Beyond Visual Words: Exploring Higher-level Image Representation for Object Categorization

Yan-Tao Zheng

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the NUS Graduate School For Integrative Sciences and Engineering

NATIONAL UNIVERSITY OF SINGAPORE

2010

Abstract

Beyond Visual Words: Exploring Higher-level Image Representation
for Object Categorization

Yan-Tao Zheng

Category-level object recognition is an important but challenging research task. The diverse and open-ended nature of object appearance makes objects, no matter from the same category or otherwise, possess boundless variation in visual looks and shapes. Such visual diversity leads to a huge gap between visual appearance of images and their semantic content. This thesis aims to tackle the issues of visual diversity for better object categorization, from two aspects: visual representation and learning scheme.

One contribution of the thesis is in devising a higher-level visual representation, visual synset. Visual synset is built on top of traditional bag of words representation. It incorporates the co-occurring and spatial scatter information of visual words to make it more descriptive to discriminate images of different categories. Moreover, visual synset leverages the "probabilistic semantics" of visual words, i.e. their class probability distributions, to group ones with similar distribution into one visual content unit. In this way, visual synset can partially bridge the visual differences of images of same class and leads to a more coherent image distribution in the feature space.

The second contribution of the thesis is in developing a generative learning model that goes beyond image appearances. By taking a Bayesian perspective,

we interpret visual diversity as a probabilistic generative phenomenon, in which the visual appearance arises from the countably infinitely many common appearance patterns. To make a valid learning model for this generative interpretation, three issues must be tackled: (1) there exist countably infinitely many appearance patterns, as the objects have limitless variation of appearance; (2) the appearance patterns are shared not only within but also across object categories, as the objects of different categories can be visually similar too; and (3) intuitively, the objects within a category should share a closer set of appearance patterns than those of different categories. To tackle these three issues, we propose a generative probabilistic model, *nested hierarchical Dirichlet process (HDP) mixture*. The stick breaking construction process in the nested HDP mixture provides the possibility of countably infinitely many appearance patterns that can grow, shrink and change freely. The hierarchical structure of our model not only enables the appearance patterns to be shared across object categories, but also allows the images within a category to arise from a closer appearance pattern set than those of different categories.

Experiments on Caltech-101 and NUS-WIDE-object dataset demonstrate that the proposed visual representation, *visual synset*, and learning scheme, *nested HDP mixture*, in the thesis can deliver promising performance and outperform existing models with significant margins.

Contents

List of Figures

List of Tables

Acknowledgments

This thesis would have not been possible, or at least not what it looks like now, without the guidance and help of many people.

Foremost, I would like to show my sincere gratitude to my advisor, Prof. Tat-Seng Chua. It was March in the year of 2006, when Prof. Chua took me into his research group. From then, I have embarked on the endeavor on multimedia and computer vision research. For the past four years, I have appreciated Prof. Chua's seemingly limitless supply of creative ideas, insight and ground-breaking visions on research problems. He has offered me with invaluable and insightful guidance that directed my research and shaped this dissertation without constraining it. As an exemplary teacher and mentor, his influence has been truly beyond the research aspect of my life.

I also like to thank my co-advisor, Dr. Qi Tian, for his encouragement and constructive feedback on my work. During my Ph.D pursuit, Dr. Tian has always been providing insightful suggestion and discerning comments to my research work and paper drafts. His suggestion and guidance have helped to improve my research work.

Many lab mates and colleagues have helped me, during my Ph.D pursuit. I like to thank Ling-Yu Duan, Ming Zhao, Shi-Yong Neo, Victor Goh, Huaxing Xu, and Sheng Tang for the inspiring brainstorming, valuable suggestion and enlightening feedbacks on my work.

Last but not least, I would like to thank all of my family, my parents Weimin and Lihua, my sister Jiejuan and my wife Xiaozhuo. For their selfless care, endless love and unconditional support, my gratitude to them is truly beyond words.

To my parents and my wife ...

Chapter 1

Introduction

Visual object categorization is a process in which a computing machine automatically perceives and recognizes objects in images at category level, such as airplane, car, boat, etc. As one of the core research problems, visual categorization has spurred much research attention in both multimedia and computer vision community. Visual categorization yields semantic descriptors for visual contents of images and videos. These semantic descriptor has profound significance in effective image indexing and search, video semantic understanding and retrieval and robot vision systems [138, 85, 73, 113, 86].

1.1 The visual representation and learning

The ultimate goal of visual categorization system is to emulate the function of Human Visual System [11] to perform accurate recognition on a multitude of object categories in images. However, due to the biological complexity of human brain, the human visual and perceptual process remains obscure. The uncertain biological and psychological processes make the machine emulation of these cognitive processes not feasible. Rather than replicating the human vision system, researchers attempt to capture the principles of this biological intelligence. The human visual system allows individuals to quickly recognize and assimilate information from the visual perception. This complicated cognitive process consists of two major steps [76], as shown in Figure 1.1. First, the lens of the eye projects an image of the surroundings onto the retina in the back of the eye. The role of retina is to convert the pattern of light into neuronal signals. At this point, the visual perception of an individual has been represented in a form that is readable by human intelligence system. Next, the brain receives these neuronal signals and processes them in a hierarchical fashion by different parts of the brain, and finally, recognizes the content of the visual surroundings.

From the computational perspective, this human visual perception can be restated as a process in which the eye, like a sensor, perceives and transforms surroundings into a set of signals and the brain, like a processor, learns and recognizes these signals. Inspired by this fact, researchers approach the visual categorization in a methodology comprising of two major modules: visual representation and learning [11, 134]. To some extent, this methodology is consistent with *Marr's Theory* [75] in 3-D object recognition setting, in which the vision process is regarded as an information processing task. The visual representation specifies the explicit interpretation of visual cues that an image contains, while the algorithm (or learning) module governs how the visual cues are manipulated and processed for visual content understanding and recognition.

Figure 1.1 shows the overall flow of this modular and sequential methodology of visual categorization. The significance of this methodology is that it sketches the contour for designing visual recognition systems. Many researchers working on visual recognition systems have organized their research effort according to this methodology, by focusing either on representation or learning or both.

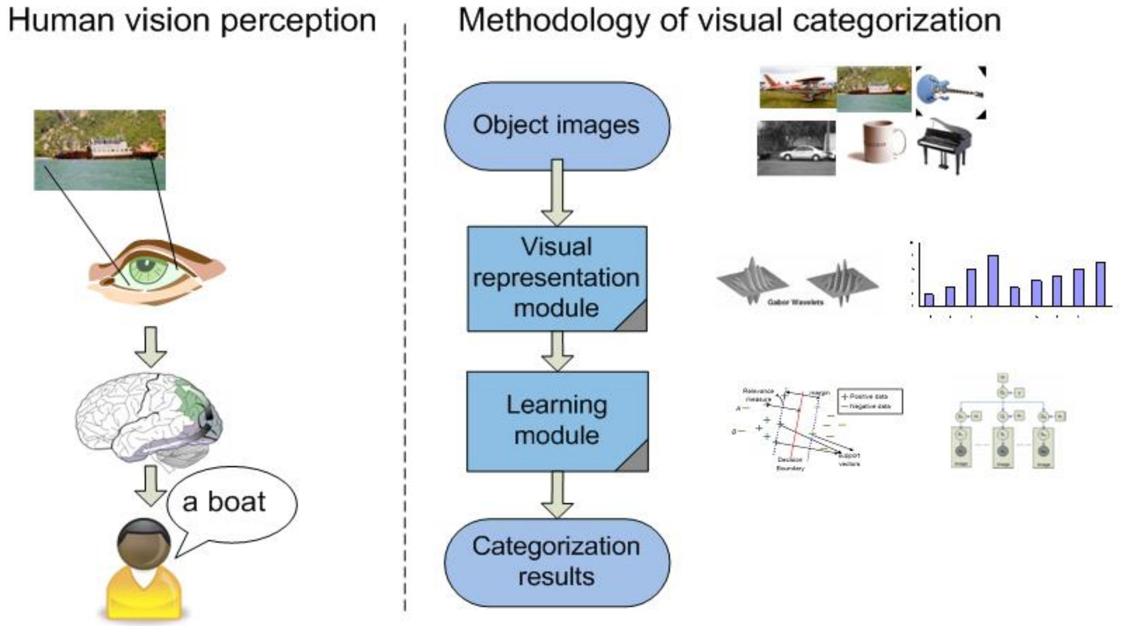


Figure 1.1: The human vision perception and the methodology of visual categorization. Similar to the human vision perception, the methodology of visual categorization consists of two sequential modules: representation and learning.

1.1.1 How to represent an image?

To identify the content of an image, the eye of human perceives and represents it in the form of neuronal signals for the brain to perform subsequent analysis and recognition. Similarly, computer vision and image processing represent the information of an image in the form of visual features. The visual features for visual categorization can be generally classified into two types: global feature representation and local feature representation. The global feature representations describe an image as a whole, while the local features depict the local regional statistics of an image [37].

Earlier research efforts on visual recognition have focused on global feature representation. As the name suggests, the global representation describes an image as a whole, in a global feature vector [62, 74, 68]. The global features are image-based or grid-based ordered features, such as the color or texture histogram over the whole image or grid [74]. Examples of global representations include a histogram of color or grayscale, a 2D histogram of edge strength and orientation, a set of responses to a group of filter banks, and so on [68]. Up to date, the global features have been extensively used in many applications, because of their attractive properties and characteristics. First, the global features produce very compact representations of images. This representation compactness enables efficiency in subsequent learning processes. Second, in general, the global feature extraction processes are efficient with reasonable computational complexity. This property make global features especially popular in online recognition systems that need to process input images on the fly. More importantly, by generalizing an entire image into a single feature vector, the global representation renders the existing similarity metric, kernel matching and machine learning techniques readily applicable on the visual categorization and recognition task.

Despite of the aforementioned strength, the global features suffer from the following drawbacks. First, the global features are sensitive to scale, pose and image capturing condition changes. Consequently, they fail to provide adequate description on an image's local structure and appearance. Second, global features are sensitive to clutter and occlusion. As a result, it is either assumed that an image only contains a single object, or that a good segmentation of the object from the background is available [68]. However, in reality, either of these two scenarios seldom exist. Third, the global representation assumes that all parts of images contribute to the representation equally [68, 37]. This makes it sensitive to the background or occlusion. For example, a global representation on an image of an airplane could be more reflective on the background sky, rather than the airplane itself.

Due to the aforementioned disadvantages of global features, much research efforts have been motivated towards some visual representation that are more resilient to scale, translation, lighting variations, clutter and occlusion. Recently, local features have attracted much research attention, as they tackle the weaknesses of global features in part, by exploiting the local regional statistics of image patches to describe an image [37, 105, 60, 58, 59, 25, 3]. The part-based local features are a set of descriptors of local image neighborhoods computed at homogeneous image regions, salient keypoints and blobs, and so on [35, 37, 111]. Compared to global features, the part-based local representations are more robust, as they code the local statistics of image parts to characterize an image [37]. The part-based local representation decomposes an image into its component local parts (local regions) and describes the image by a collection of its local region features, such as Scale Invariant Feature Transform (SIFT) [72]. It is resilient to both geometric and photometric variations, including changes in scale, translation, view point, occlusion, clutter and lighting conditions. The overlapped extraction of local regions is equivalent to extensively sampling the spatial and scale space of images, which enables the local regions to be robust to scale and translation changes. The local regions correspond to small parts of objects or background, which makes them resilient to clutter and occlusion. Moreover, the variability of small regions is much less than that of whole images [119]. This renders the region descriptor, such as Scale Invariant Feature Transform (SIFT) [72], to be capable of canceling out the effects caused by lighting condition changes.

1.1.2 Visual categorization is about learning

Paralleled by cognitive science and neuroscience studies, the visual recognition and categorization are usually formulated as a task of learning on visual representation of images. This formulation brings an essential linkage between visual categorization and the paradigm of pattern recognition and machine learning. Hence, the visual categorization research is naturally rooted in the mathematical foundations of pattern analysis and machine learning. In the setting of statistical learning, the visual categorization is cast as a supervised learning and classification task on image representation.

In general, the statistical learning methods for visual categorization can be classified into two types: discriminative and generative learning. To distinguish discriminative and generative learning, we assume an image $\mathcal{I}$ with feature X to be classified to one of m categories $\mathcal{C} = \{c_i\}_{i=1}^m$, as shown in Figure 1.2. In a Bayesian setting, this classification task can be characterized as modeling posterior probability $p(c \mid X)$. Once probability $p(c \mid X)$ are known, classifying image $\mathcal{I}$ to category c with maximum $p(c \mid X)$ gives the optimal categorization decision, in the sense that it minimizes the expected loss or Bayes risk.

To categorize the unseen images, the generative learning approach estimates the joint probability $P(X; c)$ of image feature variables and object category variable [69, 55]. This estimation can be factored to computing the category prior probabilities $p(c)$ and the class-conditional densities $p(X \mid c)$ separately, according to the Bayes' rule. The posterior probabilities $p(c \mid X)$ are then obtained using the Bayes' theorem

$$p(c \mid X) = \frac{p(X \mid c)p(c)}{\sum_c P(X \mid c)P(c)} \tag{1.1}$$

In general, the generative learning approaches assume a generative image formation process, in which the image feature variables arise from a joint probability. This generative modeling of image formation provides the possibility to explicitly identify the causal structure of image features [55]. It also helps to reveal what variables are important to emulate human vision psychophysical processes. Generally, the generative approaches characterize the inter-relation of all relevant variables in terms of a probabilistic graph. The graph also helps to interpret how the joint probability is factored into the conditional probabilities [55]. The causal

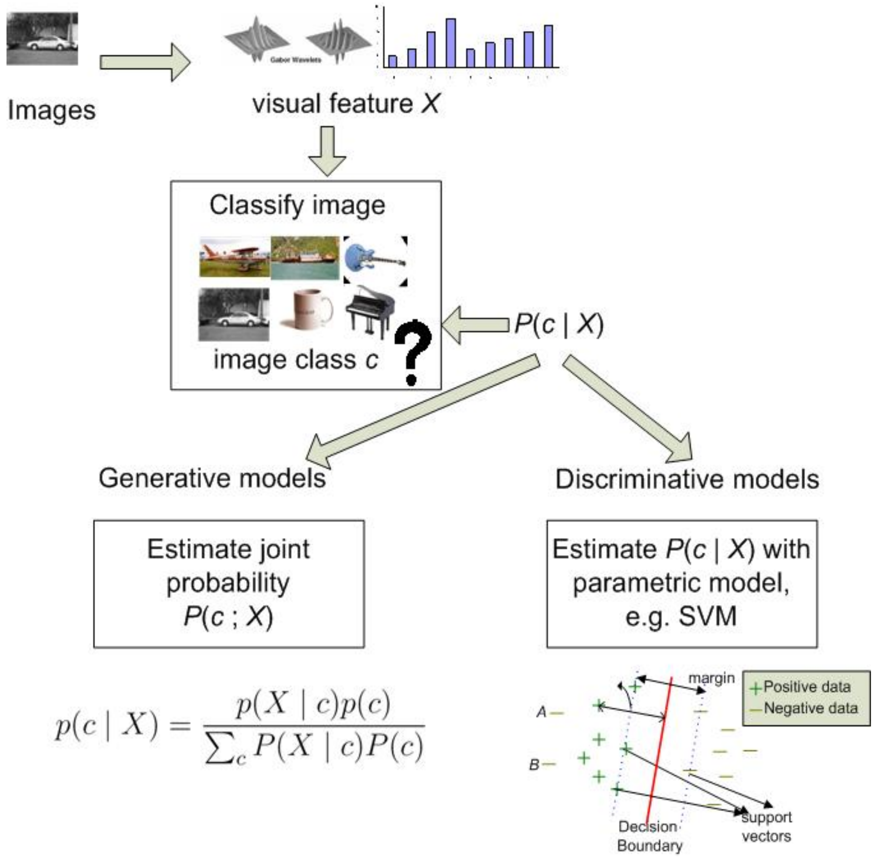


Figure 1.2: The generative learning v.s. discriminative learning. Generative learning focuses on estimating $P(X; c)$ in a probabilistic model, while the discriminative learning focuses on implicitly estimating $P(c \mid X)$ via a parametric model.

relationship defined in the graph can function as constraints to alleviate the inference computation.

In contrast to generative models, the discriminative approaches do not model the joint probability, but the posterior probability $P(c \mid X)$. Instead of explicitly estimating the density of the posterior probability, many approaches utilize a parametric model to optimize a mapping from image feature variables to object category variable. The parameters in the model can then be estimated from the labeled training data. One popular and relatively successful example is the support vector machine (SVM) [120, 59, 135]. In the task of visual categorization, SVM attempts to capture the distinct visual characteristics of different object categories, by finding the maximum margin between them in the image feature space. It tends to have good performance, when different visual categories have large inter-class variation.

Despite of their promising practical performance, the discriminative methods suffer from two major critic. First, the discriminative methods attempt to learn the mapping between input and output variables only, rather than unveiling the probabilistic structure of either the input or output domain [18]. This attempt is theoretically ill-advised, as the probabilistic structure can reveal the inter-relation among input image feature variables and output category variables, and therefore, help the system to categorize new unseen images [18]. Second, in general, the discriminative methods often require large amount of training data to produce good classifier, while the generative approaches usually need lesser supervision and manual labeling to deliver stable categorization performance [115].

In summary, the generative learning approach categorizes object images, by estimating the joint probability model of all the relevant variables, including image feature variables and object category variable [69, 55, 119]. In contrast, the discriminative approaches adopt a direct attempt to build a classifier that perform well on the training data, by circumventing the modeling of the underlying distributions [49, 69, 88].

1.2 The half success story of bag-of-words approach

Recently, one of the part-based local features, namely the bag-of-words (BoW) image representation, has achieved notably significant results in various multimedia and vision tasks. Sivic el at. [105] and Nister and Stewenius [90] demonstrated

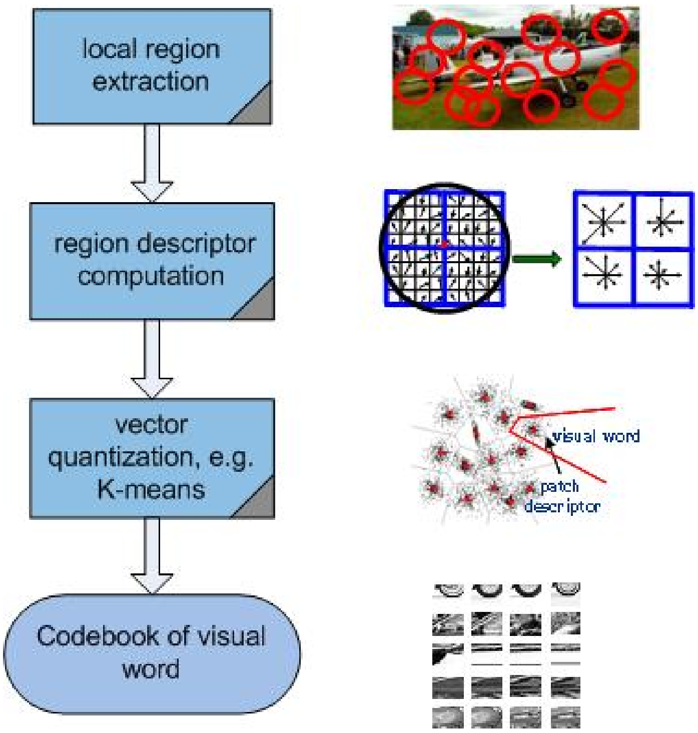


Figure 1.3: The overall flow of the bag-of-words image representation generation.

that the bag-of-words representation is able to deliver state-of-the-art performance in image retrieval, both in terms of accuracy and efficiency. Zhang el at. [136], Lazebnik el at. [58] and many other researchers [130, 25, 3] showed that the bag-of-words approaches give top performance in visual categorization evaluation, such as PASCAL-VOC. Moreover, Jiang el at. [50] and Zheng el at. [141] also exhibited that the bag-of-words approach outperforms other global or semi-global visual features in the high level feature detection in TRECVID evaluation. The simplicity, effectiveness and good practical performance of bag-of-words approach have made it one of the most popular and widely used visual features for many multimedia and vision tasks [130, 136, 59, 53]. Analogous to document representation in terms of words in text domain, the bag-of-words approach models an image as a geometry-free unordered collection of visual words.

Figure 1.3 shows the overall flow of bag-of-words image representation generatation. As shown in Figure 1.3, the first step of generating bag-of-words representation is extracting local regions in a given image $\mathcal{I}$. This step determines which part of local information will be coded to represent the image. After extraction of M local regions $\{a_i\}_{i=1}^{M}$ from image $\mathcal{I}$, the region descriptor, such as Scale Invariant Feature Transform (SIFT) [72], is computed over the region. A vector quantization process, such as k-means clustering, is then applied on the region descriptors to generate a codebook of W visual words $\mathbf{W} = \{w_1, .., w_W\}$. Each of the descriptor cluster corresponds to one visual word in the visual vocabulary. The image $\mathcal{I}$ then can be represented by a collection of visual words $\{w_{(a_1)}$,..., $w_{(a_i)}, ..\}$. The bag-of-words representation has been demonstrated to be resilient to variations in scale, translation, clutter, occlusion, and object pose, etc. The appealing properties of bag-of-words approach are attributed to its local coding of image statistics. Extensive sampling of local regions enables the bag-of-words representation to be robust to scale and translation changes. Describing local regions of an image also makes the representation resilient to clutter and occlusion. Moreover, the local region descriptor, such as Scale Invariant Feature Transform (SIFT) [72], makes the bag-of-words approach robust to lighting condition changes.

1.3 What are the challenges?

Though various systems have shown promising practical performances of bag-of-words approach [36, 124, 130, 136, 59, 53], the accuracies of visual object categorization are still incomparable to its analogue in text domain, i.e. the document categorization. The reason is obvious. The textual word possesses semantics and the documents are well-structured data regulated by grammar, linguistic and lexicon rules. In contrast, there appears to be no well-defined rule in visual word composition of images. The open-ended nature of object appearance makes objects, no matter from the same or different categories, have huge variation of visual

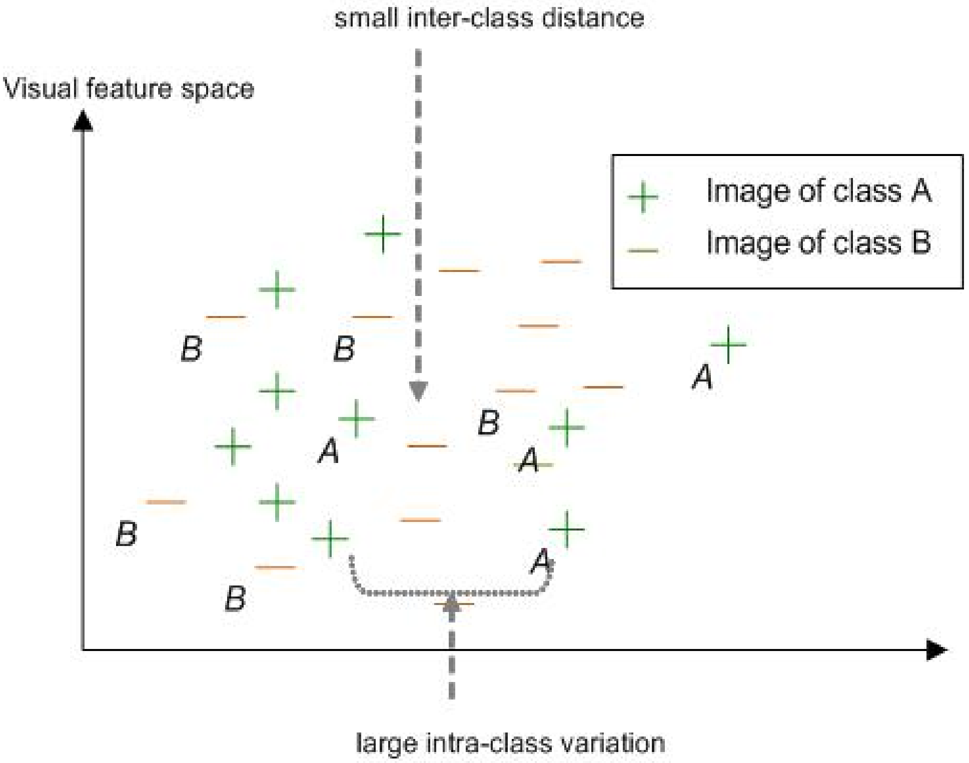


Figure 1.4: A toy example of image distributions in visual feature space. The semantic gap between image visual appearances and semantic contents is manifested by two phenomena: large intra-class variation and small inter-class distance.

looks and shapes. Such huge object appearance diversities lead to sparse correlation between visual proximity of object images and their semantic relevance. The visual features, such as bag-of-words, color histogram, wavelet texture, etc, are, therefore, not sufficiently capable to model the image semantics. This gap between visual proximity of images and semantic relevance also makes most statistical and machine learning models ineffective in visual object recognition. This gap is well known as *the semantic gap*. From the perspective of statistics, the direct consequences of this semantic gap are the large intra-class variation and small inter-class distances, as shown in Figure 1.4.

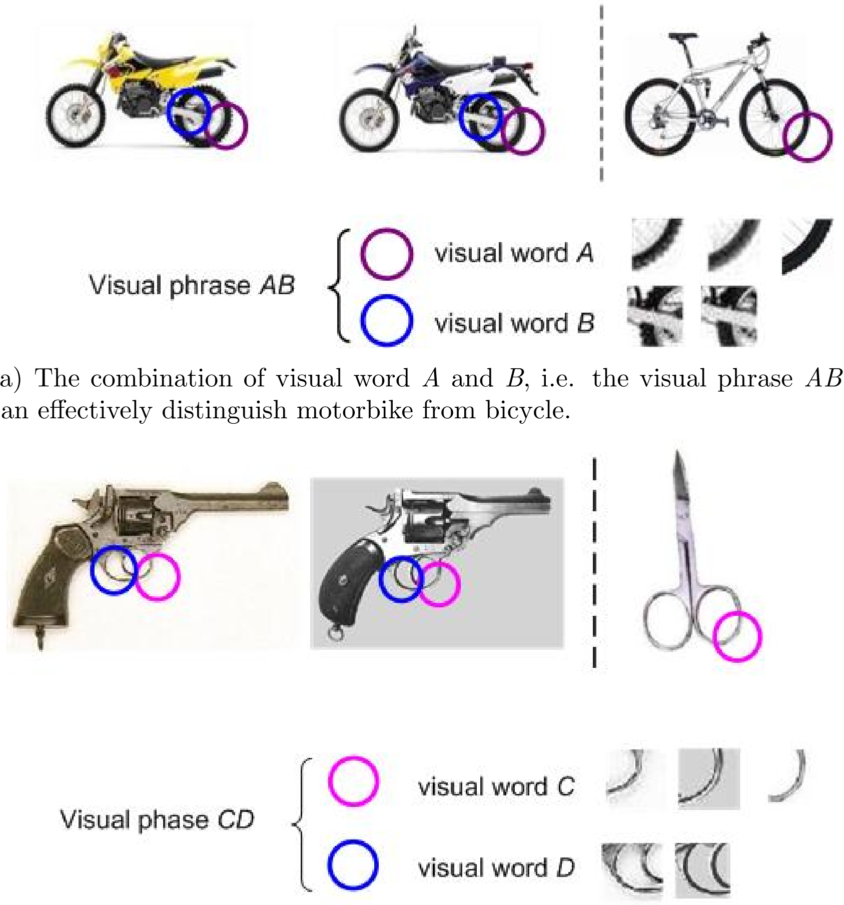
In the context of bag-of-words image representation, the gap between visual proximity of images and their semantic relevance can be regarded a form of ambiguity and uncertainty of visual information representation [132, 133]. This representation uncertainty is manifested by two phenomena: polysemy and synonymy. The polysemous visual word is a one that might represent different semantic meanings in different image context, while the synonymous words are a set of visually dissimilar words representing the same semantic meaning. By sharing a set of polysemous visual words, the semantically dissimilar images might be close to each other in feature space, while the synonymous visual words may cause the images with the same semantic to be far apart in the feature space.

1.4 A higher-level visual representation

To achieve more effective object categorization, a higher-level visual content unit is demanded so as to tackle the polysemy and synonymy issues caused by visual diversity.

Polysemy issue

Polysemy encumbers the distinctiveness of visual words and leads to under-representations [132], [133]. Its consequence is effectively low inter-class discrimination. The polysemy is rooted from two reasons. First, visual word is the result of vector quantization (clustering of region descriptors) and each visual word corresponds to a group of local regions. Due to visual diversity, it is impossible to make regions of one visual word with homogeneous appearances. Such quantization error inevitably results in ambiguity of visual word representation. Second, the regions represented in a visual word might come from the object parts with different semantics but similar local appearances. For example in Figure 1.5 (a), visual word A is not able to distinguish motorbike from bicycle, as they share visually similar tires. However, the combination of visual word A and B, i.e. the visual phrase AB, can effectively



(a) The combination of visual word A and B, i.e. the visual phrase AB, can effectively distinguish motorbike from bicycle.

(b) The combination of visual word C and D, i.e. the visual phrase CD, can effectively distinguish pistol from scissor.

Figure 1.5: The combination of visual words bring more distinctiveness to discriminate object classes.

distinguish motorbike from bicycle. The polysemy issue can, therefore, be resolved by mining inter-relation among visual words in certain neighborhood region. Yuan el at. [133] and Quack el at. [99] proposed to utilize frequently co-occurring visual word-set to address the polysemy issue. Specifically, Yuan el at. denote such visual word-set as visual phrase. The major weakness of visual phrase approach is that it merely considers the co-occurrence information among visual words but neglect spatial information among them. To tackle such issue, we propose a new visual

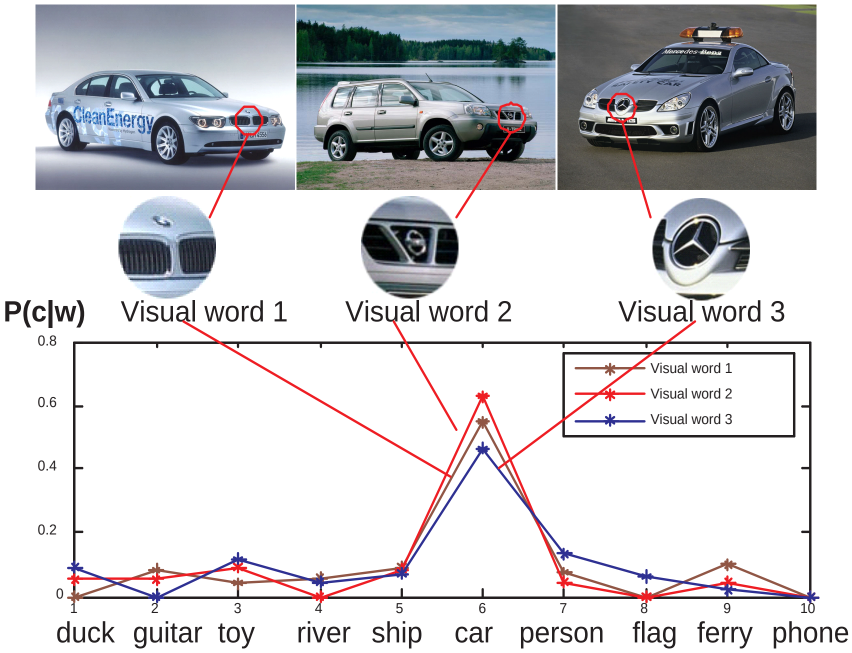


Figure 1.6: Example of visual synset that clusters three visual words with similar image class probability distributions.

descriptor - delta visual phrase, which incorporates both co-occurrence and spatial scatter information of visual words.

Synonymy issue

The synonymy is attributed to the visual diversity of objects of same semantic class. Such appearance diversity makes multiple visual words share same or similar semantic meaning. It is, in fact, an over-representation of semantics by visual words [132, 133]. The consequence is large intra-class variations. In this circumstance, both visual words and phrases become too primitive to effectively model the image semantics, as their efficacy depends highly on the visual similarity and regularity of images of same semantics. To tackle this issue, a higher level visual content unit is needed. In text domain, when documents of same topic or categories are represented by different sets of words, the word synset (synonymy set) that link words of similar semantics are robust to model them [10]. Inspired by this, we propose a novel visual content unit, *visual synset*, on top of visual words and phrases. We define *visual synset* as a relevance-consistent group of visual words or phrases with similar semantics. However, it is hard to measure the semantics of a visual word or phrase, as they are only a quantized vector of sampled regions of images. Rather than in a conceptual manner, we define the 'semantics' probabilistically as semantic inferences $P(c_i|w)$ of visual word or phrase w towards image class c_i.

Intuitively, if several visual words or phrases from different images share similar class probability distribution, like the brand logos in car images shown in Figure 1.6, then the visual synset that clusters them together shall possess similar class probability distribution and distinctiveness towards image classes. The visual synset can then partially bridge the visual differences between these images and deliver a more coherent, robust and compact representation of images.

1.5 Learning beyond visual appearances

The open-ended nature of object appearance and the resulting *semantic gap* have posed significant challenges to learning schemes for visual categorization in two aspects. First, objects of different classes can share similar visual appearances. This visual similarity leads to objects of different categories sharing similar visual features, which consequently makes them appear in close proximity in the visual feature space. In this case, the same visual feature pattern over-represents more than one semantics, which is, in essence, an ambiguity issue of visual representation [132, 143, 140]. The primary consequence is the small inter-class distance for objects of different categories. Second, the objects of the same classes can have different visual appearance. Such appearance diversity makes objects of same category have distinct visual features and distributed far apart in the visual feature space. In this case, multiple visual feature patterns may correspond to the same semantics.

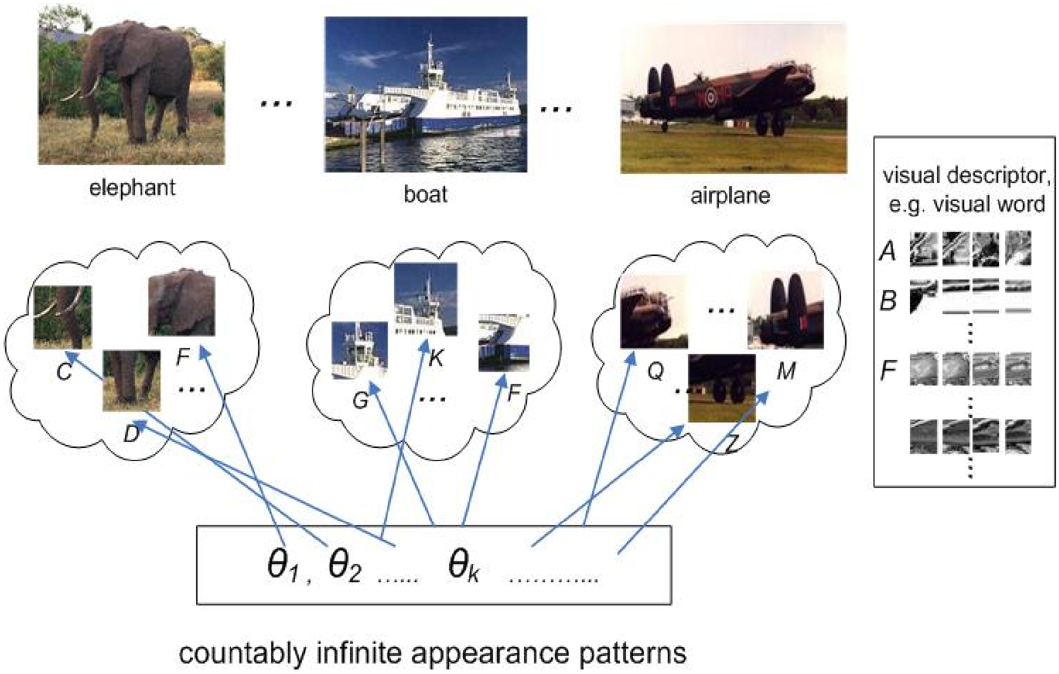

elephant
boat
airplane
visual descriptor,
e.g. visual word
A
B
F
C
F
D
G
K
F
Q
M
Z
θ_1, θ_2 ...... θ_k ....
countably infinite appearance patterns

Figure 1.7: The generative interpretation of visual diversity, in which the visual appearances arise from countably infinitely many appearance patterns.

This is an under-representation or uncertainty issue of visual feature. Hence, the objects of the same category may have a large intra-class variation [132, 143]. Consequently, the visual diversity leads to a low correlation or large gap between image proximity in the visual feature space and their semantic relevance. which, in fact, is one of the causes of the well known "semantic gap" problem.

The visual diversity of objects and its resulting semantic gap have presented a harsh reality to learning schemes: it is usually difficult to learn the visual characteristics of object categories for classification, as most object categories generally do not have any distinct visual characteristics. Therefore, rather than directly modeling object visual content, we need some learning scheme that goes beyond visual appearances. As we know, the open-ended nature of object appearance brings in the huge variation of visual appearances. We interpret the unbounded object appearance diversity as a generative phenomenon, in which the diverse visual appearances arise from countably infinitely many common *visual appearance pattern*s, as shown in Figure 4.2. In this probabilistic generative interpretation, different object categories can still be visually similar and share similar visual appearance patterns. However, the distribution and combination of appearance patterns can be distinct for different object categories. The object categorization can then be cast as a problem of analyzing the distribution and combination of appearance patterns or the visual thematic structure of object categories. Effectively, the objects of same class that are visually different can be adjacent in visual appearance pattern space. Hence, the appearance patterns can bridge the visual appearance difference of objects in part.

However, to make the aforementioned generative interpretation valid, three issues must be tackled. (1) There should exist countably infinitely many appearance patterns, as the object visual diversity is boundless. (2) All the object categories should share a universal set of visual appearance patterns, as the objects of different categories can be visually similar too. (3) Intuitively, the objects of same category should possess a closer set of appearance patterns than those of different categories. To embody the generative interpretation of object appearance, we tackle the three aforementioned issues by developing a hierarchical generative probabilistic model, named **nested hierarchical Dirichlet process (HDP) mixture**. The stick breaking construction process and Chinese restaurant franchise representation [117] in the proposed nested HDP mixture model allow the countably infinitely many appearance patterns to be shared within and across different object categories. The designed model structure also enables the images of the same category to possess a closer set of appearance patterns.

1.6 Contributions

The thesis focuses on developing a higher-level visual representation and a new generative probabilistic learning method for visual categorization. The main contributions of the thesis are as follows.

1. Visual synset: a higher-level visual representation

In order to address the polysemy and synonymy issue of visual words, we propose a novel visual content unit, *visual synsets*. To address the polysemy issue, we exploit the co-occurrence and spatial scatter information of visual words to generate a more distinctive visual compositional configuration, i.e. delta visual phrase. The improved distinctiveness leads to better inter-class distance.

To tackle the synonymy issue, we proposed to group delta visual phrase with similar 'semantics' into a visual synset. Rather than in conceptual manner, the 'semantic' of a delta visual phrase is probabilistically defined as its image class probability distribution. The visual synset is therefore a probabilistic relevance-consistent cluster of delta visual phrases, which is learned by Information Bottleneck based distributional clustering.

2. Nested HDP mixture: a learning scheme beyond visual appearances

To further recognize objects beyond their visual appearance, we adopt a generative interpretation of object appearance diversity, in which visual appearances arise from countably infinitely many common appearance patterns. To embody this interpretation, we propose a generative probabilistic model, called **nested HDP mixture**, by tackling the following three issues in the interpretation: (1) there should exist countably infinitely many appearance patterns, as the object visual diversity is boundless; (2) all the object categories should share a universal set of visual appearance patterns, as the objects of different categories can be visually similar too; (3) intuitively, the objects of same category should possess a closer set of appearance patterns than those of different categories.

1.7 Outline of the thesis

Chapter 2 introduces the background knowledge and reviews the literature on visual representation and categorization models, that are relevant to or share similar vision with the thesis.

Chapter 3 presents the proposed higher-level visual representation, *visual synset*, for visual categorization. It first delves into the process to construct the proposed compositional feature, *delta visual phrase*, based on frequently co-occurring visual word-set with similar spatial scatter. Then it presents the construction of *visual synset*, based on the probabilistic 'semantics', i.e. class probability distribution, of delta visual phrases.

Chapter 4 details the proposed generative probabilistic learning framework, *nested hierarchical Dirichlet process (HDP) mixture*, to perform image categorizations beyond visual appearances. The proposed HDP mixture model learns the common appearance patterns from diverse object appearances and performs categorization based on the pattern models.

Chapter 5 discusses the experimental observations and results on two large scale image datasets: Caltech-101 [63] and NUS-wide-object dataset [23].

Chapter 6 concludes the thesis with highlight of contributions of this thesis.

Chapter 2

Background and Related Work

This thesis is relevant to a range of research topics, including compositional feature mining, distributional clustering, generative probabilistic models, etc. This chapter serves to introduce the necessary background knowledge and concepts before delving deep into the proposed models. As some related work are also the rudimentary elements of the proposed models, this Chapter presents the related work and background together on two dimensions: image representation and statistical learning schemes for visual categorization.

2.1 Image representation

2.1.1 Global feature

From the global image feature representation in earlier research work to the more advanced part-based local feature representation in recent research efforts, the image representation for visual categorization has gone through significant evolution. The earlier global features include color, texture and shape features. Due to the simplicity and good practical performance, these visual features are still being widely used in many research tasks and systems, such as content based image retrieval

[102], visual categorization, and high level feature detection in TRECVID evaluation [109], etc. Here, we briefly review color and texture feature representation.

Color

The color feature has been one of the most widely used visual features. It has the relative advantages of robustness to background complication and invariance to image size and orientation [102]. Among color features, color histogram is the most commonly used. It depicts the pixel statistics in color spaces, which include RGB, LAB, LUV, HSV and YCrCb. From the perspective of Bayesian, color histogram denotes the joint probability of the pixel intensities of the three color channels. One variation of color histogram is the cumulated color histogram proposed by Stricker and Orengo [114], which aims to address the sparsity issue in color histogram.

Stricker and Orengo proposed the color moments approach to alleviate the quantization issue in color histogram. The rational of color moments lies in the fact that the color distribution can be characterized by its moments. Specifically, most commonly used moments are the low-order ones, such as the first moment (mean), and the second and third central moments (variance and skewness).

To capture the spatial correlation of colors, Huang et al. proposed the color correlogram [46]. Rather than simple intensity distribution, the color correlogram encodes (1) the spatial correlation of colors and; (2) the global distribution of local spatial correlation of colors. Informally, a color correlogram of an image depicts the probability of finding a pixel of a given color i at a given distance k from a pixel of a given color j. For computational simplicity, color i and j are usually set to be the same. The resulting feature is called autocorrelogram, which effectively depicts the global distribution of local spatial correlations of the pixels with the same color.

Please refer to [81, 48, 89, 126] for complete study of color visual features.

Texture

Texture denotes the visual patterns that have properties of repeatability and homogeneity, such as interwoven elements, threads of fabric, and so on [41]. It is not the consequence of single color or intensity, but the visual property of object surfaces [110]. In other words, it depicts the "structural arrangement of surfaces and their relationship to the surrounding environment".

Texture features encode several types of visual information: (1) spectral features, which include Gabor texture and wavelet texture; (2) statistical features, which cover six Tamura texture features; and (3) the wold features. Among the various texture features, the Gabor texture and wavelet texture are widely studied and used for image retrieval, visual categorization and other multimedia and vision tasks [22, 110]. Especially, the wavelet texture features have been reported to well match the perception of human vision, and therefore, wavelet transform in texture representation has been well studied in recent years [22, 110]. Smith and Chang [110] proposed a texture representation, based on the statistics (mean and variance) extracted from the wavelet subbands. Chang and Kuo [22] explored the middle-band characteristics, a tree-structured wavelet transform, to construct texture representation. For a more complete review on texture features, please refer to [102, 101, 110].

2.1.2 Local feature representation

The major drawback of global features is that they are sensitive to scale, pose and image capturing condition changes. On the other hand, the part-based local image representations, such as bag of local features, have shown robustness and resilience in photometric and geometric image variations, such as changes in scale, translation, lighting condition, viewpoint, occlusion and clutter, in part [59, 68]. In general, the local regions in part-base representation are obtained by identifying

Table 2.1: List of commonly used local region detection methods.

Method	**Description**
Difference of Gaussian (DoG)	Detect regions at local scale-space maxima of the difference-of-Gaussian. It detects blob-like local image neighborhoods [72].
Laplacian of Gaussian (LoG)	Build scale-space representation by successive smoothing of image with Gaussian based kernels and detect blob-like image structures [67].
Harris-Laplace	Detect regions via the scale adapted Harris function and the Laplacian-of-Gaussian operator in scale-space. It yields corner-like regions [78].
Hessian-Laplace	Detect regions of the local maxima of the Hessian determinant at space at and the local maxima of the Laplacian-of-Gaussian in scale [80].
Harris-Affine	Detect regions via the scale invariant Harris detector and extract affine shape of a keypoint neighborhood [78].
Hessian-Affine	Similar to Harris-Affine detector. The difference is that Hessian-Affine detector chooses interest points based on the Hessian matrix [78].
Salient region detector	Detect regions of local maxima of the entropy at scale-space. The entropy of pixel intensity histograms is measured for circular regions of various size at each image position [54].
Maximally Stable Extremal Regions (MSER)	Detect regions of homogenous color [77].
Dense random region detector	Randomly extract a large number of regions from an image [91].

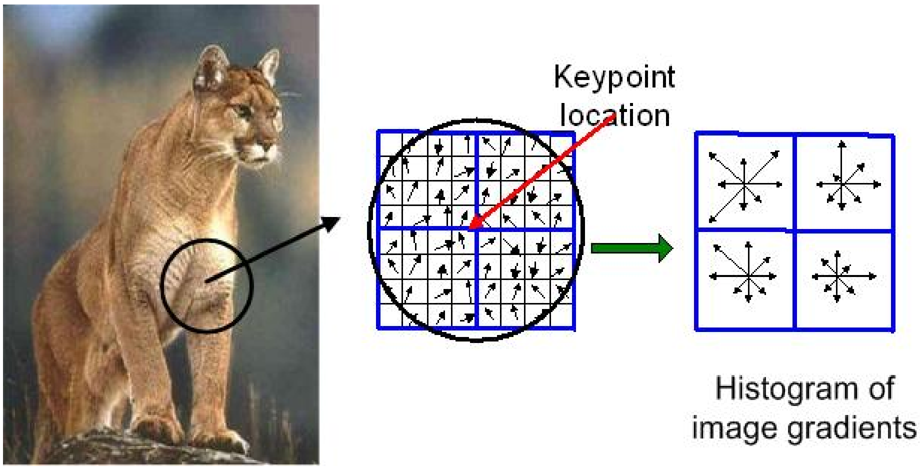


Figure 2.1: SIFT is a normalized 3D histogram on image gradient, intensity and orientation (1 dimension for image gradient orientation and 2 dimensions for spatial locations).

homogeneous image regions, local neighborhood of salient keypoints or blobs in the image. Ideally, the local region identification process should possess two properties: (1) minimizing the intra-class variations caused by geometric and photometric changes, such as different scale, lighting conditions, viewpoints, etc, (or maximizing the local similarities of images) by providing most repeatable regions among images in the same class; and (2) maximizing the inter-class variations by sampling discriminative local image regions. Towards these two goals, researchers have developed many local region extraction algorithms, such as Difference of Gaussian [72], Harris-Laplace [78], Maximally Stable Extremal Regions (MSRE) [27] , based on color or geometric saliency of keypoints or regions. Table 2.1 lists the most commonly used region detection methods and brief description of their characteristics.

For each detected local region, a feature descriptor (vector) is computed. There exist several local region descriptors, such as Gradient Location and Orientation Histogram (GLOH) [79], Scale Invariant Feature Transform [71, 72], Speeded Up Robust Features (SURF) [9], and so on. Among the various feature descriptors, the Scale Invariant Feature Transform (SIFT), developed by Lowe

[72], has been one of the most widely used descriptors. As shown in Figure 2.1, SIFT is basically a normalized 3D histogram on image gradient, intensity and orientation (1 dimension for image gradient orientation and 2 dimensions for spatial locations). The nature of image gradient (intensity difference of neighboring pixels) makes SIFT resilient to illumination changes. SIFT is also used as local feature in our model in the thesis. Among the part-base local representation, bag-of-words representation is the most widely used and has attracted much research attention, which will be introduced in the subsequent Section.

2.1.3 The bag-of-words approach

Among all the part-based local representations, bag-of-words image representation has been one of the most popular approaches and spurred much research attention due to its simplicity, computational efficiency and good practical performance [105, 60, 58, 59, 25, 3]. Following the analogy of document representation in text domain, the bag-of-words approach represent an image as an orderless bag of visual words. Though it does not incorporate any geometric structure or spatial information, the bag-of-words representation has achieved notably significant results in various multimedia and computer vision tasks, such as image retrieval [105, 90], visual categorization [136, 58, 59, 25, 3] and high level feature detection in TRECVID evaluation [50, 141].

The idea of adapting text categorization approaches to visual categorization can be traced back to the work in [144], in which Zhu et al. explored the vector quantization of small square image windows, named ”keyblocks”, to represent images. They showed that these quantized ”keyblocks” features, together with the ”well-known vector-, histogram-, and n-gram-models of text retrieval”, can deliver more ”semantics oriented” results than color and texture based approaches [25].

The bag-of-words representation has been previously utilized on texture classification [61, 122, 57]. In the texture classification task, the cluster of local features, or visual word, has another name, "texton". Recently, researchers have promoted the usage of bag-of-words approach to other tasks. Sivic and Zisserman exploited the bag-of-words representation for object and scene image (and keyframe) retrieval, by borrowing the retrieval mechanism in the text retrieval analogy. Same as the text analogy, the approach represents an image as an orderless bag of visual words that are generated via vector quantization on affine-invariant regions. The bag-of-words image representation enables the image retrieval system to utilize all the available feature weighting schemes, like tf-idf, and indexing techniques, like inverted files [131], in the text retrieval domain.

Similarly, Zhang et al. [137, 136] explored the bag-of-words representation for texture classification and object categorization. The experimentations on PASCAL VOC 06 dataset [28] and Caltech-101 dataset [63] have shown that the bag-of-words approach achieves state-of-arts performance under challenging real-world conditions, including significant intra-class variations and substantial background cluttering.

2.1.4 Hierarchical coding of local features

To construct a more efficient and robust representation, many researchers have proposed various improvements on bag-of-words approach. Nister and Stewenius have exploited a hierarchical vector quantization process on local image features to generate a hierarchical codebook, or multi-level vocabulary tree, of visual words [90]. The multi-level vocabulary tree is constructed via the hierarchical k-means clustering on local features, as shown in Figure 2.2. First, an initial k-means clustering process is applied on the training set of local features. Consequently, the training data is partitioned into k groups, where each group of local features corresponds to one cluster. Then, the same k-means clustering process recursively runs

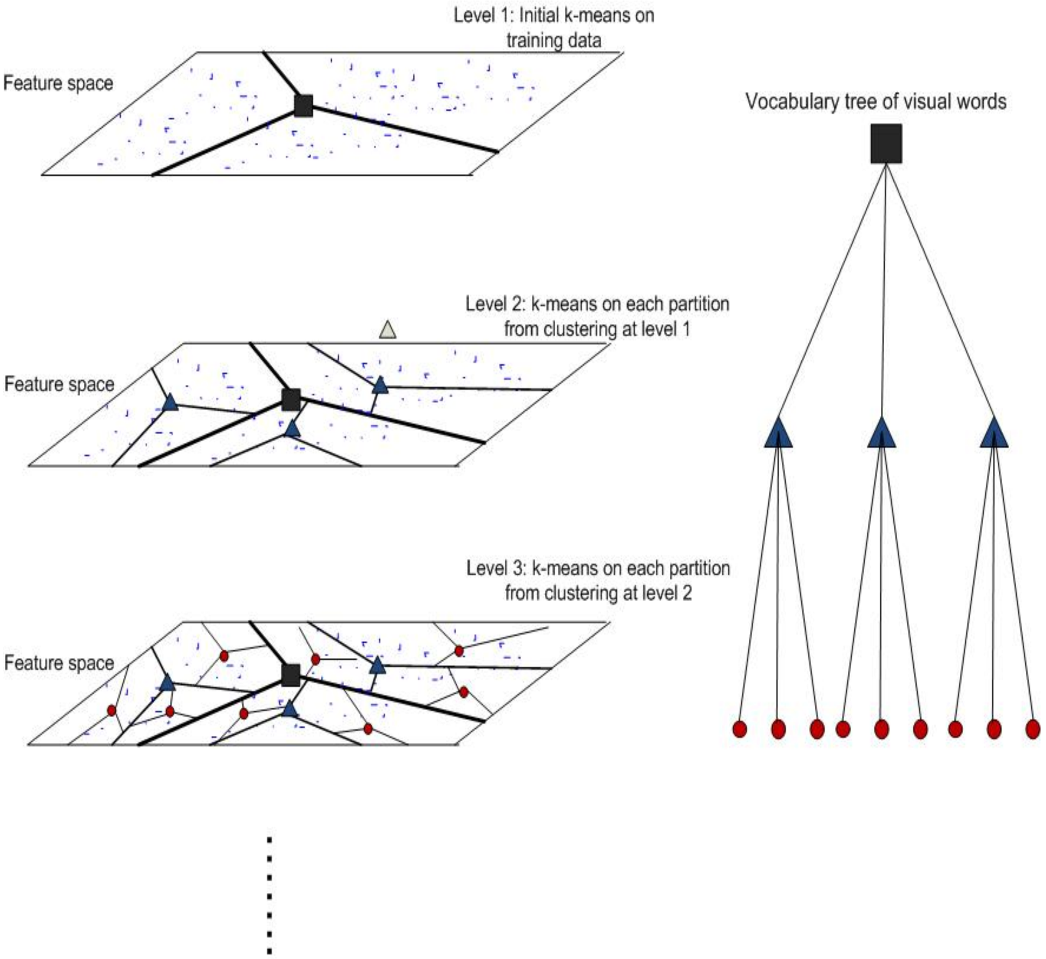

Level 1: Initial k-means on training data
Feature space
Vocabulary tree of visual words
Level 2: k-means on each partition from clustering at level 1
Feature space
Level 3: k-means on each partition from clustering at level 2
Feature space

Figure 2.2: The multi-level vocabulary tree of visual words is constructed via the hierarchical k-means clustering.

on each partition of local features, which effectively defines the "quantization cells by splitting each quantization cell into k new ones". The hierarchal representation of local features enables the new images to be efficiently inserted into the database. Moreover, the image representation can easily scale up to support efficient image indexing, retrieval and recognition.

Rather than from the perspective of visual descriptors of local regions, Agarwal and Triggs [1] proposed a multilevel hierarchal coding, by recursively incorporating spatial information of local features. The proposed approach leverages

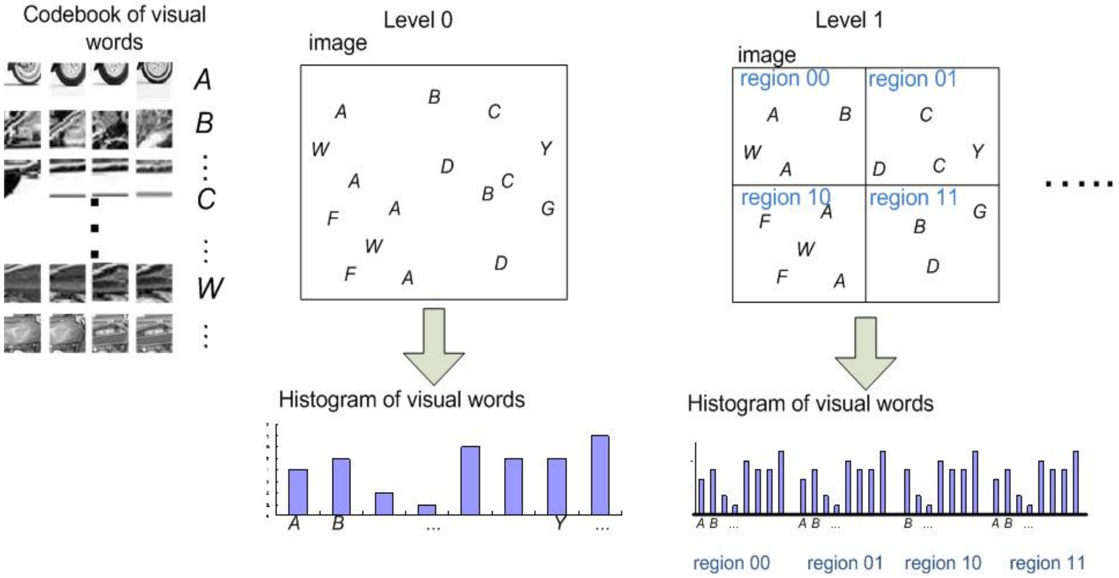

Codebook of visual words
A
B
C
W
Level 0
image
A B C W Y A D B C F A G W F A D
Histogram of visual words
A B ... Y ...
Level 1
image
region 00
region 01
region 10
region 11
A B W A C Y D C F A W F A B G D
.....
Histogram of visual words
A B ... A B ... B ... A B ...
region 00 region 01 region 10 region 11

Figure 2.3: The spatia pyrmaid is to organize the visual words in a multi-resolution histogram or a pyramid at the spatial dimension, by binning visual words into increasingly larger spatial regions.

the local histogram model to incorporate spatial information into bag-of-words representation [98]. The goal of the proposed approach is to exploit the spatial co-occurrence statistics at different spatial scales. The approach divides the images into local regions with each region being characterized by a descriptor vector, like SIFT. The base level representation contains bag of raw local descriptors. The higher levels then code the local features into visual words, by applying vector quantization on local features in the preceding level. The same process is then repeated recursively at higher levels. At each level, it generates the visual words by coding a local set of descriptor vectors from the preceding level. The resulting image representation is named hyperfeatures [1].

2.1.5 Incorporating spatial information of visual words

To make bag-of-words representation more discriminative, Lazebnik el at. proposed a spatial pyramid model to incorporate spatial information hierarchically into bag-of-words representation [59]. The proposed image representation is to organize the visual words in a multi-resolution histogram or a pyramid at the spatial dimension, by binning visual words into increasingly larger spatial regions, as shown in Figure 2.3. The advantage of this approach is that the finer spatial resolution level yields more distinctive visual words, while the coarse level gives the tolerance to the mismatch between two images. The main drawback, however, is that the increasing spatial resolution will inevitably bring in the curse of dimensionality, as a new finer spatial resolution will double the number of distinct visual words in the image representation.

2.1.6 Constructing compositional features

To further strengthen the discriminativeness of visual representation, researchers developed the compositional features that consists of several individual visual features (words) with specific spatial or locality constraints. The rational of compositional features is rooted in the compositionality principle [34, 15], which states that "in cognition in general, especially in human vision, complex entities are perceived as compositions of comparably" simple and widely usable parts [94].

Ommer and Buhmann [93, 94] proposed a category-dependent model of composition of local features to represent images with intermediate groups of features. The goal of the proposed approach is to generate compositions representative to category-distinctive subregions, so as to achieve minor intra-class variations for more effective learning afterwards. The approach first generates a codebook of visual words, by performing k-means clustering on local features. Then, by following the principle of perceptual organization, namely Gestalt laws [70], the possible candidates of visual word compositions are selected. The relevant and discriminative compositions are learned, in a weakly supervised fashion (with category label for each image). Finally, the selected relevant compositions are coupled with a shape model.

Rather than building compositional features from supervised learning, Yuan el at. [132] proposed a compositional visual configuration, visual phrase, by a unsupervised mining of frequently co-occurring visual words. The rational is that the spatially associated visual words can form more distinctive and informative visual patterns. The visual phrase learning is formulated as a frequent itemset mining task [40] in the database of the spatially adjacent visual word-sets. This approach is closely related to our proposed visual descriptor, delta visual phrase. However, different from the approaches above, the proposed delta visual phrase attempts to exploit both co-occurrence and spatial scatter information of visual words, by utilizing a series of varying support regions, so as to deliver more distinctive visual configurations.

Similar to [132], Quack el at. [99] proposed a compositional visual configuration model based data mining techniques for object detection task. The proposed model first collects a large number of spatial neighborhoods of local features. It then exploits an efficient frequent item-set mining algorithm [4] to discover association rules among the neighborhoods. The association rules that are discriminative to certain object categories are then selected for object detection task afterwards.

2.1.7 Latent visual topic representation

The performance of primitive visual features, like visual words and phrases reviewed in previous section, depends highly on the visual similarity and regularity of object categories. To mitigate such problem, researchers proposed to project images from visual feature space to an intermediate latent topic feature space. As shown in Figure 2.4, the latent topic functions as an intermediate variable that decomposes the observation between visual words and image categories.

Inspired from the document topic mining work in text domain [38, 44, 45],

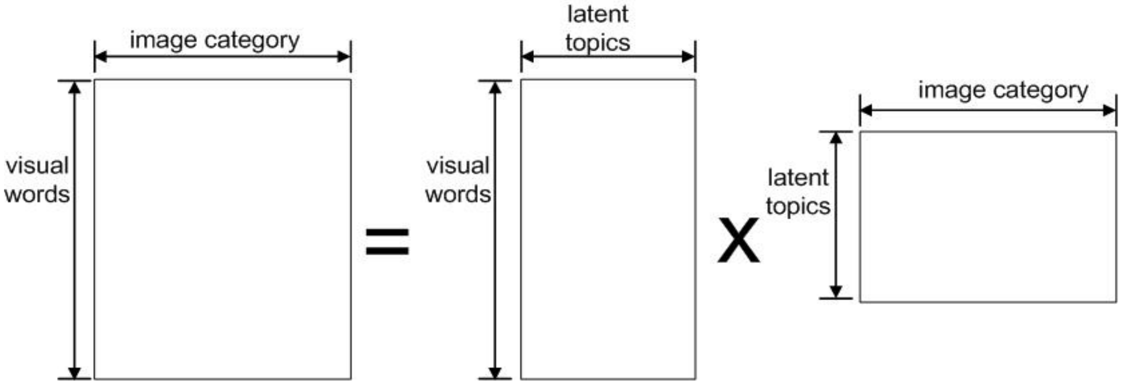


Figure 2.4: The latent topic functions as an intermediate variable that decomposes the observation between visual words and image categories.

Sivic el at. [104] proposed to model images with some higher level latent visual topic features by exploiting the probabilistic Latent Semantic Analysis (pLSA) [43] and Latent Dirichlet Allocation (LDA) [19]. pLSA and LDA are statistical models that attempts to associate a latent variable (or aspect) with each observation (occurrence of a visual word in an image) by capturing co-occurrence information between them. The proposed approach applies pLSA and LDA, topic models originally proposed for text document analysis, on visual words to project image representation from visual word space to a latent topic space. The advantage of this approach is that the visually different images of same category could share a similar set of latent topics, and therefore, be in proximity in the latent topic space. The statistical learning can, therefore, be more effective on discriminating different object categories. Agarwal and Triggs also demonstrated the effectiveness of LDA in image classification in [1].

The major drawbacks of the approach, however, are two-fold. First, the number of latent topics need to be fixed. This rigid constraint is conflicting with the open-ended nature of object appearances, which is that the objects can have limitless variation of visual appearances. Second, the model in [104] does not incorporate category knowledge and topic sharing information across different categories. Though the latent topics are mined in a principled Bayesian framework, the categorization can only be done with additional machine learning mechanism. To some extent, the model in [104] is more proper to be regarded as an unsupervised dimensionality reduction on the primitive visual features, like visual words.

2.2 Learning and recognition based on local feature representation

There exist a considerable variety of learning methods for visual categorization, from simple nearest-neighbor schemes to more complicated discriminative kernel-based classifiers, generative probabilistic Bayesian models, and so on. As the thesis focuses on part-based local image representation, this section will mainly review the learning and recognition methods based on local feature representation. It first reviews the relevant literature work on discriminative models, and then, related work on generative models.

2.2.1 Discriminative models

One advantage of using bag-of-words model is that the discriminative models used in text document categorization, such as support vector machine (SVM) [120, 84], adaptive boosting (AdaBoost) [31, 32], are readily available for visual classification. In the discriminative models, the focus usually shifts to the distance metric of local feature representation.

Berg el at. [12] explored the nearest neighbor classification framework based on local features for object recognition task on Caltech-101 dataset [63]. The approach first extracts a set of local regions and compute the geometric blur features of local regions [13] to represent the visual characteristics of an image. Then, the distance between images is computed by a correspondence algorithm that takes into account "the similarity of corresponding geometric blur point descriptors as well as the geometric distortion between pairs of corresponding" local regions [12]. By taking the correspondence measure as the similarity metric, the image classification is carried out by finding its nearest image neighbors. The advantages of nearest-neighbor schemes are: (1) they are straightforward and simple to implement, in the sense that they implicitly tackle the multi-class issue in visual categorization and; (2) they can produce reasonably good results if the similarity metric are designed carefully. The unpleasant aspect of nearest-neighbor schemes is also obvious. The classification process can be computationally expensive, as the testing image need to be compared against all the training images. This is especially so, when the number of training data is huge.

Zhang el at. [136] investigate the kernel trick in SVM classifier for visual categorization based on local features. Specifically, the work evaluated the effectiveness of different kernels on two types of local feature representations: (1) bag of local features and; (2) the quantized bag-of-words approach. The work pointed out that the Earth Mover's Distance (EMD) [100] gave satisfactory results on the bag of local features representation and the χ^2 distance delivered good performance on bag-of-words representation. SVM possesses both theoretical and empirical suitability for local feature based image classification, due to the following desirable characteristics [51]. First, SVM has good capability to handle high dimensional input space, while the local feature representation usually results in high dimensional feature space (a few thousands visual words). The overfitting protection of SVM enables it to handle such high dimensional feature space. Second, SVM has good capability to handle sparse image vectors. The image vectors of bag-of-words representation usually are sparse, due to the limited number of extracted local regions per image and high dimensionality of visual word feature space.

To combine the edges of nearest neighbor classifier and SVM, Zhang el at.

[135] proposed a discriminative nearest neighbor classification scheme, called SVM-KNN, for visual categorization. The basic idea of the proposed model is to locate k nearest neighbors of a given query image in the training database and train a local support vector machine for the query image, based on its k nearest neighbors. To some extent, this approach is similar to the local learning schemes in [21], which also utilizes k-NN together with a set of local linear classifiers with ridge regularizer. To speed up the nearest neighbor searching process, the approach in [135] first computes a crude distance between query images and training images to prune the list of nearest neighbor candidates. The accurate distance are then computed to determine the k nearest neighbors. The experiments on several datasets, like Caltech-101, showed that this approach can deliver promising results with reasonable efficiency.

Another paradigm of discriminative models is the multiple kernel learning (MKL) [7]. Rather than using a single kernel in support vector machine (SVM), the MKL schemes learn a kernel combination and the associated classifier that fuses multiple informative features and kernels [56, 121]. The recently reported work [56, 121, 66] have shown that the multiple kernel learning can deliver much superior performance than tradition SVM approach, as it can leverage different visual features simultaneously. For example, Varma and Ray [121] explored the multiple kernel learning to combine a set of base visual features, such as bag-of-words feature, global color histogram, etc, to perform visual categorization. The premise is that different base visual features tend to capture different aspects of image categories and hence share complementariness and redundancy in modeling the visual contents.

Different from the multiple kernel learning schemes, the thesis focuses on devising one higher-level local feature representation, on top of bag-of-words model, rather than combining different visual features.

2.2.2 Generative models

The discriminative learning focuses on optimizing the mapping between input image feature and output category variables. In contrast to discriminative approaches, the generative methods attempt to model the probabilistic inter-connectivity in the input and output variable domain, by estimating the joint probability of all relevant variables. The probabilistic inter-connectivity of variables reveals the structural knowledge of inter-relation among variables, and therefore, helps to achieve better generalization on categorizing new unseen images. For this reason, we approach the object categorization task via generative probabilistic learning. This section reviews the generative probabilistic work closely related to our proposed model in the thesis. Prior to presenting the relevant literature work, we briefly introduce the fundamentals of probabilistic graphical models.

Fundamentals of graphical models

As stated by Jordan [52], probabilistic graphical model are a "marriage between probability theory and graph theory" and a powerful framework for dealing with uncertainty and complexity. A probabilistic graphical model is depicted as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which nodes or vertices $\mathcal{V}$ correspond to random variables, and edges $\mathcal{E}$ represent conditional independence assumptions [82]. Graph is to represent the interrelation among variables (nodes), which are depicted by conditional independencies (edges). Effectively, graph decomposes multivariate, joint distributions into a set of local interactions among small subsets of variables. More importantly, such decomposition leads to efficient learning and inference algorithm [115].

In general, there are two types of graphical models: undirected and directed. The undirected graphical models are known as Markov networks. The directed graphical models are known as Bayesian networks, belief networks, generative models, and some other names. Our focus here is directed graphical models. In directed

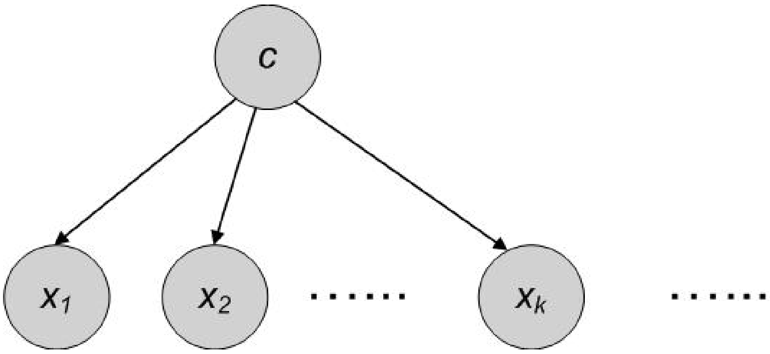


Figure 2.5: The graphical model of Naive Bayes classifier, where parent node is category variable c and child nodes are features x_k. Given category c, features x_k are independent from each other.

graphical model, an edge (i, j) connects parent node i and child node j with an arrow. The Markov property here is that a random variable x_j is conditionally independent of all the other variables, given its parent x_i.

Among various directed graphical models, Naive Bayes classifier can be regarded as the simplest one. As shown in Figure 2.5, a Naive Bayes classifier assumes that the effect of a variable (feature) x_k on a given class c is independent of other variables (features). The conditional class probability can then be computed as below.

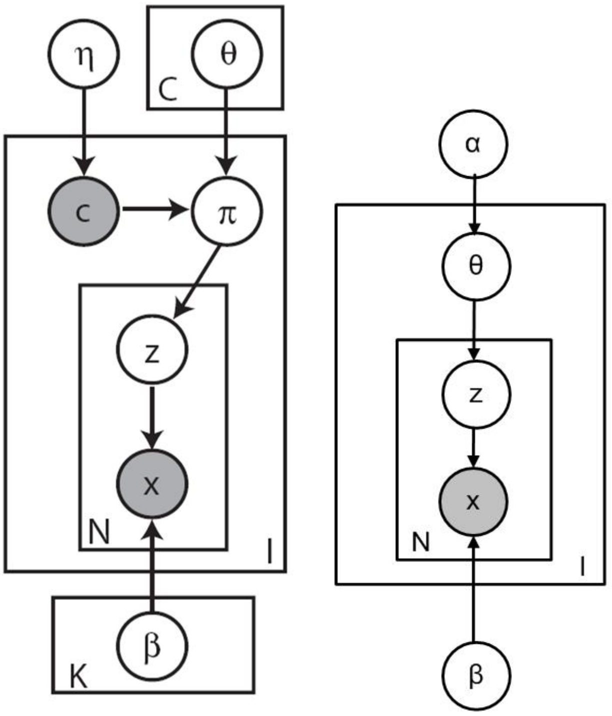
$$\begin{aligned} p(c \mid x_1, ..., x_K) &= \frac{p(x_1, ..., x_K \mid c)p(c)}{P(x_1, ..., x_K)} \\ &\doteq \Pi_k p(x_k \mid c)p(c) \end{aligned} \tag{2.1}$$

Constellation models

In the family of directed graphical models, constellation models have spurred much research attention for category-level object recognition. Similar to visual topic learning like LDA on bag-of-words representation, the constellation models [30, 29, 129, 128, 127] attempt to represent an object class by a constellation of visual parts (or topics). Constellation models take into account the geometric relationship between different parts. This is significantly different from the bag-of-words approach that neglects the geometric information completely. Specifically, the constellation models explicitly model the relative location, relative scale, and appearance of these parts for a particular object category. For this reason, they tend to be good at recognizing rigidly structured objects, such as car, airplane, people, as these object categories present strong geometric clues.

The original idea of "visual parts" model can be traced back to the work by Fischler and Elschlager [30]. They proposed to model objects as spatially deformable collections of parts [30, 127]. Weber el at. [129, 128, 127] extended this idea to learn object class models from unlabeled and unsegmented cluttered images, in an unsupervised fashion. First, the approach automatically identifies distinctive parts in the training images by utilizing clustering techniques. Then, the statistical shape model is learned via a greedy search over possible model configurations by using expectation maximization. The models of constellation of parts can be applied to discover object categories in an unsupervised fashion [127].

Fergus el at. [29] furthers the approach by Weber el at. [129] by representing visual parts in three dimensions: shape, appearance and relative scale. In the proposed approach, an object model is defined to consist of a group of parts and each part is depicted by appearance, relative scale and shape. Shape is represented by the mutual position of the parts. Appearance, scale and shape are modeled by Gaussians probability density functions. Object category learning process first detects regions and their scales, and then estimates the parameters of density functions above, such that the resulting model best explains the training data in the sense of maximum-likelihood. The recognition on a query image can then be performed by the learned model in a Bayesian manner accordingly.



(a) The graphical model of hierarchical generative framework for scene classification

(b) The graphical model of LDA

Figure 2.6: Comparison of LDA model and the modified LDA model for scene classification in [64]. Figure (a) (from [64]) shows the graphical models hierarchical generative framework for scene classification that are extended from LDA. Figure (b) shows the graphical models of LDA. x and c are image features and classes, z is the index of visual topic , π is the parameter of a multinomial distribution for choosing the visual topic, K is the total number of themes

Generative models with Bayesian priors

Fei-Fei et al [64] proposed a Bayesian hierarchical generative model for natural scene classification, by modifying the latent Dirichlet allocation (LDA) [19] to incorporate category knowledge. Figure 2.6(a) and 2.6(b) show the graphical models of the modified LDA model for scene classification and LDA model. The proposed model in [64] takes the bag-of-words representation to identify an images as a collection of visual words. It then associates the visual word of each local region to a latent visual theme to learn the distribution of visual words and the intermediate visual themes. The categorization on unseen new images is performed, via Bayesian learning on the visual theme model. The major drawback of the approach is that it inherits the weakness of LDA, which is that the number of latent visual themes must be fixed and visual themes can not grow or shrink flexibly. This structure rigidity impedes the trained model to capture more visual complexity from the new training images.

To tackle this issue, Wang et al [125] developed a variation of hierarchical Dirichlet process, by incorporating spatial information of visual descriptors. Similar to our proposed model, their approach allows different categories to share countably infinitely many appearance patterns. However, it assumes that the visual themes are independent of image categories and shared at image level only. This effectively neglects the inter-relation between visual appearance patterns and object categories and leads to an oversimplified model. In contrast, our model takes into account the inter-relation between appearance patterns, objects and categories via its hierarchical structure.

In this aspect, the model in [116] shares the same vision as our model. However, their model neglects the fact that the images within same category tend to arise from a closer set of appearance patterns than those of different category, which our proposed model has tackled with its graphical model structure.

Chapter 3

Building a Higher-level Visual Representation

3.1 Motivation

In the task of visual object categorization, the bag-of-words (BoW) approaches have achieved many significant results [36, 124, 130, 136, 59, 53]. However, compared to the analogy of text document categorization, the performance of BoW based image classification is far from satisfaction. Similarly, compared to its analogy of text document representation, the BoW image representation looks much inferior. From the perspective of statistical learning, a desirable image representation should possess the following characteristics: it is distinctive to represent images of different classes, and it is invariant to represent images of the same class. In this way, the images will be well distributed in the feature space, by concentrating into clusters according to their belonging classes. The BoW representation, however, forfeits this desired property by two phenomena: polysemy and synonymy. As introduced in Chapter 1.3, the polysemous visual word is a one that might represent different semantic meanings in different image context, while the synonymous words are

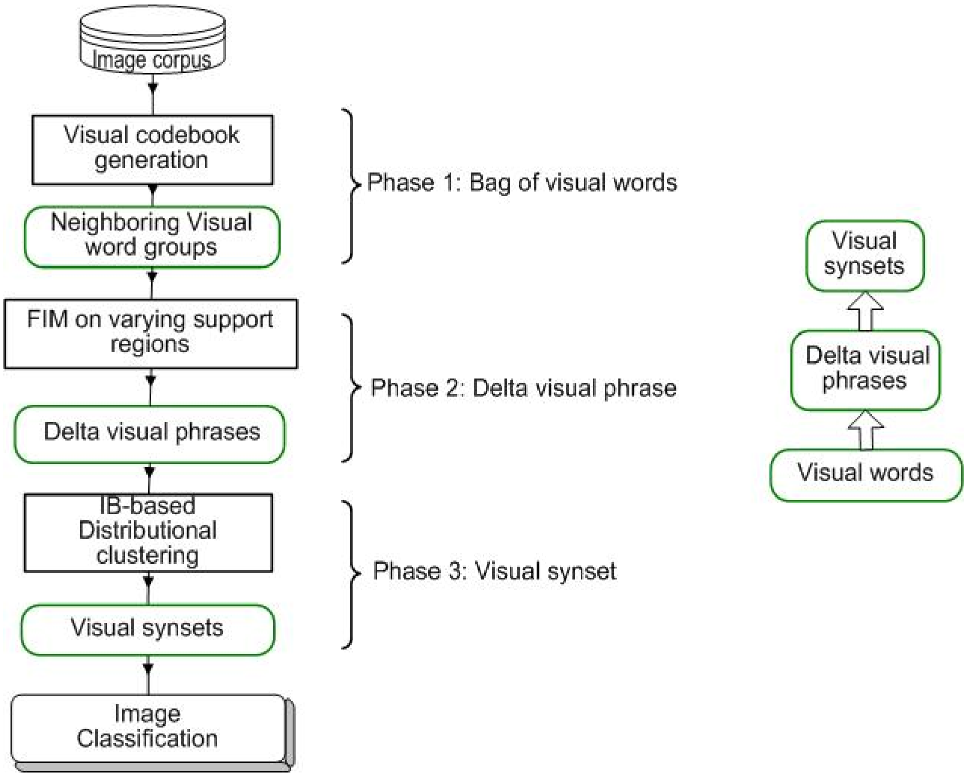


Image corpus
Visual codebook generation
Neighboring Visual word groups
Phase 1: Bag of visual words
FIM on varying support regions
Delta visual phrases
Phase 2: Delta visual phrase
IB-based Distributional clustering
Visual synsets
Phase 3: Visual synset
Image Classification
Visual synsets
Delta visual phrases
Visual words

Figure 3.1: The overall framework of visual synset generation

a set of visually dissimilar words representing the same semantic meaning. The polysemy issue impairs the distinctiveness the BoW representation and leads to small inter-class distance, while the synonymy issue encumbers the invariance of BoW representation and results in large intra-class distance. Hence, a higher-level visual content unit is needed to tackle the issues of bag-of-words representation.

3.2 Overview

Prior to delving deep into the proposed representation of visual synset, we present the overall process of building visual synset. The visual synset aims to improve the traditional bag of words representation with better discrimination and invariance power. Figure 3.1 illustrates the overall framework of building visual synsets. As shown, the overall flow of the proposed approach consists of 3 phases. Phase 1 constructs visual words or visual codebook. In phase 2, the approach strengthens the inter-class discrimination power by constructing an intermediate visual descriptor, delta visual phrase, from frequently co-occurring visual word-set with similar spatial context. In phase 3, the approach achieves better intra-class invariance power, by clustering delta visual phrases into visual synset, based on their probabilistic 'semantics', i.e. class probability distribution. Hence, the resulting visual synset can partially bridge the visual differences of images of same class.

3.3 Discovering delta visual phrase

To address the polysemy issue of visual words, we exploit the co-occurrence and spatial scatter information of visual words to generate a more distinctive visual configuration, delta visual phrase. The delta visual phrase is regarded as one kind of compositional features, as it is a combination of its constituent primitive features (visual words). The theoretical rational of compositional features is rooted in the compositionality principle [34, 15]. According to the compositionality principle, a complex expression with more semantic can be generated by "the meanings of its constituent expressions and the rules used to combine them" [94, 34]. In the context of our thesis, the "complex expression" here corresponds to the delta visual phrase, while the "constituent expressions" mean the individual visual words that compose the delta visual phrase and the "rules used to combine them" are the co-occurrence and spatial context relation of visual words. The practical rational of delta visual phrase is simple too. If two visual words co-occur frequently in the similar spatial context, their corresponding visual parts are probable to belong to the same geometric structure of objects or have strong correlation at semantic level. By combining several visual words into one visual unit, the delta visual phrase corresponds to a larger visual or structural pattern, and therefore, gives more visual distinctiveness to represent the object. Figure 3.2 illustrates examples of composition of visual words possessing more distinctiveness than individual visual words.

3.3.1 Learning spatially co-occurring visual word-sets

The first step to generate delta visual phrase is to mine the visual words that are frequently spatially co-occurring together. To learn spatially co-occurring visual word-sets, researchers have proposed several approaches, such as visual phrase mining [132, 133], frequent feature configuration mining [99], frequent co-occurring word-set learning [139, 39], etc. Here, we borrow the approach in [132] to mine spatially co-occurring visual word-sets, and furthermore, build our proposed delta visual phrase.

Here we follow the notation of [132, 143, 142]. We first extract M local regions $\{a_i\}_{i=1}^{M}$ from a given image $\mathcal{I}$. We then sample approximately 1 million local features $\{a_i\}$ and perform clustering on $\{a_i\}$ to generate visual codebook of W visual words: $\Omega = \{w_1, ..., w_W\}$, where w_i is a visual word. The image $\mathcal{I}$ is then represented by a bag of visual words $\{w_{(a_1)}, ..., w_{(a_i)}, ...\}$, where $w_{(a_i)}$ is the corresponding visual word of region a_i. For each local region $a_i \in \mathcal{I}$, its local spatial neighborhood $\mathcal{G}$ is defined as the group of its K nearest neighbor regions $\{w_{(a_i)}, w_{(a_{i_1})}, w_{(a_{i_2})} ... w_{(a_{i_K})}\}$. By processing all image, a visual word group database $\mathbf{G} = \{\mathcal{G}_i\}_{i=1}^{N}$ will be generated, in which each record in the database correspond to one group of visual words in the same spatial neighborhood. Figure 3.3 illustrates the process of generating the database $\mathbf{G}$. In the domain of data mining, the database $\mathbf{G}$ can be regarded as a transaction database [40]. The discovery of frequently co-occurring visual word-sets, i.e. visual phrases, can be reduced to a task of frequent itemset mining (FIM) in the transaction database $\mathbf{G}$ [40, 132].

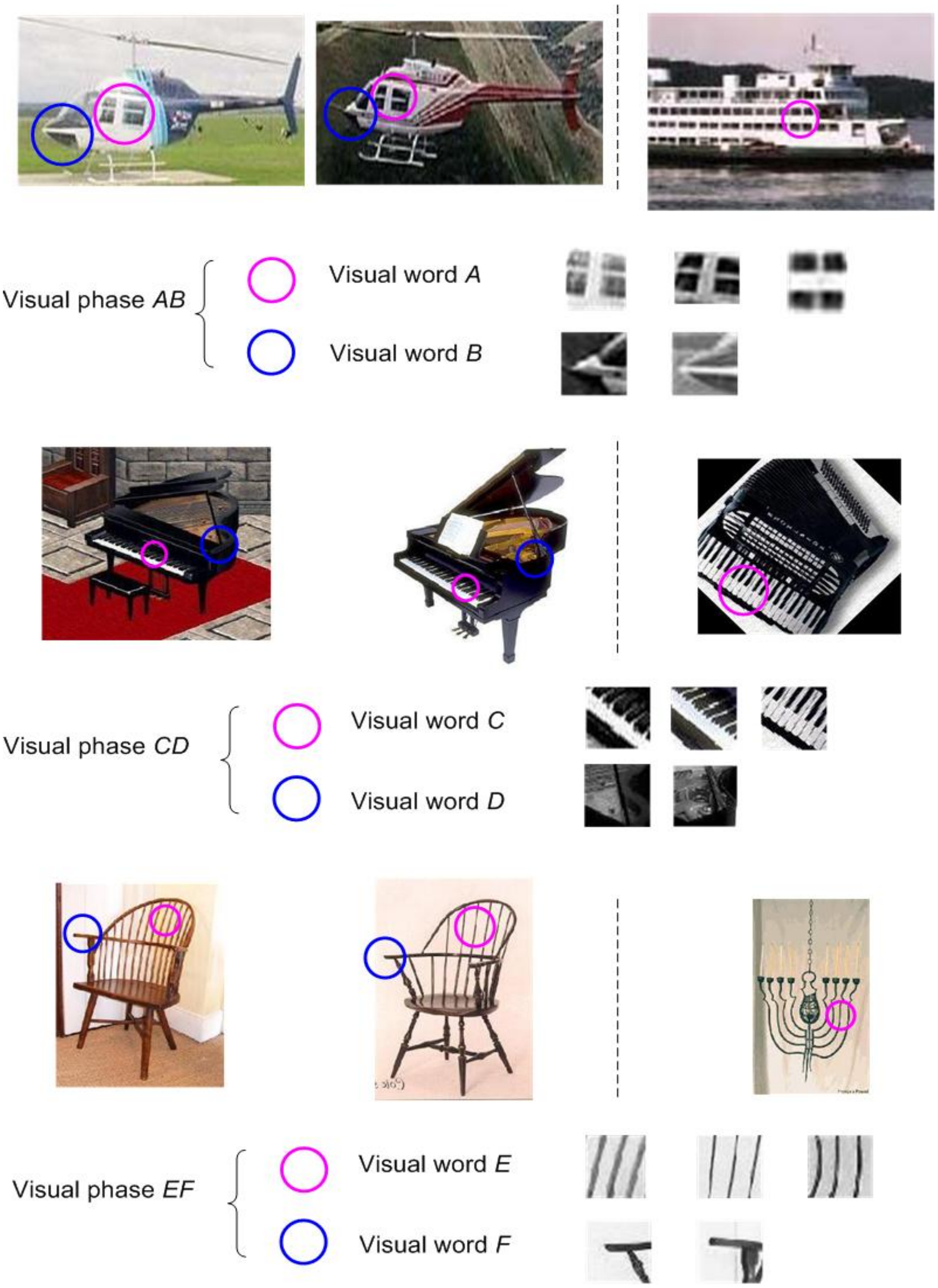


Figure 3.2: Examples of compositions of visual words from Caltech-101 dataset. The visual word A (or C) alone can not distinguish helicopter from ferry (or piano from accordion). However, the composition of visual words A and B (or C and D), namely visual phrase AB (or CD) can effectively distinguish these object classes. This is because the composition of visual words A and B (or C and D) forms a more distinctive visual content unit, as compared to individual visual words.

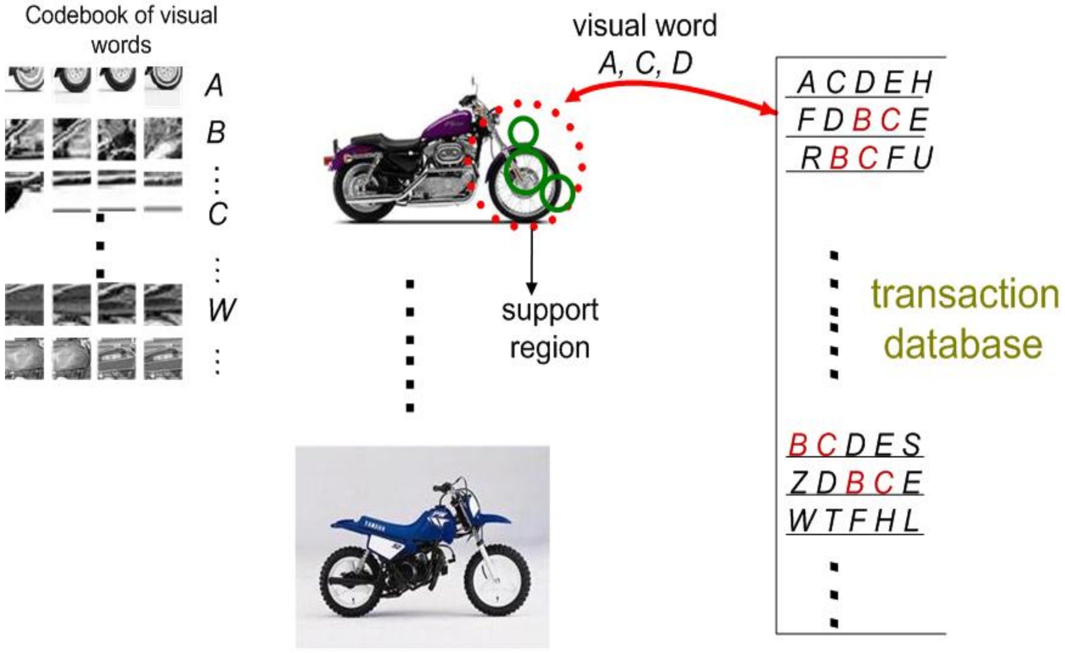


Figure 3.3: The generation of transaction database of visual word groups. Each record (row) of the transaction database corresponds to one group of visual words in the same spatial neighborhood.

3.3.2 Frequent itemset mining

Frequent itemset is defined as a group of items that are frequently co-occurring in the given transaction database. Frequent itemset mining (FIM), also known as association rule learning, is a popular and well studied method for discovering relations between variables in the databases, in the data mining paradigm. The aim of frequent itemset ming is to find association patterns and regularities in variables (visual words here) in the database. There exist various FIM techniques in the literature, such as FP-growth [2], Apriori algorithm [5], etc. We explore the FP-growth algorithm to perform the FIM task, as it is reported to be one of the most efficient algorithms for frequent itemset mining [2, 40, 20].

The FP-growth algorithm is based on an FP-tree structure, a prefix tree representation of the given database of transactions [20]. This tree shape data structure can save considerable amounts of memory for storing the transactions and allow for high efficiency. The FP-growth algorithm is basically a recursive elimination scheme. A preprocessing step first scans the transaction database to generate a list of frequent items in descending order. Based on the descending list of frequent items, FP-growth transforms the database into a frequent-itemset tree (FP-tree), in which the item association knowledge is embedded. The FP-tree mining first discovers the frequent length-1 itemsets as initial suffix itemsets. It then constructs the conditional FP-tree, based on the conditional itemset base. The conditional itemset base contains the set of prefix paths (of items) that co-occur with the suffix itemsets in the FP-tree. The recursive mining is performed on the conditional itemset base. The discovered itemsets grow by concatenating the suffix itemsets with the frequent itemsets mined from the conditional itemset base. In summary, the FP-growth algorithm casts the task of frequent long itemset mining to a task of searching for shorter ones recursively and then concatenating the suffix [2, 40].

After frequent itemset mining in the visual word-set transaction database, a visual word-set $\mathcal{P} \subset \Omega$ is counted as a frequently co-occurring set or a visual phrase, if its frequency $freq(\mathcal{P}) > \theta$. For example in Figure 3.3, the visual word B and C may compose a frequently co-occurring word-set, if they occur together frequently in the database. Specifically, the neighborhood $\mathcal{G}$ is called the support region of $\mathcal{P}$, as $\mathcal{P}$ is mined from the database of $\mathcal{G}$. We follow the notation in [132] to name the spatially co-occurring visual word-set as, *visual phrase*.

3.3.3 Building delta visual phrase

The major shortcoming of aforementioned visual phrase proposed in [132] is that it neglects the spatial inter-relation among visual words. Namely, no matter how big the spatial neighborhood $\mathcal{G}$ is, the visual words within $\mathcal{G}$ are treated equally in an orderless manner in one transaction record in the database $\mathcal{G}$. In this way, the information of their relative spatial location is completely neglected. The exclusion of such information weakens the spatial correspondence between visual words in two neighborhoods and results in visual phrases that do not incorporate the best co-occurring visual word-sets.

To tackle this issue, we propose *delta visual phrase* that does not only incorporate co-occurrence information, but also the local proximity of visual words. Such spatial proximity information defines the specificity of the visual phrase, which can be determined by the size of support region that visual phrase is mined from. Specifically, a delta visual phrase is defined in 2 dimensions: its member visual word-set $\mathcal{P}$ and its scatter $\mathcal{R}$, namely, how spread the visual phrase is across over the image.

Prior to presenting the proposed delta visual phrase, we first introduce the concept of **minimal support region**. The support region of visual phrase $\mathcal{P}$ is the visual word group $\mathcal{G}$ of size K, where K is the number of visual words in the neighborhood $\mathcal{G}$. Let $\mathcal{G}^1$, $\mathcal{G}^2$,..., $\mathcal{G}^{k-1}$, $\mathcal{G}^k$,... be a series of support regions with same centroid and growing size. The minimal support region is then defined as follows.

Definition 3.3.1. *The region $\mathcal{G}^k$ is called **minimal support region** of visual phrase $\mathcal{P}$, if any smaller region $\mathcal{G}^{k-i}, \forall i > 0$ is not large enough to discover the visual phrase $\mathcal{P}$.*

With respect to each support region $\mathcal{G}^k$, the delta visual phrase is defined as follows.

Definition 3.3.2. *The **delta visual phrase** (dVP) of region $\mathcal{G}^k$ is the visual phrase that has $\mathcal{G}^k$ as minimum support region. In other words, the delta visual phrase of region $\mathcal{G}^k$ is the newly discovered visual phrases when the support region just grows from $\mathcal{G}^{k-1}$ to $\mathcal{G}^k$. The size of $\mathcal{G}^k$ is therefore the **scatter** $\mathcal{R}$ of delta visual phrase and $\mathcal{R} = |\mathcal{G}^k|$.*

Intuitively, the delta visual phrase is mined from the changes of support regions. This is also why the word "delta" is in its name. The visual word-set $\mathcal{P}$ is deemed to be delta visual phrase [$\mathcal{P}$, $\mathcal{R}$], if it satisfies one of the following condition:

$$freq^{\mathcal{G}^k}(\mathcal{P}) - freq^{\mathcal{G}^{k-1}}(\mathcal{P}) > \theta_k, \tag{3.1}$$

where $\mathcal{R} = |\mathcal{G}^k|$, $freq^{\mathcal{G}^k}(\mathcal{P})$ is the frequency of a visual word-set $\mathcal{P}$ for support region $\mathcal{G}^k$ and θ_k is the threshold. For example in Fig. 3.4 (a), the visual word-set '*CDF*' will be considered as dVP with scatter $\mathcal{R} = |\mathcal{G}^3|$, if the number of newly discovered instances of '*CDF*' resulted from the increase of support region (from $\mathcal{G}^2$ to $\mathcal{G}^3$) is greater than the threshold. The Eq. (3.1) also ensures that the visual words of a dVP are scattered over its support region. For example in Fig. 3.4 (b), the instance of visual word-set '*AB*' will not be counted for dVP with $\mathcal{R} = |\mathcal{G}^3|$, as it lies in region $\mathcal{G}^2$ as well and will be offsetted by Eq. (3.1). If we define the size of first support region $\mathcal{G}^1$ to be 1, the resulted delta visual phrases are actually visual words with scatter $\mathcal{R} = 1$. In this way, we can combine visual words and delta visual phrases into a unified representation.

Statistical Significance Measure

Yuan el at. [132] proposed to measure the statistical significance of visual phrase based on its frequency and its component visual word frequencies. This measurement, however, neglects the coherency of component visual words in visual phrase. We measure the significance on the basis that the delta visual phrase should be a visual word-set that is frequently and coherently occurring together, with respect to certain semantic meaning. Specifically, the significance score $L([\mathcal{P}, \mathcal{R}])$ of a dVP [$\mathcal{P}$, $\mathcal{R}$] is defined as:

$$L([\mathcal{P}, \mathcal{R}])) = freq([\mathcal{P}, \mathcal{R}]) \cdot \frac{P(\mathcal{P}, \mathcal{R}|\mathbf{D}_{\mathcal{I}})}{1 + P^{-}(\mathcal{P}^{|}\mathbf{D}_{\mathcal{I}})} \tag{3.2}$$

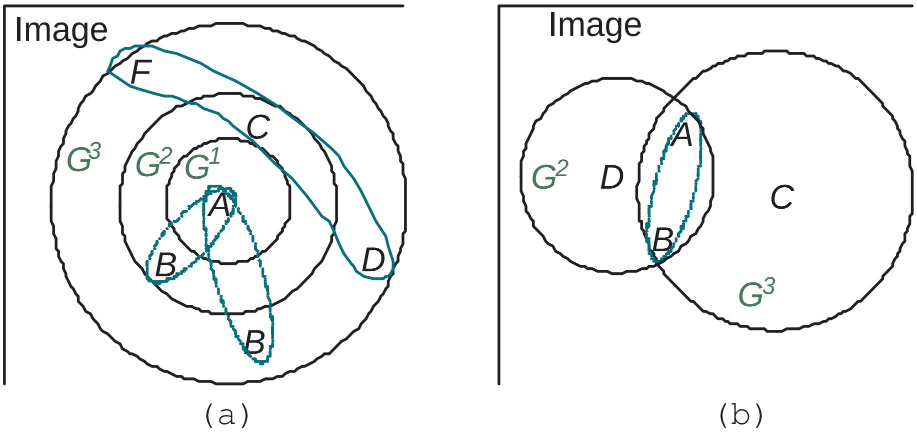


Figure 3.4: Examples of delta visual phrases. (a) Visual word-set *'CDF'* is a dVP with $\mathcal{R} = |\mathcal{G}^3|$. (b) Visual word-set *'AB'* cannot be counted as a dVP with $\mathcal{R} = |\mathcal{G}^3|$

where $P(\mathcal{P}, \mathcal{R}|\mathbf{D}_{\mathcal{I}})$ is the probability that the visual word-set $\mathcal{P}$ forms a valid dVP with scatter $\mathcal{R}$ by satisfying the condition of Eq. (3.1) and it can be approximated by $\frac{docfreq([\mathcal{P},\mathcal{R}]}{T}$, where $docfreq([\mathcal{P}, \mathcal{R}])$ is the document frequency equal to number of images containing dVP $[\mathcal{P}, \mathcal{R}]$. $P^{-}(\mathcal{P}|\mathbf{D}_{\mathcal{I}})$ is the probability that visual word-set $\mathcal{P}$ forms some random and sporadic patterns, which can be approximated by $\frac{docfreq(\mathcal{P})}{T}$. $freq([\mathcal{P}, \mathcal{R}])$ is the frequency of dVP $[\mathcal{P}, \mathcal{R}]$. Intuitively, we want to penalize the delta visual phrases whose member visual words also frequently co-occur in a random and sporadic manner. In this way, we enforce the correlation among member visual words, and therefore, ensures the coherency of delta visual phrases.

Unique Counting of Maximal Visual Word-set

The subsets of a frequent visual word-set $\mathcal{P}$ are frequent as well, and therefore, will be falsely counted as dVP. To address this problem, we exploit closed FIM algorithms to discover maximal frequent itemsets, in the way that any of its subsets will not be considered as frequent itemset, in the spirit of [132]. In the phase of FIM, a word-set might be over-counted, if it lies in the overlapping area of different neighborhood regions. To overcome this problem, we borrow the approach in [132] to re-count real instances of word-set through the original image database.

3.3.4 Comparison to the analogy of text domain

The BoW representation originates from text domain. It is therefore worth contrasting the proposed approach for polysemy issue with its analogy of text domain. Though the text categorization has reached the levels of human experts, it faces the polysemy issue too [42]. Due to langauge variability, polysemy makes the trained text classifier wrongly categorize new documents. For example, a text document containing "jaguar" with topic on luxury cars might be wrongly classified to the animal category. To resolve the aforementioned issues, the computational linguistics exploit word sense disambiguation (WSD) to identify which sense (or meaning) of a textual word in the sentence where it occurs, so as to achieve more effective information retrieval and text categorization [47]. For example, the textual word "jaguar" has two senses: (1) a large cat (panthera onca) chiefly of Central and South America that is larger than leopard; and (2) a brand of a luxury car. Example sentences with two senses can be: (1) Jaguar is a pretty mammal and; (2) sale of Jaguar sports car starts. One of the approaches of word sense disambiguation is to look at the surrounding neighbor words to determine the word sense [95]. For instance, if the word "jaguar" co-occurs with "car" in the same context, the WSD process will replace it with its more specified sense of car brand for subsequent text categorization, to avoid ambiguity in document representation.

The word sense disambiguation shares the same vision as the proposed delta visual phrase on tackling polysemy. The basic idea behind delta visual phrase is to exploit the contextual inter-relation among visual words to build more distinctive feature units, while the word sense disambiguation utilizes the context where the word occurs. Analogically, a word sense is like a delta visual phrase. The context used for word sense disambiguation is equivalent to the co-occurrence and spatial scatter information of visual words.

3.4 Generating visual synset

Though the mining on co-occurrence and spatial scatter information of visual words gives rise to a more distinctive visual configuration, delta visual phrase, the synonymy issue remains. The synonymy issue is a consequence of visual diversity of objects of the same semantic class. The objects of the same class can have arbitrarily different visual appearances and shapes, which produces synonymous visual words, with different visual appearance but same semantic meaning. The consequence of synonymy issue is the large intra-class variations. In this circumstance, both visual words and delta visual phrases become too primitive to effectively model the image semantics, as their efficacy depends highly on the visual similarity and regularity of images of the same semantic. To tackle this issue, we propose to exploit the prior available semantic knowledge, i.e. semantic class labels of training images and their distributions, to generate a higher level visual content unit, called **visual synset**, using a supervised learning process.

3.4.1 Visual synset: a semantic-consistent cluster of delta visual phrases

In text literature, the synonymous words are usually clustered into one synset (**syn**onymy **set**) to improve document categorization performance, based on word-document class distribution [10]. Such approach inspires us in solving the synonymy issue in delta visual phrases. However, it is infeasible to define the semantic meaning of delta visual phrase, as it is only a set of quantized vectors of sampled regions of images. Hence, rather than defining the semantic of a delta visual phrase

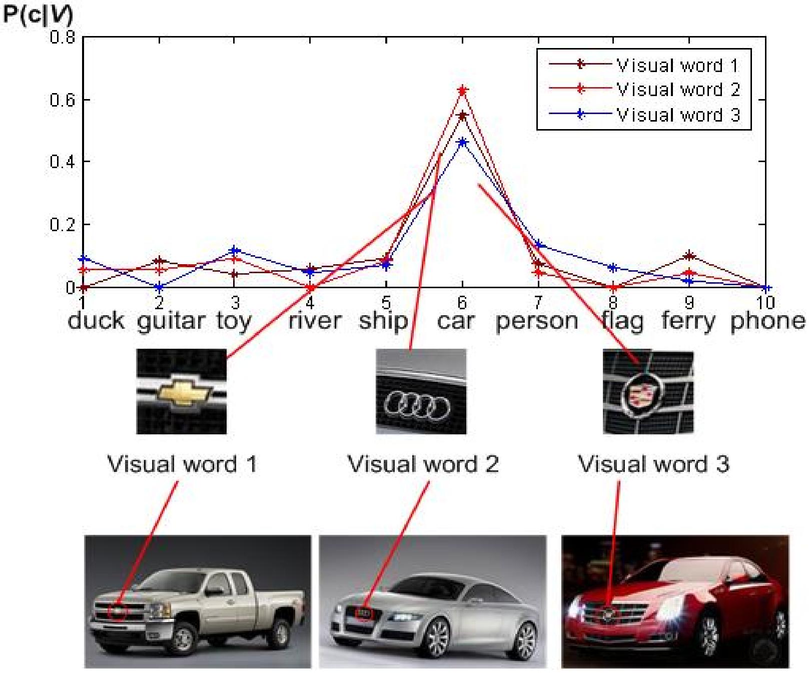


Figure 3.5: An example of visual synset generated from Caltech-101 dataset, which groups two delta visual phrases representing two salient parts of motorbikes.

in a conceptual manner, we define it probabilistically, in the spirit of [10].

Definition 3.4.1. *Given image categories $\mathcal{C} = \{c_i\}_{i=1}^{m}$, the* ***semantic*** *of a delta visual phrase v is its contribution to the classification of its belonging image, which can be approximately measured by $P(c_i|v)$.*

The probability distribution $P(c_i|v)$ implies the semantic inference of delta visual phrase v, namely how much v votes for each of the classes. Intuitively, if several delta visual phrases[1] from different images share similar class probability distribution, like the brand logos in car images shown in Figure 3.5, then the visual synset that clusters them together shall possess similar class probability distribution and distinctiveness towards image classes. The visual synsets can then partially bridge the visual differences between these images and deliver a more coherent,

[1] Visual word can be considered as a special case of delta visual phrase with only one member word and support region size = 1.

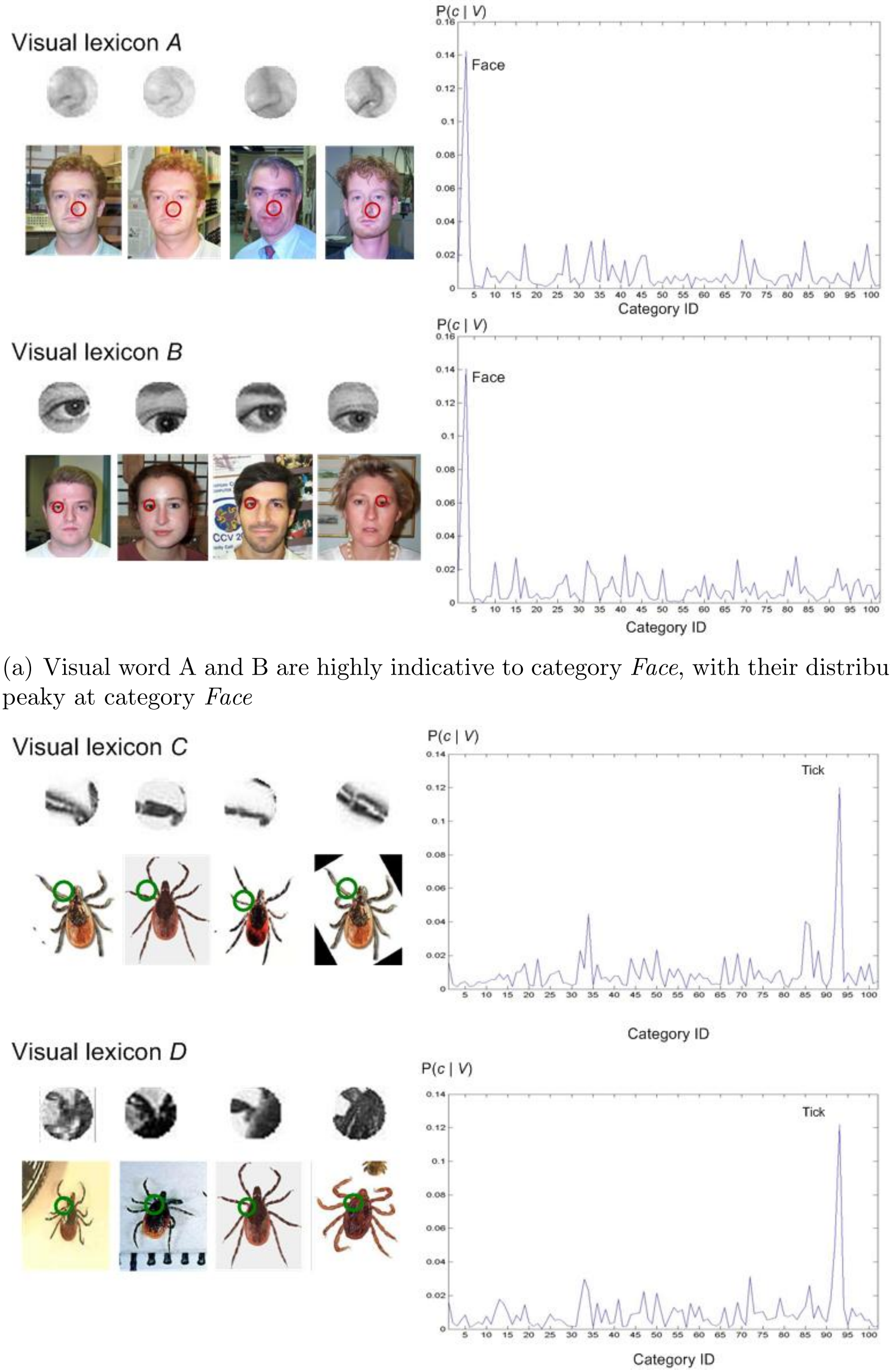
robust and compact representation of images. Specifically, we define the *visual synsets* as below.

Definition 3.4.2. *The **visual synset** is a probabilistic concept or a semantic-consistent cluster of delta visual phrases, in which the member delta visual phrases might have different visual appearances but similar semantic inferences towards the image classes.*

The rational of visual synset is that due to the visual heterogeneity and distinctiveness of objects, a considerable number of visual words/phrases are intrinsic and highly indicative to certain classes. This implies that some visual words/phrases tend to share similar probability distribution $P(c_i|v)$, which might peak around its belonging classes. Figure 3.6 shows examples of visual words/phrases with distinctive class probability distributions generated from Caltech-101 dataset. As shown, the distributions of these visual words/phrases tend to peak among their belonging categories, which make these visual words/phrases highly distinctive and indicative to their belonging categories. By grouping these highly distinctive and informative visual words/phrases into visual synsets, the visual differences of images from the same class can be partially bridged. Consequently, the image distribution in feature space will become more coherent, regular and stable. For example in Fig. 3.6 (a), if two visually salient components (visual words of *eye* and *nose*) of human face are grouped into one visual synset based on their image class probability distribution, the visually different human face images will now have some commonality in the feature space.

3.4.2 Distributional clustering and Information Bottleneck

As defined in *Definition 3.4.1* and *3.4.2*, the visual synset interprets a group of delta visual phrases with similar image class inferences. This effectively formulates the



(a) Visual word A and B are highly indicative to category *Face*, with their distributions peaky at category *Face*

(b) Delta visual phrase C and D are highly indicative to category *Tick*, with their distributions peaky at category *Tick*

Figure 3.6: Examples of visual words/phrases with distinctive class probability distributions generated from Caltech-101 dataset. The class probability distribution is estimated from the observation matrix of delta visual phrases and image categories.

visual synset construction as a clustering process on delta visual phrases, based on their class probability distributions. Clustering on distributions originates in text domain. Pereira et al. first introduced the concept of "distributional clustering" in [96]. In their work, nouns are represented as distributions over co-located verbs and nouns with similar $P(verb \mid noun)$ are clustered together distributionally. Later, Baker and McCallum also exploited the distributional clustering to group words into word-clusters for text categorization [8].

Distributional clustering gives us a clue on how to group delta visual phrases into synsets. However, visual synset construction demands a more principled method. Same as other data clustering algorithms, distributional clustering faces the issue of similarity metric selection too. The clustering results depend on the choice of similarity metric. The correctness of a metric selection relies on the application and target. Pereira et al. [96] proposed to use the relative entropy or Kullback-Leibler (KL) distance to measure the distributional similarity. The KL distance, however, is not symmetric. To address this issue, Baker and McCallum [8] proposed to utilize the average of KL divergence of each distribution as the clustering similarity metric. Such metric, however, is not well-grounded or theoretically principled.

To address the issue above, we propose to utilize the Information Bottleneck (IB) principle to guide the clustering process. The Information Bottleneck regards clustering as a process of data compression (compressing a group of data into one cluster) [106]. Given the joint distribution $P(\mathbf{V}, \mathcal{C})$ of the delta visual phrases $\mathbf{V} = \{v\}$ and image classes $\mathcal{C} = \{c\}$, the goal of IB principle is to construct the optimal compact representation of $\mathbf{V}$, namely the visual synset clusters $\mathbf{S} = \{s\}$, such that $\mathbf{S}$ preserves as much information as possible about $\mathcal{C}$. In particular, the IB principle is reduced to the following Lagrangian optimization problem to maximize

$$\mathcal{L}[P(s \mid v)] = I(\mathbf{S}; \mathcal{C}) - \beta I(\mathbf{V}; \mathbf{S}) \tag{3.3}$$

with respect to $P(s|v)$ and subject to the Markov condition $\mathbf{S} \leftarrow \mathbf{V} \leftarrow \mathcal{C}$. The mutual information $I(\mathbf{S};\mathcal{C})$ measures the information that $\mathbf{S}$ contains about $\mathcal{C}$ and is defined as below.

$$I(\mathbf{S};\mathcal{C}) = \sum p(s)p(c \mid s) \log \frac{p(c \mid s)}{p(c)}. \tag{3.4}$$

β is the lagrange multiplier controlling the tradeoff between data compression and information preservation. Intuitively, Eq. 3.3 aims to cluster or compress the delta visual phrases into visual synsets through a compact bottleneck, under the constraint that this compression keeps the information about image classes as much as possible.

The IB optimization in Eq. 3.3 yields the solution of: (1) the prior probability $P(s)$ for each visual synset cluster $s \in \mathbf{S}$; (2) the membership probability $P(s|v)$ of delta visual phrase v to its visual synset cluster s; and (3) the visual synset distribution $P(c|s)$ over image classes, which are specifically defined in the equations below:

$$\begin{cases} P(s) = \sum\limits_{v} P(s|v)P(v) \\ P(c|s) = \dfrac{1}{P(s)} \sum\limits_{v} P(s|v)P(v)P(c|v) \\ P(s|v) = \dfrac{P(s)}{Z(\beta, v)} exp(-\beta D_{KL}[P(c|v)||P(c|s)]) \end{cases} \tag{3.5}$$

where $Z(\beta, v)$ is the normalization factor, β is a lagrange parameter and $D_{KL}[P(c|v)||P(c|s)]$ is the Kulback-Libeler divergence between $P(c|v)$ and $P(c|s)$. The solutions for the self-contained equations above can be obtained by starting with a random solution and then iterating the equations. More importantly, this procedure is guaranteed to converge [118].

3.4.3 Sequential IB clustering

We adopt the sequential Information Bottleneck (sIB) clustering algorithm [107, 108] to generate the optimal visual synset clusters in our approach, as it is reported to outperform other IB clustering techniques [107].

The target principled function that sIB algorithm exploits to guide the clustering process is $\mathcal{F}(s) = \mathcal{L}[P(s|v)]$ as shown in Eq. 3.3. The sIB algorithm takes visual synset cluster cardinality $|s|$, and joint probability $P(v, c)$ as input, where $P(v, c) = \frac{N_v(c)}{N(c)}$ [2] and starts with some initial random clustering $s = \{s_1, s_2, ..., s_K\}$ on $\mathcal{V}$. At each iteration, sIB takes some $v \in \mathcal{V}$ from its current cluster $s(v)$ and reassigns it to another cluster s^{new} such that the cost (or information lost) of merging v into s^{new} is minimum. As the mutual information function in Eq. 3.3 is decomposable, the target function can be rewritten as $\mathcal{F}(s) = \sum_i \mathcal{F}(s_i)$. Thus, the merging cost $d_{\mathcal{F}}(v, s^{new})$ can be defined as the target function difference before and after reassigning v:

$$d_{\mathcal{F}}(x, s^{new}) = \mathcal{F}(\{s^{new}, v\}) - \mathcal{F}(\{s^{new}\}) - \mathcal{F}(\{v\}) \tag{3.6}$$

Specifically, $d_{\mathcal{F}}(v, s^{new})$ is defined as (cf. [107] for more details):

$$d_{\mathcal{F}}(v, s^{new}) = (P(v) + P(s)) \cdot JS(P(c \mid v), p(c \mid s)) \tag{3.7}$$

where $JS(\cdot, \cdot)$ is the *Jensen-Shannon* divergence [65].

As the new cluster of v is $s^{new} = \arg\min_{s \in s} d_{\mathcal{F}}(v, s)$, the reassignment of v either leads to a new clustering s^{new} such that $\mathcal{F}(s^{new}) > \mathcal{F}(s)$ or no changes to the original clustering $s^{new} = s$. It is, therefore, easy to verify that the sIB clustering will always converge, at least to a local maximum of target function $\mathcal{F}(s)$. Specifically, the convergence speed depends on threshold ε. The clustering is deemed to be converged, if the number of cluster assignment changes in the loop is

[2] $N_v(c)$ is the frequency of delta visual phrase v in image class c, and $N(c)$ is the total number of delta visual phrase in c

less than $\varepsilon \cdot \mid \mathbf{V}|$. In order to avoid being trapped in a local optima, several runs of clustering with different random initialization are repeated and the run with highest target function $\mathcal{F}(s)$ is chosen. Note that sIB utilized here is a "hard" clustering process, in which $P(s \mid v)$ is deterministic and one visual word belongs to only one visual synset.

3.4.4 Theoretical analysis of visual synset

Here, we theoretically analyze the properties of visual synset in various aspects and compare it with other similar intermediate features by pLSA and LDA in the literature.

Advantages of visual synset: The image representation of visual synsets possesses several appealing properties and advantages over other part-based image representation. First, the visual synset provides a means to connect images of same class but with different visual appearances (visual words), and therefore, it can reasonably address the huge intra-class image variation problem. For example in Figure 3.7, two visually different salient components (delta visual phrases) of motorbikes can be grouped into one visual synset, based on their image class probability distribution. Consequently, the visually different motorbike images will now have some commonality in the feature space.

Second, as the visual synset is built on top of bag of visual words, it inherits all the advantages of bag of visual words, such as robustness to different viewpoints, poses, lighting conditions, scale changes etc. Third, from the statistical point of view, visual synset can be regarded as a feature selection or generation process via supervised dimensionality reduction on visual words. Most well-known feature selection schemes, such as Mutual Information, Information Gain, etc, consider each feature individually, in a greedy approach. In contrast, the visual synset implicitly incorporates the inter-relation among visual words. Fourth, by fusing several visual

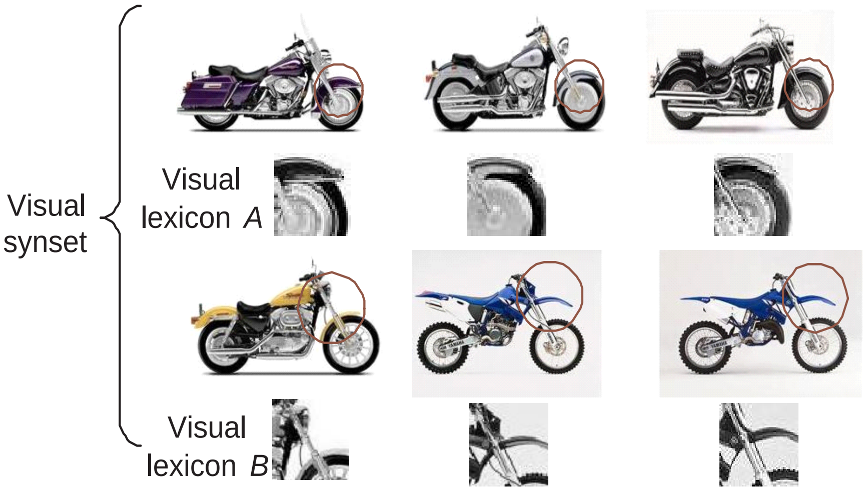


Figure 3.7: An example of visual synset generated from Caltech-101 dataset, which groups two delta visual phrases representing two salient parts of motorbikes.

words into one synset, the visual synset reasonably addresses the statistical sparseness problem that is obvious in the bag-of-words image representation. Moreover, compared to bag-of-words approach, visual synset provides a more compact image representation that results in better computational efficiency. The information bottleneck principle ensures that the compact representation has only a minor information loss, while compressing visual words to synsets. At last, the visual synset is more robust to occlusion and clutter in images. If the visual words from cluttered background follow the semantic inference distribution of meaningful visual words sampled from objects of interest, they will be absorbed into useful visual synsets respectively. If they follow the some spontaneous distributions, their negative effect in classification will also be limited by the majority of useful visual synsets

Computational complexity: The complexity of distributional clustering depends on the cost of the sequential Information Bottleneck process. At each iteration, the cost of merging a delta visual phrase to a synset is on the order of $O(|\mathcal{C}| \cdot |\mathbf{S}|)$. Hence, the time complexity of the whole process is on the order of $O(l \cdot |\mathbf{V}| \cdot |\mathcal{C}| \cdot |\mathbf{S}|)$, where l is the number of iterations in sequential IB clustering

[106]. Since in general $|\mathcal{C}|$ is small and $|\mathbf{S}|$ can be much smaller than $|\mathbf{V}|$, the computational complexity mainly relies on the number of delta visual phrases $|\mathbf{V}|$ linearly.

Comparison to pLSA and LDA: The pLSA and LDA approach are similar to the proposed visual synset in the way that they are all some kinds of intermediate features derived from primitive delta visual phrases. However, the proposed visual synset is different from pLSA and LDA in the way that visual synset is not a result of a generative model. Unlike pLSA and LDA, the proposed visual synset is not a latent or hidden semantic variable in the middle of delta visual phrases and image semantics. pLSA assumes a set of latent topic variable to tie up documents/images and words, while LDA treats a latent topic as a multinomial distribution over words and the mixture of latent topics per document/image [104]. The Markov condition in pLSA and LDA is $\mathbf{V} \leftarrow \mathbf{S} \leftarrow \mathcal{C}$ [118], where $\mathbf{V}$ denotes the delta visual phrase, like visual words, $\mathbf{S}$ denotes the latent topic variable, and $\mathcal{C}$ denotes object categories. On the contrary, the visual synset is the results of supervised data-mining process of compressing delta visual phrases via distributional clustering based on IB principle. Thus, it is only conditional on delta visual phrases, which follow the joint distribution of delta visual phrases and image classes. Consequently, the statistical causality is the Markov chain condition of $\mathbf{S} \leftarrow \mathbf{V} \leftarrow \mathcal{C}$, where $\mathbf{S}$ denotes visual synset variable, as shown in Figure 3.8.

3.4.5 Comparison to the analogy of text domain

Same as the BoW based image classification, its text analogy, automatic text categorization, suffers from synonymy too. The methods employed to address the synonymy problem in the text domain have brought much inspiration to this thesis. In this section, we survey the solutions for synonymy problems in text domain and emphasize its linkage to the proposed visual synset methodology.

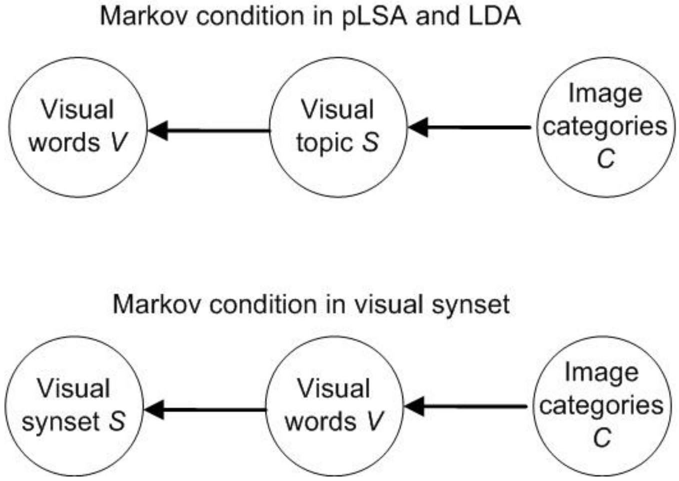


Figure 3.8: The statistical causalities or Markov condition of pLSA, LDA and visual sysnet.

By carrying the same semantic meaning, the synonymous textual words produce much ambiguity in text categorization. For instance, a text classifier trained on documents containing textual word "astronaut" for *space* category may not be able to recognize a new document about *space* topic, in which the word "cosmonaut" occurs [42]. One approach to tackle this issue is to utilize the synonymy set (synset) of textual words as terms for representation of documents. In this way, the documents with synonymous keywords can be represented in a semantic consistent manner, by a bag of synsets in the vector space. The weights of synsets in documents can then be computed using the same schemes for textual word terms [42]. WordNet is an intuitive source to obtain the synset knowledge [123]. Another approach to mine the synset knowledge is to exploit the distributional linguistic structures of textual words, from the computational perspective. Bekkerman et al. proposed a new document representation based on *word-clusters* and showed that the word-cluster representation can achieve better performance than bag of words [10]. A word-cluster is defined to be a group of words with similar distributions, but not necessarily similar semantic meaning. It is mined via the distributional clustering initially proposed by Pereira et al. [96]. This approach inspires us towards a new way of measuring the semantic of a word, which is its distribution. We borrow this idea to measure the "semantic" of a visual word by its probabilistic distribution. This measure circumvents the issue of how to represent the lexical semantics of a visual word, as it might not even exist.

3.5 Summary

In order to address the polysemy and synonymy issue of visual words, we proposed a novel image feature, *visual synsets*, for visual object categorization. To address the polysemy issue, we exploit the co-occurrence and spatial scatter information of visual words to generate a more distinctive visual configuration, i.e. delta visual phrase, for better inter-class distance. To tackle the synonymy issue, we proposed to group delta visual phrase with similar 'semantic' into a visual synset. Rather than in a conceptual manner, we define the 'semantic' of a delta visual phrase probabilistically as its image class probability distribution. The visual synset is, therefore, a probabilistic relevance-consistent cluster of delta visual phrases, which is learned by Information Bottleneck based distributional clustering. The effect of visual synset is to partially bridge the visual difference of images of the same class, and finally, reduce the intra-class variations.

Chapter 4

A Generative Learning Scheme beyond Visual Appearances

4.1 Motivation

Though the delta visual phrase and visual synset can partially alleviate the polysemy and synonymy issues caused by the visual diversity in objects, the gap between visual proximity and semantic relevance remains significant. This gap has presented us with a harsh reality: it is usually difficult to learn the visual characteristics of object categories for classification, as most object categories generally do not have any distinct visual characteristics. As shown in Figure 4.1, the open-ended nature of object appearances makes the objects possess huge variations of visual looks and shapes. Therefore, rather than directly modeling object semantics from low level visual features, we need some learning scheme that goes beyond visual appearances.

Here, we approach this problem by taking a Bayesian perspective. We interpret the huge diversity in object appearances as a generative phenomenon, in which the diverse visual appearances arise from countably infinitely many common visual *appearance pattern*s, as shown in Figure 4.2. In this probabilistic genera-

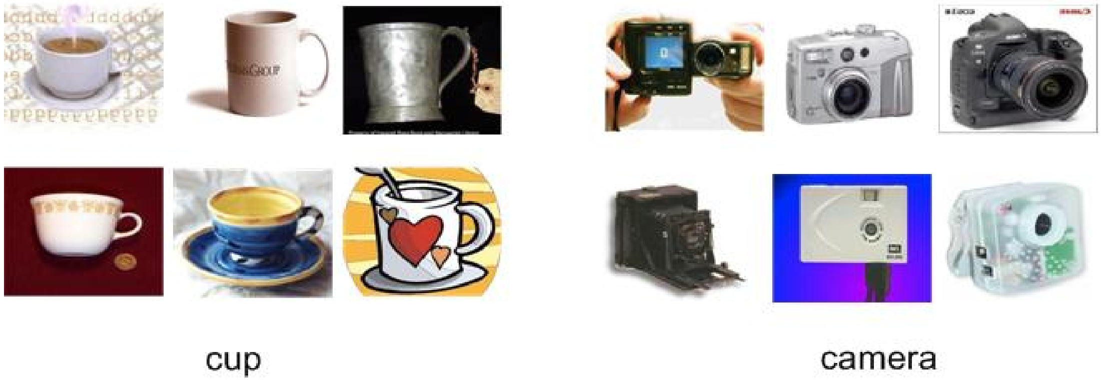


Figure 4.1: The objects of same category may have huge variations in their visual appearances and shapes.

tive interpretation, different object categories can still be visually similar and share similar visual appearance patterns. However, the distribution and combination of appearance patterns can be distinct for different object categories. The object categorization can then be cast as a problem of analyzing the distribution and combination of appearance patterns or the visual thematic structure of object categories. Effectively, objects of same class that are visually different can be adjacent in visual appearance pattern space. Hence, the appearance patterns can partially bridge the visual appearance difference of objects.

However, to make the aforementioned generative interpretation valid, three issues must be tackled. First, there should exist countably infinitely many appearance patterns, as the object visual diversity is boundless. Second, all the object categories should share a universal set of visual appearance patterns, as the objects of different categories can be visually similar too. Third, intuitively, the objects of same category should possess a closer set of appearance patterns than those of different categories. To embody the generative interpretation of object appearance, we tackle the three aforementioned issues by developing a hierarchical generative probabilistic model, named **nested hierarchical Dirichlet process (HDP) mixture**. The stick breaking construction process and Chinese restaurant

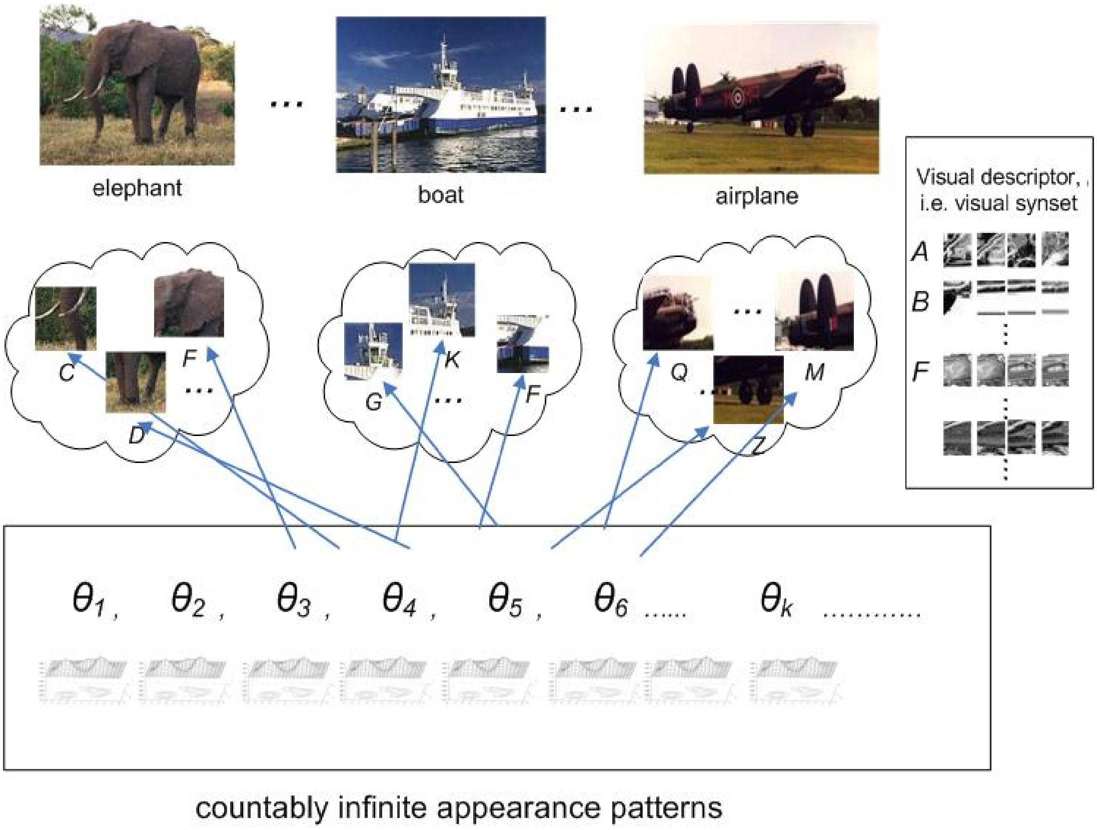

elephant
...
boat
...
airplane
Visual descriptor, i.e. visual synset
A
B
F
C
F
D
...
K
G
F
...
Q
...
M
Z
θ_1, θ_2, θ_3, θ_4, θ_5, θ_6 ...... θ_k ....
countably infinite appearance patterns

Figure 4.2: The generative interpretation of visual diversity, in which the visual appearances arise from countably infinitely many appearance patterns.

franchise representation [117] in the proposed nested HDP mixture model allow the countably infinitely many appearance patterns to be shared within and across different object categories. The designed model structure also enables the images of the same category to possess a closer set of appearance patterns.

4.2 Overview and preliminaries

Prior to presenting the proposed nested HDP mixture, we revisit the overall process of the visual category recognition in the thesis, so as to give a clear picture of the relation between visual synset and the nested HDP mixture model. Then, we

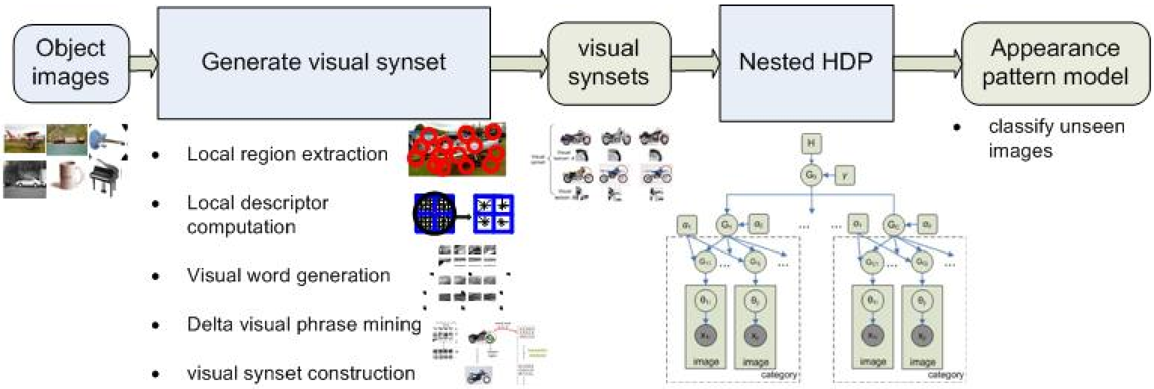


Figure 4.3: The overall framework of the proposed appearance pattern model.

introduce some background knowledge on generative probabilistic models.

Given a collection of object images of category $\mathcal{C} = \{c_j\}_{j=1}^{m}$, our target is to infer the category of a new unseen image. As shown in Figure 4.3, the first step is to extract local visual features and build image representation. We extract M regions $\{a_i\}_{i=1}^{M}$ from the image and compute visual features of regions a_i. We then perform k-means clustering on the region features to generate a codebook of W visual words $\mathbf{W} = \{w_1, .., w_W\}$. Following the method in Chapter 3, we build delta visual phrase and finally visual synset $\mathbf{S} = \{s_1, ..., s_L\}$ [143] on top of visual words w, by incorporating spatial and distributional information of visual words w_i. The image $\mathcal{I}$ is then represented by a bag of visual synsets $\{s_{(a_1)}\ , ..., \ s_{(a_i)}, ..\}$, where $s_{(a_i)}$ is the corresponding visual synset of region a_i.

By representing image $\mathcal{I}$ as a bag of visual synsets $\{s_{(a_1)}\ , ..., \ s_{(a_i)}, ..\}$, we then apply our generative interpretation and learn the appearance patterns shared by different object categories $\mathcal{C} = \{c_j\}_{j=1}^{m}$, via the proposed nested hierarchical Dirichlet process (HDP) mixture model. Furthermore, the visual thematic structure of each category c_j is determined by the learned appearance patterns and the categorization of unseen object images is then performed accordingly.

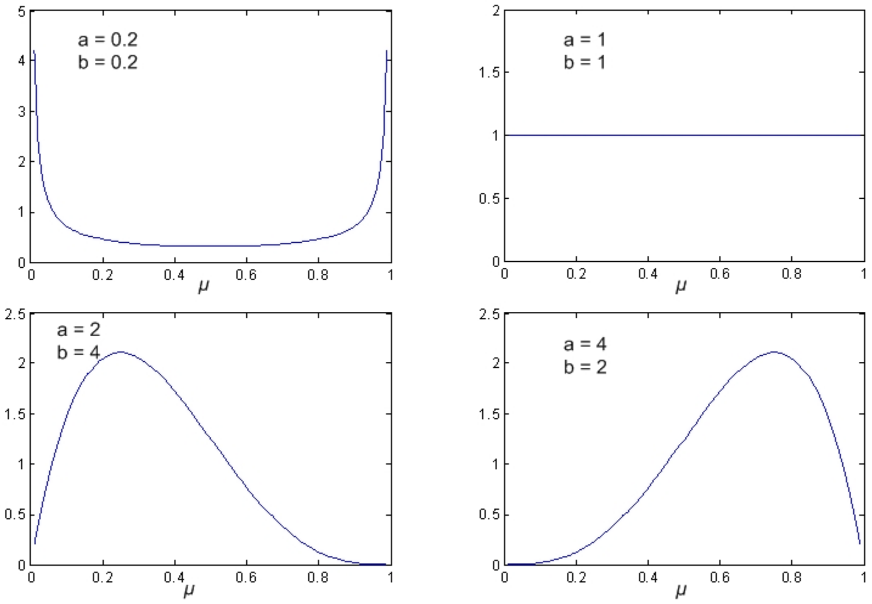


Figure 4.4: The plots of beta distributions with different values of a and b.

4.2.1 Basic concepts of probability theory

Here, we introduce some preliminary probability concepts relevant to the proposed nested HDP mixture.

Binary variables, Binomial and Beta distribution

Let random variable x denote the output of a probability experiment trial, with $x = 1$ representing "success" and $x = 0$ representing "fail". The random variable x is called binary variable, as it can take only one of two possible values of 0 and 1. Let μ denote the probability of success trial or $x = 1$. The binomial distribution is defined as the discrete probability distribution of the number of successes in a sequence of N independent trials [26, 6], as below.

$$\mathrm{Bin}(k \mid N, \mu) = \left(\begin{array}{c} N \\ k \end{array} \right) \mu^k (1-\mu)^{N-k}, \tag{4.1}$$

where k is the number of success trials, N is total number of trials and $\left(\begin{array}{c} N \\ k \end{array}\right)$ is the number of combinations of k objects out of N objects.

In the context of Bayesian learning, a beta distribution $p(\mu)$ is generally introduced to be the prior distribution of μ, as below.

$$\mathrm{Beta}(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}, \tag{4.2}$$

where $\Gamma(a)$ is the gamma function of a. One pleasant property of beta distribution is *conjugacy* [14], which means the posterior distribution of μ and the prior distribution have the same functional form [16]. Figure 4.4 shows the plots of beta distributions with different values of a and b. a and b are often called hyperparameters of μ, because they control the distribution of the parameter μ [16].

Multinomial variables, multinomial and Dirichlet distributions

Multinomial variable is a generalization of binary variable. A random variable x that denotes the output of a probability experiment trial is called a multinomial variable, if it can take one of K possible mutually exclusive values [1 2 ... K] [16, 17]. Let μ_k be the probability of x taking value of k. We then have $\sum\limits_k \mu_k = 1$. The multinomial distribution is the discrete probability distribution of the number of trials with different outputs in a sequence of N independent trials [26], as below.

$$\mathrm{Mult}(m_1, m_2, ..., m_K \mid \boldsymbol{\mu}, N) = \left(\begin{array}{c} N \\ m_1 m_2 ... m_K \end{array}\right) \prod_k \mu_k^{m_k}, \tag{4.3}$$

where m_k is the number of trials with $x = k$.

The Dirichlet distribution is a conjugate prior for the parameters $\boldsymbol{\mu}$ of the multinomial distribution above, which is defined as below.

$$\mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_k \mu_k^{\alpha_k - 1}, \tag{4.4}$$

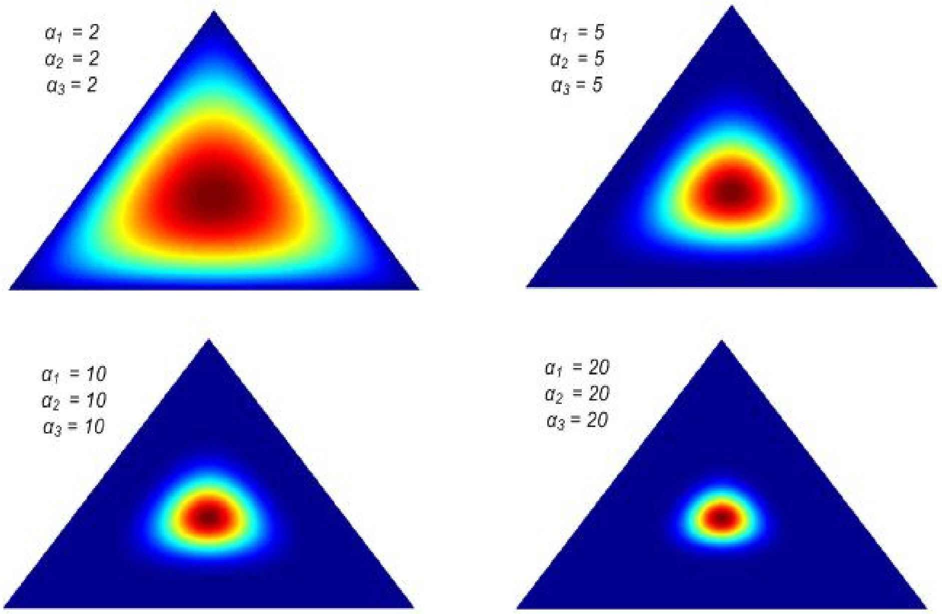


Figure 4.5: The plots of 3-dimensional Dirichlet distributions with different values of α. The triangle represents the plane where (μ_1, μ_2, μ_3) lies due to the constraint $\sum \mu_k = 1$. The color indicates the probability for the corresponding data point.

where $\boldsymbol{\alpha} = \{\alpha_k\}$ are the hyperparameters of $\boldsymbol{\mu} = \{\mu_k\}$, $\alpha_0 = \sum_{k=1}^{k=K} \alpha_k$ and $\Gamma(\alpha_k)$ is the gamma function of α_k. Because the summation of μ_k is equal to 1, the Dirichlet distribution of μ_k is a simplex of dimensionality $K - 1$. Figure 4.5 illustrates the plots of a 3-dimensional Dirichlet distribution with different parameter values.

4.3 A generative interpretation of visual diversity

We interpret the huge variation of object appearances as a generative probabilistic phenomenon, in which different visual appearances arise from a common set of countably infinitely many appearance patterns. The appearance pattern is to represent the inter-relation of primitive visual descriptors. Specifically, it can be regarded as a distribution of random variables of visual descriptors, which can correspond to any image descriptor, like color histogram, visual words, etc. Here, we utilize visual synset as the image descriptor. By describing an image $\mathcal{I}$ as a bag of visual synsets $\{s_{(a_1)}, ..., s_{(a_i)}, ..\}$, we define the **appearance pattern** as below:

Definition 4.3.1. ***The appearance pattern*** *is a cluster or distribution $F(\theta)$ of random variable visual synset s that reveals the visual or contextual relatedness of visual synsets across different object categories, where $s \mid \theta \sim F(\theta)$ and θ is the parameter associated with the appearance pattern.*

An appearance pattern can be regarded as a cluster of visual synsets with $F(\theta)$ as membership function. By representing a facet of inter-relation of primitive visual descriptors as a cluster, an appearance pattern effectively depicts one aspect of the object visual characteristics. The modeling of object categories is now based on inter-relation of visual descriptors, rather than directly on primitive visual descriptors. To learn the appearance pattern, we assume a generative process for object images as below:

- Given a pool of countably infinitely many appearance patterns, for each image $\mathcal{I}$ in corpus I:
    - For each local region a_i in image $\mathcal{I}$:

      1. Sample one appearance pattern or $F(\theta)$ from the appearance pattern pool; and

      2. Sample a region feature (visual synset s) conditioned on the appearance pattern $F(\theta)$: $s \sim F(\theta)$

The aforementioned generative process also imposes two issues for the appearance pattern modeling, which are

(a) The model should provide the possibility of a countably infinite number of appearance patterns, as the visual diversity of objects is boundless.

(b) The appearance patterns should be shared not only within but also across all object categories, as the objects from different categories can share similar visual appearance too.

Here, we add one more issue to reflect the inter-relation between appearance patterns, objects and categories, as below.

(c) Intuitively, the objects of the same category should arise from a closer set of appearance patterns than those of different categories.

For ease of reference, we list these three issues in Table 4.1.

Based on the generative process, we approach the appearance pattern modeling by leveraging a mixture model of object category, in which an appearance pattern corresponds to a mixture component. The appearance patterns can, therefore, be obtained by applying mixture model learning on the object categories. However, most of the existing mixture models, like Gaussian mixture model, latent Dirichlet allocation, etc, are not able to tackle the three issues in Table 4.1 simultaneously. Therefore, we proposed a probabilistic generative model, called **nested HDP mixture**, based on the hierarchical Dirichlet process (HDP) [116], to model the appearance patterns. The proposed nested HDP mixture is a hierarchical generalization of hierarchical Dirichlet Process (HDP) mixtures. It provides a Bayesian mixture model that learns the hidden structures of related groups of data. The stick breaking construction process, the Chinese restaurant franchise representation and its hierarchical structure enable the nested HDP mixture to tackle all the issues listed in Table 4.1. Prior to presenting nested HDP mixture, let us introduce HDP mixture first.

Table 4.1: Three issues in the generative interpretation of object appearance diversity.

(a) The model should provide the possibility of a countably infinite number of appearance patterns, as the visual diversity of objects is boundless

(b) The appearance patterns should be shared across all object categories, as the objects from different categories may share similar visual appearance.

(c) Intuitively, the objects of the same category should arise from a closer set of appearance patterns than those of different categories

4.4 Hierarchical Dirichlet process mixture

The hierarchical Dirichlet process (HDP) mixture [117] is a hierarchical organization of a number of Dirichlet process mixtures that share common global parameter atoms. In the HDP mixture model, each mixture component is assumed to correspond to one appearance pattern $F(\theta)$. As the random variable of visual synset s is discrete and can only take up one of values $\{s_1, .., s_L\}$, $F(\theta)$ can, therefore, be assumed to be a multinomial distribution of visual synset s. The task now is to learn the pattern parameter θ only. Prior to presenting the hierarchical Dirichlet process, the learning of pattern parameter θ in Dirichlet process (DP) mixture is introduced first, as the stick breaking construction process in DP facilitates the proposed nested HDP mixture in tackling the issue (a) in Table 4.1.

4.4.1 Dirichlet process mixtures

• **Dirichlet Process (DP):** Let H be a probability measure (or distribution) on the appearance pattern parameter space Θ and γ be some real positive number. Dirichlet process (DP), denoted by DP(γ, H), is defined as below [117].

Definition 4.4.1. *A **Dirichlet process** DP(γ, H) is a distribution over measures on Θ, such that for any finite partition $(Q_1, ..., Q_v)$ of appearance pattern parameter space Θ, the random vector $(G(Q_1)$,..., $G(Q_v))$ is distributed as a Dirichlet distribution as follows.*

$$(G(Q_1), ..., G(Q_v)) \sim Dir(\gamma H(Q_1), ..., \gamma H(Q_v)) \tag{4.5}$$

Here, γ is the scalar concentration parameter that controls the similarity of samples $G \sim$ DP(γ, H) from the base measure H. Intuitively, G is a measure or distribution of the appearance pattern parameter θ on the space Θ, and the Dirichlet process (γ, H) is the distribution that governs G. The sample G from DP(γ, H) is a discrete distribution with probability one. This property is ensured by the *stick breaking construction* process in DP [117] as follows:

$$\begin{aligned} &\theta_k \sim H \\ &\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l), \qquad \beta'_k \sim Beta(1,\gamma) \\ &G(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k), \end{aligned} \tag{4.6}$$

where β_k denotes the probability of drawing θ_k.

As illustrated in Figure 4.6, a metaphor of stick breaking construction is as follows. Given a stick of length 1, at first we break it and regard the stick partition as β_1. We then recursively break the remaining portion of the stick to obtain β_2, β_3,... and so on. The stick breaking construction process of DP gives rise to two implications. First, the probability $\boldsymbol{\beta} = \{\beta_1, ...\beta_k, ...\}$ is generated by using random

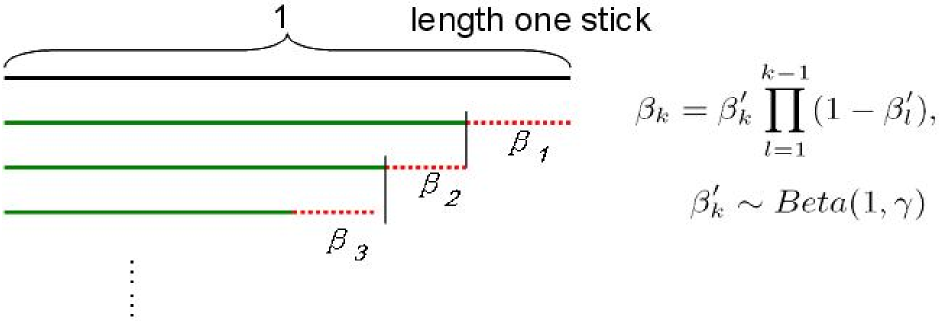

1
length one stick
β_1
β_2
β_3
$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l),$
$\beta'_k \sim Beta(1,\gamma)$

Figure 4.6: The stick breaking construction process.

variable β' to partition a unit length stick. Second, governed by $\boldsymbol{\beta}$, the parameter θ of $G(\theta)$ is actually from θ_k, which is independently drawn from the base measure H. The θ_k from the base measure H then forms the pattern parameter atoms, as defined below.

Definition 4.4.2. *The **parameter atoms** are the set of infinitely many parameters $\{\theta_k\}_{k=1}^{\infty}$ drawn independently from the base measure H. $G(\theta)$ is the probability measure or distribution of these parameter atoms based on $\boldsymbol{\beta}$.*

For terminology simplicity, we let $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$ denote a sample $G(\theta)$ drawn from the stick breaking process, with probability measure $\boldsymbol{\beta}$ on θ (GEM stands for Griffiths, Engen and McClosekey; refer to [117] for detail). By providing infinite parameter atoms, the stick breaking construction process can facilitate the proposed nested HDP mixture in handling issue (a) in Table 4.1.

• **Dirichlet Process Mixture Model:** For terminology simplicity, let x_i be an observation denoting the visual synset $s_{(a_i)}$ of local region a_i. A mixture model for the observation x_i can be built by using the Dirichlet Process as non-parametric prior on the parameters of the model. In the mixture model, the observation x_i arises in a generative process as follows:

$$\begin{aligned} &\theta_i \mid G \sim G \\ &x_i \mid \theta_i \sim F(\theta_i), \end{aligned} \tag{4.7}$$

where $F(\theta_i)$ is the appearance pattern or the probability distribution of the observation x_i with pattern parameter θ_i. Here, we let an indicator variable z_i of integer value denote the appearance pattern (or mixture component) index in $G(\theta)$ associated with observation x_i, such that $x_i \sim F(\theta_{z_i})$. According to Eq. (4.6) and (4.7), the Dirichlet process mixture follows the generative process below:

(a) $\boldsymbol{\beta}$ is distributed according to the stick breaking construction process: $\boldsymbol{\beta} \mid \gamma \sim \mathrm{GEM}(\gamma)$

(b) The pattern parameter atom θ_k is distributed according to the base measure H: $\theta_k \mid H \sim H$

(c) For each observation x_i:

- **(i)** Sample a pattern parameter index z_i: $z_i \mid \boldsymbol{\beta} \sim \boldsymbol{\beta}$; and
- **(ii)** Sample an observation x_i from $F(\theta_{z_i})$:
  $x_i \mid z_i, (\theta_k)_{k=1}^{\infty} \sim F(\theta_{z_i})$

According to the sticking breaking construction process or GEM(γ) in Eq (4.6), the DP mixture has partially handled the issue (a) in Table 4.1. To fully tackle issues (a) and (b), we add "hierarchy" into the Dirichlet process mixture.

4.4.2 Hierarchical organization of Dirichlet process mixture

As introduced in Section 4.2.1, a Dirichlet process (DP) can provide a mixture model on one group of data. An intuitive approach to model the appearance patterns of multiple object categories is to apply one DP on each object category and let all the DPs share the common base measure H. According to Definition 4.4.2, the common base measure H will make DPs of all object categories share a common set of pattern parameter atoms θ and effectively the same set of appearance patterns $F(\theta)$. The aforementioned solution is plausible but not valid. This is so

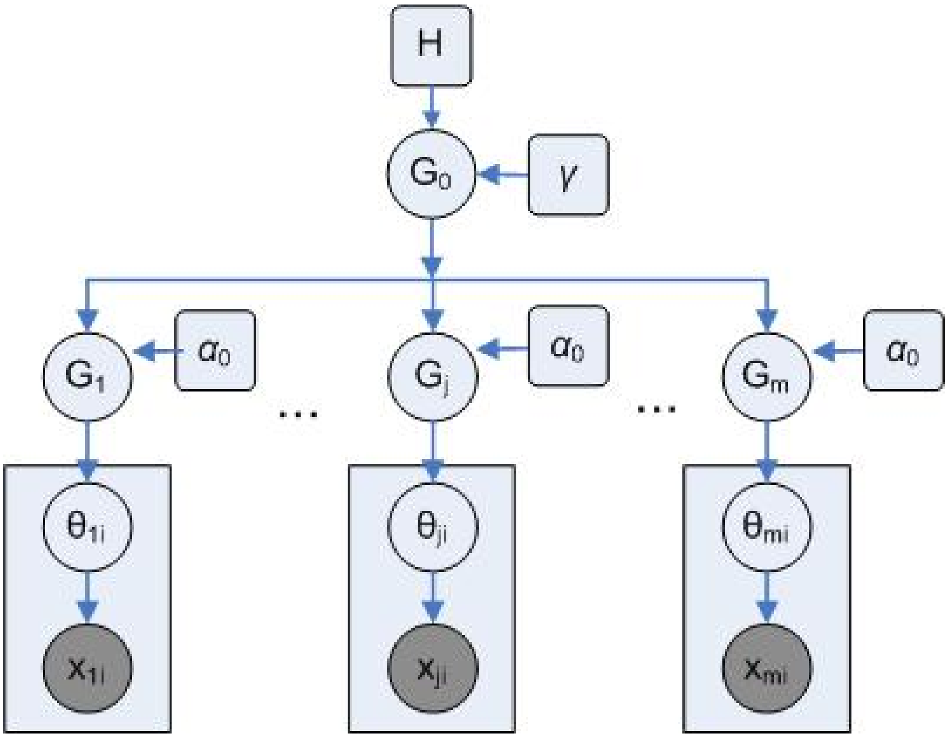


Figure 4.7: The graphical model of hierarchical Dirichlet process.

because the base measure H is a continuous distribution and its parameter atoms are uncountably infinite, while DPs have discrete distribution with probability one and their parameter atoms are countably infinite. Consequently, the parameter atoms in the base measure H and DPs will not be one-to-one corresponding. In other words, the DPs will not share exactly the same set of parameter atoms. Effectively, the object categories will not share a common set of appearance patterns, which conflicts with both issues (a) and (b).

To address the problem above, we simply put a DP on top of DPs of object categories, as the base measure. This leads to the hierarchical Dirichlet process (HDP) [117], as shown in Figure 4.7. G_0 drawn from DP(γ, H) is the global random probability measure for G_j drawn from DP of object category j.

$$\begin{aligned} G_0 \mid \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ G_0(\theta) &= \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k), \end{aligned} \tag{4.8}$$

where γ is the concentration parameter, $\theta_k \sim H$ and $\boldsymbol{\beta} = \{\beta_1, ...\beta_k, ...\} \sim \mathrm{GEM}(\gamma)$. As shown in Eq (4.8), G_0 is a discrete distribution on countably infinite pattern parameters. This countable infinity ensures the one-to-one correspondence of parameter atoms θ between base measure G_0 and its descendant DPs. Hence, the global measure G_0 enables G_js of all object categories to share a common set of parameter atoms θ, which can tackle both issues (a) and (b). The distribution G_j from DP of category j is defined as:

$$\begin{aligned} G_j \mid \alpha_0, G_0 &\sim \mathrm{DP}(\alpha_0, G_0) \\ G_j(\theta) &= \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k), \end{aligned} \tag{4.9}$$

where π_{jk} denotes the probability of θ_k in G_j.

According to Eq (4.8) and (4.9), the hierarchical Dirichlet process mixture can be summarized as the following generative process:

(a) $\boldsymbol{\beta}$ is distributed according to the stick breaking construction process: $\boldsymbol{\beta} \mid \gamma \sim \mathrm{GEM}(\gamma)$

(b) The appearance pattern atom parameter θ_k is distributed according to the base measure H: $\theta_k \mid H \sim H$

(c) For each category j:

- **(i)** $\boldsymbol{\pi}_j = \{\pi_{j1}, ..., \pi_{jk}, ...\}$ is distributed according to $\mathrm{DP}(\alpha_0, G_0)$: $\boldsymbol{\pi}_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0)$

- **(ii)** For each observation x_{ji}:

1. Sample a component parameter index z_{ji}: $z_{ji} \mid \boldsymbol{\pi}_j \sim \boldsymbol{\pi}_j$; and

2. Sample an observation x_{ji} from $F(\theta_{z_{ji}})$:
$x_{ji} \mid z_{ji}, (\theta_k)_{k=1}^{\infty} \sim F(\theta_{z_{ji}})$

• **The Chinese Restaurant Franchise:** As shown in Figure 4.8, the hierarchical Dirichlet process can be described by a metaphor called the *Chinese restaurant franchise* [117]. In the metaphor, there is a Chinese restaurant franchise, of which all the restaurants share a global menu of dishes from G_0. The restaurant j corresponds to G_j. The customer i at restaurant j corresponds to observation x_{ji} (visual synset feature of a local region), and the global menu of dishes correspond to the K parameter atoms $\theta_1, ..., \theta_K$ from G_0. The ith table t_{ji} at restaurant j has only one dish k_{ji} (appearance pattern parameter) that is shared by all the customers (observations) sitting there. The dish (appearance pattern parameter) is decided by the first customer (observation) of the table. By integrating G_0 and G_j, we can obtain the conditional distribution of table and dish assignment variables as follows:

$$p(t_{ji}|t_{j1}, ..., t_{ji-1}, \alpha_0, G_0) \sim \sum_t n_{jt.}\delta(t_{ji}, t) + \alpha_0\delta(t_{ji}, \bar{t}), \tag{4.10}$$

$$p(k_{ji}|k_{11}, k_{12}, ..., k_{21}, ..., k_{ji-1}, \gamma) \sim \sum_k m_{.k}\delta(k_{ji}, k) + \gamma\delta(k_{ji}, \bar{k}), \tag{4.11}$$

where $n_{jt.}$ is the number of customers in restaurant j at table t, and $m_{.k}$ is the number of tables in all restaurants serving dish k (or assigned with parameter atom θ_k).

Metaphorically, Eq (4.10) tells that a new customer prefers to sit at a table that has many customers already. Meanwhile, the customer can pick a new table to sit at with the probability conditioned on α_0. Eq (4.11) tells that customers tend to order popular dishes, but a new dish can be ordered too with the probability conditioned on γ. Intuitively, Eq (4.10) and (4.11) depict that in the hierarchical

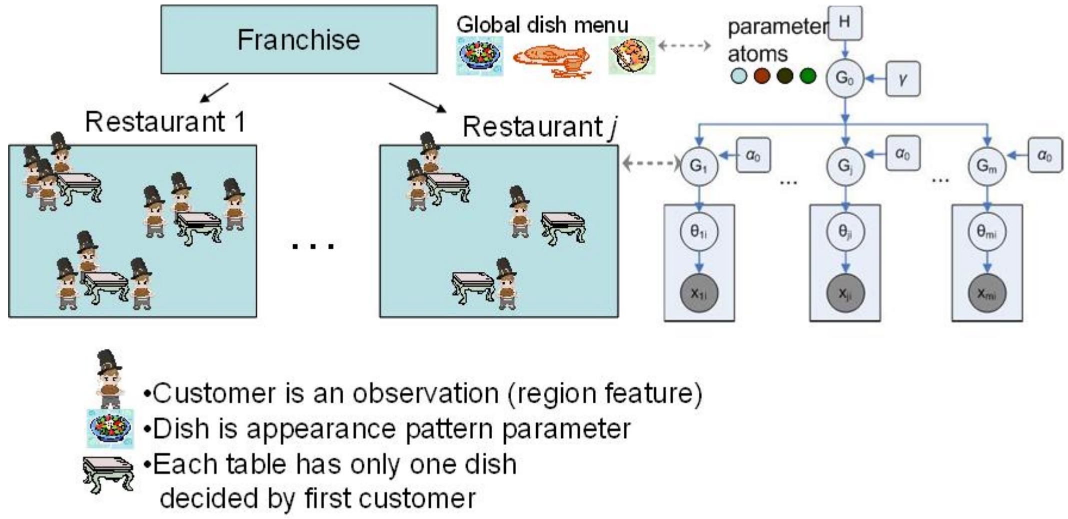

Franchise
Global dish menu
parameter atoms
H
G₀
γ
Restaurant 1
Restaurant j
. . .
G₁
α₀
Gⱼ
α₀
Gₘ
α₀
θ₁ᵢ
θⱼᵢ
θₘᵢ
x₁ᵢ
xⱼᵢ
xₘᵢ
•Customer is an observation (region feature)
•Dish is appearance pattern parameter
•Each table has only one dish decided by first customer

Figure 4.8: The Chinese restaurant franchise representation of hierarchical Dirichlet process. The restaurants in the franchise share a global menu of dishes from G_0. The restaurant j corresponds to DP G_j. The customer i at restaurant j corresponds to observation x_{ji} and the global menu of dishes correspond to the K parameter atoms $\theta_1, ..., \theta_K$ from G_0.

Dirichlet process, the object category can either reuse the existing appearance patterns or create a new appearance pattern to capture the visual diversity. Figure 4.8 illustrates the metaphorical correspondences between the components of the Chinese restaurant franchise and the hierarchical Dirichlet process.

4.4.3 Two variations of HDP mixture

Based on the HDP mixture introduced above, we have two variations of HDP mixture with different representation of the mixture model and Chinese restaurant franchise.

- **Mixture model (a):** As shown in Figure 4.9, the first variation of HDP mixture, named mixture model (a), incorporates two levels of Dirichlet process. This model is similar to the one introduced in [116]. In mixture model (a), each category is assumed to correspond to one restaurant in the Chinese restaurant

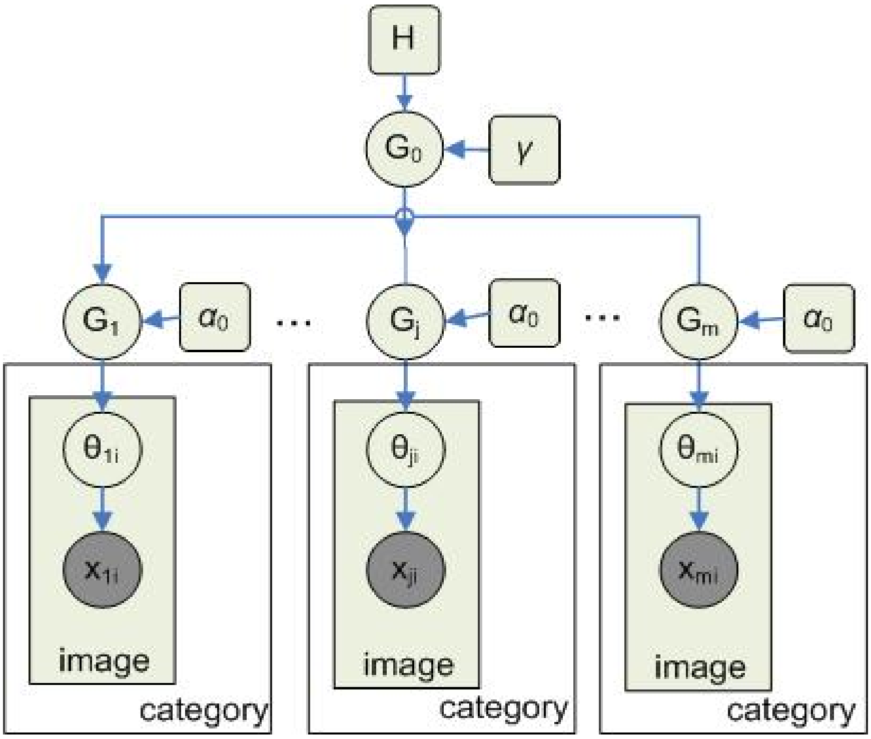


Figure 4.9: HDP mixture variation model (a): each category corresponds to one restaurant and all the images of that category share one single DP.

franchise representation and all the images of a category share one DP of that category. This arrangement allows the appearance patterns to be naturally shared within and across different object categories. In this model, the probability $p(\theta_k \mid c_j)$ of an appearance pattern θ_k in a given category j can be computed by measuring the count of assigning global appearance pattern k to the tables of category j.

$$p(\theta_k \mid c_j) = \hat{\pi}_{jk}, \tag{4.12}$$

where $\hat{\pi_{jk}}$ denotes the weight assigned to pattern k by the tables of category j.

- **Mixture model (b):** Figure 4.10 shows the second variation of HDP mixture, namely mixture model (b). This is similar to the model presented in [125]. Mixture model (b) assumes that the appearance patterns are shared at the level of images. In this model, each image is assumed to correspond to one restaurant in

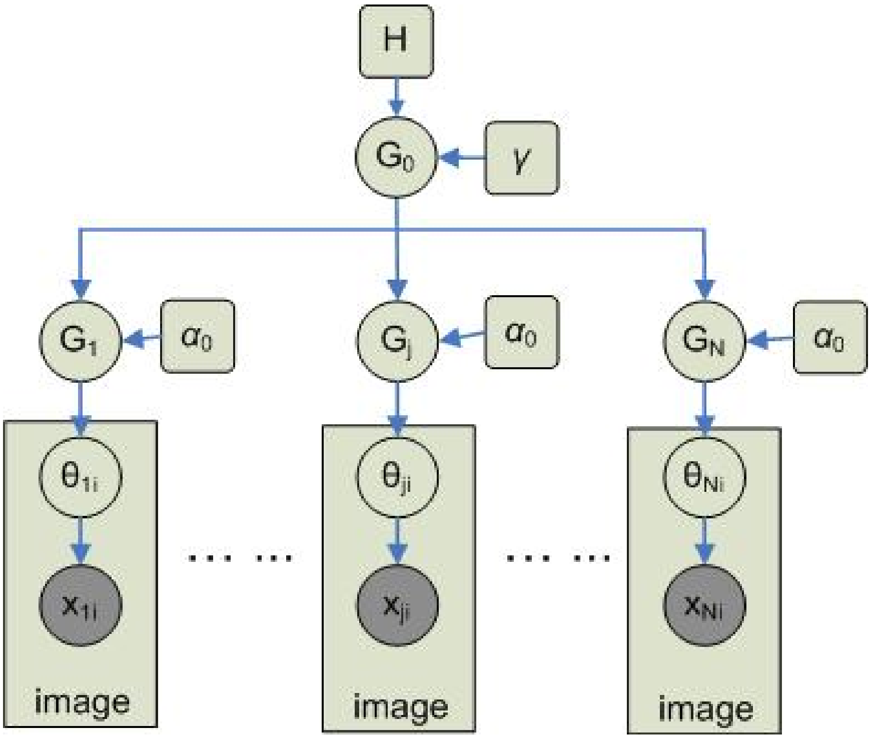


Figure 4.10: HDP mixture variation model (b): each image corresponds to one restaurant and has one DP respectively.

the Chinese restaurant franchise representation and have one DP drawn from the global base measure G_0. In this model, the probability $p(\theta_k \mid c_j)$ of a pattern θ_k in a given category j can be computed by counting frequency of pattern k assigned to the tables of all categories.

4.5 Nested HDP mixture

So far, the presented model can handle both issues (a) and (b) in Table 4.1, as it allows countably infinitely many appearance patterns to be shared within and across object categories. However, the issue (c), namely the inter-relation of appearance pattern, object and category, is neglected. To tackle this issue, we propose a generative probabilistic mixture model, called nested HDP mixture, based on a three-level tree of Dirichlet processes as shown in Figure 4.11. In this model, the categories are ensured to share a global set of appearance patterns, by drawing DPs from the base measure G_0. Moreover, each image in the category also has its own DP drawn from the category DP, as follows.

$$\begin{aligned} G_j \mid \alpha_0, G_0 &\sim \mathrm{DP}(\alpha_0, G_0), \text{ for category } j \\ G_{ji} \mid \alpha_1, G_j &\sim \mathrm{DP}(\alpha_1, G_j), \text{ for image } i \text{ in category } j \end{aligned} \tag{4.13}$$

This designed hierarchical structure tackles issue (c) in Table 4.1, in the way that the appearance patterns for images of the same category are sampled from their own category DP and the category DPs are governed by a global based measure DP G_0. The nested HDP mixture can be summarized as the following generative process.

(a) $\boldsymbol{\beta}$ is distributed according to the stick breaking construction process: $\boldsymbol{\beta} \mid \gamma \sim \mathrm{GEM}(\gamma)$

(b) The appearance pattern atom parameter θ_k is distributed according to the base measure H: $\theta_k \mid H \sim H$

(c) For each category j:

- **(i)** $\boldsymbol{\pi}_j = \{\pi_{j1}, ..., \pi_{jk}, ...\}$ is distributed according to $\mathrm{DP}(\alpha_0, G_0)$: $\boldsymbol{\pi}_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0)$
- **(ii)** For each image $\mathcal{I}_{ji}$:
  - **(a)** $\boldsymbol{\pi}_{ji}$ is distributed according to $\mathrm{DP}(\alpha_1, G_j)$: $\boldsymbol{\pi}_{ji} \mid \alpha_1, G_j \sim \mathrm{DP}(\alpha_1, G_j)$
  - **(b)** For each observation x_{jik}:
    - **1.** Sample a component parameter index z_{jik}: $z_{jik} \mid \boldsymbol{\pi}_{ji} \sim \boldsymbol{\pi}_{ji}$; and
    - **2.** Sample an observation x_{jik} from $F(\theta_{z_{jik}})$: $x_{jik} \mid z_{jik} \sim F(\theta_{z_{jik}})$

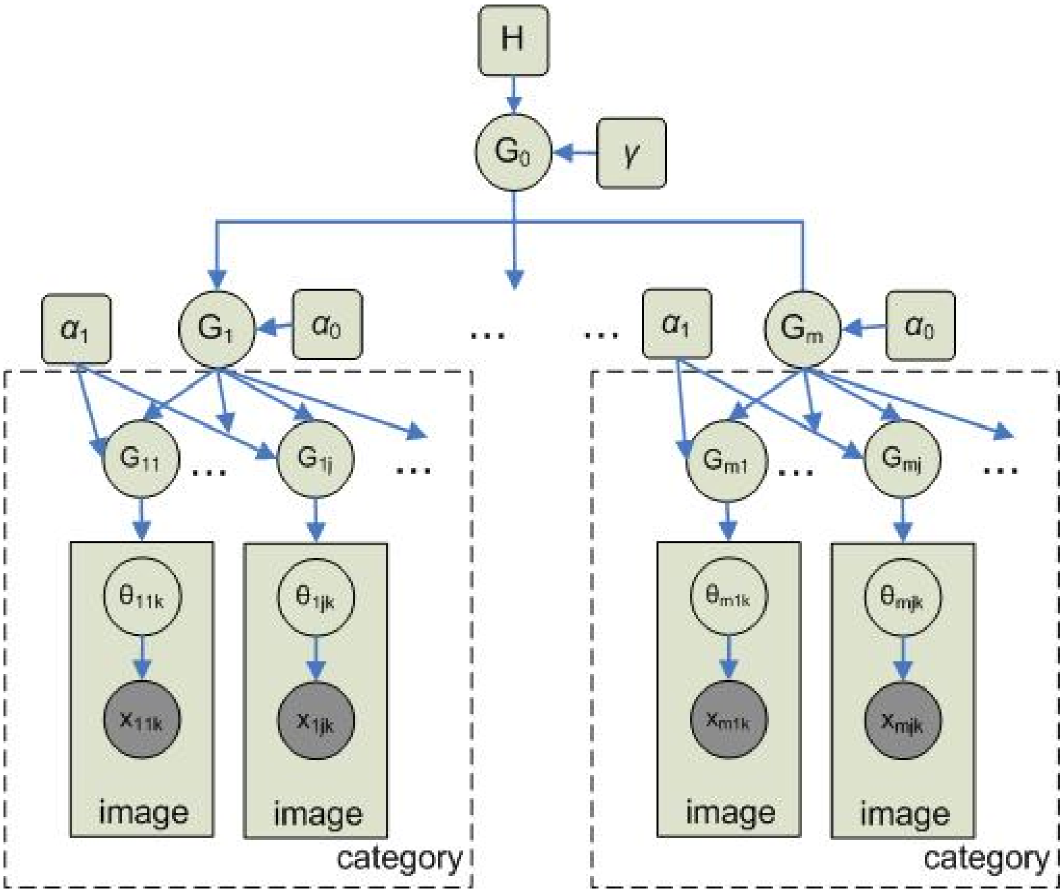


Figure 4.11: The proposed nested HDP mixture model: each category corresponds to one restaurant and has one DP. Each image corresponds to one restaurant in the next level and has one DP respectively.

Similar to mixture model (a) and (b), the probability $p(\theta_k \mid c)$ can be computed by counting frequency of pattern k assigned to the tables of the categories respectively.

4.5.1 Inference in nested HDP mixture

Based on the representation of Chinese restaurant franchise, we utilize the Gibbs sampling [33, 117] to perform the parameter inference in the nested HDP mixture model. For terminology simplicity, we only present the sampling process at one

Table 4.2: List of variables in Gibbs sampling for nested HDP mixture

Notation	**Explanation**
x_{ji}	the ith customer (observation) in restaurant j
$\boldsymbol{x}_{jt}$	the customers (observations) at table t in restaurant j
t_{ji}	the table i at restaurant j
$\boldsymbol{t}_{-ji}$	$\{t_{ji} : \text{all } j, i\}$ except t_{ji}
k_{ji}	the index of atomic parameter θ_k served at table i in restaurant j
$\boldsymbol{k}$	$\{k_{ji} : \text{all } j, i\}$
$\theta_{k_{ji}}$	the global parameter atom of table i in restaurant j.
$n_{jt.}$	the number of customers in restaurant j at table t
$m_{.k}$	the number of tables in all restaurants serving dish k
$m_{..}$	the number of tables in all restaurants
γ	the concentration parameter of parent DP
α_0	the concentration parameter of descendent DP

restaurant level. The sampling for the nested HDP mixture is simply to iterate this process for each DP from the bottom of the nested HDP structure.

The Chinese restaurant franchise representation has two types of assignments: (1) the restaurant j can have many local appearance patterns (tables) t_{ji}, which are assigned to the global atom patterns (global dish menu) k_{ji} and; (2) each observation (customer) x_{ji} is assigned with some local appearance pattern (table) t_{ji}. The Gibbs sampler, therefore, has two types of sampling: sampling the table assignment $\boldsymbol{t}$ of customers and sampling the pattern assignment $\boldsymbol{k}$ of tables. For clarity, we list the variables used in this Section in Table 4.2.

Sampling table assignment $\boldsymbol{t}$ of customers

Here, we compute the conditional posterior for t_{ji} by using the prior distribution of t_{ji} and the likelihood of observation x_{ji}. Eq (4.10) defines the prior of t_{ji}. The probability of t_{ji} taking a previously used value t is proportional to $n_{jt.}$, and the probability of t_{ji} taking a new value t^{new} is proportional to α_0. The likelihood of observation x_{ji} given t_{ji} with value of previously used t or $t_{ji} = t^{used}$ is: $f(x_{ji}|\theta_{k_{ji}})$. According to Eq (4.11), the likelihood of observation x_{ji} given $t_{ji} = t^{new}$ is then:

$$P(x_{ji} \mid t_{ji} = t^{new}, \boldsymbol{k}) = \sum_k \frac{m_{.k}}{m_{..} + \gamma} f(x_{ji}|\theta_{k_{ji}}) + \frac{\gamma}{m_{..} + \gamma} f(x_{ji}|\theta_{k^{new}}), \tag{4.14}$$

where $\boldsymbol{k} = \{k_{ji} : \text{all } j, i\}$, $m_{..}$ is the number of tables in all restaurants, and $m_{.k}$ is the number of tables serving dish k. The conditional distribution of t_{ji} is then:

$$P(t_{ji} = t|x_{ji}, \boldsymbol{t}_{-ji}, \boldsymbol{k}) \propto \begin{cases} \alpha_0 P(x_{ji} \mid t_{ji} = t^{new}, \boldsymbol{k}) & \text{if } t = t^{new} \\ n_{jt.} f(x_{ji}|\theta_{k_{ji}}) & \text{if } t = t^{used}, \end{cases} \tag{4.15}$$

where $\boldsymbol{t}_{-ji}$ denotes $\{t_{ji} : \text{all } j, i\}$ except t_{ji}. The probability for a restaurant to create a new local appearance pattern (table) is proportional to:

$$\begin{cases} \gamma f(x_{ji}|\theta_{k^{new}}) & \text{if } k = k^{new} \\ m_{.k} f(x_{ji}|\theta_{k_{ji}}) & \text{if } k \text{ is used} \end{cases} \tag{4.16}$$

Sampling dish assignment $\boldsymbol{k}$ of tables

The dish assignment k_{jt} of table j in restaurant t is shared by all the customers (observations) $\boldsymbol{x}_{jt}$ at that table, as there is only one dish for each table. Hence, the likelihood of $\boldsymbol{x}_{jt}$ given $k_{jt} = k$ is $f(\boldsymbol{x}_{jt}|\theta_{k_{jt}})$. The conditional probability of k_{jt} can then be computed as:

$$\begin{cases} \gamma f(\boldsymbol{x}_{jt}|\theta_{k^{new}}) & \text{if } k = k^{new} \\ m_{.k} f(\boldsymbol{x}_{jt}|\theta_{k_{ji}}) & \text{if } k \text{ is used} \end{cases} \tag{4.17}$$

The indicator variable z_{ji} of appearance pattern associated with observation x_{ji} can then be uniquely determined and updated by the observation x_{ji}'s table assignment t_{ji} and the corresponding table's pattern assignment k_{ji}

Insensitivity to hyperparameters and base distribution

Overall, Gibbs sampling is to infer the posterior distribution of table-customer assignment and dish-table assignment from conditional prior and observation likelihood, based on Bayes rules as below.

$$\text{posterior} \propto \text{conditional prior} \cdot \text{observation likelihood} \tag{4.18}$$

The conditional prior is jointly determined by the hyperparameters in base distribution and CRF representation, while the observation likelihood is derived from the distribution function $F(\theta)$ of visual synset and CRF representation. As Gibbs sampling is an iterative computational process, the posterior distribution is updated jointly by conditional prior and observation likelihood at each iteration. As the iteration goes on, the posterior distribution will be more and more dominated by the observation likelihood, namely the data, until convergence. Hence, the inference of posterior distribution is insensitive to the initial hyperparameters in the base distribution H.

4.5.2 Categorizing unseen images

After learning the appearance patterns, the nested HDP mixture can then perform categorization on unseen images. Assume the unseen image $\mathcal{I}$ consists of M local regions $\{a_i\}_{i=1}^{M}$ and is represented by the observation set of visual synsets $\{s_{(a_1)}, ..., s_{(a_i)}, ..s_{(a_M)}\}$. According to Definition 4.3.1, we have $p(s|\theta_k) = F(\theta_k)$. The likelihood probability $p(\mathcal{I} \mid c)$ of image $\mathcal{I}$ for category c can then be computed as:

$$p(\mathcal{I} \mid c) = \prod_i p(s_{(a_i)} \mid c) = \prod_i (\sum_k p(s_{(a_i)} \mid \theta_k) p(\theta_k \mid c)) \quad (4.19)$$

According to Bayes rules [87], $p(c \mid \mathcal{I})$ can be computed as:

$$p(c \mid \mathcal{I}) = \frac{p(\mathcal{I} \mid c) p(c)}{p(\mathcal{I})} \quad (4.20)$$

The categorization can then be made based on:

$$c = \arg\max_c p(c \mid \mathcal{I}) \quad (4.21)$$

The direct computing of Eq. (4.19) involves multiplication of a number of decimal fractions. This leads to the arithmetic underflow problem, in which the floating operations yield a result that is smaller in magnitude than the smallest quantity representable. To circumvent this problem, we compute the log probability as below:

$$\log p(\mathcal{I} \mid c) = \log \prod_i p(s_{(a_i)} \mid c) = \sum_i \log p(s_{(a_i)} \mid c) \quad (4.22)$$

4.6 Summary

Due to the open-ended nature of object appearances, the objects, no matter from the same or different categories, can have arbitrarily different visual looks. To address this visual diversity issue for object categorization, we take a probabilistic Bayesian perspective and explore a generative interpretation of object appearance diversity. The object appearance diversity is interpreted as a probabilistic generative phenomenon, in which the object visual appearance arises from countably infinitely many common appearance patterns. To make a valid model for this interpretation, three issues must be tackled: (1) the model should provide the possibility of countably infinitely many appearance patterns, as the object visual appearance is boundless; (2) the appearance patterns are shared not only within but also across object categories, as the objects of different categories can be visually similar too; and (3) intuitively, the objects within a category should share a closer set of appearance patterns than those of different categories. We propose a generative probabilistic model, named **nested hierarchical Dirichlet process (HDP) mixture**, to tackle these three issues and embody our generative interpretation. The proposed model exploits the stick breaking construction process to provide the possibility of countably infinitely many appearance patterns. The designed hierarchical structure of our model not only enables the appearance patterns to be shared across different object categories, but also takes into account the inter-relation between appearance patterns, objects and categories, thus allowing the objects within a category to share closer appearance pattern set.

Chapter 5

Experimental Evaluation

This chapter presents the experimental evaluation on the proposed visual representation, visual synset; and the proposed learning framework, nested HDP mixture, respectively.

5.1 Testing dataset

We evaluate the proposed visual synset and nested HDP mixture model using two large-scale datasets: 1) Caltech-101 dataset [63]; and 2) NUS-WIDE-object dataset [23]. The Caltech-101 dataset contains 102 image categories and a total of 9233 images. Figure 5.1 illustrates the example images of 36 categories from Caltech-101 dataset. As shown in Figure 5.1, the difficulties of Caltech-101 are its large number of classes and huge intra-class appearance variations, while its easiness is that most visual objects are dominant objects positioned at the center of their images with clean or no background. For benchmark purpose, we follow the setting of [59, 83, 103, 125, 36, 121, 136] to set the number of training images per category to 30.

To compensate the weakness of Caltech-101 dataset, we utilize another more

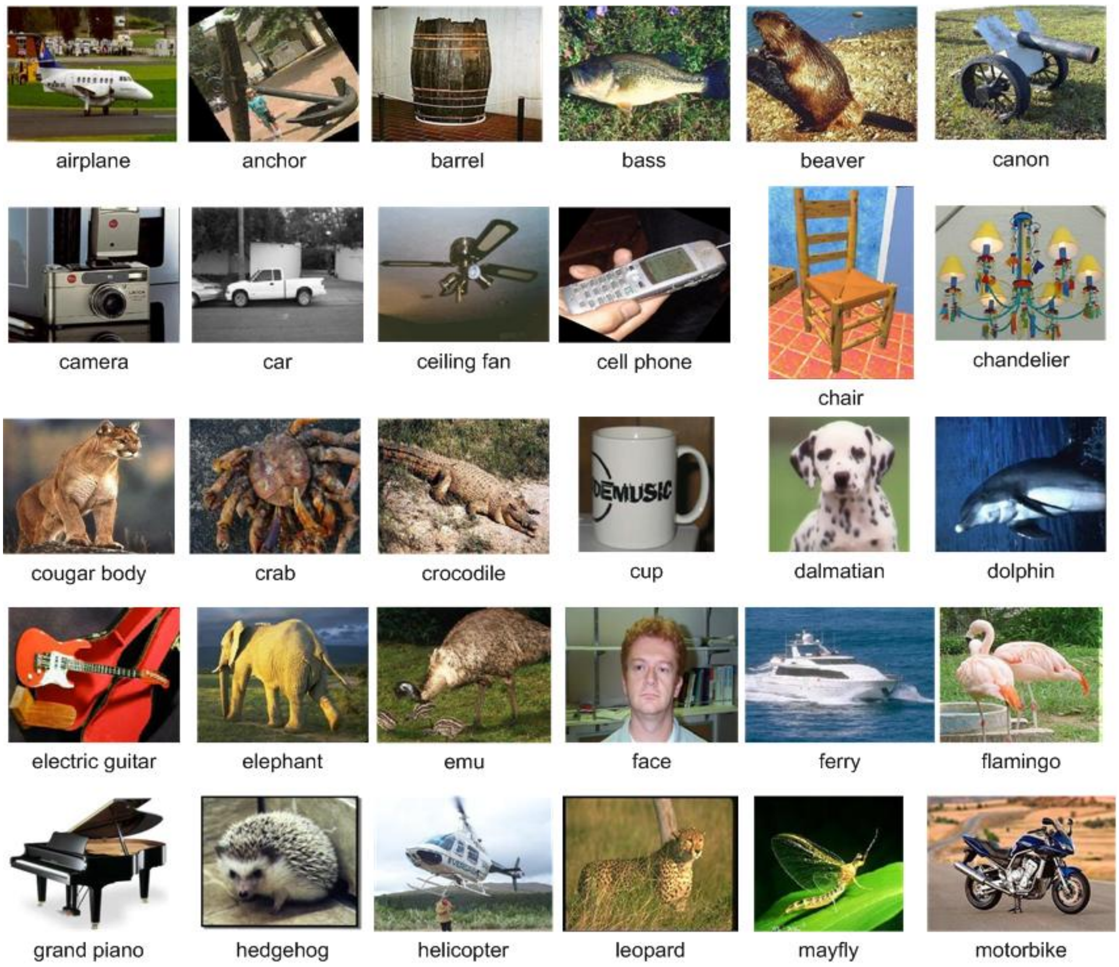


Figure 5.1: The example images of 30 categories from Caltech-101 dataset.

challenging real world image dataset, NUS-WIDE-object dataset [23]. NUS-WIDE-object is a large-scale dataset that consists of 31 object categories, such as computer, coral, dog, rock, etc, and 30,000 images in total. Figure 5.2 shows example images of 15 categories from NUS-WIDE-object dataset. Compared to Caltech-101, the difficulty of NUS-WIDE-object dataset lies in its large visual and scale variation. Figure 5.3 (a) presents the average images of 16 object categories from NUS-WIDE-object. The average image is made by averaging 100 images per category together. As shown, most average images are gray images with high entropy and low visual information. Compared to the average image of caltech-101 dataset in Figure 5.3

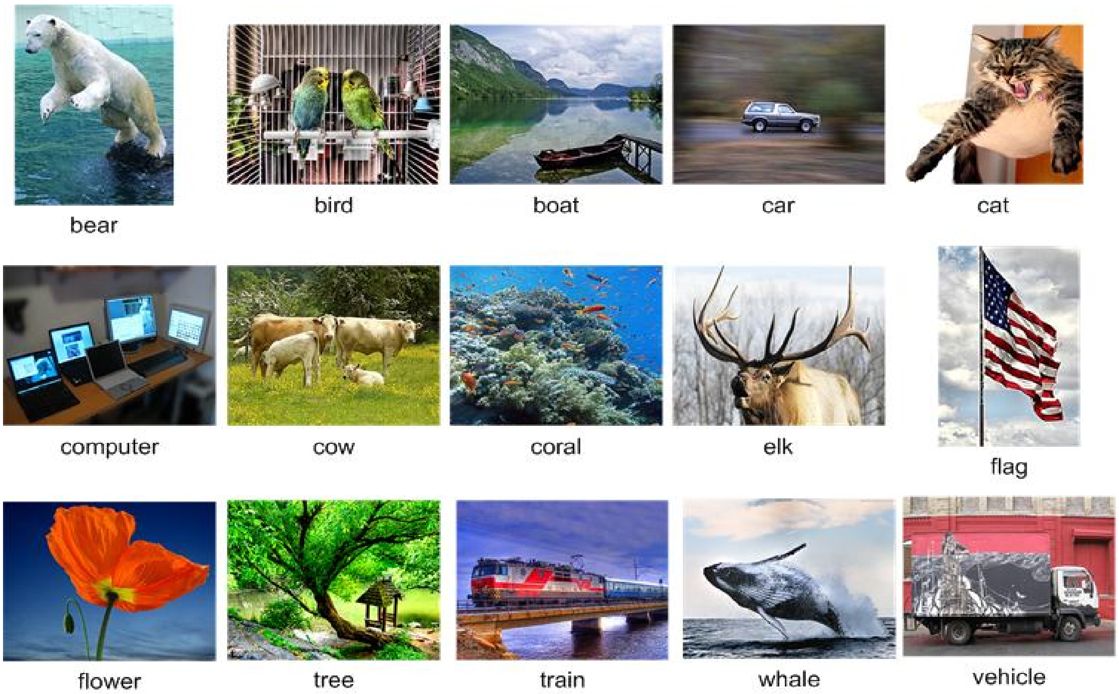


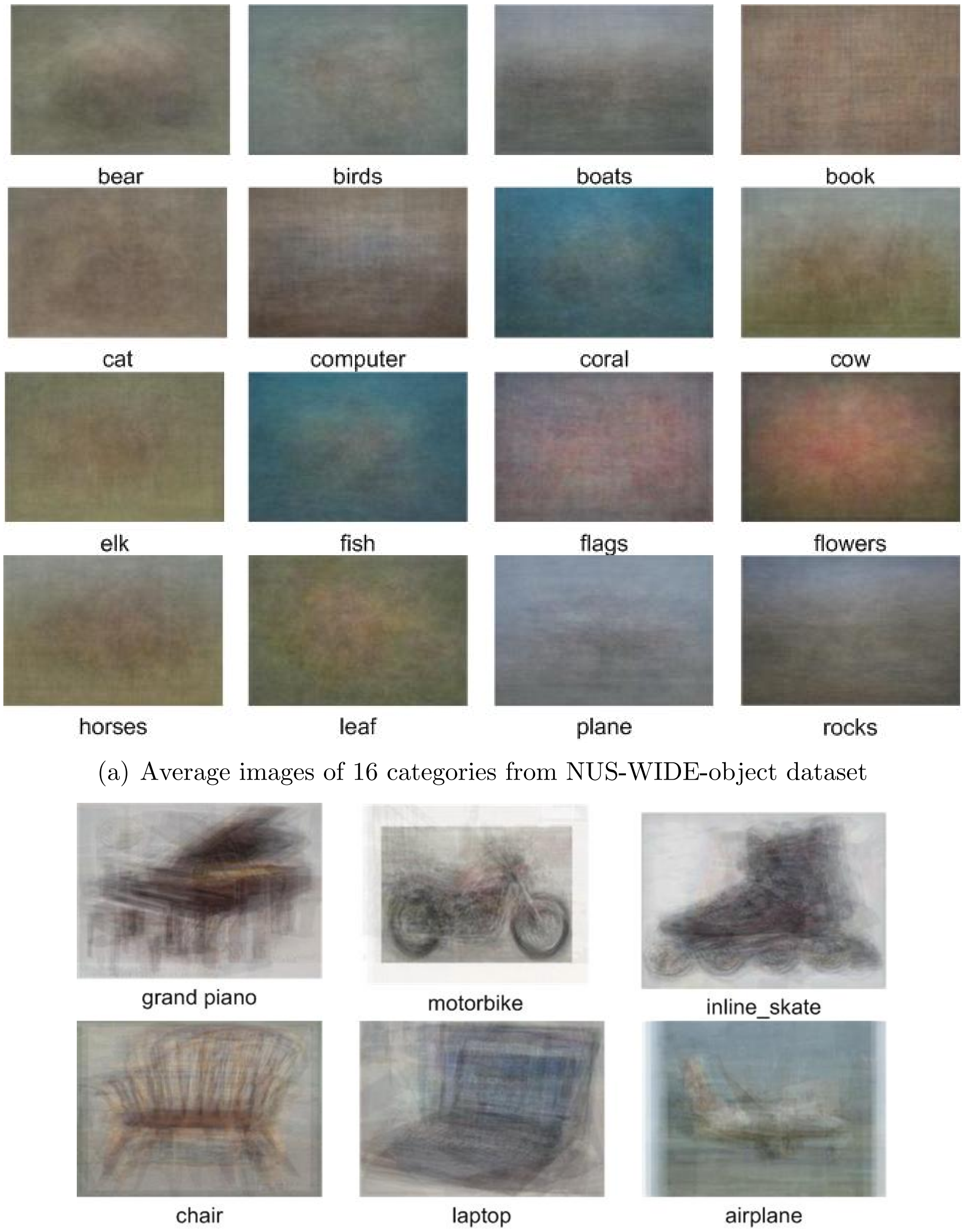
Figure 5.2: The example images of 15 categories from NUS-WIDE-object dataset.

(b), the objects of NUS-WIDE-object does not have much distinct visual or spatial characteristics.

The evaluation criteria here is the average of classification accuracy of all categories.

Generation of visual codebook

To construct the codebook of visual words, we follow the approach of [92] to randomly sample $M = 5,000$ local regions per image and compute the Scale Invariant Feature Transform (SIFT) [72] feature for each region. We then subsample approximately 1 million SIFT features from images in the dataset. A k-means clustering is performed on the SIFT features to generate a codebook of 2000 visual words. The same procedures are applied on both Caltech-101 and NUS-WIDE-object datasets to generate two visual codebooks for them respectively.



(a) Average images of 16 categories from NUS-WIDE-object dataset

(b) Average images of 6 categories from caltech-101 dataset.

Figure 5.3: Average images of Caltech-101 and NUS-WIDE-object dataset.

5.2 The Caltech-101 Dataset

In this section, we evaluate the performance of the proposed models on the Caltech-101 Dataset. The focus of this section is to examine the performance of visual synset. The experimental characteristics of nested HDP mixture is only briefly introduced. Next section will present the results of the nested HDP mixture in detail.

5.2.1 Evaluation on visual synset

Experimental setup

By using the aforementioned codebook of visual words, we build the delta visual phrases. We perform FIM on the database **G** of approximately 2 million visual word groups with the support region size of 1, 4 and 8 respectively. Based on the significance score in Equation 3.1, we construct the delta visual phrase codebook by selecting the top K delta visual phrases (dVP) with the highest scores. In the experiments, K is set to 2200, 2400, 2600, 2800, 3000, 3200, 3600 and 4000 respectively. To test the performance of delta visual phrase and visual synset, we adopt support vector machines (SVM) [120] with generalized RBF kernel as the classifier.

For benchmark purpose, we follow the setup of [59] and [136] by selecting 30 images from each category for training and 30 images per category for testing. The evaluation criteria is the mean classification accuracy, which is the average of evenly weighed recognition rate of each category.

Performance of delta visual phrase

We first perform classification, based on 2000 visual words. The bag-of-words approach yields a mean classification accuracy of 44%. This classification is used as

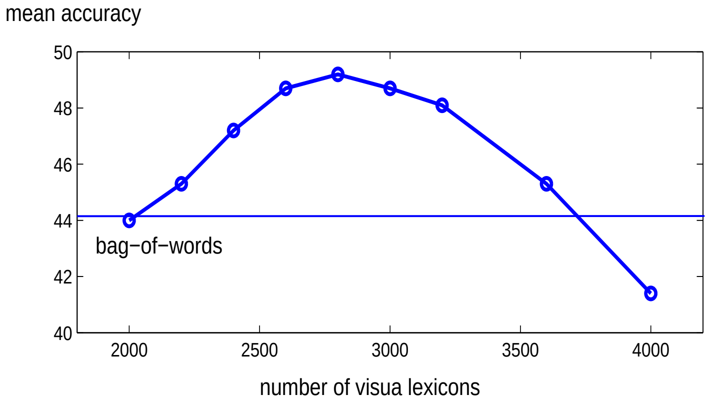


Figure 5.4: The average classification accuracy by delta visual phrases on Caltech-101 dataset.

the baseline of our experiments.

Next, we perform object categorization, based on 2200, 2400, 2600, 2800, 3000, 3200, 3600 and 4000 delta visual phrases respectively. As shown in Fig. 5.4, the performance increases as more delta visual phrases are incorporated up to 2800. In particular, the codebook with 2800 delta visual phrases gives the highest accuracy of 49.2%. This demonstrates that by incorporating co-occurrence and spatial scatter information, the delta visual phrases do carry more discriminative information than visual words. Fig. 5.5 shows some examples of delta visual phrases with different spatial scatter. As shown, when objects share some appearance similarity in a large scope, the delta visual phrase can combine the ambiguous visual words scattered in such area into one distinctive unit, which can contribute to distinguishing objects of different classes with larger inter-class distance and better classification. We also observe that a delta visual phrase (dVP) consists of on average only two member visual words . This is so because the significance measure for dVP selection favors frequent dVP. A dVP with three or more member visual words have much lower frequency than those with two member words, as it requires more visual words to be frequently co-occurring together.

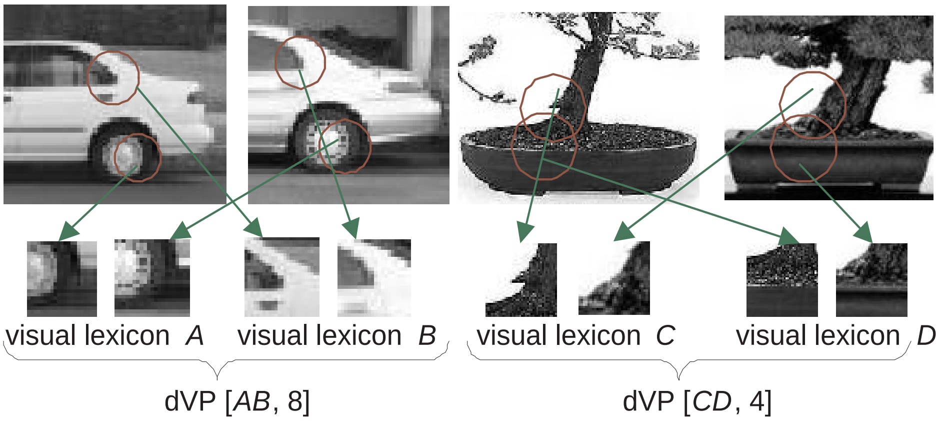


Figure 5.5: The examples of delta visual phrases generated from Caltech-101 dataset. The first dVP consists of disjoint visual words A and B with a scatter of 8 and the second has joint visual words C and D with a scatter of 4

With optimal performance at 2800 delta visual phrases, we increase the codebook size to further investigate the performance of dVP. When the number of lexicons is above 3200, the performance drops drastically and even becomes inferior to the original visual word representation. We attribute such performance degradation to two reasons. First, the newly incorporated delta visual phrases with lesser significance score might not be statistically substantial. Though these delta visual phrases might still be distinctive patterns, their statistical sparseness renders image distributions in feature space more incoherent, sporadic or even noisy. Second, the increased dimensionality inevitably leads to the curse of dimensionality and intensifies the issue of statistical sparseness of image representation.

Performance of visual synset

We evaluate the effectiveness of visual synset, by performing IB-based distributional clustering on the codebook of 2800 delta visual phrases (best run from previous section). Specifically, we set the cardinality of visual synsets $|\mathbf{S}|$ to 200, 600, 1200, 1600, 1800, 2000, 2200 and 2400. Fig. 5.6 gives the average classification accuracies. From Fig. 5.6, we observe that with proper cardinality, the visual synset represen-

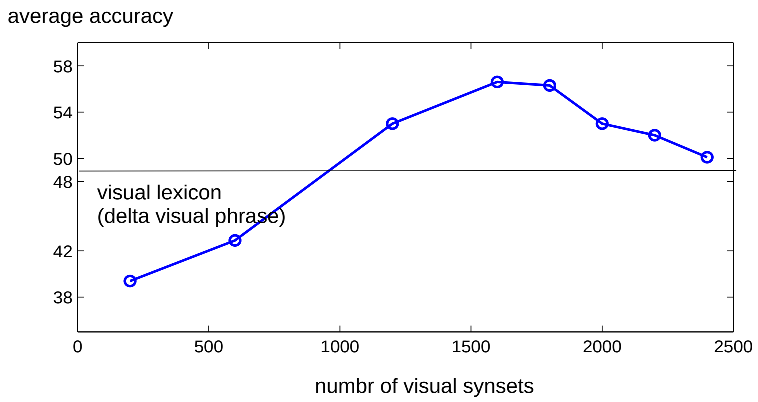


Figure 5.6: The average classification accuracy by visual synsets on Caltech-101 dataset.

tation can deliver superior results over both delta visual phrases and visual words with a more compact representation. For example, the run with only 200 visual synsets can achieve an accuracy of 39.4%, while the runs with 1200 visual synsets has achieved superior accuracies over the run with 2800 delta visual phrases. This representation compactness does not only enable high computational efficiency but also alleviate the issue of curse of dimensionality.

The best run is the one with 1600 visual synsets and it achieves an accuracy of 56.6%. We attribute such improvements to two factors: (1) by fusing semantic-consistent delta visual phrases together, the visual synset reduces the intra-class variations and renders the image distribution in feature space more coherent and manageable; and (2) the visual synset is a result of supervised dimensionality reduction and the properly reduced dimensionality can partially resolve the statistical sparseness problem of delta visual phrases and also enable better classification.

We also observe that the number of visual synsets plays an important role in its performance. A too small number of visual synsets usually gives bad performance. This is because a small number of visual synsets will force the distinctiveness-inconsistent visual words together and generate noninformative and nondistinctive

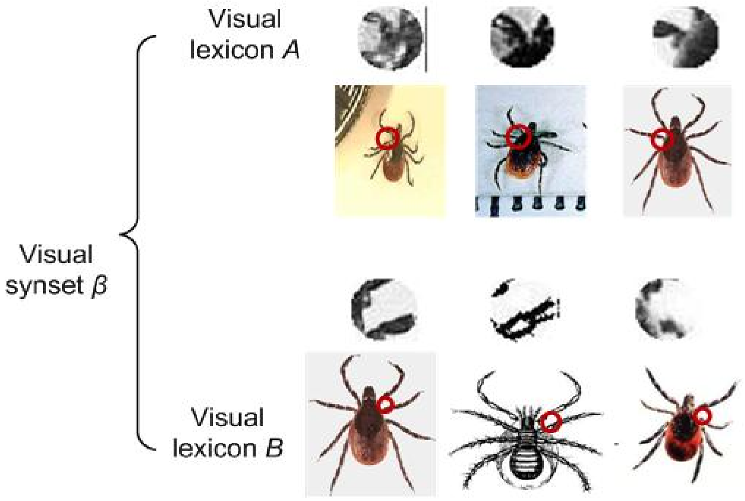


Figure 5.7: Example of visual synset generated from Caltech-101 dataset.

visual synsets. Overall, the experimental results show that the number of visual synsets between 1/3 and 2/3 of delta visual phrase codebook size usually gives a reasonably good performance. Fig. 5.7 shows examples of visual synset generated from Caltech-101 dataset.

Comparison with other visual features: Here we compare the performance of visual synset, delta visual phrase and visual words with other global and semi-global visual features, such as wavelet texture. To benchmark the performance of different visual features, we follow the same experimental setting of visual synset to perform the following runs of experiments with different visual features:

1. *CH: color histogram*
2. *CC: color correlogram*
3. *CM: color moments*
4. *TC: texture cooccurence*
5. *WT: wavelet texture*

Here, the classifier is SVM for all the runs. For CM, we compute the first 3 moments of RGB color channels over 55 grids to form a 225D feature vector. For WT, we divide a keyframe into 43 grids and compute the variance in 9 Haar wavelet sub-bands for each grid. Table 5.1 lists the average accuracy of all the visual features. As shown, the part-based local feature, i.e. bag-of-words, outperforms other global or semi-global features, like color correlgoram, with significant margins. This is consistent with that reported in [50], which further confirms the strength of local features over global features.

Compared to the bag-of-words approach, the proposed delta visual phrase improves the performance of visual words by 5.2%, by building more distinctive delta visual phrases with co-occurrence and spatial scatter information. Moreover, the visual synset further advances the performance of delta visual phrases by 7.4%, by incorporating the distributional information.

Comparison with LDA and pLSA: We also compare the visual synset to pLSA and LDA in the same setting, as all of them can be considered as some kind of dimensionality reduction. pLSA and LDA are essentially a process of unsupervised dimensionality reduction from delta visual phrase space to hidden topic space. We set the number of hidden topics to 50, 100, 150 and 200 and run pLSA and LDA on the codebook of 2800 delta visual phrases (best run of delta visual phrases). We then utilize these hidden topic features to train SVM classifiers to classify the testing images. The runs with the best accuracy are selected for comparison. Specifically, LDA and pLSA give an accuracy of 51.4% and 52.8% respectively, both of which are lower than the accuracy delivered by visual synset. This demonstrates visual synset gives rise to more effective dimensionality reduction, by leveraging the prior distributional information of visual words in a supervised fashion.

Table 5.1: Comparison of performance by visual synset (VS), delta visual phrase (dVP), bag-of-words (BoW) and other visual features with SVM classifier.

	VS	dVP	BoW	CH	CC	CM	TC	WT
Accur(%)	56.6	49.2	44	16.1	27.3	19	16.4	20.5

5.2.2 Performance of nested HDP mixture model

By taking the visual synset representation as input, we apply the proposed nested HDP mixture model to perform the categorization. Here, we ran Gibbs sampler for 20 iterations to perform the parameter inference. This gives rise to an average categorization accuracy of 64.1%. Figure 5.8 illustrates the confusion matrix of the categorization. The nested HDP mixture results in further 7.5% improvements in accuracy over SVM. This demonstrates that the generative framework of nested HDP mixture is able to model the image data better than the discriminative approach of SVM. Section 5.3 shall delve deeper into the examination of the nested HDP mixture model.

5.2.3 Comparison with other state-of-the-arts methods

Here, we benchmark the performance of our proposed models with other existing approaches on the Caltch-101 dataset. By applying the nested HDP mixture on the representation of 1600 visual synsets, we achieve an accuracy of 64.1%. Table 5.2 summarizes the accuracies of other reported systems. As shown in Table 5.2, the proposed visual synset approach outperforms most of existing systems and delivers a comparable result with the state-of-the-arts methods. One exception is the approach proposed in [121], which delivered a much higher accuracy of 78.43%. This is because their approach has utilized multiple visual features and multiple kernels in SVM. As different visual features tend to capture different aspects of visual characteristics of image categories, they tend to share complementariness

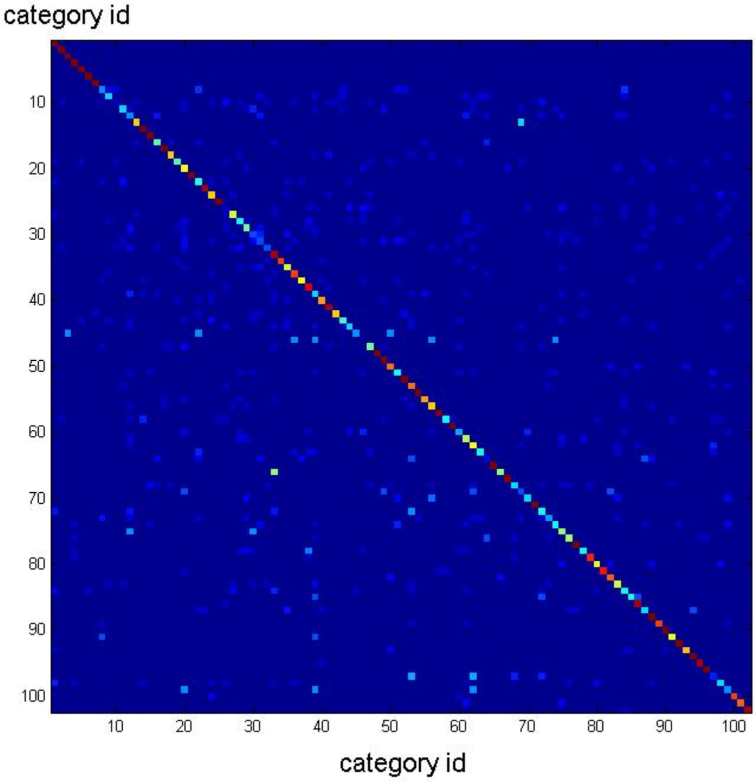


Figure 5.8: The confusion matrix of the categorization by visual synset with nested HDP as classifier on Caltech-101 dataset. The rows denote true label and the columns denote predicted label.

and redundancy in modeling the visual contents. This complementariness and redundancy result in better categorization performance. In comparison, our work utilized only one type of part-based local features.

To evaluate the effect of incorporating more features, we perform a simple experiment by combining the categorization results of the runs of VS + nested HDP, together with those obtained by using color corrlegram (CC) + SVM and wavelet texture (WT) + SVM. Here, CC and WT are the two global features with highest reported accuracies, as mentioned in the previous Section. The feature combination is done in a late fusion manner [112]. Specifically, the prediction scores of classifiers on each individual feature are normalized to the range of 0 and

Table 5.2: Benchmark of classification performance on Caltech-101 dataset. VS means visual synset and Fusion (VS + CM + WT) indicates the fusion of visual synset, color correlogram (CC) and wavelet texture (WT).

run	VS + nested HDP	VS + SVM	Fusion (VS + CC + WT)	[103]	[136]	[83]	[36]	[125]	[59]	[121]
(%)	64.1	56.6	72.8	42	53.9	56	58	63	64.6	78.43

1. The SVM outputs of CC and WT features are normalized via a sigmoid function [97]. The final classification score is the linear summation of the scores of all the classifiers on individual feature. This simple late fusion of visual synset, CC and WT features gives an average classification accuracy of 72.8%, which is comparable to the performance of the multiple kernel learning approach reported in [121]. However, compared to the late fusion, the multiple kernel learning schemes have more complicated learning models and higher computational complexity.

5.3 The NUS-WIDE-object dataset

This section tests the proposed models on the NUS-WIDE-object dataset. It briefly presents the testing on visual synset and examines the experimental prosperities of the proposed nested HDP mixture model in detail.

Experimental setup

The NUS-WIDE-object dataset contains 31 object classes, such as boat, fish, coral, etc, as shown in Figure 5.2. For efficiency purpose, we utilize 30 images per category for training and 200 for testing. We follow the same experimental setup for Caltech-101 to generate codebook of 2000 visual words and incremental number of delta visual phrases and visual synsets. The evaluation criteria adopted here is the average categorization accuracy.

Table 5.3: Average categorization accuracy of the NUS-WIDE-object dataset based on bag-of-words (BoW), best run of delta visual phrases and best run of visual synsets (VS).

	bag-of-words	delta visual phrase	visual synset
Average accuracy	11.7%	12.4%	13.8%

Performance of visual synset

Table 5.3 summarizes the results of classification based on visual words only, delta visual phrases and visual synsets. The classifier adopted here is SVM with generalized RBF kernel. The baseline classification with visual words give an average accuracy of 11.7%. Similar to Caltech-101, the mean accuracy increases as more delta visual phrases are incorporated and reaches its peak of 12.4%, when the number of delta visual phrases is 2300. The optimum number of delta visual phrases here is lesser than 2800 in Caltech-101. We attribute this to the fact that the images of same category in NUS-WIDE-object are more visually diverse and less geometrically consistent. Therefore, the resulting delta visual phrases are less statistically stable. Based on the best run of delta visual phrases, we generate visual synsets and perform the classifications. Consistent to the observation in Caltech-101, the visual synset achieves both compactness and superior performance. Specifically, the run with 1000 visual synsets delivers the best mean accuracy of 13.8%.

5.3.1 Evaluation on nested HDP

In this Section, we take visual synset as image representation and evaluate the proposed nested HDP mixture model in detail.

Appearance Pattern Learning

First, we investigate the appearance pattern learning in the proposed nested HDP mixture and the HDP variation model (a) from [116] and (b) from [125]. We ran Gibbs sampler for 20 iterations to perform the parameter inference in all three models respectively. We also set the initial number of parameter atoms θ_k to $K = 20$. For efficiency purpose, we limit the number of training images per category to $T = 30$.

• **The number of appearance patterns:** Figure 5.9 shows how the number of appearance patterns changes in the three models, as Gibbs sampler is running. As shown, the appearance patterns in all three models are free to shrink and grow, so as to best explain the visual data in the current iteration of Gibbs sampling. In each iteration, the Gibbs sampler can generate new appearance pattern, if the existing appearance patterns are not sufficient to explain the given data. Similarly, it can also eliminate existing appearance patterns, if the evolving appearance patterns make some of them redundant to interpret the data observations. Specifically, during the Gibbs sampling, the number of appearance patterns in HDP mixture model (a) and (b) goes through large fluctuation and then converge at certain point. In contrast, the appearance patterns in the proposed nested HDP mixture does not fluctuate much and converge to 24 after iteration 12. This indicates that the model structure of nested HDP mixture fits the data better than HDP model (a) and (b). We attribute the smoother change of appearance patterns in the proposed nested HDP mixture to the fact that it handles issue (c), while the other two models do not.

• **Sharing appearance pattern across object categories:** To investigate how the appearance patterns are shared across object categories, we visualize how the categories are distributed in the embedding space of appearance patterns. We use the visual appearance patterns generated by the nested HDP mixture. First, we compute the pair-wise distance of object categories in the appearance pattern

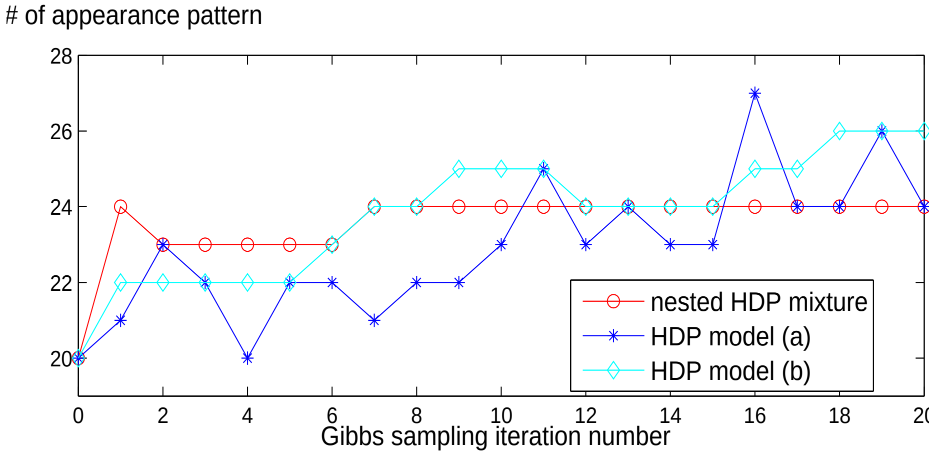


Figure 5.9: The number of appearance patterns in nested HDP mixture, HDP mixture model (a) and (b) for each iteration of Gibbs sampling.

space, by using symmetrized Kullback Leibler (KL) divergence [24] as follows.

$$KL(c_i, c_j) = \sum_k p(\theta_k \mid c_i) \log \frac{p(\theta_k \mid c_i)}{p(\theta_k \mid c_j)} + p(\theta_k \mid c_j) \log \frac{p(\theta_k \mid c_j)}{p(\theta_k \mid c_i)}, \quad (5.1)$$

where $KL(c_i, c_j)$ is the symmetrized KL divergence of category c_i and c_j. To get a sense of how the categories are distributed in the appearance pattern space, we utilize the metric multidimensional scaling (MDS) to plot the categories, based their pairwise symmetrized KL divergence. Multidimensional scaling (MDS) is a tool that can visualize how near or far the data points are from each other. Its required input is the similarity of data points, rather than their location coordinates, which fits our task here. Figure 5.10 illustrates how the object categories are distributed in the two-dimensional embedding of appearance pattern space by the metric MDS. As shown, the animal type objects are clustered on the upper left corner, while the transport type objects are concentrated on the right side. This indicates that these two types of objects arise from a fairly different set of appearance patterns. Another observation is that the categories ”sun” and ”book” are far apart from the rest of the categories. This is attributed to the fact that the visual appearances of ”sun” and ”book” are not tightly related to any of the other categories.

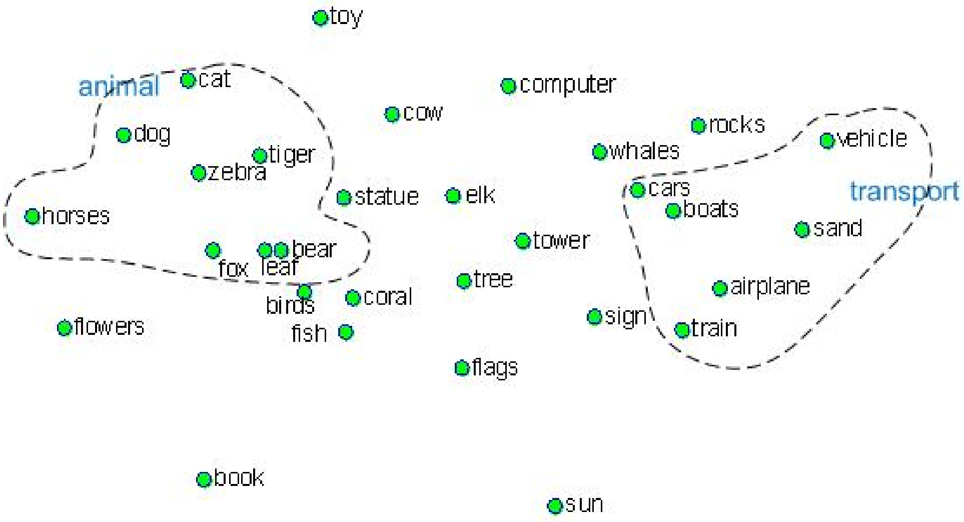


Figure 5.10: The visualization of object categories in the two-dimensional embedding of appearance pattern space by metric MDS.

Categorization Performance

Here, we evaluate the effectiveness of the generative interpretation and its embodiment, nested HDP mixture model, for object categorization. We set the number of training images per category $T = 30$ and the number of testing images to 200 per category. First, we run the proposed nested HDP mixture, which yields an average classification accuracy of 0.167. With the same experimental setting, we run the HDP model (a) from [116] and (b) from [125], which deliver average accuracy of 0.15 and 0.11 respectively. The result of the proposed nested HDP mixture is superior to that of model (a) and (b), by a margin of 11% and 52% respectively. We attribute this performance improvement to the graphical model structure of the nested HDP mixture that is more appropriate to interpret the object visual characteristics. Moreover, among three models, model (b) yields the lowest accuracy, which indicates that its assumption of appearance pattern sharing at image level is improper. Compared to model (a), the proposed nested HDP mixture can improve the average accuracy by 11% relatively. We attribute this improvement

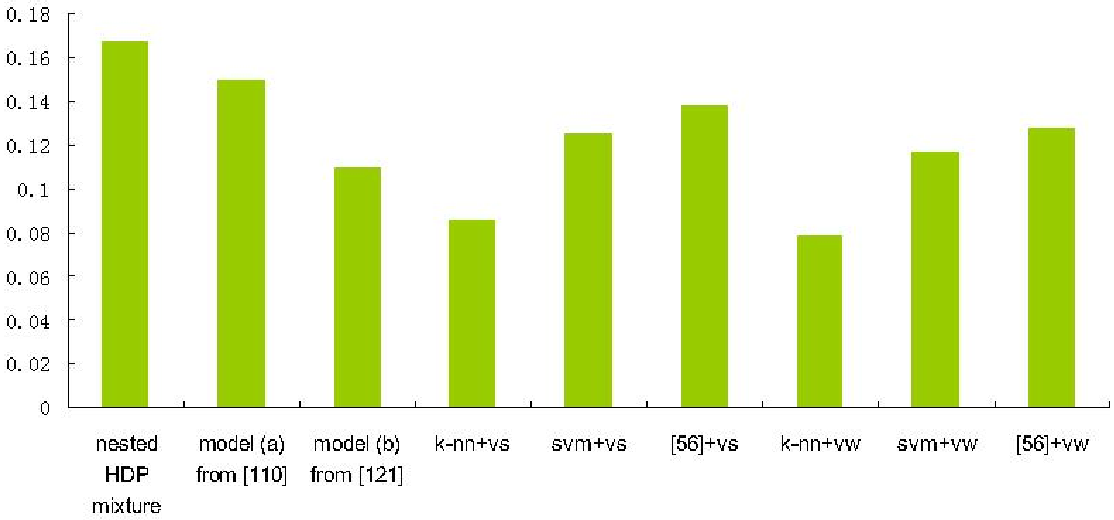


Figure 5.11: The average accuracy by proposed nested HDP mixture, k-NN, SVM, approach in on visual synsets and visual words respectively.

to the model structure difference between nested HDP mixture and model (a). In the nested HDP mixture, the objects of same category are assumed to arise from a closer set of appearance patterns, so as to tackle issue (c) in Table 4.1.

- **Benchmark:** To further evaluate the effectiveness of our generative interpretation of object appearance diversity and the proposed embodiment model, we compare the performance of nested HDP mixture to other methods, as below.

- HDP model (a) from [116] with visual synset, denoted by "model (a) from [116]".
- HDP model (b) from [125] with visual synset, denoted by "model (b) from [125]".
- k-NN on visual synset (vs), denoted by "k-nn+vs".
- SVM on visual synset (vs), denoted by "svm+vs".

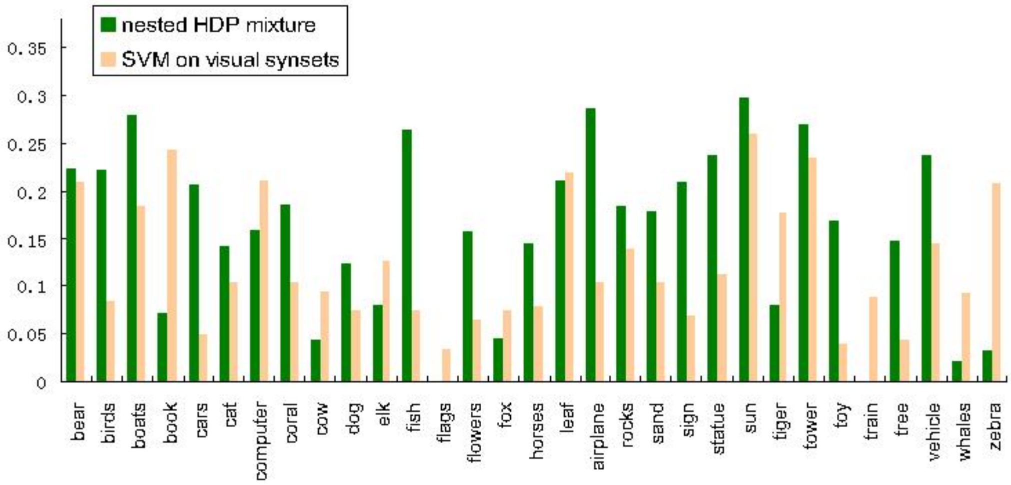


Figure 5.12: The categorization accuracy for all categories by the proposed nested HDP mixture and SVM.

- SVM with spatial pyramid kernel proposed in [59] on visual synset (vs), denoted by " [59]+vs".

To further evaluate the contribution by visual synset descriptor, we also perform the following runs, based on visual words:

- k-NN on visual words (vw), denoted by "k-nn+vw".
- SVM on visual words (vw), denoted by "svm+vw".
- SVM with spatial pyramid kernel proposed in [59] on visual words, denoted by " [59]+vw".

Here, we choose the discriminative classifiers of k-NN, support vector machine (SVM) [120] and SVM with spatial pyramid kernel [59] for benchmark, as they are reported to be robust and deliver performance comparable to state-of-the-arts approaches [59, 113, 135]. The "k" parameter in k-NN is tuned from 10 to 20 and the run with best accuracy is selected. The kernel of SVM is RBF and its gamma and cost parameter is determined by a 3-fold cross validation. For the spatial pyramid kernel, we follow the setting in [59] and set the pyramid level to 2.

Figure 5.11 displays the accuracy of all the runs. As shown, the proposed nested HDP mixture delivers the best results, outperforming other runs with substantial margins. Specifically, it outperforms k-NN on visual synset by 94% relatively, SVM on visual synset [143] by 34% relatively, SVM with spatial pyramid kernel [59] by 21% relatively, and SVM on visual words by 42% relatively.

As show in Figure 5.11, the proposed nested HDP mixture model consistently outperforms SVM based discriminative models on only 30 training images per category. We attribute this partially to the appealing prosperities of generative learning models over the discriminative ones. In contrast to discriminative models, generative models attempt to explicitly identify the causal structure of image features by modeling the probabilistic inter-relation of all the variables as a joint probability distribution. This enables generative models to have better generalization. On the other hand, discriminative models attempt to learn a mapping between input feature and output category variables only, rather than unveiling the probabilistic structure of input or output domain. Consequently, it requires a large amount of training data to produce good classifiers.

Furthermore, we carry out a detailed examination on the accuracy of each category by the proposed nested HDP mixture and SVM on visual synset. As shown in Figure 5.12, we observe that among the 31 object categories, the proposed HDP mixture outperforms SVM on 20 categories, while SVM gives better accuracy on the other 11 categories, such as leaf, zebra, whale, etc. We conjecture that the better performance of SVM is due to the fact that these categories have relatively distinct visual patterns, which suit discriminative models like SVM. For example, the category "zebra" may be dominated by its homogeneous and distinct texture pattern. In this circumstance, the discriminative approach of SVM may be more advantageous.

Chapter 6

Conclusion

6.1 Summary

The motivation for research in this thesis stemmed from the meditation on why the bag-of-words representation in visual categorization cannot perform at comparable level to its analogy of document categorization in the text domain. After a careful examination of the differing natures of image and text categorization, we realize that the main reason for the unsatisfactory performance in visual categorization is partially rooted in visual representation. Existing visual features are not capable of associating the semantic relevance of images with their proximity in visual feature space. This apartness of image semantic and its visual features is also widely known as *the semantic gap*. We, therefore, devoted our effort on developing a higher-level visual representation. As the bag-of-words image representation has consistently delivered the state-of-the-art performance, we took the bag-of-words approach as the base feature, from which we built a higher-level representation.

After careful investigation of bag-of-words representation, we found that the robustness of bag-of-words approach was hindered by two basic issues: polysemy and synonymy. The polysemous visual word is the one that might represent different semantic meanings in different image contexts, while the synonymous words are the set of visually dissimilar words representing the same semantic meaning. By sharing a set of polysemous visual words, the semantically dissimilar images might be close to each other in the feature space, while the synonymous visual words may cause the images with the same semantic to be far apart in the feature space. The direct consequences of polysemy and synonymy are small inter-class distance and large intra-class variation. In other words, the image distributions in the feature space tend to be disordered, sporadic and incoherent. This statistically explains why the image categorization based on bag-of-words approach is not comparable to its analogy in text domain. To tackle these aforementioned issues, we proposed a higher-level visual unit, named *visual synset*, on top of the bag-of-words approach. This is done, by incorporating co-occurrence, spatial scatter and distributional information of visual words.

Though the resulting visual synset is superior over the bag-of-words approach, its robustness is still limited by the open-ended nature of object visual appearances. Namely, the objects, no matter from the same class or otherwise, tend to possess huge variations of visual looks and shapes. To achieve better visual categorization, we focused on devising a learning scheme that can categorize images beyond their visual appearances. To do so, we first admitted the fact that objects from the same class are not necessarily visually similar. Then, we took a Bayesian perspective to explain the object appearance diversity in a generative interpretation. Namely, the diverse visual appearances of objects arise from a set of countably infinitely many common appearance patterns. Based on this interpretation, we devised a generative learning framework, named *nested HDP mixture model*, to perform image categorization. The testing on two large-scale datasets, Caltech-101 and NUS-WIDE-object, show that the proposed visual representation, visual synset, and the generative learning model, nested HDP mixture, can deliver promising performances.

6.2 Contributions

The work of the thesis consists of two major parts: building a higher-level visual representation and developing a new generative probabilistic learning method for visual categorization. The contributions of the thesis are listed as follows.

Visual synset: a higher-level visual representation

We proposed *visual synset*, a higher-level visual representation, built on top of visual words. Visual synset attempts to address the polysemy and synonymy issues of visual words. To address the polysemy issue, we exploited the co-occurrence and spatial scatter information of visual words to generate a more distinctive compositional visual configuration, i.e. *delta visual phrase*. The improved distinctiveness leads to better inter-class distance.

To tackle the synonymy issue, we proposed to group delta visual phrase with similar 'semantic' into a visual synset. Rather than in conceptual manner, we define the 'semantic' of a delta visual phrase probabilistically as its image class probability distribution. The visual synset is therefore a probabilistic relevance-consistent cluster of delta visual phrases, which is learned by Information Bottleneck based distributional clustering. By grouping different delta visual phrases into one unit, the visual synset can partially bridge the visual differences between these images and deliver a more coherent, robust and compact representation of images.

Nested HDP mixture: a learning scheme beyond visual appearances

To further recognize objects beyond their visual appearance, we developed a learning model that categorizes images beyond their visual appearances. By taking a

Bayesian perspective, we interpreted the object visual diversity as a probabilistic generative phenomenon, in which the visual appearance arises from the countably infinitely many common *appearance patterns*. To make this generative interpretation valid, three issues were tackled: (1) there exist countably infinitely many appearance patterns, as the objects have boundless variation of appearance; (2) the appearance patterns are shared not only within but also across object categories, as the objects of different categories can be visually similar too; and (3) intuitively, the objects within a category should share a closer set of appearance patterns than those of different categories. To embody the generative interpretation, we proposed a generative probabilistic model, named *nested hierarchical Dirichlet process (nested HDP) mixture*, to tackle the three issues above. The *stick breaking construction process* in the nested HDP mixture provides the possibility of countably infinitely many appearance patterns that can grow, shrink and change freely. The hierarchical structure of our model not only enables the appearance patterns to be shared across object categories, but also allows the images within a category to arise from a closer appearance pattern set than those of different categories.

6.3 Limitations of this research and future work

While the works in the thesis contributed to the advancement of image categorization, they do suffer from some limitations. First, the generation of delta visual phrase is a time-consuming task. The time complexity does not only lie in the frequent itemset mining, but also the extraction of delta visual phrases from new images. The generation of delta visual phrases requires the spatial neighborhood information of local regions. As the number of local regions per image is usually large (in the order of thousands), the computation of pairwise spatial distance between local regions tend to be time consuming. A more efficient algorithm is required.

One possible solution is to utilize approximate distance estimation, such as hashing or indexing of spatial locations of local regions.

Second, delta visual phrase is not guaranteed to be scale invariant. The basis of delta visual phrase generation is co-occurring and spatial scatter inter-relation of visual words. The scale invariance aspect is not taken into consideration. This renders the scale invariance degree of a delta visual phrase solely relying on its component visual words. The visual words achieve scale invariance by extracting local regions at multiple scale levels at the given image. One possible solution to this issue is to incorporate scale consistency inter-relation of visual words when constructing delta visual phrases.

Third, how the number of classes changes the semantic inference distribution of delta visual phrases and how this affects the visual synset generation and final classification have not been investigated. The number of classes indirectly determines the resolution of probabilistic semantic, as a large number of classes gives a fine class conditional distribution. The investigation of this issue is one direction of our future work.

Fourth, one limitation of the proposed nested HDP is that it is not theoretically adaptable to scene classification task. This is so because of the generative assumption of image formation that the graphical model takes. The model assumes that appearances of local parts of object arise from a universal pool of appearance patterns. This assumption is justifiable for objects, because the appearances of object's local parts are constrained by object form and structure. On the other hand, the local appearances of a scene is much more complicated and unconstrained, as objects from any categories might be locally visible in a scene. This visual complexity and randomness render the generative assumption of local appearances ill-posed.

Fifth, in the process of developing the generative learning model, we neglect the semantic ontology of image classes. However, this semantic inter-relation between image categories, such as vehicle and car, can provide much informative clues for categorization. One of our future works is on exploring how to incorporate such inter-relations into the model structure. The incorporation of semantic ontology might contribute to another future work, which is to extend the proposed categorization model to perform multi-label learning task. In the multi-label context, images are annotated with more than one labels. The semantic inter-relation of labels in the same image could provide useful information in building the learning model.

Bibliography

[1] A. Agarwal and W. Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2006.

[2] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. *J. Parallel Distrib. Comput.*, 61(3):350–371, 2001.

[3] S. Agarwal and A. Awan. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, 2004. Member-Dan Roth.

[4] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, pages 207–216. ACM Press, 1993.

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994.

[6] N. Alon and J. H. Spencer. *The Probabilistic Method.* Wiley, New York, NY, 1991.

[7] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the International Conference on Machine learning*, page 6, New York, NY, USA, 2004. ACM.

[8] L. Baker and A. McCallum. Distributional clustering of words for text classification. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of ACM SIGIR*, pages 96–103, Melbourne, AU, 1998.

[9] H. Bay, T. Tuytelaars, V. Gool, and L. Surf: Speeded up robust features. In *9th European Conference on Computer Vision*, Graz Austria, May 2006.

[10] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research g*, 3:1183–1208, 2003.

[11] M. Bennamoun and G. J. Mamic. *Object Recognition: Fundamentals and Case Studies.* Advances in Pattern Recognition. Springer, 2002.

[12] A. C. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2005.

[13] A. C. Berg and J. Malik. Geometric blur for template matching. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1:607, 2001.

[14] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer, March 1993.

[15] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychol Review*, 94(2):115–147, 1987.

[16] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[17] T. C. Black and W. J. Thompson. Bayesian data analysis. *Computing in Science and Engineering*, 3(4):86–91, 2001.

[18] D. M. Blei. *Probabilistic models of text and images*. PhD thesis, Berkeley, CA, USA, 2004. Chair-Jordan, Michael I.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[20] C. Borgelt. An implementation of the fp-growth algorithm. In *OSDM '05: Proceedings of the 1st International workshop on open source data mining*, pages 1–5, New York, NY, USA, 2005. ACM.

[21] L. Bottou and V. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

[22] T. Chang and C. C. J. Kuo. Texture analysis and classification with tree-structured wavelet transform. *IEEE Transactions on Image Processing*, 2(4):429–441, 1993.

[23] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval*, Santorini, Greece., 2009.

[24] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[25] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *Proceedings of ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[26] M. H. Degroot and M. J. Schervish. *Probability and Statistics, 3rd Edition*. Addison Wesley, 3rd edition, October 2001.

[27] M. Donoser and H. Bischof. Efficient maximally stable extremal region (mser) tracking. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 553–560, 2006.

[28] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006.

[29] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, 2003.

[30] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1):67–92, 1973.

[31] Y. Freund. Boosting a weak learning algorithm by majority. In *COLT '90: Proceedings of the third annual workshop on Computational learning theory*, pages 202–216, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.

[32] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

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

[34] S. Geman, D. Potter, and Z. Chi. Composition systems. Technical report, Division of Applied Mathematics, Brown University, Providence, RI, 1998.

[35] Y. Gong. Advancing content-based image retrieval by exploiting image color and region features. *Multimedia Systems*, 7(6):449–457, 1999.

[36] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proceedings of International Conference on Computer Vision*, pages 1458–1465, USA, 2005. IEEE Computer Society.

[37] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, 8:725–760, 2007.

[38] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl 1:5228–5235, April 2004.

[39] F. Han, Y. Shan, H. S. Sawhney, and R. Kumar. Discovering class specific composite features through discriminative sampling with swendsen-wang cut. In *Proc. of Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, U.S., 2008. IEEE Computer Society.

[40] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 14(1), 2007.

[41] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 3(6):610–621, 1973.

[42] J. M. G. Hidalgo, M. de Buenaga Rodríguez, and J. C. Cortizo. The role of word sense disambiguation in automated text categorization. In *Proceedings of Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005*, volume 3513, pages 298–309, Alicante, Spain, June 15-17 2005.

[43] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence, UAI*, Stockholm, 1999.

[44] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[45] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[46] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 762, Washington, DC, USA, 1997. IEEE Computer Society.

[47] N. Ide and J. Vronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40, 1998.

[48] M. Ioka. A method of defining the similarity of images on the basis of color information. Technical report, IBM Research, Tokyo Research Laboratory, 1989.

[49] T. Jebara. *Discriminative, generative and imitative learning.* PhD thesis, 2002. Supervisor-Pentland, Alex P.

[50] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of ACM Conference on Image and video retrieval*, pages 494–501, New York, NY, USA, 2007.

[51] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of European Conference on Machine Learning*, pages 137–142, Heidelberg et al., 1998.

[52] M. I. Jordan, editor. *Learning in graphical models.* MIT Press, Cambridge, MA, USA, 1999.

[53] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of International Conference on Computer Vision*, Washington, DC, USA, 2005.

[54] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[55] D. Kersten. Object perception: Generative image models and bayesian inference. In *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 207–218, London, UK, 2002. Springer-Verlag.

[56] A. Kumar and C. Sminchisescu. Support kernel machines for object recognition. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, October 14-20 2007.

[57] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods, 2003.

[58] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1265–1278, 2005.

[59] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006.

[60] T. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. *IEEE International Conference on Computer Vision*, 2:1010, 1999.

[61] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, June 2001.

[62] B. Li, K. Goh, and E. Y. Chang. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proceedings of ACM International Conference on Multimedia*, 2003.

[63] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach aested on 101 object categories. In *Proceedings of Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, USA, 2004.

[64] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 524–531, Washington, DC, USA, 2005.

[65] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Infor. Theory*, 37:145–151, 1991.

[66] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Local ensemble kernel learning for object category recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis, Minnesota, USA, 18-23 June 2007.

[67] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30:79–116, 1998.

[68] D. A. Lisin, M. A. Mattar, M. B. Blaschko, E. G. Learned-Miller, and M. C. Benfield. Combining local and global image features for object class recognition. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, page 47, Washington, DC, USA, 2005. IEEE Computer Society.

[69] P. M. Long, R. A. Servedio, and H. U. Simon. Discriminative learning can succeed where generative learning fails. *Inf. Process. Lett.*, 103(4):131–135, August 2007.

[70] D. G. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Norwell, MA, USA, 1985.

[71] D. G. Lowe. Object recognition from local scale-invariant features. *IEEE International Conference on Computer Vision*, 2:1150–1157 vol.2, 1999.

[72] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.

[73] J. Luo, M. R. Boutell, R. T. Gray, and C. M. Brown. Image transform bootstrapping and its applications to semantic scene classification. *IEEE Transactions on SMC*, 35(3):563–570, 2005.

[74] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8):837–842, 1996.

[75] D. Marr. *Vision*. W. H. Freeman and Company, 1980.

[76] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, March 1983.

[77] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference 2002*, Cardiff, UK, 2-5 September 2002.

[78] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[79] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[80] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005.

[81] M. Miyahara and Y. Yoshida. Mathematical transform of (r,g,b) color data to munsell (h,s,v) color data. In *SPIE Proceedings : Visual Communications and Image Processing*, volume 1001, pages 650–657, San Jose, California, U.S.A, 1988. SPIE.

[82] K. P. Murphy. An introduction to graphical models. Technical report, University of British Columbia, 2001.

[83] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *Proceedings of Conference on Computer Vision and Pattern Recognition '06*, pages 11–18, Washington, DC, USA, 2006.

[84] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data: Empirical Inference Science (Information Science and Statistics)*. Springer, March 2006.

[85] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *Proceedings of ACM International Conference on Multimedia*, pages 660–667, New York, NY, USA, 2004. ACM Press.

[86] A. P. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of ACM Conference on Multimedia*, pages 991–1000, New York, NY, USA, 2007. ACM.

[87] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[88] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, page 14, 2002.

[89] C. W. Niblack, R. J. Barber, W. R. Equitz, M. D. Flickner, D. Glasman, D. Petkovic, and P. C. Yanker. The qbic project: Querying image by content using color, texture, and shape. 1908:173–187, February 1993.

[90] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 2161–2168, Washington, DC, USA, 2006.

[91] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, Graz, Austria, 2006.

[92] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*. Springer, 2006.

[93] B. Ommer and J. M. Buhmann. A compositionality architecture for perceptual feature grouping. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume LNCS 2683, pages 275–290, June 16-22 2003.

[94] B. Ommer and J. M. Buhmann. Learning compositional categorization models. In *Proceedings of ECCV 2006, 9th European Conference on Computer Vision*, pages 316–329, Graz, Austria, May 7-13, 2006, 2006.

[95] T. Pedersen. A decision tree of bigrams is an accurate predictor of word sense. In *In Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 79–86, 2001.

[96] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of ACL*, pages 183–190, Morristown, NJ, USA, 1993.

[97] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classiers, MIT Press*, pages 61–74, 2000.

[98] J. Puzicha, T. Hofmann, and J. M. Buhmann. Histogram clustering for unsupervised segmentation and image retrieval. *Pattern Recogn. Lett.*, 20(9):889–909, 1999.

[99] T. Quack, V. Ferrari, B. Leibe, and L. Van-Gool. Efficient mining of frequent and distinctive feature configurations. In *Proceedings of IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, October 14-20 2007.

[100] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[101] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Past, present, and future. In *Journal of Visual Communication and Image Representation*, volume 10, pages 1–23, 1997.

[102] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, March 1999.

[103] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 994–1000, Washington, DC, USA, 2005. IEEE Computer Society.

[104] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005.

[105] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, page 1470, 2003.

[106] N. Slonim. *The Information Bottleneck: Theory and Applications.* PhD thesis, the Senate of the Hebrew University, 2002.

[107] N. Slonim, N. Friedman, and N. Tishby. Agglomerative multivariate information bottleneck. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[108] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization, 2002.

[109] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of ACM MIR Workshop*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[110] J. R. Smith and S.-F. Chang. Automated binary texture feature sets for image retrieval. In *ICASSP '96: Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference*, pages 2239–2242, Washington, DC, USA, 1996. IEEE Computer Society.

[111] J. R. Smith and C.-S. Li. Image classification and querying using composite region templates. *Comput. Vis. Image Underst.*, 75(1-2):165–174, 1999.

[112] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 399–402, Singapore, November 2005.

[113] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of ACM Conference on Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.

[114] M. A. Stricker and M. Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 381–392, 1995.

[115] E. Sudderth. *Graphical Models for Visual Object Recognition and Tracking.* PhD thesis, Massachusetts Institute of Technology, May 2006.

[116] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed dirichlet processes. In *Advances in Neural Information Processing Systems 18*, pages 1299–1306. MIT Press, 2005.

[117] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[118] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[119] I. Ulusoy and C. M. Bishop. Comparison of generative and discriminative techniques for object detection and classification. In *Toward Category-Level Object Recognition*, pages 173–195, 2006.

[120] V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, New York, USA, 1995.

[121] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Rio de Janeiro, Brazil, 2007.

[122] M. Varma and A. Zisserman. Classifying materials from images: to cluster or not to cluster? In *Proceedings of the 2nd International Workshop on Texture Analysis and Synthesis, Copenhagen, Denmark*, pages 139–144, May 2002.

[123] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *SIGIR '93: Proceedings of the 16th annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 171–180, New York, NY, USA, 1993. ACM Press.

[124] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proceedings of International Conference on Computer Vision*, page 257, Nice, France, 2003.

[125] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, Washington, DC, USA, 2006. IEEE Computer Society.

[126] J. Wang, W.-j. Yang, and R. Acharya. Color clustering techniques for color-content-based image retrieval from image databases. In *ICMCS '97: Proceedings of the 1997 International Conference on Multimedia Computing and Systems*, page 442, Washington, DC, USA, 1997. IEEE Computer Society.

[127] M. Weber. *Unsupervised learning of models for object recognition.* PhD thesis, Pasadena, CA, USA, 2000. Supervisor-Perona, Pietro.

[128] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. *IEEE Conference on Computer Vision and Pattern Recognition,*, 2:2101, 2000.

[129] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proceedings of European Conference on Computer Vision, Part I*, pages 18–32, Dublin, Ireland, June 26 - July 1 2000.

[130] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *In Proceedings of ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.

[131] I. H. Witten, A. Moffat, and T. C. Bell. *Managing gigabytes: compressing and indexing cocuments and images.* Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[132] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2007.

[133] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *Proceedings of Conference on Knowledge discovery and data mining*, pages 864–873, San Jose, California, USA, 2007. ACM Press.

[134] H. Zhang. *Adapting Learning Techniques for Visual Recognition.* PhD thesis, EECS Department, University of California, Berkeley, May 2007.

[135] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: discriminative nearest neighbor classification for visual category recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2126–2136, Washington, DC, USA, 2006.

[136] J. Zhang, M. Marsza, S. Lazebnik, and C. Schmid. Local features and kernels for cassification of texture and object categories: a comprehensive study. *International Journal of Computer Vision,*, 73(2):213–238, 2007.

[137] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In *Beyond Patches workshop, in conjunction with CVPR*, jun 2006.

[138] R. Zhao and W. I. Grosky. Negotiating the semantic gap: from feature maps to semantic landscapes. *, Pattern Recognition*, 35:593–600, 2002.

[139] Q.-F. Zheng, W.-Q. Wang, and W. Gao. Effective and efficient object-based image retrieval using visual phrases. In *Proceedings of ACM International Conference on Multimedia*, pages 77–80, Santa Barbara, CA, USA, 2006.

[140] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian. Object-based image retrieval beyond visual appearances. In *Proceedings of ACM Conference on Multimedia Modeling*, Kyoto, Japan, Jan 9-11 2008.

[141] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian. Probabilistic optimized ranking for multimedia semantic concept detection via rvm. In *Proceedings of ACM Conference on Image and Video Retrieal (CIVR)*, Niagara Falls, Canada, Jul 7-9 2008.

[142] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian. Visual synset: a higher-level visual representation for object-based image retrieval. *The Visual Computer*, volume25, Issue 1:page 13, 2009.

[143] Y.-T. Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, and Q. Tian. Visual synset: Towards a higher-level visual representation. In *Proc. of Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, U.S., 2008.

[144] L. Zhu, A. Rao, and A. Zhang. Theory of keyblock-based image retrieval. *ACM Transactions on Information Systems*, 20(2):224–257, 2002.

Publications

1. Yan-Tao Zheng, Ming Zhao, Shi-Yong Neo, Tat-Seng Chua, Qi Tian, ”Visual Synset: towards a Higher-level Visual Representation”, in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2008*, Achorage, Alaska, U.S., June 24-26,2008

2. Yan-Tao Zheng, Shi-Yong Neo, Tat-Seng Chua, Qi Tian, ”Object-based Image Retrieval Beyond Visual Appearances”, in P*roceedings of ACM Conference on Multimeida Modeling (MMM) 2008*, Kyoto, Japan, Jan 9-11, 2008,

3. Yan-Tao Zheng, Shi-Yong Neo, Tat-Seng Chua, Qi Tian, ”Visual Synset: a Higher-level Visual Representation for Object-based Image Retrieval”, *The Visual Computer*, Volume 25, Issue 1 (2009), page 13.

4. Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, Hartmut Neven, ”Tour the World: building a web-scale landmark recogntion engine”, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, U.S., June 20-25, 2009

5. Yan-Tao Zheng, Shi-Yong Neo, Xianyu Chen, Tat-Seng Chua, ”VisionGo: towards true interactivity”, in *Proceedings of ACM Conference on Image and Video Retrieal (CIVR) 2009*, Santorini, Greece, July 8-10, 2009

6. Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, Hartmut Neven, Jay Yagnik, "Tour the World: a Technical Demonstration of a Web-Scale Landmark Recognition Engine", in *Proceedings of ACM Conference on Multimedia (ACM MM) 2009*, Beijing, China, October 19-24, 2009

7. Ling-Yu Duan , Jinqiao Wang , Yan-Tao Zheng , Hanqing Lu , Jesse S. Jin, Digesting Commercial Clips from TV Streams, *IEEE MultiMedia*, Volume 15, Issue 1, Date: Jan 2008, pp. 28-41

8. Yan-Tao Zheng, Shi-Yong Neo, Tat-Seng Chua, Qi Tian, Probabilistic Optimized Ranking for Multimedia Semantic Concept Detection via RVM, *Proceedings of ACM Conference on Image and Video Retrieal (CIVR) 2008*, Niagara Falls, Canada, Jul 7-9, 2008

9. Huan-Bo Luan, Yan-Tao Zheng, Shi-Yong Neo, Yong-Dong Zhang, Shou-Xun Lin, Tat-Seng Chua, Adaptive Multiple Feedback Strategies for Interactive Video Search, In *Proceedings of ACM Conference on Image and Video Retrieal (CIVR) 2008*, Niagara Falls, Canada, Jul 7-9, 2008

10. Shi-Yong Neo, Huan-Bo Luan, Yan-Tao Zheng, Hai-Kiat Goh, Tat-Seng Chua, VisionGo: Bridging Users and Multimedia Video Retrieval, *Proceedings of ACM Conference on Image and Video Retrieal (CIVR) 2008*, Niagara Falls, Canada, Jul 7-9, 2008

11. Shi-Yong Neo, Yuanyuan Ran, Hai-Kiat Goh, Yan-Tao Zheng, Tat-Seng Chua, Jintao Li, The Use of Topic Evolution to help Users Browse and Find Answers in News Video Corpus, I*n Proceedings of ACM conference on Multimedia (ACM MM) 2007*, Augsburg, Germany, Sep 23-29, 2007, full paper. (pdf)

12. Yan-Tao Zheng, Shi-Yong Neo, Tat-Seng Chua, Qi Tian, ”The Use of Temporal, Semantic and Visual Parititioning Model for Efficient Near-Duplciate Keyframe Detection in Large Scale Corpus”, in Proceedings of ACM Conference on Image and Video Retrieal (CIVR) 2007, Amsterdam, Holland, July 2007 (pdf)

13. Shi-Yong Neo, Yan-Tao Zheng, Chua Tat-Seng, Tian Qi, ”News Video Search With Fuzzy Event Clustering using High-level Features”, *In Proceedings of ACM conference on Multimedia (ACM MM) 2006*, Santa Barbara, U.S.A, Nov 2006