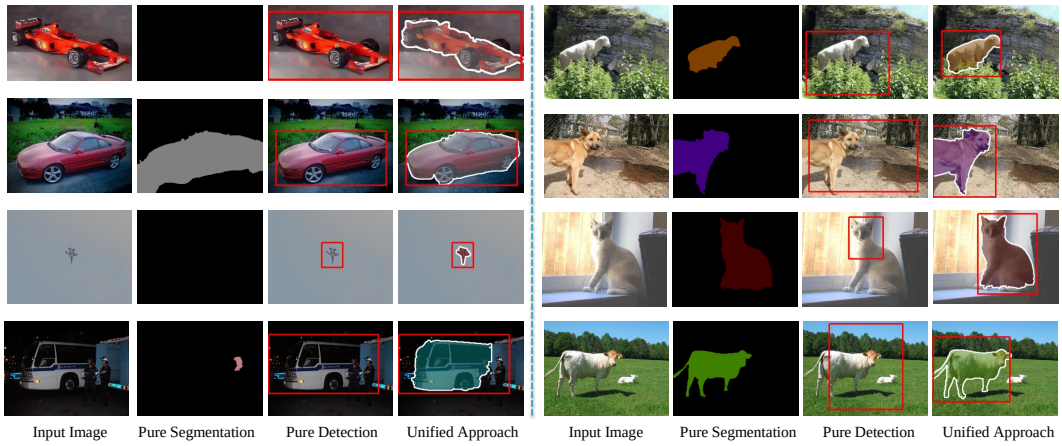


Figure 3.4: More exemplar results on VOC 2012 from the proposed UDS framework and baseline methods (DPM [44] for detection and O_2P [13] for segmentation).

complementarity of the two tasks are shown in Figure 3.4.

- The context component can further improve the performance for both tasks. By employing both the local and global context cues, the full model (S-DPM+Sg+Ct) can better distinguish ambiguous objects and thus yield the best performance.

3.5.2 Comparison with State-of-the-arts

In this subsection, we evaluate our UDS framework on the Pasval VOC test set to have a direct comparison with the state-of-the-arts. Though our framework can perform joint detection and segmentation, these two tasks are usually evaluated using different image sets. Hence, we slightly tweak the training process to allow the direct comparison with previous methods. Specifically, for the detection task, we train the model on the VOC 2010 "main-trainval" set, as many leading methods [48, 22] only reported their results on this dataset. For the segmentation task, we perform the experiments on the union of the VOC 2012 "main" and "seg" sets. The extra segmentation annotation from [57] are used for both tasks. We omit the results of VOC 2010 segmentation and VOC 2012 detection due to space limitation.

Object Detection: The detailed comparison of the proposed framework with

Table 3.3: Comparison of detection performance on VOC 2010 test set.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
DPM [44]	48.2	52.2	14.8	13.8	28.7	53.2	44.9	26.0	18.4	24.4	13.7	23.1	45.8	50.5	43.7	9.8	31.1	21.5	44.4	35.7	32.2
van de Sande *et al.* [95]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	**15.3**	41.1	31.8	47.0	44.8	35.1
Gu *et al.* [53]	53.7	42.9	18.1	16.5	23.5	48.1	42.1	45.4	6.7	23.4	27.7	35.2	40.7	49.0	32.0	11.6	34.6	28.7	43.3	39.2	33.1
NLPR [39]	53.3	**55.3**	19.2	21.0	30.0	54.4	46.7	41.2	**20.0**	31.5	20.7	30.3	48.6	55.3	**46.5**	10.2	34.4	26.5	50.3	40.3	36.8
MITUCLA [118]	54.2	48.5	15.7	19.2	29.2	55.5	43.5	41.7	16.9	28.5	26.7	30.9	48.3	55.0	41.7	9.7	35.8	30.8	47.2	40.8	36.0
ContextSVM [86]	53.1	52.7	18.1	13.5	30.7	53.9	43.5	40.3	17.7	31.9	28.0	29.5	**52.9**	56.6	44.2	12.6	36.2	28.7	50.5	40.7	36.8
FV [22]	**65.9**	50.1	23.7	**24.1**	20.4	52.6	47.1	50.9	13.2	32.8	**31.8**	41.4	43.9	55.3	29.8	14.1	**41.7**	35.6	46.7	**46.9**	38.4
	Using Extra Semantic Segmentation Annotation From [57]																				
segDPM [48]	58.7	51.4	**25.3**	**24.1**	**33.8**	52.5	49.2	48.8	11.7	30.4	21.6	37.7	46.0	53.1	46.0	13.1	35.7	29.4	52.5	41.8	38.1
Ours:UDS	60.1	54.3	23.9	22.9	31.8	**57.0**	**51.1**	**54.8**	17.6	**35.7**	26.7	**42.8**	51.2	**58.0**	41.7	**15.3**	37.8	**39.8**	**54.9**	45.6	**41.2**

current leading approaches for object detection is presented in Table 3.3. The first two methods represent two different lines of approaches for object detection. DPM [44] employed shape based templates with the sliding window strategy while van de Sande *et al.* [95] utilized the appearance based model with the selective window strategy. Gu *et al.* [53] further extended DPM with a multiple component mechanism. Despite their theoretical interest, these methods only focus on the information within the windows and thus ignore the informative context cues, which leads to inferior results compared with other competitors. All other methods are obtained through the combinations of multiple techniques in order to obtain better performance.

From Table 3.3, it can be observed that our proposed UDS outperforms all the competitors in terms of mAP. The proposed UDS framework achieves the best performance in 9 out of the 20 categories with an mAP of 41.2%, which is 3.1% higher than that of the state-of-the-arts. With our unified approach, the advantages of both object detection and semantic segmentation techniques can be leveraged to improve the overall performance. In addition, it can be noted that our method can significantly improve the performance on the categories with homogeneous appearances, such as cats and dogs. For such categories, the underlying segmentation component can easily segment the objects out for rectifying the localization errors.

Semantic Segmentation: Table 3.4 shows the detailed comparison of the proposed framework with previous approaches on the VOC 2012 segmentation chal-

lenge. Based on the basic idea behind the methods, all the competing methods can be divided into two categories. The first category (O2P-CPMC-CSI, CMBR-O2P-CPMC-LIN, O2P-CPMC-FGT-SEGM and Yadollahpour) employs the hypotheses based segmentation. The difference among them mainly lies in the hypotheses generation procedure and ranking function design. Most of them provide the results with/without extra annotation from [57]. The other category (NUS-DET-SPR-GC-SP and Xia) estimates the semantic segmentation results based on the bounding boxes from object detection. Hence, these approaches heavily rely on the detector performance and need extra annotation for object detection.

The results in Table 3.4 demonstrate that the proposed UDS framework performs the best in 8 out of the 21 categories, achieving the best average performance of 50%. As discussed above, our unified approach can leverage the advantages of both object detection and semantic segmentation techniques. One main source of the improvement for semantic segmentation comes from the successful detection of objects in cluttered backgrounds. The bottom-up segmentation techniques may not be able to extract the accurate boundary of objects in cluttered backgrounds, which makes the following ranking problem very difficult. However, the template based detection mainly focuses on the object shape and thus is robust to the cluttered backgrounds to some extent. Hence, the proposed framework can significantly improve the semantic segmentation performance of rigid objects, such as aeroplane, bus and motorbike, as verified in Table 3.4.

3.6 Chapter Summary

In this chapter, we proposed a unified framework for joint object detection and semantic segmentation. Noticing the complementarity of current detection and segmentation approaches, we explicitly enforce the consistency between their outputs to leverage the advantages of both techniques. Both local and global context information are further integrated into the framework to better distinguish the ambiguous samples. All the information is aggregated at the end of the pipeline for decision

Table 3.4: Comparison of segmentation performance on VOC 2012 test set.

Method	b/g	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	avg
O2P-CPMC-CSI [38]	85.0	59.3	27.9	43.9	39.8	41.4	52.2	**61.5**	56.4	13.6	44.5	26.1	42.8	51.7	57.9	51.3	29.8	45.7	28.8	49.9	43.3	45.4
CMBR-O2P-CPMC-LIN [38]	83.9	60.0	27.3	46.4	40.0	41.7	57.6	59.0	50.4	10.0	41.6	22.3	43.0	51.7	56.8	50.1	33.7	43.7	29.5	47.5	44.7	44.8
O2P-CPMC-FGT-SEGM [38]	85.1	65.4	29.3	51.3	33.4	44.2	59.8	60.3	52.5	13.6	**53.6**	32.6	40.3	57.6	57.3	49.0	33.5	53.5	29.2	47.6	37.6	47.0
Yadollahpour *et al.* [110]	**85.7**	62.7	25.6	46.9	43.0	54.8	58.4	58.6	55.6	14.6	47.5	31.2	44.7	51.0	60.9	**53.5**	36.6	50.9	30.1	50.2	46.8	48.1
Relying on Extra Object Detector																						
NUS-DET-SPR-GC-SP [38]	82.8	52.9	**31.0**	39.8	44.5	58.9	60.8	52.5	49.0	**22.6**	38.1	27.5	47.4	52.4	46.8	51.9	35.7	**55.3**	**40.8**	**54.2**	47.8	47.3
Xia *et al.* [107]	82.5	52.1	29.5	50.6	35.6	**59.8**	64.4	55.5	54.7	22.0	38.7	24.3	**48.3**	55.6	52.9	52.2	38.2	49.1	35.5	53.7	**53.5**	48.0
Using Extra Semantic Segmentation Annotation From [57]																						
O2P-CPMC-CSI [38]	85.0	63.6	26.8	45.6	41.7	47.1	54.3	58.6	55.1	14.5	49.0	30.9	46.1	52.6	58.2	53.4	32.0	44.5	34.6	45.3	43.1	46.8
CMBR-O2P-CPMC-LIN [38]	84.7	63.9	23.8	44.6	40.3	45.5	59.6	58.7	57.1	11.7	45.9	34.9	43.0	54.9	58.0	51.5	34.6	44.1	29.9	50.5	44.5	46.7
O2P-CPMC-FGT-SEGM [38]	85.2	63.4	27.3	**56.1**	37.7	47.2	57.9	59.3	55.0	11.5	50.8	30.5	45.0	58.4	57.4	48.6	34.6	53.3	32.4	47.6	39.2	47.5
Ours:UDS	85.2	**67.0**	24.5	47.2	**45.0**	47.9	**65.3**	60.6	**58.5**	15.5	50.8	**37.4**	45.8	**59.9**	**62.0**	52.7	**40.8**	48.2	36.8	53.1	45.6	**50.0**

making and thus hard decision is avoided to make at the early stage as in traditional pipelines. The relative importance of different components is automatically learned for each category to guarantee the overall performance. Extensive experimental results clearly demonstrated the proposed framework has achieved the state-of-the-art performance.

Chapter 4

A Deformable Mixture Parsing Model with Parselets

In this work, we address the problem of human parsing, namely partitioning the human body into semantic regions. Traditional methods usually handle this problem by a sequential or iterative approach. By reconsidering the basic representation for human parsing, we propose the novel Parselet representation. Then, we directly solve the human parsing problem without intermediate tasks to guarantee the parsing performance.

4.1 Introduction

Human parsing [111] has drawn much attention recently for its wide applications in human-centric analysis, such as person identification [50] and clothing analysis [19, 74]. The success of human parsing relies on the seamless cooperation of human pose estimation [113], segmentation [4], and region labeling [111]. However, previous works often consider solving the problem of human pose estimation as the prerequisite of human parsing [111]. We argue that these approaches cannot obtain optimal pixel level parsing due to the inconsistent targets of these tasks.

In this chapter we aim to develop a unified framework for human parsing. To this end, we reconsider the basic level representation. Although the key points [117]

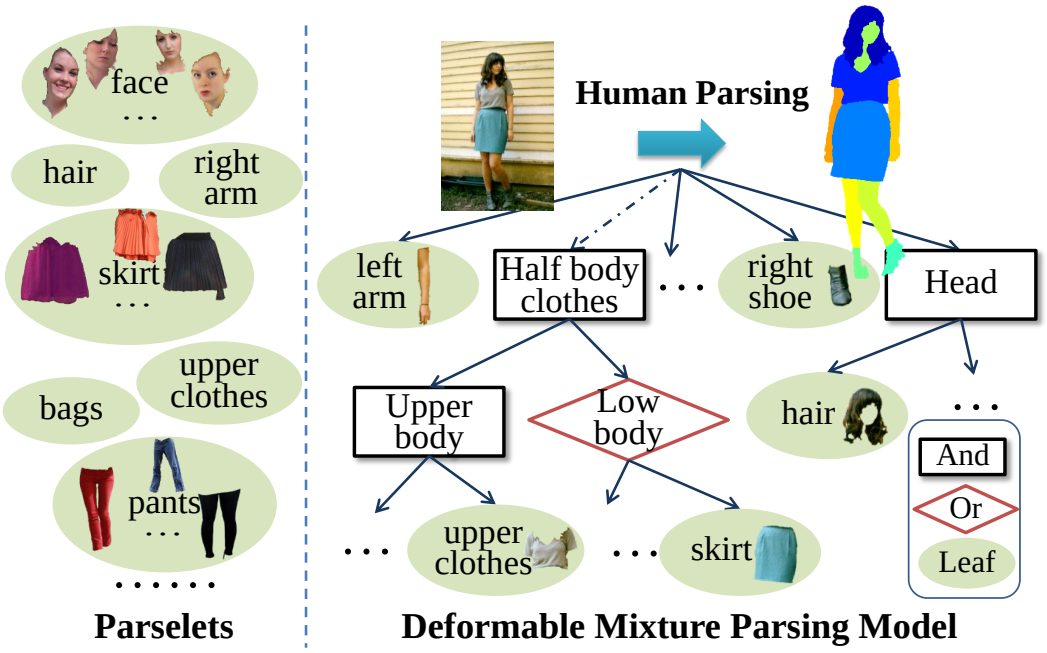


Figure 4.1: Parselets are image segments that can generally be obtained by low-level segmentation techniques and bear strong semantic meaning. The instantiated Parselets, which are activated by our Deformable Mixture Parsing Model, provide accurate semantic labeling for human parsing.

or rigid templates [113, 44] representation can facilitate the localization of human parts, leading to great success in human detection and pose estimation [113], it fails to provide accurate pixel-level labeling. This limitation hinders key points or templates to be the ideal building blocks for human parsing. On the other hand, there exists exciting progress of bottom-up region hypotheses based segmentation methods [14, 35], which have achieved the state-of-the-art performance [38]. More specifically, region hypotheses based segmentation is performed by first generating extensive object hypotheses based on bottom-up information and then ranking them, with the critical assumption that the object has a large probability to be tightly covered by at least one of the generated hypotheses. This assumption usually holds well for objects with homogeneous appearance. However, for objects with large appearance variance, finding a single region hypothesis to tightly cover the whole object is very difficult.

Based on the above observation, we propose to use **Parselets** as the building

blocks for human parsing as shown in Fig. 4.1. The Parselets are a group of semantic image segments with the following characteristics: (1) they can generally be obtained by low-level over-segmentation algorithms [5, 1], *i.e.* they are parsable by bottom-up techniques; (2) they have strong and consistent semantic meaning, *i.e.* they are parsable by the human knowledge. An object consisting of parts with large variance usually cannot be well segmented out by the low-level segmentation methods, *e.g.* a human body cannot be perfectly segmented by edge-based segmentation [5]. However, we argue that the localized semantic regions, *e.g.* the skirt or hair area of human in Fig. 4.1, often show homogeneous appearance and can be segmented out as segments. Such image segments, denoted as Parselets, explicitly encode segmentation and semantic level information.

With the Parselet representation, we propose the Deformable Mixture Parsing Model (DMPM) for human parsing. DMPM is represented as an "And-Or" graph [117] based hierarchical model to simultaneously handle the deformation and multi-modalities of Parselets. The joint learning and inference of best configuration for both appearance and structure in our DMPM guarantee the overall performance. We perform human parsing by generating extensive hypotheses for Parselets and subsequently assembling them by DMPM. The major contributions of this chapter can be summarized as follows:

- We propose the novel Parselet representation. By explicitly encoding segmentation and semantic information, Parselets serve as ideal building blocks for human parsing models. Human parsing is then performed with the Parselet representation, rather than with the key point [117] or rigid template [113, 44] representation. The instantiated Parselets directly provide accurate pixel-level semantic information. In practice, several over-segmentation techniques are utilized to ensure the high recall rate of Parselets.

- We build a novel Deformable Mixture Parsing Model (DMPM) for human parsing. The "co-occurrence" and "exclusive" modalities of Parselets are exhibited as the "And-Or" structure of sub-trees. To further solve the problem of

Parselet occlusion or absence, we directly add the "visibility" property at the corresponding nodes. Joint learning and inference of appearance and structure parameters guarantee the overall performance. In addition, the tree structure of our DMPM allows efficient inference.

- In order to verify the effectiveness of the proposed framework, we construct a high resolution human parsing dataset consisting of 2,500 images. All the pixels in the images are thoroughly annotated with 18 types of Parselets. As far as we know, this is the largest human dataset with full parsing labels. It could serve as the benchmark for segmentation-based human analysis in the research community.

4.2 Related Work

Selective Search for Recognition: Selective search approaches for object recognition have achieved great success in the past few years [83, 35, 96, 14, 4, 13]. This line of works first generate a set of object hypotheses based on bottom-up information and then convert the recognition problem into a ranking problem. Compared with exhaustive sliding window scanning [27, 44], selective search usually enables more expensive and potentially more powerful recognition techniques [97, 96]. Our work differs from the above works significantly as we focus on parts instead of whole objects. We claim that region hypotheses are better hypotheses for parts than for objects toward categories with heterogeneous appearance. Gu et al. [54] also addressed the problem of segmenting and recognizing objects based on their parts. They generated part hypotheses and then formulated the problem in the generalized Hough transformation framework. Our work differs from this work significantly as their work focuses on the segmentation and is unable to exploit the hierarchical structure of the object.

Part Based Model: Hierarchical part based models can better grasp the complicated structure than rigid models and thus usually achieve better performance for articulated objects [44, 113, 120]. Pictorial Structure (PS) based meth-

ods [45, 44, 113] are the most common approaches for pose estimation and object recognition. However, unlike our DMPM, part templates are usually spread in all nodes of PS based models, which makes it inconvenient to model complicated composite relation. The stochastic image grammar model [117, 21] is also effective for modeling the hierarchical structure. However, these models rely on complex learning and inference procedures which can only be made tractable using approximate algorithms [87]. On the contrary, despite the sophisticated structure of DMPM, we show that a tractable and exact inference algorithm exists.

Human Parsing: Human parsing plays an important role in many human-centric applications [19, 74, 100, 73]. Our method differs from previous methods [111] as previous research on human parsing tends to first align human parts [113] due to the large pose variations or the complexity of the models. However, such sequential approaches may fail to capture the correlations between human appearance and structure, leading to unsatisfactory results. The proposed DMPM, which can solve human parsing in a unified framework, significantly distinguishes our work from others.

4.3 Parselets

Parselets lie at the heart of our human parsing framework. In this section, we first give the definition of human Parselets. Then we present the details of hypothesis generation and feature representation for Parselets. And finally, we briefly introduce the modalities of Parselet ensembles.

4.3.1 Parselet Definition

We notice that the classical part-based models [45, 113] usually divide body into parts based on joints. However, such decomposition is unsuitable for segment hypotheses because joint-based parts usually do not correspond to the segments from bottom-up cues. Considering the left image in Fig. 4.2, the whole dress is likely to be captured by a single segment from the bottom-up techniques. But for the right

Figure 4.2: Human decomposition based on different basic elements. The original image, Parselet based decomposition and joint based decomposition are shown sequentially.

image, the upper clothes, coat and pants should intuitively correspond to three separate segments. This difference is hard to be grasped by joint based decomposition. To overcome this limitation, we propose the Parselets to serve as the building elements for our parsing model. Formally, the **Parselets** are a group of semantic image segments which have the following characteristics: (1) they can generally be obtained by low-level segmentation algorithms [5, 1, 14], *i.e.* they are parsable by the bottom-up techniques. This characteristic guarantees that Parselets can be retrieved with high possibility by the bottom-up hypothesis generation schemes. (2) They bear strong and consistent semantic meaning, *i.e.* they are parsable by the human knowledge. Since our ultimate goal is to perform human parsing, the basic elements of the parsing model should have clear semantic meaning.

We now decompose human body into homogeneous regions based on low-level cues. The homogeneous regions, which have clear semantic meaning and appear in many different images, are defined as Parselets. Through careful design, each defined Parselet will have high probability to form a single segment. Specifically, we define 18 types of Parselets as described in Table 4.1. These Parselets are representative and can properly cover most of human body. They engage about 98.4% of human body in our labeled datasets and can be obtained with high recall rate using the method introduced in Section 4.3.2. Detailed statistics are shown in the experiment section. It is worth noting that the Parselet definition is flexible to be redesigned for different applications. The only assumption here is that those semantic regions

Table 4.1: 18 types of Parselets for human

	Parselets		
Head	hat	hair	sunglasses
Body	upper clothes skirt	coat pants	full body clothes
Foot	left/right shoe		
Skin	face	left/right arm	left/right leg
Accessory	bag	scarf	belt

can be segmented out with high probability.

4.3.2 Hypothesis Generation for Parselets

In order to obtain the Parselet hypotheses with high recall rate, we combine several low-level segmentation methods. As Parselets usually appear in different scales, the hierarchical segmentation algorithm should be a natural way to generate hypotheses. Here, we choose Ultrametric Contour Map (UCM) [5], which works well to preserve the boundary information. However, the merging scheme of UCM proceeds by removing the edge with smallest probability and thus only neighboring super-pixels can be merged. This may prevent non-adjacent segments from merging as a single segment and lead to unsatisfactory results for some Parselets, which are separated by noise segments. For example, the dress in the left image of Fig. 4.2 is split into separate segments by the stripe pattern with strong edges. Hence UCM fails to merge them in the early stage. In addition, some garments, such as a belt, may also divide a Parselet into separate segments. To handle these difficulties, we add another appearance based segmentation and merging scheme. Specifically, we first use the fast appearance based over-segmentation method [1] and sequentially merge the nearby (not necessarily adjacent) regions with the smallest similarity score in a similar manner as in [96]. We define the similarity score S between segments a and b as $S(a,b) = S_{size}(a,b) + S_{appearance}(a,b)$, both of which are normalized to [0,1]. $S_{size}(a,b)$ is defined as the fraction of the image that the region a and b jointly occupy. This factor encourages small regions to be merged early. $S_{appearance}(a,b)$

is defined as the χ^2 distance of the color and SIFT [75] histogram of segments a and b [99]. Finally, we utilize another complementary scheme, namely CPMC [14], which directly generates many segments of different scales. The segments from the above three methods are combined into the final Parselet hypothesis.

4.3.3 Feature Representation

Compared with exhaustive sliding window scanning [27, 44], our Parselet based representation enables complex and expensive feature design. It has been shown that the bag of words feature performs better than the rigid template for categories with large pose and view variance [97, 96, 14]. As our Parselet categorization is essentially a classification problem, we follow the state-of-the-art feature extraction-coding-pooling classification pipeline [49, 18, 13]. In this work, we adopt the Fisher Kernel (FK) + average pooling [49] and enhanced feature + second order pooling [13], which have been shown with the best performance among current BoW encoding methods. In addition, as our algorithm only employs the size and appearance features which can be efficiently propagated throughout the hierarchical structure embedded in the pools of segments, the feature extraction is reasonably fast.

4.3.4 Parselet Ensemble

Parselets serve as the building blocks of our human parsing model. The Parselets are low-level parts from the definition. In practice, several Parselets are often grouped together in order to form the middle-level human body part, *e.g.* head, body, etc. Those middle-level parts cannot be represented by a single type of Parselets but can be modeled by the ensembles of Parselets. More specifically, the ensembles of Parselets show two kinds of modalities as follows: (1) **Co-occurrence**. The modality of co-occurrence represents the relation that several types of Parselets coexist and are merged to form a larger middle-level human part. This is the most typical modality of Parselet ensembles. For example, the "hair" usually comes with "face" to form the "head". (2) **Exclusivity**. The modality of exclusivity models the relationship of different types of Parselets that cannot coexist logically. For

example, for the "lower-body" area, there are two possible Parselets, *i.e.* "skirts" and "pants". However, "skirts" and "pants" usually cannot coexist. The exclusivity for the middle-level concept "lower-body" means that only one of the two exclusive Parselets, *i.e.* "skirts" and "pants", can exist for the "lower-body".

The middle level concepts formed from Parselet ensembles can be further merged with Parselet(s) or other middle level concepts. They also exhibit co-occurrence or exclusivity modalities to form an even higher level concept. This higher level concept thus inherits all the information from its sub-components. This inheritance property guarantees that we can model complex objects (*e.g.* human) with multiple levels of concepts.

4.4 Human Parsing over Parselets

With the Parselets and their ensembles, we propose the Deformable Mixture Parsing Model (DMPM) for human parsing. Specifically, we propose to employ an "And-Or" graph [117] based hierarchical model to simultaneously handle the deformation and multi-modalities of Parselets. The "co-occurrence" modality is modeled as the "And" relation while "exclusivity" modality is modeled as the "Or" relation in the graph. The deformation is modeled as pairwise parent-child distance. We construct a hierarchical model, as hierarchical models have been shown to be effective for grasping the structure of objects in part based approaches [113, 117]. In addition, absence/occlusion is common for some Parselets. Hence we explicitly model this by utilizing a special structure call virtual "Leaf" node. Fig. 4.3 shows a subgraph from our human graph, while the full graph of our parsing model is listed in the supplemental file. In the next subsections, we will introduce our DMPM followed by the inference and learning algorithms.

4.4.1 Deformable Mixture Parsing Model

We first define the notations used in the following section. P represents the Parselet hypothesis segments in an image generated according to Section 4.3.2. For a hy-

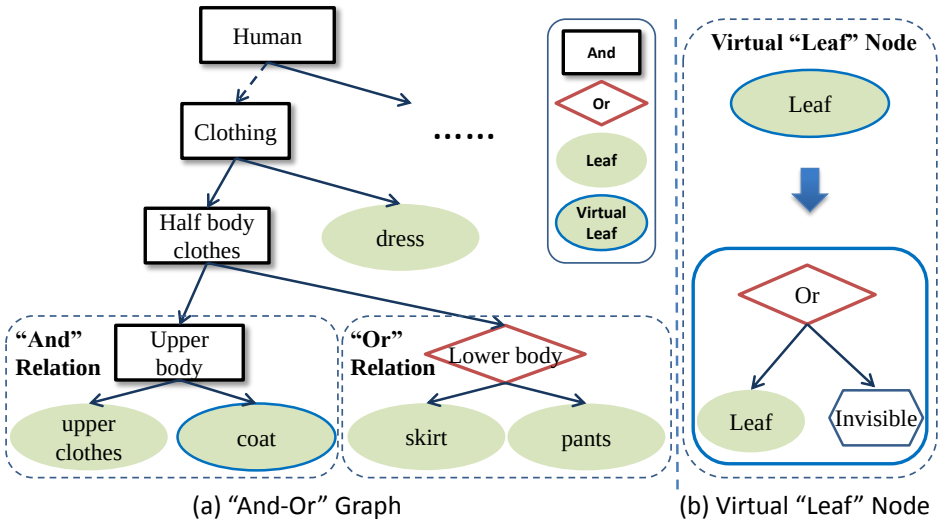


Figure 4.3: The subgraph from our human "And-Or" graph. The diamonds, rectangles, eclipses and eclipses with boundary represent "Or" nodes, "And" nodes, "Leaf" nodes and virtual "Leaf" nodes, respectively.

pothesis segment with index i, its scale (the square root of its area) and centroid are donated as s_i and $c_i = (x_i, y_i)$. Formally, a DMPM model is represented as a graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. The edges are defined by the parent-child structure and kids(ν) denote the children of node ν. There are three basic types of nodes, "And","Or" and "Leaf" nodes which specify different parent-child relationships as depicted in Fig. 4.3 by diamonds, rectangles and eclipses respectively. Each "Leaf" node corresponds to one type of Parselets.

The state variables of the graph specify the graph configuration. Specifically, the graph topology is instantiated by a switch variable t at "Or" nodes, which indicates the set of active nodes $V(t)$. Starting from the top level, an active "Or" node $\nu \in V^O(t)$ selects a child $t_\nu \in$ kids(ν). The active "And" or "Or" nodes have the state variables $g_\nu = (s_\nu, c_\nu)$ which specify the virtual scale and centroid of the node. The active "Leaf" nodes $\nu \in V^L(t)$ have the state variables d_ν which specify the index of the segments for Parselets. In summary, we specify the configuration of the graph by the states $z = \{(t_\nu, g_\nu) : \nu \in V^O(t)\} \bigcup \{g_\nu : \nu \in V^A(t)\} \bigcup \{d_\nu : \nu \in V^L(t)\}$ where the active nodes $V(t)$ are determined from the $\{t_\nu : \nu \in V^O(t)\}$. We then let $z_{\text{kids}(\nu)} = \{z_\mu : \mu \in \text{kids}(\nu)\}$ denote the states of all the child nodes of an "And"

node $\nu \in V^A$ and let z_{t_ν} denote the state of the selected child node of an "Or" node $\nu \in V^O$.

Invisibility Modeling: Some Parselets, such as bags and scarfs, have high probability to be absent or occluded, namely invisible. In other words, these "Leaf" nodes should be with the visibility property. We explicitly model these notes by using a special structure, denoted as virtual "Leaf" node. More specifically, we introduce an auxiliary "Invisible" type of nodes which have no appearance representation. Then the virtual "Leaf" node is represented as a structure consisting of an "Or" node, an ordinary "Leaf" node and an "Invisible" node, as shown in Fig. 4.3. The activated nodes in the virtual "Leaf" node structure thus explicitly suggest whether the corresponding "Leaf" node (Parselet) is visible or not. For standard "Leaf" node μ, the corresponding score is $w_\mu^L \cdot \Phi^L(P, z_\mu)$, where $\Phi^L(P, z_\mu)$ is the feature vector extracted from the segment d_μ as described in Section 4.3.3. For the virtual "Leaf" node with "Or" node ν, "Leaf" node μ and "Invisible" node ρ, the score is $w_\mu^L \cdot \Phi^L(P, z_\mu) + w_{\nu,\mu}^O$ or $w_{\nu,\rho}^O$ depending on the visibility of the corresponding Parselet. $w_{\nu,\mu}^O$ and $w_{\nu,\rho}^O$ are the learned weights for the visibility property, which are embedded in the "Or" node of the virtual "Leaf" node. It is worth noting that the state of the "Invisible" node fully depends on its weight in the "Or" node and its own score is always 0.

We can now write the full score associated with a state variable z:

$$\begin{aligned} S(P, z) = & \sum_{\mu \in V^L(t)} w_\mu^L \cdot \Phi^L(P, z_\mu) + \sum_{\mu \in V^O(t)} w_{\mu, t_\mu}^O \\ & + \sum_{\mu \in V^A(t)} w_\mu^A \cdot \Phi^A(z_\mu, z_{\text{kids}(\mu)}). \end{aligned} \tag{4.1}$$

The first term in Eqn. (4.1) is an appearance model that computes the local score of assigning the segment d_μ as Parselet μ. The last two terms are independent of the data and can be considered as priors of occurrence and the spatial geometry. Based on the graph structure, we can further decompose the last term of Eqn. (4.1)

as follows:

$$S(P,z) = \sum_{\mu \in V^L(t)} w^L_\mu \cdot \Phi^L(P, z_\mu) + \sum_{\mu \in V^O(t)} w^O_{\mu,t_\mu} + \sum_{\mu \in V^A(t)} \sum_{\nu \in \text{kids}(\mu)} w^A_{\mu,\nu} \cdot \psi(d_\mu, d_\nu). \tag{4.2}$$

$\psi(d_\mu, d_\nu) = [dx \;\; dx^2 \;\; dy \;\; dy^2 \;\; ds]^T$ measures the geometric difference between part μ and ν, where $dx = (x_\nu - x_\mu)/\sqrt{s_\nu \cdot s_\mu}$, $dy = (y_\nu - y_\mu)/\sqrt{s_\nu \cdot s_\mu}$ and $ds = s_\nu / s_\mu$ are the relative location and scale of part ν with respect to μ.

Compared with the most prevalent hierarchical modeling approaches [113, 44], the proposed model has the following distinctive characteristics:

- We use Parselets as the basic elements for our parsing model. The parsing problem is now transferred as searching the best configuration of the hierarchical model. Once the maximization is obtained, we can directly get the accurate pixel-level segmentation and semantic labels from the corresponding Parselets.

- The "And-Or" graph structure allows both co-occurrence and exclusivity relations between different parts. Unlike previous methods [113, 44], which often use "Or" node to model the multi-view properties of the same part, the "Or" node here plays the role of selecting the best configuration among mixture of subgraphs, which is more flexible.

- We explicitly model the visibility property of the "Leaf" node, which is practical and critical for some Parselets. The introduction of a special node, *i.e.* the Invisible node, brings the flexibility for the real-life situation without adding extra model complexity.

4.4.2 Inference

Inference corresponds to maximizing $S(P, z)$ from Eqn. (4.2) over z. As graph G = (V, E) is a tree, inference can be done efficiently with dynamic programming.

More specifically, we can simply iterate over all subparts starting from the leaves and moving "upstream" to the root. The message from children to their parent can be computed by the following:

$$\text{score}^I_\tau(z_\tau) = 0, \tag{4.3}$$

$$\text{score}^L_\tau(z_\tau) = w^L_\tau \cdot \Phi^L(P, z_\tau), \tag{4.4}$$

$$\text{score}^O_\nu(z_\nu) = \max_{\rho \in \text{kids}(\nu)} [m_\rho(z_\nu)], \tag{4.5}$$

$$m_\rho(z_\nu) = \max_{z_\rho} [\text{score}_\rho(z_\rho)] + w^O_{\nu,\rho}, \tag{4.6}$$

$$\text{score}^A_\mu(z_\mu) = \sum_{\rho \in \text{kids}(\mu)} n_\rho(z_\mu), \tag{4.7}$$

$$n_\rho(z_\mu) = \max_{z_\rho} [\text{score}_\rho(z_\rho) + w^A_{\mu,\rho} \cdot \psi(d_\mu, d_\rho)]. \tag{4.8}$$

At the bottom level, the scores of "Invisible" nodes and "Leaf" nodes are calculated as in Eqn. (4.3) and Eqn. (4.4). "Or" node selects the maximal response from its children for its score as in Eqn. (4.5) and Eqn. (4.6). The score of "And" node is calculated by accumulating the scores of its children plus the corresponding deformation as in Eqn. (4.7) and Eqn. (4.8). The above equations suggest that we can express the energy function recursively and hence find the optimal z using dynamic programming. In addition, the maximization over z can be partially accelerated by generalized distance transformation, which makes the whole algorithm more efficient [46, 44].

4.4.3 Learning

Given the labeled examples $\{P_i, z_i\}$, the max-margin framework is arguably preferable to maximum-likelihood estimation as our final goal is discrimination. Note that the scoring function of Eqn. (4.2) is linear in model parameters $w = (w^L, w^O, w^A)$, and can be written compactly as $S(P, z) = w \cdot \Phi(P, z)$. Thus both appearance and structure parameters can be learned in a unified framework, which is critical for achieving the state-of-the-art performance for many applications [44, 113]. Here, we

formulate the structured learning problem in a max-margin framework as in [44]:

$$\begin{aligned} &\min_{w} \|w\|^2 + C\sum_{i} \xi_i \\ &\text{s.t.} \quad w \cdot (\Phi(P_i, z_i) - \Phi(P_i, z)) \geq \Delta(z_i, z) - \xi_i, \forall z; \end{aligned} \tag{4.9}$$

where $\Delta(z_i, z_j)$ is a loss function which penalizes incorrect estimate of z. This loss function gives partial credit to states which differ from the ground truth slightly. The loss function is defined as follows:

$$\Delta(z_i, z_j) = \sum_{\nu \in V^L(t_i) \bigcup V^L(t_j)} \delta(z_i^\nu, z_j^\nu), \tag{4.10}$$

where $\delta(z_i^\nu, z_j^\nu) = 1$, if $\nu \notin V^L(t_i) \bigcap V^L(t_j)$ or $\text{sim}(d_i^\nu, d_j^\nu) \leq \sigma$. $\text{sim}(\cdot, \cdot)$ is the intersection over union ratio of two segments d_i^ν and d_j^ν, and σ is the threshold, which is set as 0.8 in the experiments. This loss function penalizes both configurations with "wrong" topology and leaf nodes with wrong segments. The optimization problem Eqn. (4.9) is known as a structural SVM, which can be efficiently solved by the cutting plane solver of SVMStruct [61] and the stochastic gradient descent solver in [44].

4.5 Experiments

4.5.1 Experimental Settings

Evaluation Criterion: The parsing result is evaluated based on two complementary metrics. The first one is Average Pixel Accuracy (APA) [111], which is defined as the proportion of correctly labeled pixels in the whole image. This metric mainly measures the overall performance over the entire image. Since most pixels are background, APA is greatly affected by mislabeling a large region of background pixels as body parts. The second metric is Intersection over Union (IoU) [38], which is widely used in evaluating segmentation and suitable for measuring the performance of each Parselet separately. We also devise two variants of IoU for Parselets to make

Table 4.2: Comparison of Parselets versus objects in terms of the best IoU score on FS and DP datasets.

	dataset	CPMC [96]	SLIC [1]	UCM [5]	Combined
Obj IoU	FS	0.830	0.559	0.430	0.831
Par mIoU	FS	**0.895**	**0.725**	**0.604**	**0.917**
Par wIoU	FS	0.844	0.621	0.546	0.860
Obj IoU	DP	0.815	0.534	0.443	0.816
Par mIoU	DP	**0.896**	**0.722**	**0.638**	**0.928**
Par wIoU	DP	0.831	0.614	0.608	0.862

Parselets comparable with objects. The first one is the "Merging IoU" (mIoU) which merges the hypothesis for each Parselet into an object hypothesis to obtain the object level IoU. The second one is the "Weighted IoU" (wIoU) which is calculated by accumulating each Parselet's IoU score weighted by the ratio of its pixels occupying the whole object. Note that generally mIoU is higher than wIoU.

Implementation Details: We extract dense SIFT [75], HOG [27] and color moment as low-level features for Parselets. The size of Gaussian Mixture Model in FK is set to 128. The training:testing ratio is 2:1 for both datasets. The penalty parameter C is determined by 3-fold cross validation in the training set.

4.5.2 Hypotheses Comparison: Parselets vs. Objects

We first validate the assumption that segmentation can provide better hypotheses for Parselets than for objects with heterogeneous appearance (*e.g.* human) by comparing the best IoU scores of Parselets and objects. The best IoU score for a segmentation method is defined as the maximal IoU score between the segments produced by that method and the ground truth segments. The same hypothesis segments, which are generated through the methods introduced in Section 4.3.2, are used for both Parselets and objects. We calculate the best IoU of Parselets and objects for different method on two datasets. The comparison results are displayed in Table 4.2, from which it can be observed that the best IoU of Parselets is much higher than that of objects. This trend is consistent among different algorithms and

Table 4.3: The best IoU scores for each type of Parselets on the FS and DP datasets.

dataset	FS	DP
hat	84.0	83.5
hair	78.2	81.0
s-gls	56.6	58.8
u-cloth	84.1	88.6
coat	null	71.9
f-cloth	90.8	93.9
skirt	91.6	89.3
pants	92.8	92.5
belt	65.7	71.0
l-shoe	72.4	73.2
r-shoe	71.9	73.8
face	83.4	85.6
l-arm	79.8	93.4
r-arm	79.8	92.9
l-leg	79.2	86.7
r-leg	79.9	86.5
bag	81.8	84.8
scarf	76.1	78.2

datasets, which makes the usage of segments as Parselet hypotheses more convincing. In addition, combining all three complementary algorithms leads to the best performance and we use this setting thereafter. The detailed best IoU for each type of Parselets based on combined hypotheses are shown in Table 4.3 .

4.5.3 Evaluation for Human Parsing

Human Parsing: We now compare our proposed framework with the work of Yamaguchi *et al.* [111] for human parsing. This baseline works by first estimating the human pose and then labeling the super-pixel based on the pose estimation results. We use the public available implementation of version 0.2 and carefully tune the parameter according to [111]. The baseline method achieves 83% for FS dataset and 82% for DP dataset in terms of APA, which are inferior to 86% and 87% of our framework. Though APA is good at measuring the overall performance of human parsing, it fails to distinguish the performance of separate Parselets and has bias towards background. More specifically, naively assigning all segments as background

Figure 4.4: Comparison of parsing results. Original images, our results and baseline's results [111] are shown sequentially.

Figure 4.5: More exemplar results from our parsing framework.

results in a reasonably good APA of 78% for DP and 77% for FS. Therefore, we further employ the more discriminative IoU criterion for comparison. The detailed comparison results on all types of Parselets are reported in Table 4.4. It can be seen that our method performs much better than the baseline method, especially for the Parselet level results. This mainly verifies the stability of our algorithm. Unlike our method, the baseline method does not model the exclusive relation of different labels, which leads to unstable results as shown in Fig. 4.4. Note that their method can achieve good performance with the prior information specifying what type of Parselets appears in the image. However, such information is usually difficult to obtain for real-world applications. In addition, it can be observed that the results from our model are more robust to uncommon poses and absent/occluded parts. The baseline method estimates the human pose and labels the region separately. This non-unified nature omits the strong correlation of appearance and structure for human. On the contrast, by employing the low-level visual cues and high-level structure information in a unified framework with explicit invisibility modeling, our model is much more robust to these difficult examples. More exemplar results from our framework are shown in Fig. 4.5.

Parsing as Segmentation: As human parsing results in pixel-level segment labeling, our framework implicitly provides human segmentation results. We thus

Table 4.4: Comparison of human parsing IoU scores on FS and DP datasets.

	Baseline [111]	DMPM	Baseline [111]	DMPM
dataset	FS	FS	DP	DP
hat	2.5	5.6	1.3	28.9
hair	47.2	67.9	43.5	74.8
s-gls	0.8	2.8	0.6	9.6
u-cloth	36.4	56.3	21.3	42.5
coat	null	null	19.5	39.4
f-cloth	23.2	56.6	21.8	61.0
skirt	21.6	55.3	12.2	50.3
pants	19.1	40.0	28.7	66.3
belt	8.9	18.2	4.8	16.6
l-shoe	27.6	58.6	25.6	57.0
r-shoe	25.2	53.4	21.7	51.8
face	59.3	72.4	52.6	78.1
l-arm	33.0	52.7	32.4	62.7
r-arm	30.5	45.4	28.3	59.3
l-leg	32.6	48.8	23.5	52.6
r-leg	24.1	41.6	18.4	35.5
bag	9.5	20.6	8.5	12.7
scarf	0.9	1.2	1.2	9.3
wIoU	29.9	51.7	24.6	53.0
mIoU	77.6	83.1	76.6	84.6

further compare the segmentation results between our human parsing method and the state-of-the-art image segmentation method [13], to demonstrate the effectiveness of our framework. The baseline method [13] employs the bottom-up segments as the object hypotheses and only achieves the IoU score of 73% for FS dataset and 70% for DP dataset, which is much lower than the result of Merging IoU of 83.1% and 84.6% as shown in Table 4.4. Some exemplar results are shown in Fig. 4.6, from which we can observe obvious defects for the baseline segmentation results in column (d). Such defects are avoidless for the baseline method as a single segment from the bottom-up segmentation can hardly cover the whole body tightly. On the contrary, our framework can employ the top-down knowledge and assemble several homogeneous segments into an object, which leads to much more accurate segmentation.

Figure 4.6: Comparison of human segmentation results. (a)-(d) are input images, our human parsing results, segmentation results by merging (b) and results from the segmentation method [13], respectively

4.5.4 Human Parsing for High Level Applications

Parselets provide a middle-level representation and well bridge the gap between the low-level segments and the high-level concepts. Hence, our Parselet based parsing framework can serve as the basis for many high-level applications. Here, we build a prototype system to retrieve visually similar person as a representative. More specifically, given a query image, we first filter images in the database based on the Parselet types. For each pair of corresponding Parselets, the similarity is calculated based on the Euclidean distance of the extracted features. Then the similarity between images is defined as the sum of Parselet-level similarities weighted by the faction of their pixels occupying the object. Such a system can be extended for clothing retrieval, person identification and many other human centric analysis. Fig. 4.7 shows some top retrieval results for Parselets such as upper clothes + coat and pants, respectively. It can be observed that the visually similar persons are successfully retrieved independent of pose and uninterested regions. Here, we do not pursue this further for the space limitation.

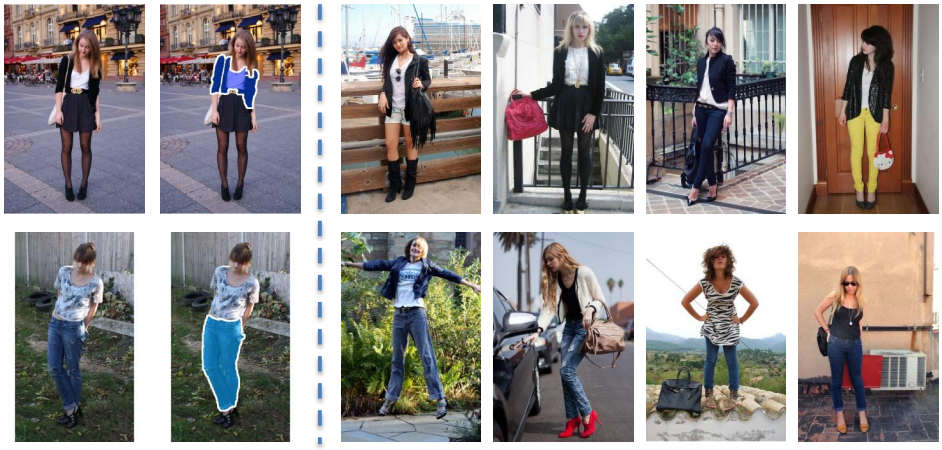


Figure 4.7: Top retrieval results from our visually similar person retrieval system. The retrieval results (right columns) are visually similar to the query human for the highlighted Parselets (the second column) independent of pose and uninterested regions.

4.6 Chapter Summary

In this chapter, we proposed an effective framework for human parsing. By reconsidering the human parsing problem, we utilized the novel Parselets as the basic elements. A unique Deformable Mixture Parsing Model (DMPM) was built to jointly learn and infer the best configuration for both appearance and structure effectively. Extensive experimental results clearly demonstrated the effectiveness of the proposed framework.

Chapter 5

Towards Unified Human Parsing and Pose Estimation

Human pose estimation and human parsing are two strongly correlated tasks. However, correlation between them has rarely been explored. In this chapter, we show how to jointly solve human parsing and pose estimation in a unified framework. By utilizing Parselets and Mixture of Joint-Group Templates as the representations for semantic parts, we seamlessly formulate the human parsing and pose estimation problem within a united framework via a tailored And-Or graph to boost the performance of each other.

5.1 Introduction

Human parsing (partitioning the human body into semantic regions) and pose estimation (predicting the joint positions) are two main topics of human body configuration analysis. They have drawn much attention in the recent years and serve as the basis for many high-level applications [9, 113, 32]. Despite their different focuses, these two tasks are highly correlated and complementary. On one hand, most works on pose estimation usually divide the body into parts based on joint structure [113]. However, such joint-based decomposition ignores the influence of clothes, which may significantly change the appearance/shape of a person. For example, it

is hard for joint-based models to accurately locate the knee positions of a person wearing long dress as shown in Figure 5.1. In this case, the human parsing results can provide valuable context information for locating the missing joints. On the other hand, human parsing can be formulated as inference in a conditional random field (CRF) [89, 32]. However, without top-down information such as human pose, it is often intractable for CRF to distinguish ambiguous regions (*e.g.* , the left shoe v.s. the right shoe) using local cues as illustrated in Figure 5.1. Despite the strong connection of these two tasks, the intrinsic consistency between them has not been fully explored, which hinders the two tasks from benefiting each other. Only very recently, some works [111, 91] began to link these two tasks with the strategy of performing parsing and pose estimation sequentially or iteratively. While effective, this paradigm is suboptimal, as errors in one task will propagate to the other.

In this chapter, we aim to seamlessly integrate human parsing and pose estimation under a unified framework. To this end, we first unify the basic elements for both tasks by proposing the concept of "semantic part". A semantic part is either a region with contour (*e.g.* , hair, face and skirt) related to the parsing task, or a joint group (*e.g.* , right arm with wrist, elbow and shoulder joints) serving for pose estimation. For the representation of semantic regions, we adopt the recently proposed Parselets [32]. Parselets are defined as a group of segments which can be generally obtained by low-level over-segmentation algorithms and bear strong semantic meaning. Unlike the raw pixels used by traditional parsing methods [89], which are not directly compatible with the template based representation for pose estimation, Parselets allow us to easily convert the human parsing task into the structure learning problem as in pose estimation. For pose estimation, we employ joint groups instead of single joints as basic elements since joints themselves are too fine-grained for effective interaction with Parselets. We then represent each joint group as one Mixture of Joint-Group Templates (MJGT), which can be regarded as a mixture of pictorial structure models defined on the joints and their interpolated keypoints. This design ensures that the semantic region and joint group representation of the semantic parts are at the similar level and thus can be seamlessly

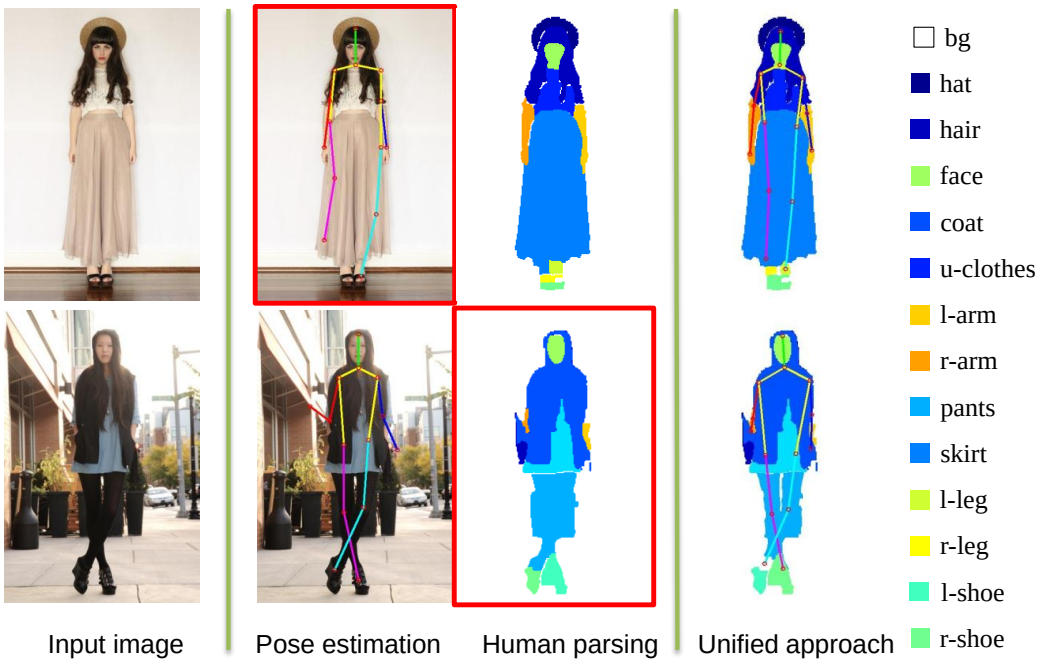


Figure 5.1: Motivations for unified human parsing and pose estimation. The images in top row show the scenario where pose estimation [113] fails due to joints occluded by clothing (*e.g.* , knee covered by dress) while the human parsing works fine. The images in bottom row show the scenario where human parsing [32] is not accurate when body regions are crossed together (*e.g.* , the intersection of the legs). Thus, the human parsing and pose estimation may benefit each other, and more satisfactory results (the right column) can be achieved for both tasks using our unified framework.

connected together.

By utilizing Parselets and MJGTs as the semantic parts representation, we propose a Hybrid Parsing Model (HPM) for simultaneous human parsing and pose estimation. The HPM is a tailored "And-Or" graph [117] built upon these semantic parts, which encodes the hierarchical and reconfigurable composition of parts as well as the geometric and compatibility constraints between parts. Furthermore, we design a novel grid-based pairwise feature, called Grid Layout Feature (GLF), to capture the spatial co-occurrence/occlusion information between/within the Parselets and MJGTs. The mutually complementary nature of these two tasks can thus be harnessed to boost the performance of each other. Joint learning and inference of best configuration for both human parsing and pose related parameters guarantee

the overall performance. The major contributions of this chapter include:

- We build a novel Hybrid Parsing Model for unified human parsing and pose estimation. Unlike previous works, we seamlessly integrate two tasks under a unified framework, which allows joint learning of human parsing and pose estimation related parameters to guarantee the overall performance.
- We propose a novel Grid Layout Feature (GLF) to effectively model the geometry relation between semantic parts in a unified way. The GLF not only models the deformation as in the traditional framework but also captures the spatial co-occurrence/occlusion information of those semantic parts.
- HPM achieves the state-of-the-art for both human parsing and pose estimation on two public datasets, which verifies the effectiveness of joint human parsing and pose estimation, and thus well demonstrates the mutually complementary nature of both tasks.

5.2 Related Work

Human pose estimation has drawn much research attention during the past few years [9]. Due to the large variance in viewpoint and body pose, most recent works utilize mixture of models at a certain level [113, 84]. Similar to the influential deformable part models [44], some methods [84] treat the entire body as a mixture of templates. However, since the number of plausible human poses is exponentially large, the number of parameters that need to be estimated is prohibitive without a large dataset or a part sharing mechanism. Another approach [113] focuses on directly modeling modes only at the part level. Although this approach has combinatorial model richness, it usually lacks the ability to reason about large pose structures at a time. To strike a balance between model richness and complexity, many works begin to investigate the mixtures at the middle level in hierarchical models, which have achieved promising performance [21, 87, 88, 81]. As we aim to perform simultaneous human parsing and pose estimation, we tailor the above

techniques for the proposed HPM by utilizing the mixture of joint-group templates as basic representation for body joints.

Dong *et al.* proposed the concept of Parselets for direct human parsing under the structure learning framework [32]. Recently, Torr and Zisserman proposed an approach for joint human pose estimation and body part labeling under the CRF framework [91], which can be regarded as a continuation of the theme of combining segmentation and human pose estimation [63, 47, 98]. Due to the complexity of this model, the optimization cannot be carried out directly and thus is conducted by first generating a pool of pose candidates and then determining the best pixel labeling within this restricted set of candidates. Our method differs from previous approaches as we aim to solve human parsing and pose estimation simultaneously in a unified framework, which allows joint learning of all parameters to guarantee the overall performance.

5.3 Unified Human Parsing and Pose Estimation

In this section, we introduce the framework of the proposed Hybrid Parsing Model and detail the key components.

5.3.1 Unified Framework

We first give some probabilistic motivations for our approach. Human parsing can be formally formulated as a pixel labeling problem. Given an image I, the parsing system should assign the label mask $L \equiv \{l_i\}$ to each pixel i, such as face or dress, from a pre-defined label set. Human pose estimation aims to predict the joint positions $X \equiv \{x_j\}$, which is a set of image coordinates x_j for body joints j. As human parsing and pose estimation are intuitively strongly correlated, ideally one would like to perform MAP estimation over joint distribution $p(X, L|I)$. However, previous works either estimate $p(X|I)$ and $p(L|I)$ separately [113] or estimate $p(X|I)$ and $p(L|X, I)$ sequentially [111]. The first case obviously ignores the strong correlation between joint positions X and parsing label mask L. The second approach may also

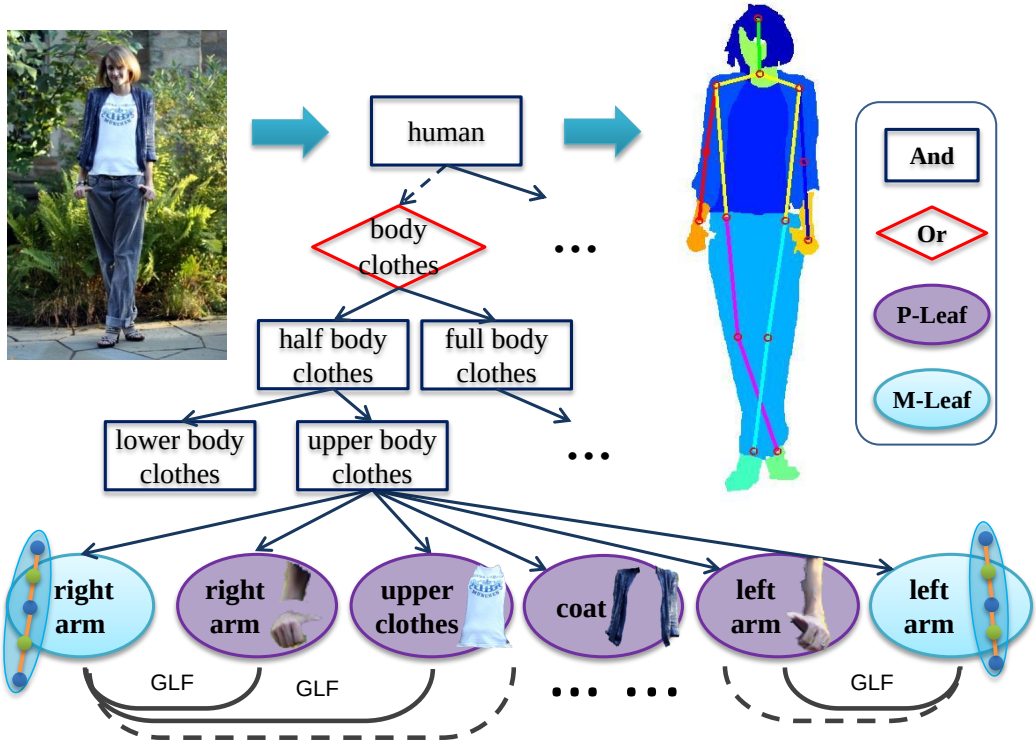


Figure 5.2: Illustration of the proposed Hybrid Parsing Model. The hierarchical and reconfigurable composition of semantic parts are encoded under the And-Or graph framework. The "P-Leaf" nodes encode the region information for parsing while the "M-Leaf" nodes capture the joint information for pose estimation. The pairwise connection between/within "P-Leaf"s and "M-Leaf" is modelled through Grid Layout Feature (GLF). HPM can simultaneously perform parsing and pose estimation effectively.

be suboptimal, as errors in estimating X will propagate to L.

To overcome the limitations of previous approaches, we propose the Hybrid Parsing Model (HPM) for unified human parsing and pose estimation by directly estimating MAP over $P(X, L|I)$. The proposed HPM uses Parselets and Mixture of Joint-Group Templates (MJGT) as the semantic part representation (which will be detailed in Section 5.3.2) under the "And-Or" graph framework. This instantiated "And-Or" graph encodes the hierarchical and reconfigurable composition of semantic parts as well as the geometric and compatibility constraints between them. Formally, an HPM is represented as a graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. The edges are defined according to the parent-child relation and "kids(ν)" denotes the children of node ν. Unlike the traditional And-Or

graph, we define four basic types of nodes, namely, "And","Or", "P-Leaf" and "M-Leaf" nodes as depicted in Figure 5.2. Each "P-Leaf" node corresponds to one type of Parselets encoding pixel-wise labeling information, while each "M-Leaf" node represents one type of MJGTs for joint localization. The graph topology is specified by the switch variable t at "Or" nodes, which indicates the set of active nodes $V(t)$. $V^O(t), V^A(t), V^{L_P}(t)$ and $V^{L_M}(t)$ represent the active "Or", "And", "P-Leaf" and "M-Leaf" nodes, respectively. Starting from the top level, an active "Or" node $\nu \in V^O(t)$ selects a child $t_\nu \in \text{kids}(\nu)$. P represents the set of Parselet hypotheses in an image and z denotes the state variables for the whole graph. We then define $z_{\text{kids}(\nu)} = \{z_\mu : \mu \in \text{kids}(\nu)\}$ as the states of all the child nodes of an "And" node $\nu \in V^A$ and let z_{t_ν} denote the state of the selected child node of an "Or" node $\nu \in V^O$.

Based on the above representation, the conditional distribution on the state variable z and the data can then be formulated as the following energy function (Gibbs distribution): The "P-Leaf" component $E^{L_P}(.)$ links the model with the pixel-wise semantic labeling, while the 'M-Leaf" component $E^{L_M}(.)$ models the contribution of keypoints. The "And" component $E^A(.)$ captures the geometry interaction among nodes. The final "Or" component $E^O(.)$ encodes the prior distribution/compatibility of different parts. It is worth noting that there exists pairwise connection at the bottom level in our "And-Or" graph as shown in Figure 5.2. This ensures that more sophisticated pairwise modeling can be utilized to model the connection between/within "P-Leaf" and "M-Leaf" nodes. We approach this by designing the Grid Layout Feature (GLF). The detailed introduction of each component and GLF are given below.

5.3.2 Representation for Semantic Parts

In this subsection, we give details of the representation for the semantic parts. More specifically, we utilize Parselets and Mixture of Joint-Group Templates (MJGT) as the representation for regions and joint groups.

Region Representation with Parselets

Traditional CRF-based approaches for human parsing [47, 81] are inconsistent with structure learning approaches widely used for pose estimation. To overcome this difficulty, we employ the recently proposed Parselets [32] as building blocks for human parsing. In a nutshell, Parselets are a group of semantic image segments with the following characteristics: (1) can generally be obtained by low-level over-segmentation algorithms; and (2) bear strong and consistent semantic meanings. With a pool of Parselets, we can convert the human parsing task into the structure learning problem, which can thus be unified with pose estimation under the "And-Or" graph framework.

As Parselet categorization can be viewed as a region classification problem, we follow [32] by utilizing the state-of-the-art classification pipelines [49, 13] for feature extraction. The parsing node score can then be calculated by

$$E^{L_P}(I, z_\mu) = w_\mu^{L_P} \cdot \Phi^{L_P}(I, z_\mu), \tag{5.1}$$

where $\Phi^{L_P}(.)$ is the concatenation of appearance features for the corresponding Parselet of node μ.

Mixture of Joint-Group Templates

The HoG template based structure learning approaches have shown to be effective for human pose estimation [113, 81, 84]. Most of these approaches treat keypoints (joints) as basic elements. However, joints are too fine-grained for effective interaction with Parselets. Since joints and Parselets have no apparent one-to-one correspondence (*e.g.* , knee joints may be visible or be covered by pants, dress or skirt), direct interaction between all joints (plus additional interpolated keypoints) and the Parselets is almost intractable. Hence, we divide the common 14 joints for pose estimation [113, 81] into 5 groups (*i.e.* left/right arm, left/right leg and head), as shown in Figure 5.3. Each joint group is modeled by one Mixture of Joint-Group Templates (MJGT). MJGT can be regarded as a mixture of pictorial structure models [45, 113] defined on the joints and interpolated keypoints (blue points and green

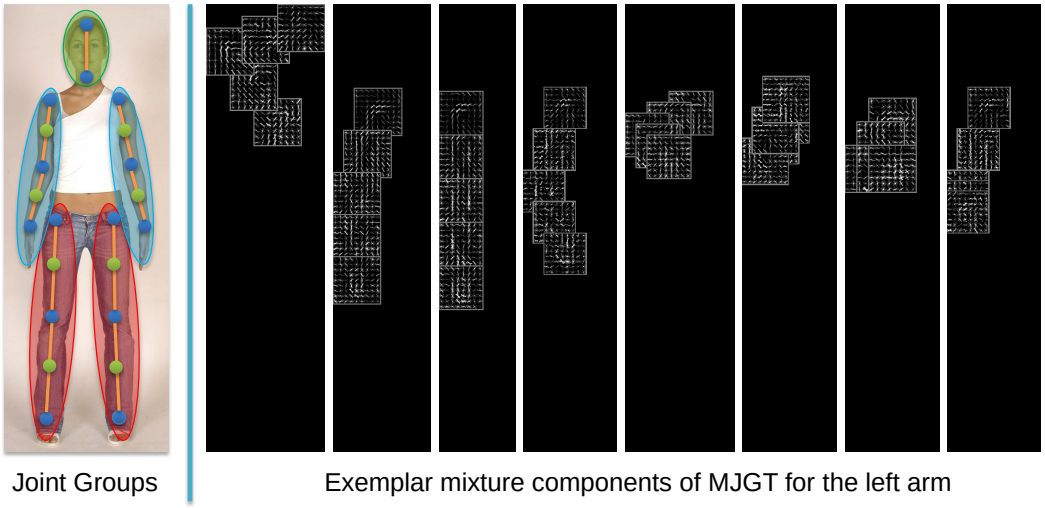


Figure 5.3: The left image shows our joint-group definition (marked as ellipses). Each group consist of several joints (marked as blue dots) and their interpolated points (marked as green dots). We represent each group as one Mixture of Joint-Group Templates (MJGT). Some exemplar mixture components of the MJGT for the right arm are shown on the right side.

points in Figure 5.3). We choose MJGT defined on joint groups as the building block for modeling human pose mainly for three reasons: (1) there are much fewer joint groups than keypoints, which allows more complicated interaction with Parselets; (2) with the reduced complexity in each component brought by the mixture models, we can employ the linear HoG template + spring deformation representation for pictorial structure modeling [113, 84] to ensure the effectiveness of pose estimation; and (3) each component of an MJGT can easily embed mid-level status information (*e.g.* , the average mask).

In practice, we set the number of mixtures as 32/16/16 for MJGT to handle the arms/legs/head group variance respectively. The training data are split into different components based on the clusters of the joint configurations. In addition, an average mask is attached to each component of MJGTs to unify the interaction between Parselet and MJGT, which will be discussed in Section 5.3.3. The state of the instantiated mask for a component of an MJGT is fully specified by the scale and the position of the root node.

For an MJGT model μ, we can now write the score function associated with a

configuration of component m and positions c as in [113, 84]:

$$S_\mu(I,c,m) = b_m + \sum_{i\in\mathcal{V}_\mu} w_i^{\mu,m} \cdot \mathbf{f}_i(I,c_i) + \sum_{(i,j)\in\mathcal{E}_\mu} w_{(i,j)}^{\mu,m} \cdot \mathbf{f}_{i,j}(c_i,c_j), \tag{5.2}$$

where $\mathcal{V}_\mu$ and $\mathcal{E}_\mu$ are the node and edge set, respectively. $\mathbf{f}_i(I,c_i)$ is the HoG feature extracted from pixel location c_i in image I and $\mathbf{f}_{i,j}(c_i,c_j)$ is the relative location $([dx, dy, dx^2, dy^2])$ of joint i with respect to j. Each M-Leaf node can be seen as the wrapper of an MJGT model. Hence the score of M-Leaf is equal to that of the corresponding MJGT model. As the state variable z_μ contains the component and position information for M-Leaf node μ, the final score can be written more compactly as follows:

$$E^{L_M}(I,z_\mu) = w_\mu^{L_M} \cdot \Phi^{L_M}(I,z_\mu), \tag{5.3}$$

where $\Phi^{L_M}(.)$ is the concatenation of the HoG features and the relative geometric features for all the components within the joint group.

5.3.3 Pairwise Geometry Modeling

According to our "And-Or" graph construction, there exist three types of pairwise geometry relations in the HPM: (1) Parselet-Parselet, (2) Parselet-MJGT, and (3) parent-child in "And" nodes. Articulated geometry relation, such as relative displacement and scale, is widely used in the pictorial structure models to capture the pairwise connection. We follow this tradition to model the parent-child interaction (3) as in [113]. However, the pairwise relation of (1) and (2) is much more complex. For example, as shown in Figure 5.4, the "coat" Parselet has been split into two parts and its relation with the "upper clothes" Parselet can hardly be accurately modeled by using only their relative center positions and scales. Furthermore, as Parselets and MJGTs essentially model the same person by different representations, a more precise constraint than the articulated geometry should be employed to ensure their consistency.

To overcome the above difficulties, we propose a Grid Layout Feature (GLF) to model the pairwise geometry relation between two nodes. More specially, as a region mask can be derived from each Parselet or MJGT (the average mask is utilized

for MJGTs), the relation between two nodes can be measured by the pixel spatial distribution relation of their corresponding masks. As illustrated in Figure 5.4, to measure the GLF of mask A with respect to mask B, we first calculate the tight bounding box of A and then divide the whole image into 12 spatial bins, denoted by $R_i, i = 1, \cdots, 12$. The 12 spatial bins consist of 8 cells outside of the bounding box and 4 central bins inside it. We then count the pixels of mask B falling in each bin ($|B \cap R_i|$). Besides the spatial relation, we also model the level of overlap between mask A and B, which has two main functions, *i.e.* (1) to avoid the overlap between Parselets and (2) to encourage the overlap between corresponding Parselets and MJGTs. This is achieved by further counting pixels of the insertion region between A and B in the four central bins ($|A \cap B \cap R_i|$) as shown in Figure 5.4 (c). The resultant 16 dimension feature is normalized by the total pixel number of mask B ($|B|$). By swapping mask A and mask B, we can get another complementary feature centered at the mask B, which is then concatenated with the original one to form the final 32 dimension sparse vector. Formally, we define the Grid Layout Feature as follows:

$$\begin{aligned} PG(A,B) &= \begin{bmatrix} \frac{|B\cap R_i|}{|B|}, i = 1, \cdots, 12; \\ \frac{|A\cap B\cap R_i|}{|B|}, i = 9, \cdots, 12 \end{bmatrix}, \\ \psi_G(A,B) &= [PG(A,B); PG(B,A)], \end{aligned} \tag{5.4}$$

where $\psi_G(A, B)$ is the GLF between mask A and B. With GLF, the interaction between Parseles, such as "coat" and "upper clothes", can be effectively captured. Furthermore, as each mixture component of an MJGT is attached with an average mask, interaction (1) and (2) can be easily unified with the help of GLF.

We can then write out the score of the "And" node, whose child nodes consist of multiple Parselets/MJGTs, as follows:

$$E^A(z_\mu, z_{\text{kids}(\mu)}) = \sum_{\nu \in \text{kids}(\mu)} w^A_{\mu,\nu} \cdot \psi(\mu,\nu) + \sum_{\omega,\upsilon \in \text{kids}(\mu),(\omega,\upsilon)\in E} w^A_{\omega,\upsilon} \cdot \psi_G(\omega,\upsilon), \tag{5.5}$$

where $\psi_G(\omega, \upsilon)$ is the GLF feature between Parselet/MJGT ω and υ. $\psi(\mu, \nu) = [dx \quad dx^2 \quad dy \quad dy^2 \quad ds]^T$ is the articulated geometry feature to measure the geometric difference between part μ and ν, where $dx = (x_\nu - x_\mu)/\sqrt{s_\nu \cdot s_\mu}$, $dy =$

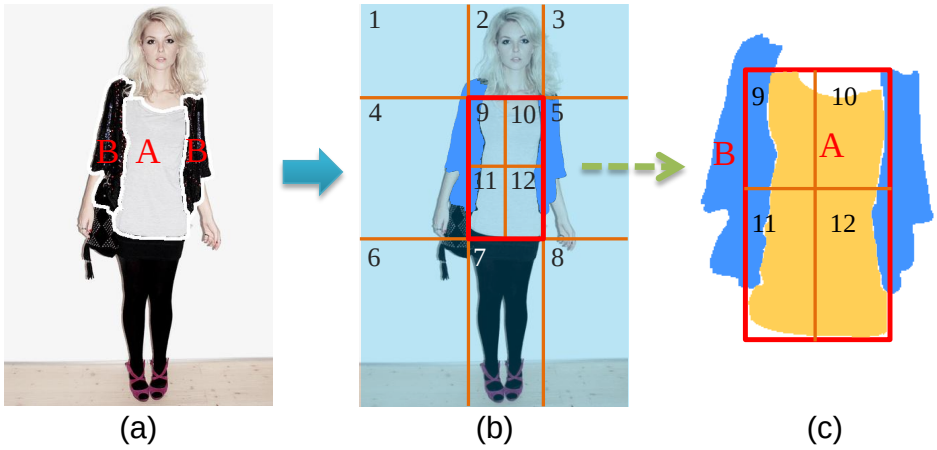


Figure 5.4: Grid Layout Feature (GLF): GLF measures the pixel spatial distribution relation of two masks. To calculate GLF of mask B with respect to mask A, the image is first divided into 12 spatial bins based on the tight bounding box of A as shown in (b), which includes 8 surrounding and 4 central bins. GLF consists of two parts: (1) the ratio of pixels of mask B falling in the 12 bins , and (2) the ratio of pixels of the interaction of mask A and B falling in the 4 central bins as shown in (c).

$(y_\nu - y_\mu)/\sqrt{s_\nu \cdot s_\mu}$ and $ds = s_\nu/s_\mu$ are the relative location and scale of part ν with respect to μ. As the horizontal relations (Parselet-Parselet, Parselet-MJGT) only exist between the "Leaf" nodes under a common "And" node, the GLF term will be removed for those "And" nodes not connected to "Leaf" nodes. By concatenating all geometry interaction features, the score can be written compactly as:

$$E^A(z_\mu, z_{\text{kids}(\mu)}) = w_\mu^A \cdot \Phi^A(z_\mu, z_{\text{kids}(\mu)}). \tag{5.6}$$

5.3.4 Summary

Finally, we summarize the proposed HPM model. For a Parselet hypothesis with index i, its scale (the square root of its area) and centroid can be directly calculated. The switch variable t at "Or" nodes indicates the set of active nodes $V(t)$. The active "And", "Or" and "M-Leaf" nodes have the state variables $g_\nu = (s_\nu, c_\nu)$ which specify the (virtual) scale and centroid of the nodes. The active "P-Leaf" nodes $\nu \in V^{L_P}(t)$ have the state variables d_ν which specify the index of the segments for Parselets, while the active "M-Leaf" nodes $\nu \in V^{L_M}(t)$ have the state variables d_ν which

specify the active component index of the MJGTs. In summary, we specify the configuration of the graph by the states $z = \{(t_\nu, g_\nu) : \nu \in V^O(t)\} \bigcup \{g_\nu : \nu \in V^A(t)\} \bigcup \{d_\nu : \nu \in V^{L_P}(t)\} \bigcup \{(d_\nu, g_\nu) : \nu \in V^{L_M}(t)\}$. The full score associated with a state variable z can now be written as:

$$\begin{aligned} S(I,z) = \sum_{\mu \in V^O(t)} w^O_{\mu,t_\mu} + \sum_{\mu \in V^A(t)} w^A_\mu \cdot \Phi^A(z_\mu, z_{\text{kids}(\mu)}) + \\ \sum_{\mu \in V^{L_P}(t)} w^{L_P}_\mu \cdot \Phi^{L_P}(I, z_\mu) + \lambda \sum_{\mu \in V^{L_M}(t)} w^{L_M}_\mu \cdot \Phi^{L_M}(I, z_\mu), \end{aligned} \tag{5.7}$$

where w^O_{μ,t_μ} measures priors of occurrence for different parts and λ controls the relative weight of the pose and parsing related terms.

5.4 Inference

The inference corresponds to maximizing $S(I,z)$ from Eqn. (5.7) over z. As our model follows the summarization principle [119], it naturally leads to a dynamic programming type algorithm that computes optimal part configurations from bottom to up. As the horizontal relation only exists between the "Leaf" nodes under a common "And" node, if we have already calculated the states of all nodes in the second layer, the following inference can be performed effectively on a tree due to the Markov property of our model. In other words, if we regard all cliques containing an "And" in the second layer and all its child "Leaf" nodes as super nodes, the original model can be converted to a tree model. Hence, the maximization over positions and scales for upper level nodes can be computed very efficiently using distance transforms with linear complexity as in [44].

Since the cycles only exist in the first and second layers, the main computation cost for the proposed model lies in passing the message from "Leaf" nodes to their parent "And" node. However, there are only a limited number of "Leaf" nodes under each "And" node. Furthermore, with the filtering through appearance and spatial constraints, there are usually less than 30 hypotheses for each type of Parselets. Hence, though there are cycles at the bottom level, the algorithm is still reasonably fast.

5.5 Learning

We solve the unified human parsing and pose estimation under the structural learning framework. We follow the setting of [32] to perform the Parselet selection and training. As pose annotation contains no information about mixture component labeling of joint-groups, we derive these labels using k-means algorithm based on joint locations as in [113, 84]. Though such assignment is derived heuristically, it is usually found that treating these labels as latent variables will not improve the performance as these labels tend not to change over iterations [113, 84]. We thus directly use the cluster membership as the supervised definition of mixture component labels for training examples.

As the scoring function of Eqn. (5.7) is linear in model parameters $w = (w^{L_P}, w^{L_M}, w^O, w^A)$, it can be written compactly as $S(I, z) = w \cdot \Phi(I, z)$. Then both pose and parsing related parameters can be learned in a unified framework. Thus we learn all the parameters simultaneously rather than learn local subsets of the parameters independently or iteratively to guarantee the overall performance. Given the labeled examples $\{(I_i, z_i)\}$, the structured learning problem can be formulated in a max-margin framework as in [44]:

$$\begin{aligned} &\min_w \|w\|^2 + C \sum_i \xi_i \\ &\text{s.t.} \quad w \cdot (\Phi(I_i, z_i) - \Phi(I_i, z)) \geq \Delta(z_i, z) - \xi_i, \forall z, \end{aligned} \tag{5.8}$$

where $\Delta(z_i, z_j)$ is a loss function which penalizes the incorrect estimate of z. This loss function should give partial credit to states which differ from the ground truth slightly, and thus is defined based on [81, 32] as follows:

$$\Delta(z_i, z_j) = \sum_{\nu \in V^{L_P}(t_i) \bigcup V^{L_P}(t_j)} \delta(z_i^\nu, z_j^\nu) + \lambda \sum_{\nu \in V^{L_M}(t_i)} \min(2 * \text{PCP}(z_i^\nu, z_j^\nu), 1), \tag{5.9}$$

where $\delta(z_i^\nu, z_j^\nu) = 1$, if $\nu \notin V^L(t_i) \bigcap V^L(t_j)$ or $\text{sim}(d_i^\nu, d_j^\nu) \leq \sigma$. $\text{sim}(\cdot, \cdot)$ is the intersection over union ratio of two segments d_i^ν and d_j^ν, and σ is the threshold, which is set as 0.8 in the experiments. This loss term penalizes both configurations with "wrong" topology and leaf nodes with wrong segments. The second term penalizes the derivation from the correct poses, where $\text{PCP}(z_i^\nu, z_j^\nu)$ is the average

PCP score [47] of all points in the corresponding MJGT. The optimization problem Eqn. (5.8) is known as a structural SVM, which can be efficiently solved by the cutting plane solver of SVMStruct [61] and the stochastic gradient descent solver in [44].

5.6 Experiments

5.6.1 Experimental Settings

Dataset: Simultaneous human parsing and pose estimation requires annotation for both body joint positions and pixel-wise semantic labeling. Traditional pose estimation datasets, such as the Parse [113] and Buffy [47], are of insufficient resolution and lack the pixel-wise semantic labeling. Hence we conduct the experiments on two recently proposed human parsing datasets as in the previous chapters.

Evaluation Criteria: There exist several competing evaluation protocols for human pose estimation throughout the literature. We adopt the probability of a correct pose (PCP) method described in [113], which appears to be the most common variant. Unlike pose estimation, human parsing is rarely studied and with no common evaluation protocols. Here, we utilize two complementary metrics (APA and the same IoU based metrics as the previous chapter) to allow direct comparison with previous works [111, 32].

Implementation Details: We use the same definition of Parselets and settings for feature extraction as in [32]. The dense SIFT, HoG and color moment are extracted as low-level features for Parselets. The size of Gaussian Mixture Model in FK is set to 128. For pose estimation, we follow [113] by using the 5×5 HoG cells for each template. The training : testing ratio is 2:1 for both datasets as in [32]. The penalty parameter C and relative weight λ are determined by 3-fold cross validation over the training set.

5.6.2 Experimental Results

To the best of our knowledge, there are few works handling human parsing and pose estimation simultaneously. Hence, besides the recent representative approach [111], which performs parsing and pose estimation iteratively, we also compare the proposed method with the state-of-the-art methods designed for each task separately.

Human Pose Estimation: For human pose estimation, as the experiments are conducted on these two new datasets, we only compare with several state-of-the-art methods with publicly available codes for retraining [113, 111]. The comparison results are shown in Table 5.1. Method [111] utilizes the results of [113] as initial estimation of pose for human parsing. The parsing results are then fed back as additional features to re-estimate the pose. However, the improvement of [111] over [113] is marginal probably because of its sequential optimization nature. As the error from initial pose estimation results will propagate to parsing, it is difficult for the re-estimation step to rectify the initial pose results from error-propagated parsing results. On the contrary, we perform human parsing and pose estimation simultaneously, which significantly improves the state-of-the-art performance [113, 111]. We also evaluate the raw MJGT baseline which only utilizes the MJGT representation and removes the Parselet from the "And-Or" graph. The worse results compared with the full HPM model verify the advantages of joint parsing and pose estimation.

Figure 5.5 shows some qualitative comparison results. It can be seen that all other methods fail in cases where joints are occluded by clothing, *e.g.* , wearing long dress or skirt. By contrast, with the help of Parselets and the pairwise constraints brought by the GLF, the proposed method can still obtain reasonable joint positions.

Human Parsing: For human parsing, we compare the proposed framework with the works [111] and [32]. In terms of APA, our method achieves 87% for FS dataset and 88% for DP dataset, which are superior to 86% and 87% of the current leading approach [32]. The improvement is not significant as APA metric is dominated by the background. Even naively assigning all segments as background results in a reasonably good APA of 78% for DP and 77% for FS. Therefore, the more

Table 5.1: Comparison of human pose estimation PCP scores on FS and DP datasets.

method	[113]	[111]	raw MJGT	HPM	[113]	[111]	raw MJGT	HPM
dataset	FS	FS	FS	FS	DP	DP	DP	DP
torso	100.0	99.6	100.0	99.5	99.8	99.8	99.8	99.8
ul leg	94.2	94.1	91.9	95.3	91.2	92.0	90.0	95.5
ur leg	93.0	95.1	91.6	95.6	93.9	94.2	92.3	96.4
ll leg	90.9	89.6	83.9	92.2	90.3	90.9	89.0	93.3
lr leg	90.1	91.9	82.5	92.7	90.0	90.0	88.7	92.7
ul arm	86.5	85.8	80.4	89.9	89.1	89.5	85.6	92.4
ur arm	85.2	86.9	81.1	90.9	88.8	88.7	85.7	91.7
ll arm	62.3	62.1	54.7	69.6	66.9	68.2	60.4	72.8
lr arm	61.9	63.6	58.2	69.7	61.7	62.6	48.0	69.3
head	99.2	99.3	97.5	99.1	99.5	99.5	99.6	99.7
avg	86.3	86.8	82.2	**89.5**	87.1	87.5	83.9	**90.4**

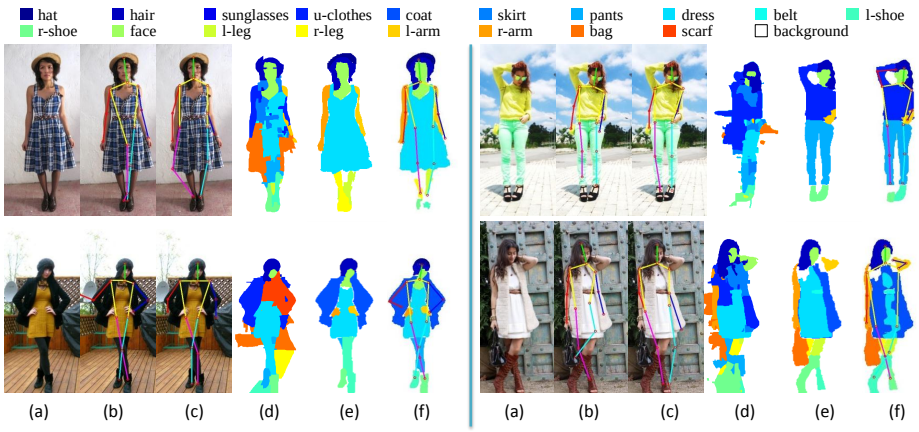


Figure 5.5: Comparison of human parsing and pose estimation results. (a) input image, (b) pose results from [113], (c) pose results from [111], (d) parsing results from [111], (e) parsing results from [32], and (f) our HPM results are shown sequentially.

discriminative IoU criterion is more suitable to measure the real performance of each algorithm. The detailed comparison results in terms of IoU are shown in Table 5.2. It can be seen that our framework is consistently better than other methods across different datasets and metrics. This significant improvement mainly comes from the complementary nature of two tasks and the strong pairwise modeling, which verifies the effectiveness of our unified parsing and pose estimation framework.

Some example human parsing results are shown in Figure 5.5. It can be observed

Table 5.2: Comparison of human parsing IoU scores on FS and DP datasets.

method	[111]	[32]	HPM	[111]	[32]	HPM
dataset	FS	FS	FS	DP	DP	DP
hat	2.5	5.6	7.9	1.3	28.9	26.4
hair	47.2	67.9	70.8	43.5	74.8	74.2
s-gls	0.8	2.8	2.6	0.6	9.6	8.3
u-cloth	36.4	56.3	59.5	21.3	42.5	47.9
coat	null	null	null	19.5	39.4	43.6
f-cloth	23.2	56.6	58.0	21.8	61.0	64.7
skirt	21.6	55.3	56.3	12.2	50.3	53.6
pants	19.1	40.0	48.3	28.7	66.3	70.7
belt	8.9	18.2	16.6	4.8	16.6	17.2
l-shoe	27.6	58.6	58.9	25.6	57.0	59.7
r-shoe	25.2	53.4	51.8	21.7	51.8	53.0
face	59.3	72.4	76.1	52.6	78.1	78.9
l-arm	33.0	52.7	56.7	32.4	62.7	67.9
r-arm	30.5	45.4	50.3	28.3	59.3	64.7
l-leg	32.6	48.8	52.6	23.5	52.6	55.1
r-leg	24.1	41.6	41.5	18.4	35.5	39.9
bag	9.5	20.6	17.7	8.5	12.7	16.2
scarf	0.9	1.2	2.3	1.2	9.3	6.6
aIoU	23.8	41.0	**42.8**	20.3	44.9	**47.1**
wIoU	29.9	51.7	**54.3**	24.6	53.0	**56.4**

that the sequential approach [111] performs much worse than ours. This may be owing to the errors propagated from the inaccurate pose estimation results as well as the lack of the ability to model the exclusive relation of different labels, which usually leads to cluttered results. Though this method can achieve much better performance with the additional information about the type of clothes in the target image as illustrated in [111], such information is usually difficult to obtain for real applications. Our method also outperforms the baseline [32], which has obvious artifacts for persons with joint crossed (*e.g.* , legs and foot). The lack of top-down information makes it difficult for the method [32] to distinguish the left shoe from the right shoe. On the contrary, by jointly modeling human parsing and pose estimation, our model can achieve reasonably good results for these cases. In addition, as the method [32] does not explicitly model the overlap between Parselets, the resultant Parselets may occlude each other seriously. For example, the "dress" Parselet is

badly occluded by the "coat" Parselet in the right-bottom image. With the help of GLF, our unified model can effectively avoid the severe overlap of Parselets and thus leads to more promising results.

Finally, we want to emphasize that our goal is to explore the intrinsic correlation between human parsing and pose estimation. To achieve this, we propose the HPM which is a unified model built upon the unified representation and the novel pairwise geometry modeling. Separating our framework into different components leads to inferior results as demonstrated in Table 5.1 and 5.2. Though we use more annotations than methods for individual tasks, the promising results of our framework verify that human parsing and pose estimation are essentially complementary and thus performing two tasks simultaneously will boost the performance of each other.

5.7 Chapter Summary

In this chapter, we present a unified framework for simultaneous human parsing and pose estimation, as well as an effective feature to measure the pairwise geometric relation between two semantic parts. By utilizing Parselets and Mixture of Deformable Templates as basic elements, the proposed Hybrid Parsing Model allows joint learning and inference of the best configuration for all parameters. The proposed framework is evaluated on two benchmark datasets with superior performance to the current state-of-the-arts in both cases, which verifies the advantage of joint human parsing and pose estimation.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

This thesis explored the intrinsic consistency among different tasks for visual recognition. In the previous chapters, we have been through several topics that expand the frontier of recognition along four directions, contextualizing object classification and detection with subcategory awareness, performing joint object detection and semantic segmentation, exploring human parsing via a unified approach and conducting joint human parsing and pose estimation. Below, we will summarize the main content and contributions of the thesis.

In Chapter 2, Looking Inside Category: Subcategory-aware Object Recognition, we designed a system to integrate the state-of-the-art object classification and detection techniques for joint object detection and classification. The detailed experiments have revealed that the proposed contextualized framework significantly outperformed the current leading approaches designed for individual tasks [34, 18]. This performance improvement partially comes from the complementary properties of two tasks. The success in one task can be employed to rectify the ambiguous results in the other task. It was also found that subcategory structure was common for most categories in current object recognition datasets. However, previous works usually ignored such informative subcategory structure and thus represented each category by a monolithic model. To fully utilize the embedded subcategory struc-

ture for each category, we proposed an ambiguity guided subcategory mining. Ambiguity guided subcategory mining results was then seamlessly integrated into the subcategory-aware detection assisted object classification framework. The overall system has achieved the state-of-the-art performance on the Pascal VOC benchmark dataset, which clearly demonstrated the effectiveness of proposed subcategory-aware contextualized strategy [34].

In Chapter 3, Towards Unified Object Detection and Semantic Segmentation, we presented a unified approach for joint object detection and segmentation. The experiments have shown that the proposed approach achieved promising performance in the popular Pascal VOC benchmark [33]. The main contributions of this paper are three-fold. First, our holistic model is able to improve performance for both tasks, which verifies that the two core tasks for visual recognition are highly correlated. By properly integrating classical algorithms designed for different levels of recognition, the resulting pipelines should further improve the state-of-the-art visual recognition system. Second, we have provided detailed quantitative and qualitative analysis for the role of each component, which explains why these tasks are completely and how they benefit each other. Finally, our unified approach has provided an invaluable way to understand visual recognition in a bigger picture.

In Chapter 4, A Deformable Mixture Parsing Model with Parselets, we proposed an effective framework for human parsing. By reconsidering the human parsing problem, we utilized the novel Parselets as the basic elements for our parsing model. Then, a unique Deformable Mixture Parsing Model (DMPM) was built to jointly learn and infer the best configuration for both appearance and structure effectively over the generated pool of Parselets. Extensive experimental results have clearly demonstrated the effectiveness of the proposed framework [32]. Besides providing an elegant solution for the important human parsing problem, we also showed how to tailor the leading techniques of recognition for real applications in a principled way.

In Chapter 5, Towards Unified Human Parsing and Pose Estimation, we present a unified framework for simultaneous human parsing and pose estimation. By uti-

lizing Parselets and Mixture of Deformable Templates as basic elements for human pose estimation and human parsing respectively, both tasks can be effectively unified under the proposed Hybrid Parsing Model. To better measure the pairwise geometric relation between two semantic parts, we further proposed an effective Grid Layout Feature. Thanks to the unification of human parsing and pose estimation, the resulting Hybrid Parsing Model allows joint learning and inference of the best configuration for all parameters to guarantee the overall performance. The proposed framework is evaluated on two benchmark datasets with superior performance to the current state-of-the-arts in both cases, which verifies the advantage of joint human parsing and pose estimation.

6.2 Future Works

Though great success has been achieved for core visual recognition tasks by the proposed systems, it should be noticed that the works in this thesis still have several limitations:

- First, all the experiment results are obtained on the itemized datasets. Though these benchmark datasets are well designed, they are still far from enough to represent the real world. Hence, the performance of the proposed framework for real world applications has not been fully tested.

- Second, the proposed frameworks focus on the general problems of visual recognition. However, the specific applications may bear particular features that are not well captured by current framework. Thus, possible modification may be needed to adapt to specific applications.

- Finally, due to the prevalence of depth sensor, such as Kinect, the depth information can now be obtained easily. However, this thesis focused on the traditional RGB images only, as depth information is still impossible to obtain for many applications, such as web image retrieval.

Based on the limitations observed, there are several directions that can be further explored:

- First, current frameworks highly depend on hand-crafted feature for representation of various tasks. However, the deep architecture has shown to achieve great success for automatic feature learning during the past few years. Hence, it might be promising to embed the automatic feature learning, which can naturally generate the feature suitable for the target task, into the current framework.

- Though the proposed approaches have achieved promising performance for many applications, they usually rely on complicated feature extraction/inference algorithms and may be too slow for real-time applications or mobile devices. Hence, besides continuing to improve the final performance, we also plan to improve the efficiency of current framework.

- Finally, with exciting advantages achieved in benchmark datasets, it might be the right time to touch the real world applications. Thus, we would like to build several customer oriented systems, such as visual product search and clothes retrieval, based on current visual object recognition techniques.

These directions are worthwhile to take for both research and industry.

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[2] O. Aghazadeh, H. Azizpour, J. Sullivan, and S. Carlsson. Mixture component identification and learning for visual recognition. In *European Conference on Computer Vision*. 2012.

[3] O. Aghazadeh, H. Azizpour, J. Sullivan, and S. Carlsson. Mixture component identification and learning for visual recognition. In *European Conference on Computer Vision*. 2012.

[4] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[5] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[6] A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[7] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. Serrat, and

J. Gonzàlez. Harmony potentials. *International Journal of Computer Vision*, 2012.

[8] I. M. Bomze. Branch-and-bound approaches to standard quadratic optimization problems. *Journal of Global Optimization*, 2002.

[9] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*, 2009.

[10] L. D. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision*, 2010.

[11] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *Conference on Computer Vision and Pattern Recognition*, 2011.

[12] T. Brox, L. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *Conference on Computer Vision and Pattern Recognition*, 2011.

[13] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision*. 2012.

[14] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[15] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. International Conference on Computer Vision, 2013.

[16] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[17] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

[18] K. Chatfield, V. Lempitsky, and A. Vedaldi. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[19] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European Conference on Computer Vision*. 2012.

[20] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan. Hierarchical matching with side information for image classification. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[21] S. chun Zhu and D. Mumford. A stochastic grammar of images. In *Foundations and Trends in Computer Graphics and Vision*, 2006.

[22] R. G. Cinbis, J. Verbeek, C. Schmid, et al. Segmentation driven object detection with fisher vectors. In *International Conference on Computer Vision*, 2013.

[23] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[24] J. Dai, J. Feng, and J. Zhou. Subordinate class recognition using relational object models. In *International Conference on Pattern Recognition*, 2012.

[25] J. Dai, S. Yan, X. Tang, and J. T. Kwok. Locally adaptive classification piloted by uncertainty. In *International Conference on Machine Learning*, 2006.

[26] Q. Dai and D. Hoiem. Learning to localize detected objects. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[27] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[28] S. K. Divvala, A. A. Efros, and M. Hebert. How important are "deformable parts" in the deformable parts model? In *European Conference on Computer Vision Workshops*, 2012.

[29] S. K. Divvala, A. A. Efros, and M. Hebert. Object instance sharing by enhanced bounding box correspondence. In *British Machine Vision Conference*, 2012.

[30] J. Dong, Q. Chen, J. Feng, K. Jia, Z. Huang, and S. Yan. Looking inside category: subcategory-aware object recognition. *Transactions on Circuits and Systems for Video Technology*, 2014.

[31] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. 2014.

[32] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *International Conference on Computer Vision*, 2013.

[33] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and segmentation. In *European Conference on Computer Vision*. 2014.

[34] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan. Subcategory-aware object classification. In *Conference on Computer Vision and Pattern Recognition*, 2013.

[35] I. Endres and D. Hoiem. Category independent object proposals. *European Conference on Computer Vision*, 2010.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes

Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2010.

[40] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008.

[41] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 2007.

[42] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[43] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[44] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[45] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005.

[46] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 2012.

[47] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[48] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *Conference on Computer Vision and Pattern Recognition*, 2013.

[49] J. S. Florent Perronnin and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *European Conference on Computer Vision*, 2010.

[50] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[51] R. B. Girshick, P. Felzenszwalb, and D. Mcallester. Object detection with grammar models. In *Neural Information Processing Systems*. 2011.

[52] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[53] C. Gu, P. A. Arbeláez, Y. Lin, K. Yu, and J. Malik. Multi-component models for object detection. In *European Conference on Computer Vision*, 2012.

[54] C. Gu, J. J. Lim, P. Arbeláez, and J. Malik. Recognition using regions. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[55] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *European Conference on Computer Vision*, 2010.

[56] H. Hajishirzi, M. Rastegari, A. Farhadi, and J. Hodgins. Understanding of proffesional soccer commentaries. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

[57] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic

contours from inverse detectors. In *International Conference on Computer Vision*, 2011.

[58] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *International Conference on Computer Vision*, 2009.

[59] M. Hoai and A. Zisserman. Discriminative sub-categorization. In *Conference on Computer Vision and Pattern Recognition*, 2013.

[60] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European Conference on Computer Vision*. 2012.

[61] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 2009.

[62] T.-K. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[63] P. Kohli, J. Rihan, M. Bray, and P. H. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *International Journal of Computer Vision*, 2008.

[64] M. P. Kumar, P. Ton, and A. Zisserman. Obj cut. In *Conference on Computer Vision and Pattern Recognition*, 2005.

[65] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[66] L. Ladickỳ, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr. What, where and how many? combining object detectors and crfs. In *European Conference on Computer Vision*. 2010.

[67] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Conference on Computer Vision and Pattern Recognition*, 2006.

[68] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *International Conference on Computer Vision*, 2009.

[69] C. Li, D. Parikh, and T. Chen. Extracting adaptive contextual cues from unlabeled regions. In *International Conference on Computer Vision*, 2011.

[70] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[71] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[72] H. Liu and S. Yan. Robust graph mode seeking by graph shift. In *International Conference on Machine Learning*, 2010.

[73] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACM Multimedia*, 2012.

[74] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[75] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[76] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *International Conference on Computer Vision*, 2011.

[77] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[78] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 1996.

[79] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *European Conference on Computer Vision*, 2010.

[80] O. M. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman. The truth about cats and dogs. In *International Conference on Computer Vision*, 2011.

[81] B. Rothrock, S. Park, and S.-C. Zhu. Integrating grammar and segmentation for human pose estimation. 2013.

[82] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *European Conference on Computer Vision*, 2012.

[83] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Conference on Computer Vision and Pattern Recognition*, 2006.

[84] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2013.

[85] F. H. Sinz, O. Chapelle, A. Agarwal, and B. Schölkopf. An analysis of inference with the universum. In *Neural Information Processing Systems*, 2007.

[86] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *Conference on Computer Vision and Pattern Recognition*, 2011.

[87] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *International Conference on Computer Vision*, 2011.

[88] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *European Conference on Computer Vision*. 2012.

[89] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *International Journal of Computer Vision*.

[90] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[91] P. H. Torr and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. 2013.

[92] L. Torresani, M. Szummer, and A. W. Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision*, 2010.

[93] M. Toussaint and S. Vijayakumar. Learning discontinuities with products-of-sigmoids for switching between local models. In *International Conference on Machine Learning*, 2005.

[94] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *European Conference on Computer Vision*. 2010.

[95] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.

[96] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *International Conference on Computer Vision*, 2011.

[97] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *International Conference on Computer Vision*, 2009.

[98] H. Wang and D. Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2011.

[99] J. Wang, J. Yang, K. Yu, F. Lv, and T. Huang. Locality-constrained linear coding for image classification. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[100] X. Wang and T. Zhang. Clothes search in consumer photos via color matching and attribute learning. In *ACM Multimedia*.

[101] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *Conference on Computer Vision and Pattern Recognition*, 2011.

[102] Y. Wang, D. Tran, Z. Liao, and D. Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *Journal of Machine Learning Research*, 2012.

[103] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, June 2014.

[104] J. Weibull. *Evolutionary game theory*. MIT press, 1997.

[105] J. Weston, R. Collobert, F. H. Sinz, L. Bottou, and V. Vapnik. Inference with the universum. In *International Conference on Machine Learning*, 2006.

[106] W. Xia, C. Domokos, L. F. Cheong, and S. Yan. Background context augmented hypothesis graph for object segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014.

[107] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan. Semantic segmentation without annotating segments. 2013.

[108] W. Xia, Z. Song, J. Feng, L. F. Cheong, and S. Yan. Segmentation over detection by coupled global and local sparse representations.

[109] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[110] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. In *Conference on Computer Vision and Pattern Recognition*, 2013.

[111] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[112] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes. Layered object models for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[113] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Conference on Computer Vision and Pattern Recognition*, 2011.

[114] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *International Conference on Machine Learning*, 2009.

[115] J. Yuen, C. L. Zitnick, C. Liu, and A. Torralba. A framework for encoding object-level image priors. Technical report, Microsoft Research Technical Report.

[116] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 2003.

[117] L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille. Max margin and/or graph learning for parsing the human body. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[118] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman. Latent hierarchical structural learning for object detection. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[119] L. L. Zhu, Y. Chen, C. Lin, and A. Yuille. Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation. *International Journal of Computer Vision*, 2011.

[120] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Conference on Computer Vision and Pattern Recognition*, 2012.

[121] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection? In *British Machine Vision Conference*, 2012.

Publication List

[1] J. Dong, Q. Chen, J. Feng, K. Jia, Z. Huang, and S. Yan. Looking inside category: subcategory-aware object recognition. *Transactions on Circuits and Systems for Video Technology*, 2014.

[2] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and segmentation. In *European Conference on Computer Vision*. 2014.

[3] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *Conference on Computer Vision and Pattern Recognition*, 2014.

[4] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *International Conference on Computer Vision*, 2013.

[5] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan. Subcategory-aware object classification. In *Conference on Computer Vision and Pattern Recognition*, 2013.

[6] J. Dong, B. Cheng, X. Chen, T. Chua, S. Yan, and X. Zhou. Robust image annotation via simultaneous feature and sample outlier pursuit. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2013.

[7] J. Dong, Y. Ni, J. Feng, and S. Yan. Purposive hidden-object game (P-HOG) towards imperceptible human computation. In *ACM Multimedia*, 2011.

[8] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan. Contextualiz-

ing object detection and classification. *Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[9] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan. Semantic segmentation without annotating segments. 2013.

[10] J. Feng, Y. Ni, J. Dong, Z. Wang, and S. Yan. Purposive hidden-object-game: Embedding human computation in popular game. *Transactions on Multimedia*, 2012.

[11] M. Xu, B. Ni, J. Dong, Z. Huang, M. Wang, and S. Yan. Touch saliency. In *ACM Multimedia*, 2012.

[12] S. Liu, Q. Chen, J. Dong, S. Yan, C. Xu, and H. Lu. Snap & play: auto-generate personalized find-the-difference mobile game. In *ACM Multimedia*, 2011.

[13] Y. Mu, J. Dong, X. Yuan, and S. Yan. Accelerated low-rank visual recovery by random projection. In *Conference on Computer Vision and Pattern Recognition*, 2011.

Rewards

- 2014, Winner Prize of the detection task in ImageNet ILSVRC2014 Challenge
- 2013, President Graduate Fellowship
- 2013, Runner-up Prize of the classification task in ImageNet ILSVRC2013 Challenge
- 2012, Winner Prize of the classification task in PASCAL VOC2012 Challenge
- 2012, Winner Prize of the segmentation task in PASCAL VOC2012 Challenge
- 2012, Winner Prize of the ICPR-HARL Human activities recognition and localization competition