state-of-the-art：当前最高水平；一流的，最先进的；目前为止最好的（算法）

compressively：压缩的？精简的？

**Discover Novel Visual Categories from Dynamic Hierarchies Using Multimodal Attributes**

**基于动态层次结构和多模态属性的新可视类别发现**

Abstract—Learning novel visual categories from observations and experiences in unexplored environment is a vitally important cognitive ability for human beings. A dynamic category hierarchy that is an inherent structure in human mind is a key component for this ability. This paper develops a framework to build dynamic category hierarchy based on object attributes and a topic model. Since humans trend to utilize multimodal information to learn novel categories, we also develop an algorithm to learn multimodal attributes from multimodal data. The new multimodal attributes can describe objects efficiently and can generalize from learned categories to novel ones. By comparison with state-of-the-art unimodal attribute, the multimodal attributes can achieve average 4-19% improvements. We also develop a constrained topic model, which can accurately construct category hierarchies for large scale categories. Based on them, the novel framework can effectively detect novel categories and relate them with known categories for further category learning. Extensive experiments are conducted using a public multimodal dataset, i.e. color and point cloud data, to evaluate the multimodal attributes and the dynamic category hierarchy. The experimental results show the effectiveness of multimodal attributes to describe objects and the satisfactory performance of the dynamic category hierarchy to discover novel categories. By comparison with state-of-the-art methods, the dynamic category hierarchy achieves 7% improvements.

摘要 -- 在未知环境里，从观察和经验中学习新的可视化类别是人类一种极其重要的认知能力。在人的心中是内在结构的动态类别层次结构是这种认知能力的一个关键组成部分。本文构建了一个框架，来建立基于对象属性和主题模型的动态类别层次结构。我们首先提出了一种从多模态数据中学习多模态对象属性的算法。新的多模态属性可以有效地描述对象，并能够从学习的类别中归纳出新类别。对比目前为止最好的单模态属性，多模态属性平均能够达到4-19%的提升。我们还提出了一种约束主题模型，对于大规模的类别数量，该模型可以准确的构建出类别层次结构。基于对象属性和约束主题模型，本文提出的框架可以有效地检测出新的类别，并且能够将这些类别跟已知的类别联系起来。本文在公共多模态数据集中（包括彩色图和点云数据）进行了翔实的实验，评估本文提出的多模态属性和动态类别层次结构。实验结果表明多模态属性描述对象的有效性以及动态类别层次结构在发现新的类别上具有令人满意的性能。通过对比目前最好的方法，动态类别层次结构达到了7%的提升。

Index Terms—Multimodal Sensor, Dynamic Category Hierarchies, Constrained Topic Model, Multimodal Object Attributes, Novel Visual Category Discovery, RGB-D data.

索引词汇—多模态传感器， 动态类别层次结构， 约束主题模型，多模态对象属性，新视觉类别探索，RGB-D数据

**1. Introduction**

In an unexplored environment, there must be some unfamiliar objects to an artificial intelligent system. How to efficiently discover these novel categories is a key issue to explore new environments. Here the term ‘discover’ means to describe, and find the relationship between learned and novel categories. Human beings have an amazing ability to quickly discover visual categories from their experiences and observations when they are in an unexplored environment. The category hierarchy plays a key role for this visual cognition. Through organizing categories into a hierarchical structure, humans can remember and recognize thousands of categories, and categories unseen objects correctly. When people are 2-4 years old, they are taught to build category levels to help them to remember more and more objects [1]. Flexible hierarchies of categories of abstraction have become one of the most important issues in categorization and recognition [2] [3]. Researchers have also discovered neurophysiological evidence for category hierarchies [4] [5]. The category hierarchy is important to represent category relationships, store a large number of categories, and furthermore create novel categories [1] [2].

在未知环境中，对于一个人工智能系统，必定存在一些未知类别的对象。如何有效地发现这些新的类别是探索新环境的一个关键问题。这里的“发现”的意思是去描述新的类别，并且找寻与已经学习的类别之间的关系。当处于一个未知环境中，人类有着很惊人的认知能力，可以从他们的经验和观察中快速发现可视化类别。其中，类别层次结构在可视化认知中扮演着非常关键的角色。通过将类别组织成一个层次结构，人类能够正确记忆与识别成千上万种类别和那些未知对象的类别。当人类在2-4岁的时候，他们被教会建立类别的层次，可以帮助他们记住越来越多的对象[1]。 如何建立灵活的抽象对象类别层次结构已经成为分类和识别领域中最重要的问题之一[2][3]。 研究人员已经发现了类别层次结构的神经生理学证据[4][5]。类别层次结构在体现类别关系，存储大量类别以及认知新类别等方面十分重要 [1][2]。

We argue that a basic characteristic of a category hierarchy is dynamics. There are two aspects in the dynamics of a category hierarchy. First, a category will be assigned to different superordinates in different cognitive environments. For instance, a vase on a dining table is always looked at as a container for arranging flowers, while it will often be regarded as an artwork when people see it in a museum. Second, when novel categories emerge, a category hierarchy can dynamically change its structures, branching off automatically at an appropriate node and generating new nodes to represent novel categories. This process can be found in human cognition for learning objects of a novel category (NC). People will analyze properties of the new objects and find similar categories which have some of these properties. People can then categorize the new objects into an NC that has the same superordinates with similar categories.

我们认为类别层次结构的一个基本特征是具有动态性。类别层次结构的动态性体现在两个方面。首先，一个类别在不同认知环境中会被归属到不同的上级层次。例如，餐桌上的花瓶一般被作为插花容器，然而在博物馆人们看到它的时候往往被视为一件艺术品。其次，当新的类别出现的时候，类别层次结构可以动态的改变结构，自动分配到合适的中间层次，并生成表示新类别的节点。我们可以在人类学习新类别的对象的认知过程中发现这样的过程：人会分析新对象的特性和找到同样有着这些特性的相似类别，然后把新的对象分配到与相似类别相同的上级层次中。

In this study, motivated by human visual cognition, we proposed a novel framework to build a dynamic category hierarchy. This novel hierarchy can simultaneously satisfy two dynamic requirements. Our framework is based on object attributes [9] [10] and topic models [11] [12] [13]. Because in general multimodal information can improve the robustness and accuracy for visual perception system (e.g. [14] [15]), we therefore extend the object attributes algorithm in [9] to multimodal ones by integrating 2D and 3D data, which are obtained by ever-increasingly popular multimodal sensor. Thus the new object attributes can describe object more compressively and be better generalized from learned categories to novel categories. We develop a constrained topic model based on the hierarchical Latent Dirichlet Allocation (hLDA) [11] to build a more accurate category hierarchy. The constrained hLDA can guarantee that object belonging to one leaf node are in the same category, so the built category hierarchy can then accurately classify novel categories. The category hierarchy constructed by the constrained hLDA is data-driven, so it can more precisely represent the relationship among categories in a current environment. Furthermore, the constrained hLDA can clustering novel categories when more than one novel categories are detected by inserting them into appropriate position in the hierarchy. Therefore, based on multimodal attributes and the constrained hLDA, the proposed framework can efficiently discover novel categories.

受人类的视觉认知过程的启发，在本文中我们提出一种新的框架构建动态类别层次结构。新的层次结构可以同时满足前述两个动态性的要求。我们的框架是基于对象属性[9][10]和主题模型[11][12][13]。一般情况，由于多模态信息能够提升视觉感知系统的鲁棒性和准确性(比如[14][15])，因此我们扩展对象属性算法[9]为结合2D和3D数据的多模态对象属性，这些数据是从越来越广泛使用的多模态传感器中得到的。因此，新的对象属性可以更加精简的描述对象，更好的从学习的类别中归纳出新的类别。我们还提出一种基于层次隐式狄利克雷分配（hLDA）的约束主题模型[11], 去构建一个更加准确的类别层次结构。约束hLDA可以保证属于一个叶子节点下的对象都属于相同的类别，据此构建的类别层次结构可以准确的分类出新对象。通过约束hLDA构建的类别层次结构是数据驱动的，因此它可以更加精准的表示当前环境中类别之间的关系。此外，在检测到多个新类别时，通过将它们插入到层级结构中的适当位置中，约束hLDA可以聚类出新类别。因此，基于多模态属性和约束hLDA，我们提出的框架可以有效发现新类别。

In industrial applications, autonomous robots can assist people to execute a lot of tasks. However, existing autonomous robot can only execute limited learning tasks [31] [32] to learn some specific objects. When they encounter novel objects, they need the ability of discovering and learning novel categories. The proposed method can be applied to autonomous robots to further improve their autonomous abilities.

在工业应用中，自主型机器人能协助人们执行很多任务。然而，已有的自主型机器人只能执行有限的学习任务[31][32]，它们只能学习一些指定的对象。当遇到新对象，它们需要具有发现和学习出新类别的能力。我们提出的方法可以应用到自主型机器人，未来可以提升机器人的自主能力。

To summary, the main contributions in this study are three folds:

1. Novel multimodal object attributes are proposed to more accurately describe objects;
2. A novel constrained hLDA is developed to restrict a path of category hierarchy to correspond to one category;
3. And a novel framework is developed to build dynamic category hierarchy and efficiently discovery novel categories.

综上所述，本次研究中的主要贡献有三方面：

1. 提出新的多模态对象属性，能够更准确的描述对象；
2. 提出一种新的约束hLDA，限制类别层次结构的每一条路径都只对应的某一个类别；
3. 构建动态类别层次结构，提出一种新框架，能有效地发现新的类别。

By comparison with relative state-of-the-art work, extensive experiments in this study show that the proposed methods can achieve 7% improvements in terms of NC discovery.

通过对比目前相关的最先进工作，本文通过大量实验显示我们提出的方法在新类别的发现过程中能提高7%的准确度。

The remaining of this paper is organized as follows. After briefly introducing some related work in the next section, we describe how to train multimodal attributes in section III. Then we explain the proposed dynamic category hierarchy as well as the constrained hLDA in section IV. In section V, we show our experiments and the results are analyzed. Finally, we draw some conclusions in section VI.

本文的剩余部分组织如下。在紧接着的一节简要的介绍一些相关的工作之后，我们在第3节中介绍如何训练多模态属性。在第4节中解释我们提出的动态类别层级结构和约束hLDA模型。我们在第5节展示我们的实验以及对实验结果的分析。最后，在第6节描述本文的结论。

**II. Related Work**

Similarly to a general hierarchical model (e.g. [17]), a hierarchical category model can also simplify the complexity of a system and achieve higher performance. A category hierarchy organizes categories as tree-like structures where each node corresponds to one concrete category (a leaf node) or one superordinate of several concrete categories (a middle level node). The hierarchy can be built manually [18], or from existing semantic networks [7], or directly from image data [6] [8]. The first two methods are obviously not flexible enough to represent dynamic hierarchies. In [6], methods can build different hierarchies for different training image sets to improve object recognition. In [8], each node in the built hierarchy has a semantic tag which provides a certain semantic meaning. However, this method does not take into account how to change this tag when the hierarchy is changed as novel categories emerge. These methods can build category hierarchies, but they do not aim at discovering and learning novel categories but improving object recognition.

类似于一般的层次结构模型（比如[17]）,一个层次结构类别模型也可以简化系统的复杂性并达到更高的性能。一种类别层次模型把类别组织成树形结构，其中每个节点对应一个离散的类别（一个叶子节点）或者若干离散类别的上级层次（中间层节点）。层级结构可以手动构建[18]，或者通过已有的语义网络，或者直接来自图像数据 [6] [8]。用前两种方法表示动态层次结构显然不够灵活。在[6]中的方法可以针对不同的训练图像集构建不同的层次结构去提高对象的识别。[8]中的方法，在构建的层级结构中每个节点有一个语义标签，它可以提供一定的语义内容。然而，这个方法并没有考虑到当新类别出现时，层级结构改变的情况下，如何去修改这个标签。这些方法虽然能够构建类别层次结构，但是他们的目标不是发现和学习新类别，而是提高对象的识别。

In the computer vision community, only little work has been proposed recently to discover unknown object categories [10] [19] [20] [21] [22]. Lampert et al. [10] utilize semantic properties of objects to recognize novel categories which have semantic descriptions but no image samples. The method proposed in [19] also entirely relies on semantic properties to recognized novel categories which exist in the knowledge base. However, it is not always possible to obtain sufficient semantic properties to distinguish and recognize all object categories because of the limited human knowledge on object properties [9]. Furthermore, this method does not take category hierarchy into account and cannot find an entirely unknown category. Based on the work of [10] and further employing non-semantic properties of objects, the algorithms in [20] are also used to discover novel categories, yet they do not take category hierarchy into account. Sivic et al. employed a hierarchical Latent Dirichlet Allocation (hLDA) model to construct the category hierarchy for category discovery [21]. However, the hLDA is an unsupervised topic model, therefore the category hierarchy cannot guarantee that each path corresponds to one category, which inevitably leads to inaccurate hierarchies and lowers the accuracy of NC detection. Recently, in [22] a different perspective of organizing category relationship is utilized to discover novel categories. It models the interactions among objects in one image. This model can benefit from the object co-occurrence to efficiently discover categories.

在计算机视觉领域，最近只有少数工作提出是去发现未知的对象类别[10][19][20][21] [22]。Lampert等[10]利用对象的语义属性识别有语义描述却没有图像样本的新类别。论文[19]中提出的方法也完全地依赖于语义属性去识别知识库中存在的新类别。但是，这不总能够得到足够多语义属性去区分和识别所有对象类别，因为人类在对象属性上只有很有限的一些知识[9]。此外，这个方法没有考虑类别层次结构并且不能找到一个完全未知类别。基于[10]的工作以及之后采用的对象的非语义属性，论文[20]中的算法也被用于发现新的类别，但他们仍然没有考虑类别层次结构。对于类别的发现，Sivic等人采用层次隐式狄利克雷分配(hLDA)模型去构建类别层次结构[21]。然而hLDA模型是一种无监督主题模型，因此类别层次结构不能保证每条路径对应一种类别，这样必然导致不准确的层次结构并降低了新类别检测的准确度。最近，论文[22]中利用组织类别关系这一不同角度去发现新的类别。文章中构建了一幅图像中对象间的交互模型。这个模型从对象共现中受益，因此能够有效的发现类别。

In the field of machine learning, there are several methods that aim at novelty detection, such as [23] [24]. All of these works regard all known samples as belonging to one category and build a corresponding model to represent this category. A new input sample will be regarded a novel pattern, if the distance between this sample and the built model is larger than a certain value. Although these methods can detect the novelty efficiently, they cannot distinguish two different novel categories directly. It is obvious that these methods cannot be directly used for NC discovery, when more than one NC needs to be detected and recognized.

在机器学习领域，存在很多目标是新颖性检测的方法，比如[23][24]。所有这些方法把所有已知样本认为是属于某个类别的，并且构建一个对应的模型去表示这个类别。如果一个新的样本和构建的模型之间的距离是大于某个值的话，这个样本将会被认为是一种新的模式。这些方法虽然能够有效检测新颖性，但是不能直接区分两个不同的新类别。当很多新类别需要被检测和识别的时候，显然，这些方法不能直接用于发现新类别。

III. Multimodal object attributes 多模态对象属性

In this study, object attributes are employed to describe objects. Since object attributes can be shared by different categories, they have excellent generalizability from learned categories to novel ones [9] [10]. These methods are based on only 2D images. However, humans always utilize multimodal information. In particular, when they learn visual categories, 2D and 3D information are the most frequently used. Therefore, for the attributes, this is also the case that unimodal attributes are not sufficient to describe object accurately and multimodal attributes should be employed. Therefore, we develop the multimodal attributes to describe objects by integrating multimodal base-features (BFs) in this study. The multimodal attributes are extended from [9], where original 2D attributes are trained from 2D BFs. Since we use RGB-D data in this study, for each image, a corresponding point cloud can be obtained. Then we combine 3D base-features with 2D BFs to obtain the multimodal BFs.

本文中，对象属性用于描述对象。因为不同的类别对象能够共享同一对象属性，所以对象属性具有极好的从已知类别扩展到新类别特性[9][10]。目前对象属性方法都是基于2D图像。然而，人们总是利用多模态信息。特别地，当他们学习视觉类别的时候，2D和3D信息是最频繁使用的。因此对于对象属性来说，单模态属性不足以准确地描述物体，需要采用多模态属性。因此，在本研究中我们提出通过合成多模态基础特征，使用多模态属性来描述对象。我们扩展 [9]中的方法训练多模态属性。在本研究中，我们采集每幅图像的RGB-D数据，这样我们能够得到2D图像和对应的3D点云图。之后我们结合3D基础特征和2D基础特征训练得到多模态的基础特征。

1. Extraction of Multimodal Base-Features 多模态基础特征的抽取

Generally, the 3D data is represented by point clouds which are textureless data and only suitable for describing 3D shape characteristics. Two kinds of invariant 3D shape descriptors, the 3D shape context (3DSC, [25]) and the fast point feature histogram (FPFH, [26]) are employed to construct multimodal BFs. The original features, normal and curvatures are also used to form 3D BFs.

一般情况下，3D数据通过点云表示，而点云数据是无纹理数据，只适合描述3D形状特征。我们采用两种不变3D形状描述符，3D形状上下文（3DSC，[25]）以及快速点特征直方图(FPFH, [26])构建多模态基础特征。同时，一些原始特征，比如法线和曲率也用于组成3D基础特征。

The 3DSC constructs a multi-scale multi-sector sphere containing the 3D surface of an object, and counts the numbers of points for all bins to form a histogram. Thus this histogram models the distribution over relative positions of points as a robust and compact, yet discriminative descriptor. Particularly, a sphere is divided into 6 shells with logarithmically increasing radius:

3DSC 首先建立一个球体，然后将它分割为不同半径和不同扇区的一种结构，之后统计每个半径和扇区内部的点的数目，构建用于描述形状的直方图，形成一个稳健的，简洁的，具有很好区分能力的描述符。本文中，一个大球体被分成6个不同半径，这样最小的半径形成一个小球体，其余5个半径形成5个球壳。半径是随着对数增加的：



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

其中第i个半径r\_i取决于不同半径的数目和以a为底的对数。在本文中，a取值为2，这可以保证小球体和每个不同半径的球壳有同样的体积。球体的经度和纬度被分成12和角度间隔各自都是30度的6个部分。最终，每个点的3DSC描述符就是一个432维的直方图。

To robustly describe an object shape, the descriptor should be invariant. The 3DSC is natural scale invariant since the size of sphere is relevant to the size of the entire object. The 3DSC naturally holds the translation invariance since it uses the relative positions among points to construct the histogram. Unnormalized 3DSC is not rotation invariant since the sphere is divided into multi-sectors. Following the same strategy as used in [25], the Principal Axes Transform [27] method is also employed to perform rotation invariant normalization.

为了能够鲁棒的描述一个物体的形状，描述符应该具有不变性。3DSC天然就是尺度不变的，因为球的大小是跟每个对象的大小相关。同样的，3DSC本身就具有平移不变性，因为它使用点之间的相关位置去构建直方图。由于球体被分成了多扇形，未归一化的3DSC不是旋转不变的。遵循[25]中同样的策略，主轴变换[27]方法也被用于进行旋转不变归一化。

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

为了使用3DSC，我们构建“词包”类型特征。一些对象被选为训练数据，对每个对象中的所有点都计算3DSC描述符。然后通过K-means算法将这些描述符聚类成256个中心。再将对象中每个点的3DSC描述符量化为256个K-means中心中离得最近那个。因此对于每个对象，它的3D点云能够通过3DSC描述符形成一个256维的向量。

An FPFH consist of 81 bins, forming a descriptor with 81 dimensions. Unlike the 3DSC that utilizes relative positions of points in an object, the FPFH models the relative normal directions of points to form an invariant descriptor. The FPFH has been proved to efficiently represent object shape and to be used for 3D object recognition [26]. In this study, similarly to the 3DSC, 256 cluster centers are also computed by the K-means algorithm. For a point of an object, its FPFH is computed and subsequently quantized to the nearest one of 256 K-means centers. Thus one more 256 dimensional vector feature is obtained to describe object’s 3D shapes.

一个FPFH包含81个箱柱，构成一个81维的描述符。不同于3DSC利用一个对象中点的相关位置，FPFH描述符构建了多个点的相对法线方向的模型，形成一个不变的描述符。FPFH描述符已经被证明可以有效表示对象形状并且用于3D物体识别[26]。在本文中，也通过K-means算法计算256个聚类中心。对于对象点云中的某一点，先计算其FPFH描述符，随后量化为256个K-means中心中最近的一个。因此得到了又一个256维的向量特征，用于描述物体的3D形状。

The point normal are quantized into a 72-bin histogram by dividing the zenith angle into 12 parts and the azimuth angle into 6 parts with identical 30°interval. The curvatures are quantized into a 128-bin histogram by averagely dividing the difference between the maximal curvature and the minimal one. i.e. the curvature values in the i-th bin

通过把顶角均分为12个部分，将方位角均分为6个部分，能够将物体的点云法线方向形成一个72维的向量直方图用于描述物体。再通过将物体点云的曲率最大值和最小值之间的差值平均分为128分，得到128维向量的直方图，用于描述物体。比如第i维的曲率值如下：

,

where  is the curvature interval between two adjacent bins, and max(c) and min(c) denote the maximal curvature and the minimal one respectively.

其中是两个相邻维之间的曲率间隔，max(c)和min(c)分别表示最大曲率和最小曲率。

A 3D object is first divided into 8 boxes with the same size. For each box, the four kinds of aforementioned features are computed. Then for the whole object, we also compute these four kinds of features. They are all concatenated to form a 6408 dimensional feature which is referred as the 3D BFs.

一个3D物体首先分成8个同样大小的盒子。对于每个盒子，计算四种前面提到的特征。然后对于整个对象，我们计算这四个不同种类的特征。所有这些被连接起来形成一个6408维的特征，称为3D基础特征。

The 2D BFs are obtained by the same approach used in [9], where a 9751 dimensional vector is extracted from an object. Finally, there is a 16159 dimensional BF for each object by stacking 2D and 3D BFs together.

2D基础特征通过[9]中采用的相同方法得到的，即从对象中抽取出一个9751维的向量。最终，通过将2D和3D基础特征叠加到一起，对于每个对象都有了一个16159维的的基础特征。

1. Training Attributes训练属性

After choosing some objects possessing an attribute as positive samples and some objects without this attribute as negative samples, the L1-regularized logistic regression is employed to select a valid set of BFs. Then a linear Support Vector Machine (SVM) is employed to train the attribute classifiers. In this study, 32 multimodal attribute classifiers that are frequently possessed by indoor objects are trained. Among these attributes, sixteen attribute classifiers are used in [9] where they are trained by 2D BFs. In the experiment section, we will show that most of these attribute classifiers have better performance when trained from multimodal BFs. Ten attribute classifiers are shape attributes that are more suitable to be trained by 3D BFs. Finally, the remaining six attribute classifiers have more complex semantic meanings and should be trained by multimodal BFs. In the experiment section, we also train these attribute classifiers by only 2D BFs, and compare them to evaluate the performance of the new multimodal attribute classifiers.

将具有某种属性的物体作为正样本，而将没有这种属性的物体作为负样本，再通过使用L1正则化的logistic回归来选择一个有效的基础特征集。之后通过线性支持向量机(SVM)进行属性分类器训练。在本文中，我们为室内对象训练了32个常见的多模态属性分类器。在这些属性中，16个属性分类器在[9]中是通过2D基础特征训练得到。但是如果通过多模态属性训练这16属性，大部分属性具有更好的区分性，这一点我们将在实验中展示。其次，10个形状属性通过3D基本特征进行训练。最后剩余的六个属性分类器有更复杂的语义内容，因此我们通过多模态基础特征进行训练。在实验中，我们也使用2D基础特征训练这6个属性，通过和基于多模态训练的属性分类器进行比较，发现多模态属性性能更好。

Moreover, similarly to [9] the auxiliary discriminative attributes are also employed in this study to study to describe objects more accurately. In fact, one thousand discriminative attributes are trained from the multimodal BFs by using the same strategy with [9].

此外，类似于论文[9]，我们也使用了辅助的区分性属性，以便更准确的描述对象。事实上，1000有判别能力的属性通过使用[9]中相同策略从多模态基础属性中训练得到。

IV. Dynamic category hierarchies动态类别层次结构

The proposed framework includes four parts as outlined below. First the attributes of annotated objects in training image datasets are computed based on extracted BFs. Given these attributes, a constrained hierarchical Latent Dirichlet Allocation (chLDA) model is developed to build a hierarchy representing the relationship among current categories. For each node in the built hierarchy a classifier can then be trained based on those object samples assigned to this node. For new object samples, their attributes are also computed first, then node classifiers are used to determine if they belong to known or unknown categories. If the new object samples are in novel categories, this framework can indicate their superordinates and the hierarchy will branch off at appropriate nodes to generate new paths to represent new categories.

本文提出的框架包括如下四个部分。首先我们从训练图像集中提取标注的对象的属性。第二部分，在给定这些属性之后，我们提出一种约束层次隐式狄利克雷分配(chLDA)模型来构建一个表示当前类别之间关系的层次结构。第三部分，对于构建的层次结构中的每个节点，我们基于分配给这个节点的样本对象训练一个分类器。第四部分，对于一个新的样本对象，我们首先计算其属性，然后使用各个节点分类器确定它们是属于已知类别，还是未知的新类别。如果新的对象样本是属于新的类别，这个框架可以指出他们的上一层节点并且这个层次结构将会产生新的分支和新路径表示这一新的类别。

A. Constrained Hierarchical Latent Dirichlet Allocation

The hLDA is originally described by the terms ‘documents’ and ‘words’ in [11]. Actually, the terms ‘documents’ and ‘words’ are the same as the terms ‘objects’ and ‘attributes’ in this study. In this section, we use them interchangeably.

hLDA模型最初是通过“文档”和“单词”进行描述的[11]。事实上，“文档”和“单词”是跟本文中的 “对象”和“属性”相似。在本节中，我们交叉的使用它们。

The hLDA model can construct a tree-like structure where each node corresponds to one topic. A topic is a distribution of words, where the words that have high probabilities can reflect the property of this topic. It is worth nothing that nodes located at the upper level correspond to more general topics and nodes located at the lower level correspond to more specific topics. Therefore, different documents that are assigned different paths may share the same superordinates if they have the same general but different specific topics.

hLDA模型能够构建一个树形结构，每个结点对应一个主题。一个主题是单词的分布，在分布中具有很高概率的单词可以反映主题的特性。值得注意的是，位于上层的节点对应更一般主题，而位于下层的节点对应更为特定的主题。因此，指派到不同路径的不同文档，如果它们具有相同的普通主题，而具有不同的特定主题，那么它们可以共享相同的上层节点而具有不同的下层节点。

When using attributes to describe an object, the object can analog a document where, and if we treat an attribute as a word, the object can also form several topics that can represent different aspects of an object with different generalities. For instance, in Fig. 4 the ‘chair’ and ‘sofa’ have similar furniture part attributes; they can be categorized into one superordinate at the third level. But they are different categories and have different category-specific attributes, so as the fourth level they are assigned to different leaf nodes. Similar examples can be found from other categories.

当使用属性去描述对象的时候，如果我们将一个属性看作是一个单词，则对象可以看成一篇文档。那么对象也可以形成多个主题，代表对象的不同方面和不同的共性。比如，在图4中，“椅子”和“沙发”有相似的家具部分属性；因此他们在第三层上被分配到相同的节点上。但是它们是不同的类别，有不同的特定类别属性，因此在第四层，它们被指向了不同的叶子节点。相同的例子还可以从其他的类别中找到。

Therefore, the hLDA can be used to build category hierarchies. However, the hLDA is an unsupervised algorithm; it cannot guarantee that each leaf node corresponds to exactly one category. In our experiment, only 20-80 percent (the average is 53 percent) of samples in one category can be assigned to the same leaf node. We therefore extend the hLDA to a constrained hLDA. A category label is specified for each object in the training data, thus every leaf node is constrained to only contain one category label.

因此，hLDA能够被用于构建类别层次结构。但是，hLDA是一个无监督算法；它不能保证每个叶子节点对应于一个类别。在我们的实验中，同一种类别中只有20%到80%（平均53%）的样本是被分配到相同的叶子节点。因此，我们扩展hLDA为约束的hLDA。对于训练数据中的每个对象，我们指定一种类别标签，每个叶子节点被限制只能包含一种类别标签。

Suppose there is a dataset composed of a corpus of M documents, {w1, w2 ... wM}. Each document contains a collection of words from a vocabulary of V terms, divided into several different topics defined as the distribution words. The formally generative process of the constrained hLDA is:

假设有一个包含M篇文档语料库的数据集，{w1, w2 ... wM}。每篇文档包含一个单词集合，都是来自有V个词项的词汇表，被分成多个不同的定义为单词分布的主题中。则本文提出的chLAD的生成过程如下：

1. For each node k∈T in the infinite tree , draw a topic ,
2. For each document, d∈{1,2,…,D},
3. Draw a path for document d  ,
4. Draw a distribution over levels in the tree, ，
5. For each word,

i) Choose level,

ii) Choose word, which is parameterized by the topic in position  on the path.

1. For each document, draw a label .
2. 在无限树中，对于每个节点k∈T ，得到一个主题,
3. 对于每篇文档 d∈{1,2,…,D},
4. 根据分布采样文档 d 的路径,
5. 根据从数中采样一个层次上的分布，
6. 对于每个单词，

i) 根据分布采样该单词所在的层次，

ii) 根据分布采样单词，这个分布被参数化为在路径上位置处的主题分布。

1. 对于每篇文档，根据分布采样其标签.

The first two steps are the generative process of the original hLDA.  denotes a Dirichlet distribution with hyperparameter.  denotes the topic assigned to a node k in the hierarchy T. The symbolmeans that the left hand variables obey the distribution at the right hand.  denotes the path for a document d, which consists of several nodes and is sampled according to the nest Chinese Restaurant Process with hyperparameter.  defines the distribution of a document d over levels in the hierarchy, which obey a two-parameter variant of the GEM distribution with hyperparameters m and.  denotes the assignment of the n word in the d document, according to a multinomial distribution given the document-specific parameter. Then  denotes which word in the vocabulary is chosen for the n word in the d document, according to a multinomial distribution given parameters. Note here that these hyperparameter  and  can be set manually or sampled from prior distribution. For details readers can refer to [11].

前两步是原始hLDA的生成过程。表示带有参数的Dirichlet分布。表示在层次结构中分配到节点k处的主题。符号表示左边的变量服从右边的分布。表示文档d的路径，包含多个节点，根据带有参数的嵌套中国餐馆过程采样得到。定义了每个文档d在层次模型中其单词如何分配到不同的层次上的这样一个分布，服从两个参数m和的GEM分布。表示文档d中第n个单词的分配，这个分配服从指定参数的多项式分布。然后表示词汇表中被选中到文档d中的第n 个单词，服从给定参数的多项式分布。这里需要注意参数和能手动设置或者从先验分布中采样得到。对于细节，读者可以参考论文[11]。

The step (3) is an extra step different from original hLDA, which just constrains the assignments of object samples. The parameter  of the multinomial distribution is set to 0.1 priorly. By this extra step, the built hierarchy can guarantee that each leaf node corresponds to only one category. But it is worth nothing that one category may correspond to more than one leaf nodes. It is reasonable because there may be a situation where the difference in the object samples is large enough for these object samples to be treated as two sub-categories. This characteristic is especially useful when distinguishing more than one NC when they are detected in the same middle node by the method in section IV-B. The level of the built category hierarchy, L, is fixed. The large level parameter can lead to a more complicated hierarchical structure which can more accurately represent the relationship of all categories. However, it will also result in more isolated nodes without any branches and more time-cost. Therefore, the L is set to proportion to, where C is the number of categories.

第三步是不同于原始hLDA的额外一个步骤，限制了对象样本的分配。多项式分布的参数优先设为0.1。通过这额外的一步，构建的层次结构能保证每个叶子节点对应唯一一种类别。但是值得注意的是一个类别可能会对应多个叶子结点。这是合理的，因为可能存在这样一种情形，如果同一类的对象样本具有足够大的差异，我们可以考虑将这样一个类分为两个或多个子类。当我们通过第四节B中的方法检测到多个新类别属于相同的一个中间节点是，chLDA模型的这样的特性在区分多个新类别特别有用，这一点将在稍后详细解释。chLDA中构建的类别层次结构的层级数目L必须首先确定。大的层级数目能形成更复杂的层次结构，可以更准确的表示所有类别的关系。然而，这也将导致更多没有分支孤立的节点，从而需要话费更多的运行时间构建对象层次。因此，本文中我们限制L跟成正比，其中C是类别数。

Probabilistic Inference: The Gibbs sampling is employed to perform posterior inference. There are three main steps in posterior inference: the sampling of level allocations, the sampling of path assignments and the sampling of category responses. The last two steps can merge into one step. There are two situations when sampling the path. When the sampled path is a full path from the root to the leaf node, because the leaf node only corresponds to one category response, the path also corresponds to the same category response. When the sampled path ends at the middle level node, a new branch will be generate and a new category response will be created. The leaf node generated in this new path will be fixed to correspond to a new category response. For both situations, assigning a path to a document also determines the category response of this document. Therefore, the last two steps can be executed in one step and only two sets of probabilities need to be computed.

概率推断：我们采用Gibbs采样执行后验概率推断。在后验推断中有三个主要步骤：采样层级分配，采样路径分配，以及采样类别响应。其中，后面两步可以合并为一步。当采样路径时，存在两种情况。首先，当采样的路径是从根节点到叶子节点的一条全路径时，因为叶子结点只对应一个类别响应，所以整个路径也对应相同的类别响应。其次，当采样路径停在中间节点时，将会产生一个新的分支，并创建一个新的类别响应。在这条新路径中产生的叶子节点将从此对应一个固定的新的类别响应。对于这两种情况，指派一条路径到一篇文档也是确定了这篇文档的类别响应。因此，后两步能够在一步中执行，并且只有两个概率需要计算。

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

1：当文档的路径已经指定时，给定当前所有其它变量的值，我们根据下式为来自文档d中的第n个单词采样其层级分配变量：

 (2)

其中**z**和**w**分别表示所有采样的层级的集合以及所有单词的集合，表示除文档d中第n个单词之外的所有文档的单词的所分配的层级的集合；表示除了第d篇文档中的第n个单词之外的所有单词的集合；**c**表示所有采样的路径的集合。这个等式的跟hLDA中采用层级分配的那一步相同。更多细节可以在[11]中找到。

2： 当给定所有单词的层次分配之后，和文档d关联的路径按下式条件概率分布进行采样。此条件概率分布给定所有其它文档所采样的路径，所有文档的响应以及所有观测到的单词为条件：

 （3）

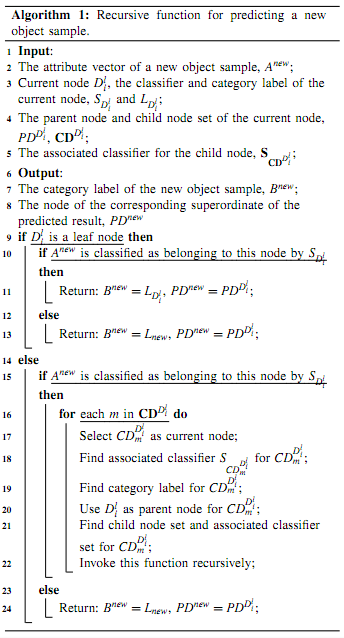
其中表示所有除了第d篇文档外的路径的集合。

公式右手边前两项是跟hLDA相同。它们的计算方法能够在[11]中找到。第三项是跟hLDA不同，因为类别响应是新加的。这一项可以像如下计算得到：

其中是当前采样迭代的响应数，P表示采样的路径。#[.]表示满足括号中条件的一个数组的元素个数。值得注意的是，一条路径必须作为一个整体进行采样，因为它在每个层级处的概率依赖于之前层级的概率[11]。如果一条通过一个叶子结点所表示的全路径（从根节点到存在的叶子节点中的一个）采样得到，那么类别相应就被确定了。因此概率将等于1因为文档采样了正确的响应，或者会是0因为文档采样了错误的响应。另一种情况，如果采样的路径的尾部是一个中间节点节点（通过这个中间层节点表示），概率是作为采样一个新类别响应来计算的。

1. 训练分类器和预测新的对象

在构建了一个类别层级结构后，我们能用这个层级结构去识别对象并找到新的类别。假设一个当前的类别层次结构是并且位于第*l*层的第*i*个结点表示为，它的父节点和子节点集合表示为和。对于根节点和中间节点，我们用“一类”SVM算法训练这个分类器，其中只有属于这个节点的对象样本认为是正样本，而其他对象样本不作考虑。对于叶子节点，我们使用“两类”SVM算法训练分类器，其中属于这个节点的对象样本认为是正样本，并且其他对象样本看作是负样本。我们用表示节点的分类器。同样的，节点的父节点和子节点集合的分类器分别表示为和。实际中，我们使用LibSVM[28]去训练这些分类器。在得到当前类别层次结构中所有节点的分类器后，列在算法1中的一个递归函数是用于预测一个新对象样本的类别标签。



从层次结构的根节点开始，节点分类器用于预测一个测试对象样本的标签。如果这个样本属于一个当前的节点分类器（15行），对应这个节点的子节点的分类器用于预测这个采样标签的（17行，21行），直到一个叶子节点（9行-13行）或者不能找到这个样本属于的分类器（24行）。如果这个样本被预测是属于一个叶子节点，表示这个样本属于一个已知类别，则这个算法将会返回这个已知类别标签和它的上一层级（11行）。如果这个样本属于一个中间节点但不是属于这个节点的任何子节点，那么这是一个未知类别的对象并且算法能输出“new”作为它的类别标签，同时也将会指出它的上一层级（13行和24行），意味着这个层次结构将会分叉到对应它的上一层级的节点。

1. 确定新对象的类别

当多个对象样本被检测为属于新的类别时，这些样本可能属于同一个上级层次中的不同新类别。然而，先前几步只给所有这些样本一个“new”标签但不能存在确定多少个新类别未来也不能确定哪个样本属于哪个新类别。实际上，这是一个以无监督的方式聚类对象类别的问题。许多方法能用于解决这个问题[29]，包括隐变量方法，以及频谱聚类方法。约束hLDA模型也是一种隐变量方法，能用于聚类新的类别。

约束hLDA模型有一个重要的特性，如果对象间的差异足够大的话，它能为一个类别创建多个叶子节点。因此我们通过分配给这些属于同一个新的叶子节点的样本一个统一的“new”类别响应，然后重复之前的生成过程，就能够将这些新对象样本分配到不同的新的叶子节点中，从而达到自动聚类这些对象类别的目的。因为属于不同的新类别的对象有足够的差异，所以它们将能够被分成不同的节点，之后我们再给每个新类别分配一个新类别标签。

相比其它的聚类方法，本文提出的基于类别层次结构的框架在从未知类别的对象中聚类出对象类别这一问题，至少具备如下的优势：只有属于相似的新的类别的对象样本会被分配到同一个新的节点，因此减少了同时需要进行聚类的新类别的数，从而可以提升新类别聚类的性能。

V.实验

实验中的数据都是通过多模态传感器（例如微软Kinect）同时收集对象的颜色和深度信息（RGB-D）。然后从深度图中生成对应的点云数据。为了使得实验更加有说服力，我们使用文［16］中提供的公共RGB-D数据集进行实验。这个数据集包含了四个不同的场景：厨房，会议室，书桌以及餐桌。从这些四个场景，我们抽取了25个类别，每个类别包括60个。对于每个对象的2D图像和点云中，我们使用约束框约束这些对象。一些例子如图1所示。

我们首先评估多模态属性的性能。然后验证构建类别层次结构的准确性。最后，我们评估新类别检测和发现的准确度。

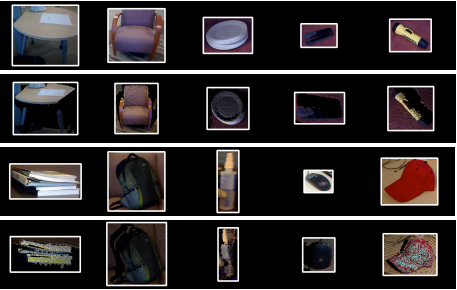


图1. 在实验中使用的对象的一些例子。

第一行是2D图像，第二行是带颜色的3D点云。

1. 多模态属性的估计

在本次研究中，我们使用32个语义属性。为了训练这些属性，我们选择一半的对象样本作为训练集。

在前16个属性已经在[9]中用到，但是在我们的实验中，他们是从多模态基础特征中训练的。因此我们首先通过对比从2D基础特征中训练出来的属性，来评估这16个多模态属性的性能。我们使用每个属性分类器的ROC曲线面积来表示分类的精确度。16个属性的ROC曲线面积显示在图2(a)中。从中我们可以观察到的，对比与完全从2D基本特征中训练得到的属性，所有这些通过多模态基础特征训练得到的16个属性具有更高或者相似的精确度，并且存在平均4%的提升。

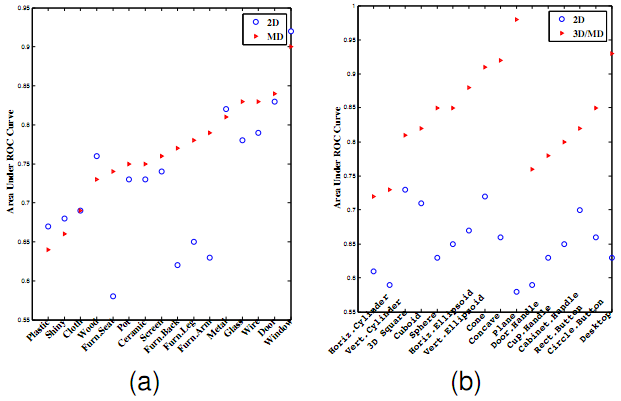


图2. 对于16种属性精确度的比较，

分布通过2D基础特征（圆圈）和多模态基础特征（三角形）

对于后面16中语义属性，我们也从2D基础特征中训练它们的2D版本作为对比。同样是使用这些属性分类器的ROC曲线面积表示分类精确度。比较结果如图2(b)所示。在这16种属性中，前10种属性是有更适合从3D数据中训练的3D形状的属性（指的是3D属性），以及后6种属性需要从2D和3D数据（指的是多模态属性）中训练的一些更复杂的属性。因此这里要比前16种属性的性能要提升的更多。所有这些属性都比那些只从2D基础特征中训练的对应属性具有更好的的区分性。其中“平面”属性和“桌面”属性提升效果最明显，意味着这两种属性更适合通过多模态基础属性来表示。最后我们计算得到通过使用多模态属性的平均性能提升各自是19%（中间10种属性）和18%（最后6种属性）。

1. 评估类别层次结构构建的精确度

在这一小节中我们评估使用chLDA构建类别层次结构的精确度。用于构建层次结构的类别是随机变化的。实际上，我们的实验中一共构建了50个层次结构。每一次我们都随机选择12-20个类别构建类别层次结构，其中每个类别有60个样本。剩余的类别则用于在下一节实验中测试新类别检测和发现的性能 。

每个层次结构的层级都设为4，根据第四节A中的描述，当C∈[12,20]时的值四舍五入后都等于4。chLDA模型的其它参数设为η={8.0,6.0,3.0,3.0}，γ=0.5，m=0.35，π=100，以及φ=0.1。通过使用这些参数，chLDA模型能在几次迭代后收敛（我们的实验中，平均迭代次数是15次）。因为不存在标准的类别层次结构作为参考可以用于评估我们构建的层次结构的正确性，所以我们邀请10个成人（5个男人和5个女人，年龄都是在20到50岁之间）来确定在一个层次结构中每种类别的正确性。一种目前为止最好的方法[6]作为对比方法。根据[6]中的建议，这个方法的一个减弱参数设为α=0.5。图3(a)显示了平均的准确度，其中所有类别上的平均准确度在图示中显示。在图3a中，“chLDA”表示本文提出的框架， “Ref”指的是参考方法。

从图3a中很明显看出chLDA模型的性能是令人满意的。本文提出的方法平均的准确度比参考的方法要高出6%。相比参考方法使用层次式归一化割方法，本文方法采用非参数贝叶斯技术，能够获得更为自然的属性分布。因此一个中间级节点不是它的子节点的简单汇总，而是具有它自己属性分布。更高概率的属性能精确的影响跟这个节点相关的主题的特征。正如图3a中显示的，chLDA能构建所有类别的更准确的层次结构。

C． 估算检测和发现新类别

在动态的环境中，人类能够识别一种新对象属于一个新类别并且区分新对象的不同的新类别的这样一种认知能力是很重要的。这也是本文提出的框架所要达成的最重要的目标。在本小节中，我们将评估新类别检测和发现精确度。我们认为新类别的检测任务是一个对象被检测为不同于所有学习的类别的一种新类别，而不管新类别对象是否属于不同的新类别。在对象样本被检测为属于新的类别后，层次结构将会产生新的路径表示他们。但是可能存在这样一个问题：如果两个对象样本被同一个中间层节点分类器检测为属于新类别，即使它们是属于不同的新类别，它们也将会被分配到当前层次结构中的相同路径。例如，图4中属于“书桌”和“餐桌”类别的对象样本通过使用节点分类器不能被检测为两种不同的类别。幸运的是，chLDA模型能将这两种对象聚类到两个不同的新类别中，正如我们在第4节的C中解释的那样，我们称之为对象的发现。

基于在最后一个实验中构建的类别层次结构，我们估计新类别的检测。此外，一个目前为止对于新颖性检测方面最好的方法[23]用于跟我们提出的框架作对比。在提出的框架中，只有当一个对象样本被分类为属于一个非叶子结点，且不属于这个节点的任何子节点，那么这个对象样本被检测为属于一种新类别。

对于参考的方法，所有用于构建一个层次结构的类别被认为是已知的类别，剩余的类别作为是将要检测的一种新类别。新类别中的所有对象样本输入到参考方法中并且计算检测正确的准确率。每种类别的平均准确率如图3(b)所示，其中两种方法的平均正确率显示在框中。

正如图3b中所示，本文提出的方法的新类别检测准确率是高于参考的方法，平均准确率提升7.5%。这是因为我们提出的框架利用了层次结构而参考的方法没有采用。本文方法是基于类别层次结构的，检测一个对象样本是否属于新的类别是从根节点到叶子节点层次式地进行判断，每次判断只需要区分同一个层次中的类别，而不是在所有类别，因此这提升了新类别检测的性能。

在检测到新类别后，我们现在需要判断属于相同的新路径的新对象样本是否属于不同的新类别。我们通过对比目前为止最好的光谱聚类方法以及全局核k-means算法[30]来评估本文提出的方法用于发现新类别的性能。基于最后实验的检测结果，检测到是属于一种新类别的新对象样本是被设置了相同的新类别标签。因此可以得到一个新的临时层次结构，。在本文提出的框架中，我们对于所有已知对象和新的对象，重复利用在第4节的A中提到的生成过程，构建一个新的层次结构，。如果属于不同新类别但指派到新的临时层次结构中的相同叶子节点的对象能分成新的层次结构中，结果认为是正确的。比如图4中的对象样本“书桌”和“餐桌”可以被区分为两个不同的新类别。

对于参考的方法，不需要指出聚类的数目，因为它的全局特性。因此这个方法能直接用在所有的叶子节点上。准确率能直接通过统计准确聚类的对象数来计算得到。每一种类别的平均准确率如图3.(c)所示，它们的平均准确率同时也显示在图中。

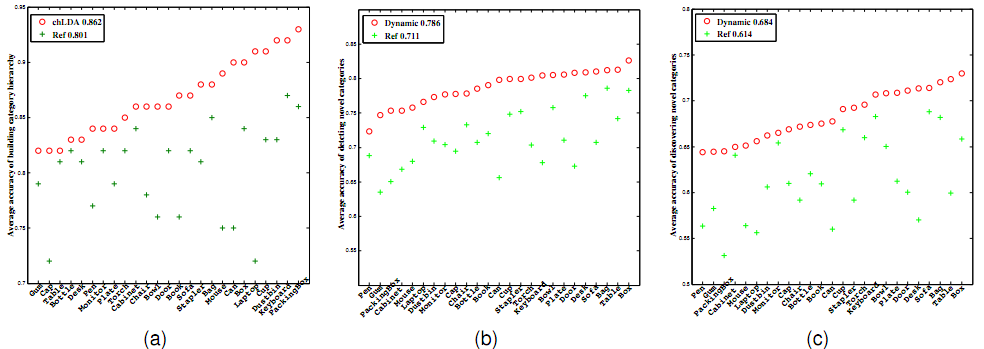


图3.构建类别层次结构的正确率

因为这个估计是基于新类别检测的准确结果，准确率要比最后一个试验中的结果要低。然而，仍然很明显我们提出的方法的性能要比参考的方法的性能好，因为提出的方法是基于层次结构。平均准确率是提升了7%。

一个层次式结构的结果显示在图4中。在图4a中，我们的方法首先构建了一个13种类别的精确层次结构。当给定了6种新类别，层次结构产生新的路径去表示他们，正如图4b所示。六种新的类别被插入到新的层次结构的适当位置。“显示器”类跟“笔记本”类有相同的属性，因此他们是在同一个上级层次下面。“杯子”类别跟相同二层节点下的相邻节点由一些相似的属性，但它也是有跟一些特殊的属性，比如“杯子的把柄”，它有一个在第三层节点中的独一无二的上级层次。相同的情况能从“钢笔”和“书”，“书桌”和“餐桌”类别中找到。正如这个例子中显示的那样，我们的框架不只能描述新的类别，也能表示跟学习的类别相关的关系。

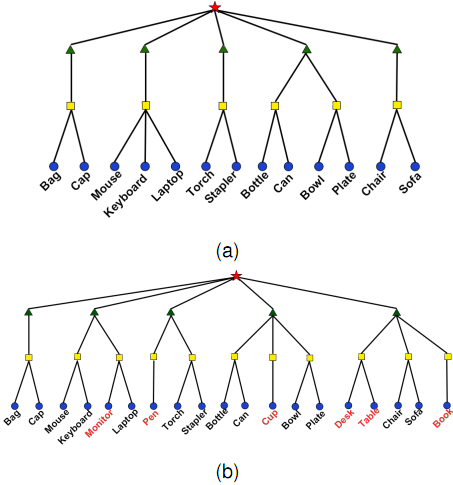


图4动态类别层次结构的例子。

新类别输入之后（(b)中的红色的单词），

1. 中旧的类别层次结构变成(b)中新的类别层次结构。
2. 注意大多数新类别能被插入到层次结构适当的位置中，除了书的类别。

VI． 结论

在我们的本次研究中，我们首先提出了从多模态传感器获取到的多模态属性。新的属性能更准确的描述对象。然后我们提出一种约束hLDA模型，更准确的构建类别层次结构。基于这两个，我们提出一种构建动态类别层次结构的新框架。通过这个框架，能获得在符合当前情况下的类别层次结构，并当这种情况改变的时候，能动态的改变他们的结构。因此属于先前没有被训练过的新类别的对象能准确的区分出来，并且对应的类别能准确的添加到当前的层次结构中。这些结论可以通过我们的扩展实验来证明。

因为我们提出的方法的优点，能应用到许多场景，例如机器人可以对未知环境进行探索。机器人遇到没有训练过的对象是无法避免的。因此，机器人，只能根据它们的经验去确定这些未知的对象。通过使用传统的对象识别技术，机器人将会把未知的对象分类到跟它们最相近的已知类别，这将导致错误并影响后续的任务。然而，通过我们的方法机器人将会推出未知的对象会属于新的类别并通过构建类别层次结构将新的类别跟已知的类别联系起来，这对于接下来的推理步骤是很有有用的，比如这些类别的功能和特征。