DOI: 10.16451/j.cnki.issn1003-6059.2010.02.009
第 23卷 第 2期 模式识别与人工智能 Vol 23 No 2
2010年 4月 PR & AI Apr 2010

图像语义分析与理解综述*

高 隽 谢 昭 张 骏 吴克伟

(合肥工业大学 计算机与信息学院 合肥 230009)

摘 要 语义分析是图像理解中高层认知的重点和难点,存在图像文本之间的语义鸿沟和文本描述多义性两大关键问题.以图像本体的语义化为核心,在归纳图像语义特征及上下文表示的基础上,全面阐述生成法、判别法和句法描述法 3种图像语义处理策略.总结语义词汇的客观基准和评价方法.最后指出图像语义理解的发展方向.

关键词 图像理解,语义鸿沟,语义一致性,语义评价

中图法分类号 TP 391.4

Image Semantic Analysis and Understanding: A Review

GAO Jun, XIE Zhao, ZHANG Jun, WU Ke-Wei

(School of Computer and Information, Hefei University of Technology, Hefei 230009)

ABSTRACT

Semantic analysis is the importance and difficulty of high-level interpretation in image understanding, in which there are two key issues of text-image semantic gap and text description polysemy. Concentrating on semantization of images ontology, three sophisticated methodologies are roundly reviewed as generative, discriminative and descriptive grammar on the basis of concluding images semantic features and context expression. The objective benchmark and evaluation for semantic vocabulary are induced as well. Finally, the summarized directions for further researches on semantics in image understanding are discussed intensively.

Key Words Image Understanding, Semantic Gap, Semantic Consistency, Semantic Evaluation

1 引 言

图像理解 (Image Understanding, IU)就是对图像的语义解释.它是以图像为对象,知识为核心,研究图像中何位置有何目标 (what is where)、目标场景之间的相互关系、图像是何场景以及如何应用场景的一门科学.图像理解输入的是数据,输出的是知识,属于图像研究领域的高层内容[1-3].语义 (Se-

* 国家自然科学基金资助项目 (No 60875012, 60905005)
收稿日期: 2009-12-21;修回日期: 2010-01-27

作者简介 **高隽**,男, 1963年生,教授,博士生导师,主要研究方向为图像理解、智能信息处理、光电信息处理等. E-mail: gaojun@hfut.edu.cn. **谢昭**,男, 1980年生,博士,讲师,主要研究方向为计算机视觉、智能信息处理、模式识别. **张骏**,女, 1984年生,博士研究生,主要研究方向为图像理解、认知视觉、机器学习. **吴克伟**,男, 1984年生,博士研究生,主要研究方向为图像理解、人工智能.

mantics)作为知识信息的基本描述载体,能将完整的图像内容转换成可直观理解的类文本语言表达,在图像理解中起着至关重要的作用.

图像理解中的语义分析在应用领域的潜力是巨大的.图像中丰富的语义知识可提供较精确的图像搜索引擎(Searching Engine),生成智能的数字图像相册和虚拟世界中的视觉场景描述.同时,在图像理解本体的研究中,可有效形成"数据 知识"的相互驱动体系,包含有意义的上下文(Context)信息和层状结构(Hierarchical-Structured)信息,能更快速、更准确地识别和检测出场景中的特定目标(如,识别出场景中的"显示器",根据场景语义知识可自动识别附近的"键盘").

尽管语义分析在图像理解中处于非常重要的位置,但传统的图像分析方法基本上全部回避了语义问题,仅针对纯粹的图像数据进行分析.究其原因主要集中于两方面:1)图像的视觉表达和语义之间很难建立合理关联,描述实体间产生巨大的语义鸿沟(Semantic Gap);2)语义本身具有表达的多义性和不确定性(Ambiguity).目前,越来越多的研究已开始关注上述"瓶颈",并致力于有效模型和方法以实现图像理解中的语义表达.

解决图像理解中的语义鸿沟需要建立图像和文本之间的对应关系,解决的思路可大致分为三类.第一条思路侧重于图像本身的研究,通过构建和图像内容相一致的模型或方法,将语义隐式地(Implicitly)融入其中,建立"文本→图像"的有向联系,核心在于如何将语义融于模型和方法中.采用此策略形成的研究成果多集中于生成(Generative)方式和判别(Discriminative)方式中.第二条思路从语义本身的句法(Grammar)表达和结构关系入手,分析其组成及相互关系,通过建立与之类似的图像视觉元素结构表达,将语义描述和分析方法显式地(Explicitly)植入包含句法关系的视觉图中,建立"图像→文本"的有向联系.核心在于如何构建符合语义规则的视觉关系图.第三条思路面向应用,以基于内容的图像检索(Image Retrieval)为核心,增加语义词汇规模,构建多语义多用户多进程的图像检索查询系统.

解决语义本身的多义性问题需要建立合理的描述规范和结构体系.Princeton大学的认知学者和语言学家早在20世纪80年代就研究构建了较合理统一的类树状结构.如今已被视为视觉图像研究领域公认的语义关系参考标准,用于大规模图像数据集的设计和标记中,有效归类统一了多义性词语.此外,一些客观的语义检索评价标准也在积极的探索过程中.

本文将对上述两个图像语义理解中的问题进行方法提炼和总结.针对语义鸿沟问题,介绍已有模型和方法的处理策略.还采用较完备的图像语义"标尺"(Benchmark)解决语义的主观多义性.

2 图像内容的语义分析

图像内容描述具有"像素 区域 目标 场景"的层次包含关系,而语义描述的本质就是采用合理的构词方式进行词汇编码(Encoding)和注解(Annotation)的过程.这种过程与图像内容的各层描述密切相关,图像像素和区域信息源于中低层数据驱动,根据结构型数据的相似特性对像素(区域)进行"标记"(Labeling),可为高层语义编码提供有效的低层实体对应关系.目标和场景的中层"分类"(Categorization)特性也具有明显的编码特性,每一类别均可视为简单的语义描述,为多语义分析的拓展提供较好的原型描述.本节将针对前述的语义鸿沟问题介绍常用的图像语义表示方法和分析策略.

2.1 语义化的图像特征

图像内容的语义分析借鉴文本分析策略.首先需要构建与之相对应的对象,整幅图像(Image)对应整篇文档(Document),而文档中的词汇(Lexicon)也需要对应相应的视觉词汇(VisualWord).视觉词汇的获取一般通过对图像信息的显著性分析提取图像的低层特征,低层特征大多从图像数据获取,包括简单的点线面特征和一些特殊的复杂特征,再由鲁棒的特征表达方式生成合适的视觉词汇,视觉词汇一般具有高重用性和若干不变特性.

点特征提取以图像中周围灰度变化剧烈的特征点或图像边界上高曲率的点为检测对象,根据灰度或滤波函数确定区域极值点(如Harris角点[4]等),并拓展至不同掩膜下的尺度空间中(如高斯 拉普拉斯、高斯差分等),分析极值点的稳定特性,得到仿射不变的Harris二阶矩描述符[5].线特征描述图像中目标区域的外表形状和轮廓特性,这类轮廓线特征以Canny算子等经典边缘检测算法为基础,集中解决边缘曲线的描述、编组以及组合表达等问题.边缘上的双切线点和高曲率点可连接形成有效的边缘链或圆弧,根据聚类策略或某些规则完成线片段编组,形成线特征的视觉词汇[6-8].区域是图像上具有灰度强相关性的像素集合,包含某种相似属性(如灰度值、纹理等),相对于点线特征,面特征有更丰富的结构信息.区域特征以点特征为中心,采用拉普

拉斯尺度下的 Harris或 Hessian仿射区域描述,对特征尺度上的椭圆仿射区域内的初始点集进行参数迭代估计,根据二阶矩矩阵的特征值测量点邻的仿射形状[4,9].另一种策略分析视觉显著区域对象 (如直方图、二值分割图等)的熵值统计特性,得到最佳尺度下的最稳定区域,满足视觉词汇的高重用性[10-11].

鲁棒特征表达对提取的特征进行量化表示.点特征一般仅具有图像坐标.线特征则充分考虑邻域边缘点的上下文形状特性,以边缘上采样点为圆心,在极坐标下计算落入等距等角间隔区域的边缘像素直方图.椭圆形面特征描述主要以尺度不变特征变换 (Scale Invariant Feature Transform, SIFT)[12-13]为主, SIFT特征对每个高斯窗口区域估计方向直方图,选择峰值作为参考方向基准,计算 4×4网格区域内 8个方向的梯度直方图,任何区域均可转换为 4×4×8=128维特征向量.该特征对图像尺度、旋转具有不变性,对亮度和视角改变也保持一定稳定性.通过对特征向量的聚类,得到最原始的特征词汇,形成的语义化图像特征也称为"码书"(Codebook)[14].

2.2　图像语义的上下文表达

图像的语义信息描述主要包含外观位置信息和上下文信息,前者如 2.1节所述,可表示成"码书".上下文信息不是从感兴趣的目标外观中直接产生,而来源于图像邻域及其标签注解,与其他目标的外观位置信息密切相关.当场景中目标外观的可视程度较低时,上下文信息就显得尤为重要.

Biederman将场景中不相关目标关系分为 5种,即支撑 (Support)、插入 (Interposition)、概率 (Probability)、位置 (Position)和大小 (Size)[15-16].五类关系均包含"知识",不需要知道目标信息就可确定支撑和插入关系,而后三类关系对应于场景中目标之间的语义交互关系,可缩短语义分析时间并消除目标歧义,通常称为"上下文特征"(Context Features),譬如一些相对复杂的特征描述 (如全局 Gist特征[17-18]、语义掩码特征等)融入场景上下文信息,本身就包含语义 (关联)信息,是语义分析的基础.如今有很多研究开始挖掘 Biederman提出的三类语义关系,可分为语义上下文、空间上下文和尺度上下文[19].

语义上下文表示目标出现在一些场景中,而没有出现在其他场景中的似然性,表示为与其他目标的共生 (Co-Occurrence)关系,可采用语义编码方式[20-21],也可由共生矩阵判断两类目标是否相关[22-23],此类上下文对应 Biederman关系中的"概率"关系.空间上下文表示目标相对于场景中其他目标出现在某个位置上的似然性,对应于"位置"关系.空间上下文隐式地对场景中目标的"共生"进行编码,为场景结构提供更加具体的信息,只需确定很少的目标,就可通过合理的目标空间关系降低目标识别的误差,消除图像中的语义歧义[24-25].尺度上下文表示目标在场景中可能的相对尺度范围,对应于"大小"关系.尺度上下文需处理目标之间的特定空间和深度关系,可缩小多尺度搜索空间,仅关注目标可能出现的尺度.尺度上下文在二维图像中较为复杂,目前仅用于简单的视觉分析系统中[26-27].

目前大多数上下文方法主要分析图像中的语义上下文和空间上下文.语义上下文可从其他两种上下文中推理获取,与场景中的目标共生相比.尺度和空间上下文的变化范围较大,而共生关系的知识更易获取,处理计算速度更快.融入上下文特征的图像语义形成了全局和局部两种分析策略,即基于场景的上下文分析和基于目标的上下文分析.前者从场景出发[15,27],将图像统计量看作整体,分析目标和场景之间的高频统计特性,获取全局上下文信息,如马路预示着汽车的出现.后者从目标出发[25,28],分析目标间的高频统计特性,获取局部上下文信息,如电脑预示着键盘的出现.总之,上下文特征包含了更丰富的知识,有助于为图像理解提供更准确的语义信息.

2.3　语义分析的生成方法

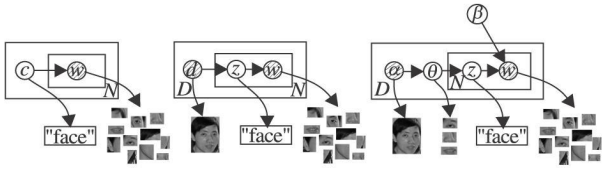
生成方法基于模型驱动,以概率统计模型和随机场理论为核心,遵循经典的贝叶斯理论,定义模型集合 M,观察数据集合 D,通过贝叶斯公式,其模型后验概率 p(M|D)可以转换为先验概率 p(M)和似然概率 p(D|M)的乘积.生成方法一般假设模型遵循固定的概率先验分布 (如高斯分布等),其核心从已训练的模型中"生成"观察数据,测试过程通过最大似然概率 (Maximize Likelihood)得到最符合观察数据分布的模型预测似然 (Predictive Likelihood).

图像语义分析的生成方法直接借用文本语义分析的图模型结构 (Graphical Models),每个节点定义某种概念,节点之间的边表示概念间的条件依赖关系,在隐空间 (Latent Space)或随机场 (Random Field)中建立文本词组和视觉描述之间的关联,生成方法无监督性明显,具有较强的语义延展性.

2.3.1　层状贝叶斯模型

图模型的节点之间由有 (无)向边连接,建立视觉词汇和语义词语之间的对应关系.朴素贝叶斯理论形成的经典 Bags-of-Words模型是层状贝叶斯模

型的雏形,该模型将同属某类语义的视觉词汇视为“包”,其图结构模型和对应的视觉关系描述如图 1(a)所示,其中灰色节点为观察变量,白色节点为隐变量, N为视觉词汇的个数,通过训练建立类别语义描述 c和特征词汇 w之间的概率关系,选取最大后验概率 p(c | w)对应的类别作为最终识别结果.



(a)朴素贝叶斯 (b)概率隐语义分析 (c)隐狄利克雷分配

(a)Naïve bayes

(b)Probabilistic latent semantic analysis

(c) Latent Dirichlet allocation

图 1 有向图语义描述

Fig 1 Semantic interpretation of directed graphs

朴素贝叶斯模型试图直接建立图像和语义之间的联系,但由于视觉目标和场景的多样性导致这种稀疏的离散分布很难捕捉有效的概率分布规律,因此 Hofmann借鉴文本分析中的概率隐语义分析(Probabilistic Latent Semantic Analysis pLSA)模型[29-30],将“语义”描述放入隐空间 Z中,生成相应的“话题”(Topic)节点,其基本描述如图 1(b)所示. D为 M个图像 d组成的集合, z表示目标的概念类别(称为“Topics”),每幅图像由 K个 Topics向量凸组合而成,通过最大似然估计进行参数迭代,似然函数为 p(w | d)的指数形式,与语义词汇和图像的频率相关. 模型由期望最大化(Expectation Maximization, EM)算法交替执行 E过程(计算隐变量后验概率期望)和 M过程(参数迭代最大化似然).

决策过程的隐变量语义归属满足

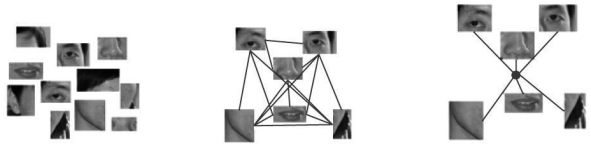
$$z^{*} = \arg\max_{z} P(z \mid d),$$

pLSA模型通过隐变量建立特征与图像间的对应关系,每个文本单元由若干个语义概念按比例组合,本质上隐空间内的语义分布仍然是稀疏的离散分布,很难满足统计的充分条件. 隐狄利克雷分配(Latent Dirichlet Allocation, LDA)模型[31-32]在此基础上引入参数 θ建立隐变量 z的概率分布.在图像语义分析中,变量 z反映词汇集合在隐空间的聚类信息,即隐语义概念,参数 θ(通常标记为 π)则描述隐语义概念在图像空间中的分布,超参 α(通常标记为 c)一般视为图像集合 D中已知的场景语义描述.如图 1(c)所示,由参数估计和变分(Variational)推理,选取

$$c = \arg\max_{c} P(w \mid c, \pi, \beta)$$

作为最终结果.

LDA中不同图像场景以不同的比例 π重用并组合隐话题空间全局聚类(Global Cluster),形成“场景 目标 部分”的语义表达关系. LDA中的隐话题聚类满足 De Finetti可交换原理,其后验分布不受参数次序影响,不同隐话题聚类相互独立,无明显的结构特性.一种显而易见的策略就是在此模型基础上融入几何或空间关系,即同时采用话题对应的语义化特征的外观描述和位置信息,这样不同话题的分布大体被限定于图像场景的某个区域,如天空总是出现在场景的上方等,减小模型决策干扰.如 Li等人[14, 33]在 LDA模型中融入词汇的外观和位置信息,并将语义词汇描述 c划分为视觉描述词汇(如 sky)和非视觉描述词汇(如 wind)两类,由词汇类别转换标签自动筛选合适的词汇描述.模型采用取样(Sampling)策略对从超参先验中生成的视觉词汇和语义标签进行后验概率学习,模型中包含位置信息的语义特征显式地体现了空间约束关系,具有更好的分析效果.



(a) 无结构 (b) 全互连结构 (c) 星状结构

(a)Unstructured (b)Full structure (c)Star structure

图 2 Part-based模型表示图

Fig 2 Representation for Part-based models

LDA模型已明确地将隐空间的“话题”语义进行合理聚类,建立与视觉词汇聚类的对应关系.隐话题聚类隐式地对应场景或目标的某些部分(parts),是一种较原始的 part-based模型.真正的 part-based模型侧重“目标 部分”之间的语义关联表达,不仅具有较强的结构特性,而且直接概念化隐空间的语义聚类,每个 part直接显式对应语义描述(如人脸可分为眼睛、鼻子、嘴等不同部分).如图 2所示,一般通过人工设定或交叉验证的方式固定重要参数(如隐聚类个数、part个数等)并混合其概率密度,其中固定参数的 Dirichlet生成过程是一种有限混合.“星群”(Constellation)模型[34-35]是其中的典型,根据不

同区域的外观位置信息描述,确定 P个部分的归属及其概率分布,将目标和背景似然比分解为外观项、形状项、尺度项以及杂项的乘积,依次计算概率密度值(一般是高斯分布或均匀分布),并 EM迭代更新参数,最后通过似然比值判断目标的语义属性.部分间的约束关系体现于形状项中,可以假设为全互连结构(Full Structure)或星状结构(Star Structure),其结构信息体现于高斯分布的协方差矩阵中(满秩或稀疏矩阵),有助于提高语义分析的准确性.

固定参数的 Dirichlet生成过程是无限混合模型的一种特例,可通过合适的随机过程,很好表达无限混合(Infinite Mixture)模型,自动确定混合个数.这种"非参"(Non-Parametric)模型可捕捉到概率空间的隐性分布,不受特定的概率密度函数形式表达限制.整个 Dirichlet过程可拓展至层次结构(Hierarchical Dirichlet Process HDP).HDP具有明显的结构特性,可以很容易对应于图像中的"场景-目标-部分"层次结构,其混合组成很显式地表达了不同目标实体间的语义包含关系. Sudderth在 HDP的基础上,引入转换函数(Transformed Function),生成转换 Dirichlet过程(Transformed Dirichlet Process TDP),每组的局部聚类不再直接"复制"全局聚类参数,而是通过不同转换函数生成变化多样的局部变参,更符合目标多变特性[36-37].

层状贝叶斯模型是当前处理图像语义问题的关注热点,其模型特有的参数化层次结构信息参照文本处理直接对应图像中的语义实体,通过图模型的参数估计和概率推理得到合适的语义描述.模型本身的发展也具有一定的递进关系,即"Bags-of-Word模型→pLSA模型→LDA模型→part-based模型→HDP模型→TDP模型"等,分析得到的结果具有层次语义包含关系.

2.3.2　随机场模型

随机场模型以均值场(Mean Field)理论为基础,图中节点变量集合 $\{x_i \mid i \in V\}$通常呈 4-邻域网格状分布,节点之间的边 $\{(x_i, x_j) \mid i\ j \in V; (x_i, x_j) \in E\}$体现隐性关联,由势函数 $\psi_{ij}(x_i, x_j)$表示,一般具有含参数 θ的近高斯指数分布形式,每个隐节点 x_i一般对应一个观察变量节点 y_i,由势函数 $\psi_i(x_i, y_i)$表示.如图 3所示,观察节点可对应图像的像素点,也可对应图像中的某个区域或目标语义化特征描述(如 2.1节所述),隐变量则对应语义"标记"或"标签"[1].

随机场模型具有丰富的结构场信息,节点间上下文关联很强,通常分析像素标记解决图像分割问题.近年来,其特定的约束关系(如桌子和椅子经常关联出现)也被用于图像区域化语义分析中,隐节点集的语义标签对应不同的语义化特征和势函数取值,最大化随机场的能量函数得到的标记赋值,就是最终的区域语义标记属性.随机场模型具有较成熟的计算框架,融合其上下文关联信息的层次贝叶斯"生成"模型是分析图像语义的主流趋势[14, 33-35, 38-40].

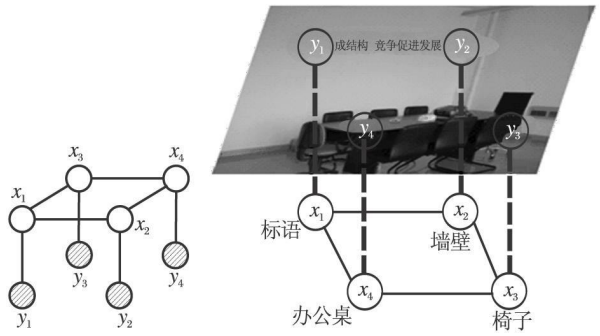


图 3　随机场模型及其图像语义描述

Fig 3　Random field model and its semantic description

2.4　语义分析的判别方法

判别方法基于数据驱动,根据已知观察样本直接学习后验概率 $p(M \mid D)$,主要通过对训练样本的(弱)监督学习,在样本空间产生合适的区分函数,采用形成的分类器或结构参数,完成对特定的特征空间中点的划分(或闭包),形成某些具有相似特性的点的集合.这些共性可直接显式对应图像理解中的若干语义信息,如目标和场景的属性、类别信息等,通常以主观形式体现于观察样本中,其本质就在于学习并获取区分不同语义信息的知识规则(如分类器等).由于语义信息主观设定(如判别几种指定类别),因此判别方法主要侧重观察样本(语义)的处理分析,而非观察样本(语义)的获取.判别方法是包含经典的机器学习方法,精确度较高且易于实现,常用于目标检测识别识别.其策略主要包括最近邻分析、集成学习和核方法.

2.4.1　最近邻方法

最近邻(k-Nearest Neighbor kNN)方法是基于样本间距离的一种分类方法.其基本思想是在任意空间中、某种距离测度下,寻找和观测点距离最接近的集合,赋予和集合元素相似的属性集合.在图像理解中,就是在图像特征空间寻找和近似的特征描述集,将已知的语义作为分析图像的最终结果.最近邻方法非常简单,但对样本要求较高,需要很多先验知

识,随着大规模语义标记图像库的出现 (如后 3.2节所述),最近邻方法有了广阔的应用前景, Torralba等人[41]建立 80万幅低分辨率彩色图像集合和相应的语义标记,图像集涵盖所有的视觉目标类别,以WordNet语义结构树 (如后 3.1节所述)的最短距离为度量,采用最近邻方法分别对其枝干进行投票,选取最多票数对应最终的语义标签输出.也可直接在图像空间中计算像素点的欧式距离,得到与分析图像相类似的语义空间布局 (Configuration). Russell等人[42]利用最近邻方法找出与输入图像相似的检索集,通过含有标记信息的检索图像知识转化到输入图像中,完成场景到目标的对齐任务.语义聚类法还被用于视频数据库中[43],具有较好的结果.

2.4.2 集成学习

集成学习将各种方法获得的模型在累加模型下形成一个对自然模型的近似[44-45],将单一学习器解决问题的思想转换为用多个学习器来共同解决问题. Boosting是集成学习方法的典型.其基本思想是每次迭代 t生成一个带权重 α_t的弱分类器 (Weaker Classifier) h_t,加大误分样本的权重,保证后续学习对此类样本的持续关注,权重 α_t表示该弱分类器 h_t的重要性,分类效果好的权重大,效果差的权重小.其集成学习的结果就是弱分类器的加权组合

$$\sum_{t=1}^{T}\sum_{x_i\in D}\alpha_t h_t(x_i)$$

构成一个分类能力很强的强分类器 (Strong Classifier),完成简单的二值或复杂的多值分类[46-47].

集成学习方法经常用于图像理解的语义分类中,其样本数据集既可以是区域块也可以是滤波后的基元乃至包括上下文和空间布局信息.其分类结果具有很明显的语义区分度.多语义分类中经常出现多类共享的情况,因此,联合 Boosting的提出极大地减少了分类器的最佳参数搜索时间,使单一弱学习器具有多类判别能力[48-51].同时,近年来多标签多实例 (Multi-Instance Multi-Label Leaming MIML)的集成学习策略[52]也倍受学者关注,图像理解中的语义划分问题可通过 MIML转化为单纯数据下的机器学习问题,其输出的分类结果就是对既定语义的编码结果.

2.4.3 核方法

核方法 (Kernel)是在数据集中寻找合适的共性"基",由"基"的混合组成共性空间,与图像理解中的低层基元表示异曲同工.使用核方法可将低维输入空间 R^n样本特征映射到高维空间中 H,即 $\Phi: R^n \to H$,将非线性问题转换为线性问题.其关键是找到合适的核函数 K保持样本在不同空间下的区分关系,即 $K(x_i, x_j) = \Phi(x_i)\cdot\Phi(x_j)$.它能够在学习框架和特定知识之间建立一种自然的分离来完成图像有意义的表达[53-54].

支持向量机 (SVM)是常用的核方法之一.它以训练误差作为优化问题的约束条件,以置信范围值最小化作为优化目标,在核函数特征空间中有效训练线性学习分类器,通过确定最优超平面 (Hyper Plane)及判别函数完成高维空间点的分类. SVM方法在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,在图像理解中,能有效解决不同环境、姿态以及视角下的广义目标识别分类问题,是目前最为通用的分类模型[55-58].针对多语义分类问题, Farhadi等人[59]将目标的语义属性细分为部分、形状及材质等,相同或相似的语义对应的样本集表明了某种特有的共性关系,采用 L1测度对数回归和线性 SVM方法学习不同语义类别的判别属性,其多语义属性的不同划分决定了指定目标的唯一描述,具有很强的语义可拓展性.

判别模型是通过模型推理学习得出的后验概率,对应不同类别目标的后验概率或对应图像前景和背景的不同后验概率来划定判决边界,进而完成目标识别,指导图像理解.判别模型在特征选取方面灵活度很高,可较快得出判别边界.

2.5 图像句法描述与分析

人对图像场景理解的本质就是对图像本身内在句法 (Grammar)的分析.句法源于对语句结构研究,通过一系列的产生式规则将语句划分为相互关联的若干词汇 (组)组合,体现句法内词汇之间的约束关系.图像句法分析直接研究图像语义,随着 20世纪70年代句法模式识别的提出, Otha就试图构建统一的基于视觉描述的知识库系统,利用人工智能相关策略进行场景语义推理.但由于视觉模型千变万化,方法针对性很强,句法分析方法曾一度没落.当前图像语义分析的一部分研究重心又重新转向图像句法.由于句法分析本身已较为成熟,因此如何建立和句法描述相对应的图像视觉描述非常关键.

2.5.1 图像与或图表达

图像 I内的实体具有一定的层次结构,可用与或图 (And-Or Graph)的树状结构表示,即解析树 pg,如图 4所示,同属一个语义概念的实体尽管在外观上具有很大差异,但与或图表达相似,与节点表示实体的分解 (Decomposition),如"场景 → 目标","目标 → 部分"等,遵循 A → BCD…的句法规则,或节点表示可供选择的结构组成,遵循 A → B | C | D…

的句法规则.同层节点间的水平连接虚线表示视觉实体间的上下文关系 R,连接包含如图 5所示的 3种类型:1)基元连接,即原始的点和线遵循一定的连接方式构成更高层的基元 (Texton);2)关节连接,即更高层的部分之间的连接方式;3)目标相互作用关系,即目标实体之间的相互关系,包含空间关系和功能关系.图像与或图的表达统一了图像中的语义规范,一般分为四层,即"场景 -目标 -部分 -基元",任何场景都可用与或图表示,每层均包含点线面的视觉词汇Δ既具有语义属性,又体现实体间的语义关系.

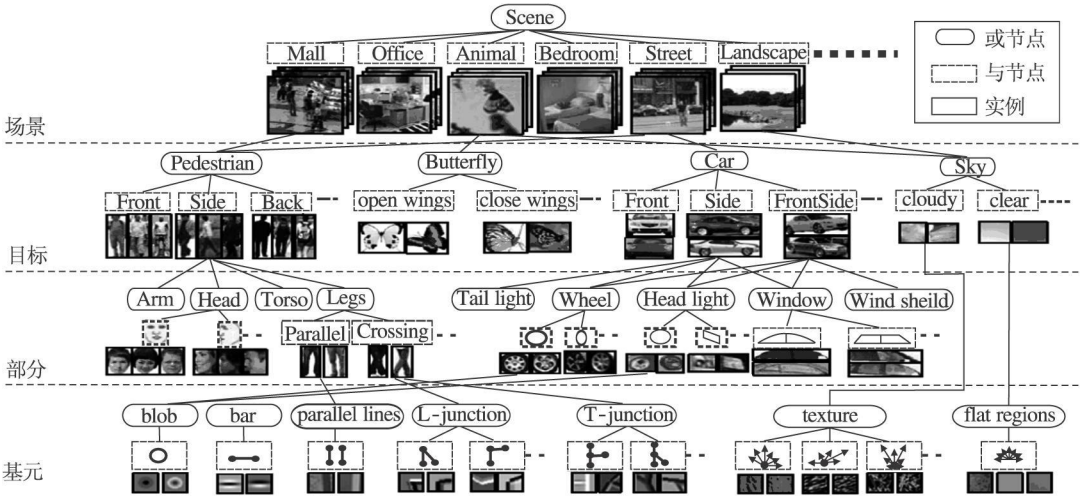
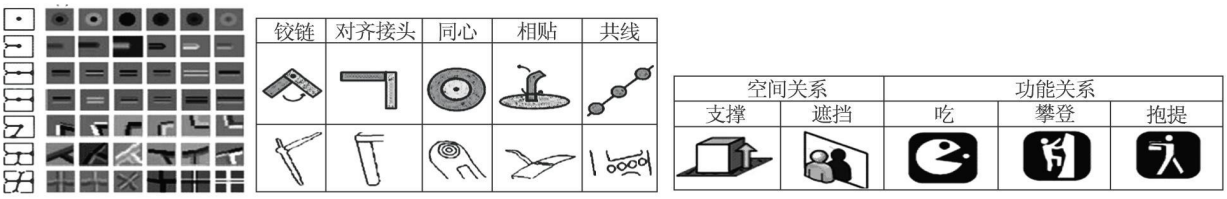


图 4　图像与或图的层次表示

Fig 4　Hierarchical representation of and-or graph



(a)基元间关系　(b)部分间关系　(c)目标间关系

(a)Relationship among textons　(b)Relationship among parts　(c) Relationship among objects

图 5　图像实体结构关系

Fig 5　Structural relationship of image entities

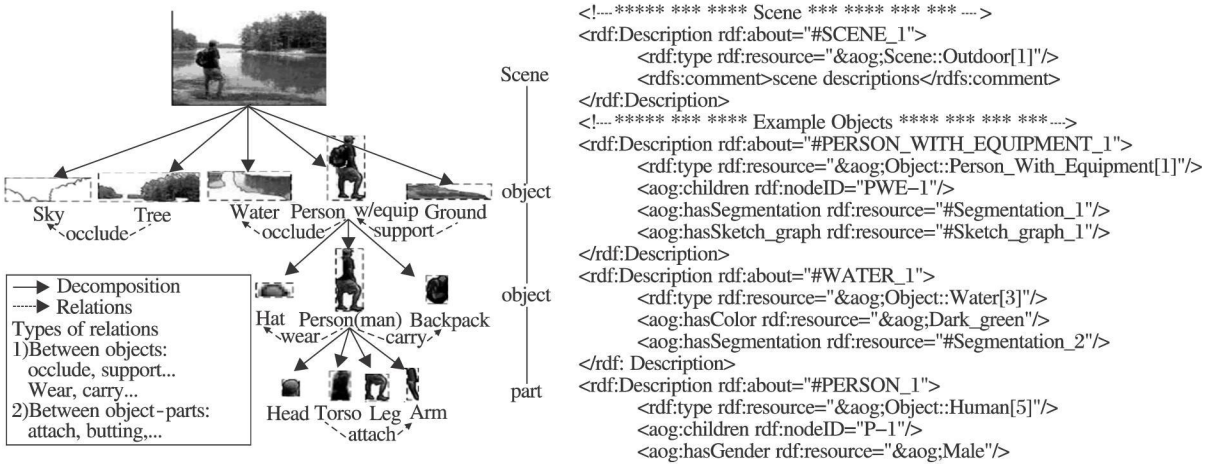


图 6　图像与或图与其 RDF格式的本体语言表达

Fig 6　And-or graph and its ontology language expression in RDF format

2.5.2 句法学习与推理

与或图的句法学习推理以贝叶斯统计概率为基础.学习过程由最大似然估计指导并遵循最小最大熵学习机制,从指定集合中根据目标分布 f取样观察样本集

$D^{obs}=\{(I_i^{obs}, pg_i^{obs}): i=1,2,\cdots,N\}\sim f(I,pg)$,

解析图 pg_i^{obs}由真实数据集 Ω_{pg}获取,概率模型 p通过最小化 KL距离向目标概率 f逼近.该过程等价于对词汇 Δ、关联 R和参数 Θ的最优最大似然估计.其学习过程主要分三步:1)对给定的 R和 Δ,从训练观测样本 D^{obs}中估计参数 Θ;2)在与或图 pg中给定 Δ,学习和捕捉关联集合 R;3)自动发现并绑定词汇 Δ和层次与或树.

与或图句法推理通常采用自顶向下(Top-Down)和自底向上(Bottom-Up)策略[60-61].在图像句法推理分析中,可通过整合 bottom-up的检测结果和 top-down的假设找到最好的结构表达方式.由于与或图是递归定义的,因此推理算法也具有递归性质.这种性质在分析大量目标类别时可采用启发式搜索,大大简化算法的复杂度.对输入图像 I,最大化后验概率计算解析图 pg得到最终结果.

2.5.3 图像句法的语义表示

基于 Web本体语言(Web Ontology Language OWL)的语义网络技术的发展为图像与或图的句法表示提供丰富的结构化语义描述,不同本体间具有显式的知识依赖关系,不同的 OWL文本通过本体映射相连,具有高度的重用性.借鉴 OWL的思想,与或图通常转换为 RDF的格式,如图 6所示.图像句法的结构化语义表示为实现"图像→文本"的规范化输出成为可能,但目前的研究仅处于初始阶段.

3 图像语义标准化描述

虽然语义是图像理解最重要的高层知识,但至今尚未形成一套公认的有效的研究体系.如何描述图像数据信息与知识语义信息间的对应关系,是图像理解中层次间相互衔接的关键所在(如第 2节所述).同时,由于语义表达本身的歧义性和不确定性,会产生大量多变的词汇短语描述相同的目标类别,如"car"目标类可用"car frontal"、"automobile"、"taxi"等词汇描述等,同时还可能需要扩展注解,合并上层的归属类别,如目标标签"car"、"person"、"tree"可能合并至归属类别"vehicle"、"animal"、"plant"中.这种语义词汇表述的广泛性受主观因素影响较大,给目标类别的标记过程带来极大的困难,因此需要一个客观的描述标准.从实体论的角度,在对人类心理科学和认知科学分析的基础上,可采用词汇语义网面向知识进行语义建模,WordNet是最常用的一个英文语义网.

3.1 WordNet图像语义结构

WordNet是 Princeton大学的心理学家、语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典[62].WordNet根据词汇含义组成"关系网络"用于描述词汇间的关系.WordNet包含语义信息,区别于一般意义上的字典,通常按照词汇含义进行分组,将具有相同或相似含义的词汇按词性分别组成"同义词集合"(Synset)网络,每个同义词通过感知语义和词汇关联相连.WordNet中词语之间的语义关联将词语组成语义层次树,层次树是唯一的,反映了一定的客观标准特性.

WordNet为图像理解中的语义提供统一的描述基准和关系结构,近年来已有不少学者将其用于图像理解的语义描述中.WordNet层次结构的语义关系可方便语义搜索和分类,选择出与查询图像相匹配的词汇表达.同时,每个可观察的视觉词汇在WordNet中特定节点上初始化搜索,也可转化为修剪(Pruning)或属性(Attribute)项,约束语义搜索.例如,目标(Object)是实体(Entity)的下属级,自然物、人造物和生物是目标的下属级,如果需要找到对人(person)这个词汇更细致的描述,就可通过 WordNet结构直接搜索出 person的子节点确定下属词,得到图像或视频中特定目标的属性概念(如不同的动作等).此外,WordNet中不同词汇间的距离反应词汇间的融合程度.距离较短的词语,其描述的相似度越接近,共现的几率较大,其图像中的上下文语义关系也越强,通过定义合适的匹配函数或距离代价函数[63],并最优化分析得到符合 WordNet构词规则的最佳的语义输出,每个单词都有形式自由的、与字典相似的原文定义,这些定义可存储为字符串,而不需要任何索引或结构化的内容.单词可是名词,也可是动词或形容词,常用于图像检索[64]、运动行为检测[65]、场景分类[41]、边缘轮廓检测[63]、目标识别[41,65]中,这些都包含图像理解的语义分析内容,采用 WordNet不仅可消除语义的歧义也可提高理解的精度.

图像语义表达类似于人类的视觉认知.由于人类获取的视觉信息非常庞大,因此对计算机而言,获取视觉信息源(即图像数据集)并建立合适的语义结构描述非常重要、能否构建较为完整、通用的训练集合已成为图像理解能否拓展泛化的核心.

3.2 WordNet语义标准图像集

WordNet网提供标准的语义描述方式,在图像语义理解中应包含符合 WordNet结构和语法规则的先验知识,即图像中包含的固有的语义信息.当前的若干图像库已达到此标准,符合 WordNet的类层次结构,具有广泛的目标类别信息和关联信息.图像中的语义信息较为广泛,准确度高.下述图像库均包含大规模的语义信息.

小规模图像集是语义标记图像库的雏形,包括 Caltech101/256[66], MSRC[67], PASCAL[68]等在内的图像库已经成为目标识别算法评价的基准库,包含上百类的不同表现形式的目标,具有简单的目标语义特性. TinyImage[41]包含从网络中搜集的 80万幅 32×32的低分辨率图像,每个 WordNet词汇对应近千幅图像,由于图像尺寸小,因此图像间相似度的计算速度很快,可发现海量图像数据中的语义关系. ESP数据集[69]源于网络标记游戏,图像库中的所有语义标记均取自标记者,通过对相同图像不同标记的分析,可发现不同层次的图像语义对应的认知关系,如归属关系、相似关系等.

LabeMe和 Lotus Hill数据集[70-71]分别有 3万和 5万个标记分割图像,包含近 200个类别,可以采用在线交互方式在图像场景中勾画目标轮廓并进行标记,具有查询功能,包含近 3 000个 WordNet词汇. ImageNet[72]完全依照 WordNet的结构构建的巨型图像库集合,约 3.2万幅图像,包含 5 247个 WordNet词汇和 12个语义关系子树,其图像库中的语义关系层次深度已达到 12级,是目前最能反映图像语义关系特性的图像集.

3.3 语义评价

语义的客观评价是衡量算法优劣的重要过程.传统方法一般针对有限语义类别进行查全率/查准率评价,判断场景中的目标出现与否,两个评价指标形成的查全率/查准率曲线 (Recall-Precision Curve RPC)一般作为基本的评价对象.也可采用接受者操作特性曲线 (Receiver Operating Characteristic Curve ROC)描述.客观评价基准可通过计算曲线下方的区域面积占总面积 (即单位 1)的比例得到曲线下面积 (Area Under Curve AUC)值. AUC值越大则全部数据的错误接受率较低,算法性能越好.也可通过计算曲线与过 (1,0)、(0,1)直线的交点坐标作为平均错误率 (Equivalence Error Rate EER).交点越靠近原点则全部数据的错误接受率较低,算法性能越好.

在图像检索中依据不同语义词汇查询关键词所搜索的图像结果,可依据图像语义标记的排序结果进行投票评价,也可计算图像中原型向量 (Prototype)和真实语义场景之间的距离得到图像级的语义标记评价结果[73]. Barnard等人[74]则在更细的区域层给出语义分析的评价标准,同时需兼顾区域分割效果.区域标记的词汇来源于 WordNet的中层表示,区域标记遵循严格的规则,可处理区域多标记和空标记的情况.根据 WordNet中不同词汇关系 (包含、同义等)形成的有向图进行广度优先搜索 (Breadth-First-Searching)得到最短距离的 BFS树.考虑包含词汇的区域面积比例、词汇出现频率以及词汇关联个数,将其相乘作为 BFS树中词汇边的权重,此时语义评价标准就转化为当前得到的语义标记和真实语义标记间最短距离的权重之和.王煜等人[75]针对表达能力 (如对象、时间、它们的属性及相互关系)、语义信息获取能力及查询服务能力 (如提供查询语言、根据对象、事件、属性等进行查询)3个方面,建立 22条评价丰富语义模型的准则.

4 结束语

图像语义是图像理解中高层认知的核心对象,其表达方式极具挑战,在经典的 Marr视觉理论中,图像语义是三维世界场景描述的最终输出对象.本文针对图像理解中的语义鸿沟和图像语义多义性、歧义性两个关键问题,全面综述当前已有的研究方法和策略,从发展趋势上看,图像理解语义分析逐渐倾向体系化、规范化、应用化研究,具有广泛的理论和应用前景.

本文主要总结二维图像的图像语义理解,而更多维的图像集合 (如视频图像、三维立体图像等)则包含更丰富的语义关联知识,可充分提取有价值的信息实现语义化的图像理解,拓展图像理解的研究领域,目前已经成为图像语义分析的发展方向和研究热点.

参 考 文 献

[1] Gao Jun, Xie Zhao. Image Understanding: Theory and Approach. Beijing, China: Science Press, 2009. (in Chinese)
(高 隽,谢 昭.图像理解理论与方法.北京:科学出版社, 2009)

[2] Xie Zhao, Gao Jun. A Novel Method for Scene Categorization with Constraint Mechanism Based on Gaussian Statistical Model. Acta Electronica Sinica, 2009, 37(4): 733-738. (in Chinese)
(谢 昭,高隽.基于高斯统计模型的场景分类及约束机制新方法.电子学报, 2009, 37(4): 733-738)

[3] Zhang Yujin. Contented-Based Visual Information Retrieval. Bei-

jing China: Science Press, 2003 (in Chinese)
(章毓晋.基于内容的视觉信息检索.北京:科学出版社, 2003)

[4] Moravec H P. Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover. Technical Report CMU-RI-TR-80-03, Pittsburgh, USA: Carnegie Mellon University, Robotics Institute, 1980

[5] Mikolajczyk K, Schmid C. Scale & Affine Invariant Interest Point Detectors. International Journal of Computer Vision, 2004, 60(1): 63—86

[6] Rothwell C A, Zisserman A, Forsythe D A, et al. Planar Object Recognition Using Projective Shape Representation. International Journal of Computer Vision, 1995, 16(1): 57—99

[7] Nelson R C, Selinger A. A Cubist Approach to Object Recognition // Proc of the IEEE International Conference on Computer Vision. Bombay, India, 1998: 614—621

[8] Jurie F, Schmid C. Scale-Invariant Shape Features for Recognition of Object Categories // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004, Ⅱ: 90—96

[9] Mikolajczyk K, Schmid C. Indexing Based on Scale Invariant Interest Points // Proc of the 8th International Conference on Computer Vision. Vancouver, Canada, 2001: 525—531

[10] Xie Zhao. Researches for Key Issues and Methods in Image Understanding. Ph.D Dissertation. Hefei, China: Hefei University of Technology, School of Computer and Information, 2007 (in Chinese)
(谢 昭.图像理解的关键问题和方法研究.博士学位论文.合肥:合肥工业大学.计算机与信息学院, 2007)

[11] Xie Zhao, Gao Jun. Object Localization Based on Visual Statistical Probabilistic Models. Journal of Image and Graphics, 2007, 12(7): 1234—1242 (in Chinese)
(谢 昭,高 隽.基于视觉统计概率模型的目标定位.中国图象图形学报, 2007, 12(7): 1234—1242)

[12] Lowe D G. Object Recognition from Local Scale Invariant Features // Proc of the IEEE International Conference on Computer Vision. Kerkyra, Greece, 1999, Ⅱ: 1150—1157

[13] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004, 60(2): 91—110

[14] Li Feifei, Perona P. A Bayesian Hierarchical Model for Learning Natural Scene Categories // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2001, Ⅱ: 524—531

[15] Biederman I. On the Semantics of a Glance at a Scene // Kubovy M, Pomerantz J R, eds. Perceptual Organization. Hillsdale, USA: Lawrence Erlbaum, 1981: 234—242

[16] Biederman I, Mezzanotte R J, Rabinowitz J C. Scene Perception: Detecting and Judging Objects Undergoing Relational Violations. Cognitive Psychology, 1982, 14(2): 143—177

[17] Oliva A, Torralba A. Modeling the Shape of the Scene a Holistic Representation of the Spatial Envelope. International Journal in Computer Vision, 2001, 42(3): 145—175

[18] Oliva A, Torralba A. Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research: Visual Perception, 2006, 155: 23—36

[19] Galleguillos C, Belongie S. Context Based Object Categorization: A Critical Survey. Technical Report UCSD CS2008-0928, San Diego, USA: University of California, Department of Computer Science and Engineering, 2008

[20] Wolf L, Bileschi S. A Critical View of Context. International Journal of Computer Vision, 2006, 69(2): 251—261

[21] Dietterich T G, Bakiri G. Solving Multiclass Learning Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Research, 1995, 2(1): 263—286

[22] Rabinovich A, Vedaldi A, Galleguillos C, et al. Objects in Context // Proc of the IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1—8

[23] Torralba A. Contextual Priming for Object Detection. International Journal of Computer Vision, 2003, 53(2): 169—191

[24] Bar M, Ullman S. Spatial Context in Recognition. Technical Report CS93-22, Rehovot, Israel: Weizmann Institute of Science, Department of Applied Mathematics & Computer Science, 1993

[25] Singhal A, Luo Jiebo, Zhu Weiyu. Probabilistic Spatial Context Models for Scene Content Understanding // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA, 2003, Ⅰ: 235—241

[26] Strat T M, Fischler M A. Context-Based Vision: Recognizing Objects Using Information from Both 2D and 3D Imagery. IEEE Trans on Pattern Analysis and Machine Intelligence, 1991, 13(10): 1050—1065

[27] Torralba A, Murphy K, Freeman W T. Contextual Models for Object Detection Using Boosted Random Fields // Saul L K, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2004, XIII: 1401—1408

[28] Galleguillos C, Rabinovich A, Belongie S. Object Categorization Using Co-Occurrence, Location and Appearance // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1—8

[29] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning, 2001, 42(1/2): 177—196

[30] Hofmann T. Probabilistic Latent Semantic Indexing // Proc of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Netherlands, 1999: 35—44

[31] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 2003, 3: 993—1022

[32] Shi Jing, Hu Ming, Shi Xin, et al. Text Segmentation Based on Model LDA. Chinese Journal of Computers, 2008, 31(10): 1865—1873 (in Chinese)
(石 晶,胡 明,石 鑫,等.基于LDA模型的文本分割.计算机学报, 2008, 31(10): 1865—1873)

[33] Li Feifei, Forgus R, Perona P. A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories // Proc of the 9th International Conference on Computer Vision. Nice, France, 2003, 1134—1141

[34] Fergus R. Visual Object Category Recognition. Ph.D Dissertation. Oxford, UK: Oxford University, Department of Engineering Sci-

ence. 2005

[35] Li Feifei, Fergus R., Perona P. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories // Proc of the Workshop on Generative-Model Based Vision in Computer Vision and Pattern Recognition. Washington, USA, 2004: 59—70

[36] Sudderth E B., Torralba A., Freeman W T., et al. Describing Visual Scenes Using Transformed Objects and Parts. International Journal of Computer Vision, 2007, 77(1/2/3): 291—330

[37] Sudderth E B. Graphical Models for Visual Object Recognition and Tracking. Ph. D Dissertation. Cambridge, USA: Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2006

[38] Verbeek J., Triggs B. Region Classification with Markov Field Aspect Models // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1367—1373

[39] Li Lijia, Socher R., Li Feifei. Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 2036—2043

[40] Xie Zhao, Gao Jun, Wu Xindong. Regional Category Parsing in Undirected Graphical Models. Pattern Recognition Letters, 2009, 30(14): 1264—1272

[41] Torralba A., Fergus R., Freeman W T. 80 Million Tiny Images: A Large Dataset for Non-Parametric Object and Scene Recognition. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1958—1970

[42] Russell B C., Torralba A., Liu Ce, et al. Object Recognition by Scene Alignment // Proc of the Conference on Neural Information Processing Systems. Vancouver, Canada, 2007: 1—8

[43] Shi Zhiping, Hu Hong, Li Qingyong, et al. Cluster-Based Index Method for Video Database. Chinese Journal of Computers, 2007, 30(3): 397—404 (in Chinese)
(施智平,胡 宏,李清勇,等.视频数据库的聚类索引方法.计算机学报, 2007, 30(3): 397—404)

[44] Wang Jue. Machine Learning and Application. Beijing, China: Tsinghua University Press, 2006 (in Chinese)
(王 珏.机器学习及其应用.北京:清华大学出版社, 2006)

[45] Dietterich T G. Ensemble Learning // Arbib M A., ed. Handbook of Brain Theory and Neural Networks. 2nd Edition. Cambridge, USA: MIT Press, 2002

[46] Freund Y., Schapire R. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. Journal of Computer and System Sciences, 1997, 55(1): 119—139

[47] Friedman J., Hastie T., Tibshirani R. Additive Logistic Regression: A Statistical View of Boosting. Annals of Statistics, 2000, 28(2): 337—374

[48] Torralba A., Murphy K., Freeman R. Sharing Features: Efficient Boosting Procedures for Multiclass Object Detection // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004, Ⅱ: 762—769

[49] Gao Jun, Xie Zhao, Wu Xindong. Generic Object Recognition with Regional Statistical Models and Layer Joint Boosting. Pattern Recognition Letters, 2007, 28(16): 2227—2237

[50] Shotton J., Winn J., Rother C., et al. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout and Context. International Journal of Computer Vision, 2009, 81(1): 2—23

[51] Shotton J., Winn J., Rother C., et al. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation // Proc of the 9th Europeon Conference on Computer Vision. Graz, Austria, 2006: 1—15

[52] Zhou Zhihua, Zhang Minling. Multi-Instance Multi-Label Learning with Application to Scene Classification // Schölkopf B., Platt J., Hofmann J., eds. Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2007, XIX: 1609—1616

[53] Li Yufeng, Guo Tianyou, Zhou Zhihua. Combo-Dimensional Kernels for Graph Classification. Chinese Journal of Computers, 2009, 32(5): 946—952 (in Chinese)
(李宇峰,郭天佑,周志华.用于图分类的组合维核方法.计算机学报, 2009, 32(5): 946—952)

[54] Teytaud O., Sarrut D. Kernel Based Image Classification // Proc of the International Conference on Artificial Neural Networks. Vienna, Austria, 2001: 369—375

[55] Sahbi H., Geman D. A Hierarchy of Support Vector Machines for Pattern Detection. The Journal of Machine Learning Research, 2006, 7: 2087—2123

[56] Fleuret F., Geman D. Fast Face Detection with Precise Pose Estimation // Proc of the 16th International Conference on Pattern Recognition. Québec, Canada, 2002, Ⅰ: 235—238

[57] Dalal N., Triggs B. Histograms of Oriented Gradients for Human Detection // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005, Ⅱ: 886—893

[58] Ying Zilun, Tang Jinghai, Li Jinwen, et al. Support Vector Discriminant Analysis and Its Application to Facial Expression Recognition. Acta Electronica Sinica, 2008, 36(4): 725—730 (in Chinese)
(应自炉,唐京海,李景文,等.支持向量鉴别分析及在人脸表情识别中的应用.电子学报, 2008, 36(4): 725—730)

[59] Farhadi A., Endres I., Hoiem D., et al. Describing Objects by Their Attributes // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1511—1519

[60] Zhu Songchun, Zhang Rong, Tu Zhuowen. Integrating Top-Down/Bottom-Up for Object Recognition by Data-Driven Markov Chain Monte Carlo // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head, USA, 2000, Ⅰ: 738—745

[61] Zhu Songchun, Mumford D. A Stochastic Grammar of Images. Foundations and Trends in Computer Graphics and Vision, 2006, 2(4): 259—362

[62] Cruse D. Lexical Semantics. Cambridge, UK: Cambridge University Press, 1986

[63] Hoiem A., Collins R. Object Boundary Detection in Images Using Semantic Ontology // Proc of the 21st International Conference on Artificial intelligence. Boston, USA, 2006: 956—963

[64] Lu Hanqing, Liu Jing. Image Annotation Based on Graph Learning. Chinese Journal of Computers, 2008, 31(9): 1629—1639 (in Chinese)
(卢汉清, 刘 静. 基于图学习的自动图像标注. 计算机学报, 2008, 31(9): 1629—1639)

[65] Hoogs A, Rittscher J, Stein G, et al. Video Content Annotation Using Visual Analysis and a Large Semantic Knowledgebase // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA, 2003, Ⅱ: 327—334

[66] Li Feifei, Fergus R, Perona P. One-Shot Learning of Object Categories. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(4): 594—611

[67] Winn J, Criminisi A, Minka T. Object Categorization by Learned Universal Visual Dictionary // Proc of the 10th IEEE International Conference on Computer Vision. Beijing, China, 2005, Ⅱ: 1800—1807

[68] Ponce J, Berg T L, Everingham M, et al. Dataset Issues in Object Recognition // Tean P, Hebert M, Schmid C, eds. Toward Category-Level Object Recognition. New York, USA: Springer, 2006: 29—48

[69] von Ahn L, Dabbish L. Labeling Images with a Computer Game // Proc of the SIGCHI Cconference on Human Factors in Computing Systems. Vienna, Austria, 2004: 319—326

[70] Russell B C, Torralba A, Murphy K P, et al. Labelme: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision, 2008, 77(1/2/3): 157—173

[71] Yao B, Yang Xiong, Zhu Songchun, et al. Introduction to a Large-Scale General Purpose Ground Truth Database: Methodology, Annotation Tool and Benchmarks // Proc of the International Symposium on Energy Optimization Algorithm in Computer Vision and Pattern Recognition. Ezhou, China, 2007: 169—183

[72] Deng Jie, Dong Wei, Socher R, et al. ImageNet: A Large-Scale Hierarchical Image Database // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 710—719

[73] Vogel J, Schiele B. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. International Journal of Computer Vision, 2007, 72(2): 133—157

[74] Barnard K, Fan Quanfu, Swaminathan R, et al. Evaluation of Localized Semantics: Data, Methodology, and Experiments. International Journal of Computer Vision, 2008, 77(1/2/3): 199—217

[75] Wang Yi, Zhou Lizhu, Xing Chunxiao. Video Semantic Models and Their Evaluation Criteria. Chinese Journal of Computers, 2007, 30(3): 337—351 (in Chinese)
(王 煜, 周立柱, 邢春晓. 视频语义模型及评价准则. 计算机学报, 2007, 30(3): 337—351)