图像语义分析

郭平 尹乾 周秀玲 著

科学出版社

图像语义分析

郭 平 尹 乾 周秀玲 著

科学出版社
北京

内 容 简 介

图像语义分析是计算机视觉领域中的基础研究课题，也是近年来的热点研究方向。本书结合作者多年来在该领域的研究成果，对图像语义分析的理论和方法进行比较系统和全面的阐述。主要内容包括图像表示与特征提取、分类判别模型与生成模型、图像中的目标检测与识别、图像语义标注、场景中的图像语义、深度学习在图像语义分析中的应用以及图像语义分析的应用。

本书可作为计算机科学与技术、信息科学与技术、电子科学与技术等专业研究生、高年级本科生的教材，也可供从事图像语义分析、计算机视觉、模式识别和机器学习等研究和应用的相关科技人员阅读参考。

图书在版编目（CIP）数据

图像语义分析/郭平，尹乾，周秀玲著. —北京：科学出版社. 2015.5

ISBN 978-7-03-044267-3

Ⅰ. ①图… Ⅱ. ①郭… ②尹… ③周… Ⅲ. ①图像分析–语义分析 Ⅳ. ①TP391.41

中国版本图书馆 CIP 数据核字（2015）第 094834 号

责任编辑：孙 芳 余 丁 董素芹 / 责任校对：郭瑞芝

责任印制：张 倩 / 封面设计：蓝 正

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http：//www.sciencep.com

中国科学院印刷厂印刷

科学出版社发行 各地新华书店经销

*

2015 年 5 月第 一 版 开本：720 × 1000 1/16

2015 年 5 月第一次印刷 印张：17 3/4

字数：343 000

定价：90.00 元

（如有印装质量问题，我社负责调换）

前　言

人类对环境的感知方式是多模态的，包括视觉、听觉、触觉和嗅觉等。而视觉是非常重要的感知外界的一种手段，人类所获取的信息约 80%来自视觉。随着计算机视觉与图像处理技术的发展，基于视觉特性的图像目标识别技术日益受到重视。人们主要根据图像的含义，而不是图像的颜色、纹理、形状等低层特征，直观地进行分类并判别图像满足自己的需要程度。这些图像的含义就是图像的高层语义，它包含人类对图像内容的理解。要使计算机能够具有和人类一样的理解能力，需要借鉴人类的认知机理来对图像语义进行研究，包括分析图像中的关键内容、提取图像中的语义等。对图像进行语义分析和研究是计算机视觉领域中极具挑战性的研究课题，属于覆盖范围很广的交叉研究，涉及图像工程、计算机视觉、计算智能和认知科学等诸多学科领域。

从机器学习的角度来看，图像语义分析可以看成有监督学习或无监督学习问题。从图像中提取语义信息，实际上是利用先验知识将低层视觉特征通过某种函数关系映射到高层语义。这种映射可以看成一种将图像分类到相应的语义类别的问题，涉及图像表示模型和低层视觉特征、图像分类学习模型和优化算法等问题。利用图像高层语义进行目标检测与识别、图像语义标注、场景中图像语义分析以及图像和视频检索等成为近年来国内外关注的热点研究内容。对于图像语义分析的研究和应用，已经取得了一系列令人鼓舞的研究成果。为了进一步推动该领域的研究与应用，尤其是反映该研究领域最新的研究成果，满足广大教学和科研人员对相关资料的迫切需要，作者撰写了本书。

本书从图像语义的概念和框架出发，并将其与图像工程、计算智能、机器学习的相关方法有机结合起来，逐步阐述图像语义分析过程中涉及的理论方法，以图像语义分析的研究内容为引导，详细介绍目标识别、图像标注、场景理解中的图像语义分析方法和图像语义分析的应用。本书依托作者多年来在该领域的研究与实践，可以帮助读者理解图像语义分析的内容、过程和方法，并取得一定的应用能力。

本书对图像语义分析的各方面进行了比较系统和全面的阐述和讨论，全书共 9 章。第 1 章概括介绍图像语义分析的基本概念、研究内容、研究方法及应用。第 2～4 章介绍图像语义分析过程中涉及的主要理论方法，包括图像表示与特征提取、判别模型和生成模型等分类模型以及相关优化算法。第 5～7 章介绍图像语义

分析主要的研究内容，包括图像中的目标检测与识别、图像语义标注和场景中的图像语义。第 8 章介绍深度学习在图像语义分析中的应用。第 9 章对图像语义分析的应用进行介绍。全书较系统总结了现有图像语义分析涉及的理论和方法，并尽可能地对其应用进行分析与讨论。同时，撰写过程中融入了近年来作者在该领域的研究成果，尤其是关于图像语义特征提取、分类模型和优化以及目标识别等前沿问题的研究成果。

本书是作者及其研究团队成员多年来从事图像语义分析研究的集体智慧的结晶，是国家自然科学基金项目（60275002、60675011、90820010 和 61375045）的部分研究成果总结。杨栋、王静、徐冰心、王元龙、林松、司马海峰、李彦军、姚垚、姜子恒和北京师范大学图形图像与模式识别实验室的吴鹏、余健、陈步等共同努力，一起协助作者完成了本书的撰写工作，在此表示诚挚的感谢。

由于作者水平有限，书中难免出现不妥之处，欢迎广大读者给予批评指正。

作　者

2014 年 12 月于北京

目　录

第1章 绪 论

1.1 图像语义分析的基本概念

图像语义，就是图像内容的含义。图像语义可以通过语言来表达，包括自然语言和符号语言（数学语言）。但图像语义的表达并不限于自然语言，其外延对应于人类视觉系统对于图像的所有理解方式。例如，对于一幅小狗的图像，其图像语义可以包括自然语言单词“小狗”，也可以是一个表示该幅图像中的小狗图像的符号，该符号指的是与该幅图像中的小狗具有相同品种、性别等属性的“小狗”，即此时自然语言的表达和符号语言的表达并不相同。而人的视觉系统不仅将这幅图像理解为语言单词“小狗”及其属性“品种”、“性别”等语言文字，也将其理解为一种抽象的“印象”，该“印象”使得人可以将这幅图像与其他小狗的图像区分开，在遇到该小狗的其他图像时可以将其回忆起来。但限于对人的视觉系统的认知程度，目前图像语义分析仍然主要通过语言来表达图像语义，特别是使用语言单词（称为“关键词”）来表达图像语义[1, 2]。图像视觉基本特征的提取与选择、目标识别和图像内容理解等均与高层语义特征相关。基于图像语义的图像理解成为近年来的研究热点之一[3, 4]。

图像语义分析是对图像和图像语义之间的关系进行分析的过程。图像和图像语义都可以作为该过程的输入。若输入为图像，则输出是该图像对应的图像语义；若输入为图像语义，则输出为包含该图像语义的图像。前者包括识别图像中包含的各种目标的图像识别技术，也包括标注图像对应的各种语义标签的图像标注技术，而后者则包括输入检索关键词从图像数据库中检索出与该关键词相关图像的图像检索技术。

图像语义分析的主要内容包括语义体系构建、图像语义标注、场景分析与理解、图像语义推理等。

1.1.1 图像语义分析与计算智能

计算智能与人工智能一样，是当前信息科学研究的重要领域之一。不同于人工智能中的符号智能主要以知识为基础进行推理，计算智能主要以数据为基础，利用已知数据进行训练建立联系。计算智能包括人工神经网络、模糊系统、演化

计算算法、粒度计算和群体智能等。

图像语义分析研究的是图像和图像语义之间的关系，一般依据已知图像和相应的图像语义的数据库进行研究。而计算智能在处理大规模数据、复杂关系方面具有优势，很多研究人员正在研究基于计算智能的图像语义分析方法。

基于模糊系统的图像语义分析，主要借助模糊系统中的概念来表达图像和图像语义之间的关系。例如，对图像语义“非常”、“中性”、“几乎不”这几种情感程度的模糊量化，可以建立起表达情感的形容词与图像低层特征之间的联系[5]。

图像语义分析可以利用遗传算法或蚁群算法优化基于用户反馈的图像语义标注中的用户选择。例如，使用交互式遗传算法将用户的选择作为适应度函数，通过不断地迭代选择适应度函数较高的用户选择，就可以在下一代选择更好的图像[6]。基于蚁群算法的图像语义分析将每个用户视为一只蚂蚁，顺着前人的信息寻找自己想要的图像，当用户完成检索之后留下新的信息素，随着用户反馈的积累逐渐形成图像之间的语义网络[7]。

基于人工神经网络的图像语义分析，常使用神经网络来建立图像和图像语义的关系模型。例如，将低层视觉特征作为神经网络的输入，将语义期望值作为网络的输出，训练神经网络对自然图像进行分类[8]。

基于粒度计算的图像语义分析，一般使用粒度来表达图像语义的层次关系。例如，基于图像语义的粒度特性（层次特性），将图像的视觉内容，如图像像素的空间关系、亮度、形状的规则程度、纹理的类型，从粒度角度进行形象化表示[9]。

1.1.2　图像语义分析与认知科学

图像语义分析是模拟人类的认知过程，分析图像中能被人类认知到的含义。认知科学是运用信息技术研究人类思维的一门科学。认知科学试图用计算机模拟人类学习，研究人的知觉、记忆、思维等过程。人类视觉系统是认知中重要的信息来源。参考人类对于视觉信息的认知机理，对于理解图像语义分析的过程是非常重要的。

认知科学将视觉认知分为三个层次：第一层是感知，第二层是思维，第三层是认知。感知层是直接获取图像的层次，对应着人类视觉系统获取图像的过程。思维层是对图像进行初步分析的层次，将图像转换为符号数据的过程。不同于图像是一种由所有像素定义的具体的数据表示方式，符号数据是一种抽象的数据表示形式。认知层是对图像进行高级分析的层次，将符号数据进一步转换为知识数据的过程。知识是视觉认知的结果和核心，可以表示输入图像中有什么目标、图像是何种场景、目标场景之间的相互关系等图像语义知识。图像语义分析和视觉认知的过程一样，输入的是图像，输出的是知识。在对图像的获取过程中起到重

要作用的是人类视觉系统，其主要结构是人眼构成的光学系统、进行成像的视网膜和传递视觉信号的视觉通路。人眼光学系统的物理模型是一个凸透镜系统，外界目标反射的光线经过该凸透镜系统在视网膜上成像。视网膜上具有视锥细胞和视杆细胞，视锥细胞在中央凹分布密集，而在视网膜周边区相对较少。视锥细胞负责感知光度和色彩，视杆细胞仅能感知光度，不能感知颜色，但其对光的敏感度是视锥细胞的10000倍。视锥细胞有三种，分别对红（570nm）、绿（535nm）、蓝（445nm）光最敏感，它们有重叠的频率响应曲线，但响应强度有所不同，共同决定了色彩感觉。视觉传导通路由三级神经元组成。第一级神经元为视网膜的双极细胞，其周围支与形成视觉感受器的视锥细胞和视杆细胞形成突触，中枢支与节细胞形成突触。第二级神经元是节细胞，其轴突在视神经盘处集合向后穿巩膜形成视神经。视神经向后经视神经管入颅腔，形成视交叉后，延为视束。第三级神经元的胞体在外侧膝状体内，它们发出的轴突组成视辐射，经内囊后肢，终止于大脑距状沟周围的枕叶皮质（视区）。视觉信息只有传到脑的视皮质并经过处理、分析，才能最后形成主观的视觉感受[10]。

在图像语义分析的过程中，可有效形成数据—知识的相互驱动体系，其中也包括了认知科学中的心理学部分。格式塔心理学中包含一个重要的与图像语义分析相关的内容——知觉组织，其中一个重要的概念是知觉的组织规律是整体先于部分而存在的。人类视觉系统中也符合这个规律，例如，人眼会将断续排列的线段自动拟合为一条直线或曲线，值得一提的是，人类和所有哺乳动物一样，对于人脸的图像是非常敏感的，甚至可以从与人脸完全无关的图案中自动识别出人脸来。人类视觉系统的这种聚合模式能力，可以用来进行图像的模式识别。格式塔心理学包括五种组织原则：前景和背景、接近性、相似性、连续性和封闭性[11]。现阶段认知心理学仍然处在初级研究阶段，还需要大量的研究才能将认知心理学和图像语义分析真正地衔接起来并进入实用阶段。

1.2 图像语义分析研究内容

图像语义分析的研究内容主要包括语义体系构建、图像语义标注、场景分析与理解、图像语义推理等。

1.2.1 图像理解与高层语义

图像工程综合了各种图像技术，是图像技术的一个整体框架。图像工程将各种图像技术分为三个层次：图像处理、图像分析、图像理解。图像处理主要涉及图像的采集、变换、编码等技术；图像分析主要涉及图像分割，图像表达，图像

的颜色、纹理、形状等低层图像特征；图像理解主要涉及图像的三维表达、立体视觉、图像的广义匹配、多传感器融合等技术[12]。

图像语义分析应归入图像理解的范畴，但图像语义分析技术离不开图像处理和图像分析技术。图像语义分析研究图像和图像语义之间的关系，但由于图像包含成千上万像素，一般不直接使用图像本身来表征图像，而是提取图像的低层图像特征来表征图像。构建低层图像特征和图像语义之间的关系，来代替构建图像和图像语义之间的关系。

图像的低层特征可以分为全局特征和局部特征。图像低层特征的全局特征提供了对图像的总体描述，如图像所有像素的平均颜色等。而图像低层特征的局部特征则描述图像中的不同组成部分和组成部分之间的关系，如局部特征向量袋（bag of local feature）对局部特征进行聚类，属于每个聚类中心的局部特征的个数作为每个聚类中心对应的码字，从而将图像描述为一个直方图向量。

与低层图像特征相比，图像语义是一种高层的图像特征[13]。图像语义具有一定的层次性，最基本的是表达图像中各种目标的目标语义，其次是表达目标所在场景的场景语义，然后是表达图像中的动态行为和情感属性的语义。对于低层图像特征和高层图像语义之间的巨大差异，人们常将其形象地称为“语义鸿沟”[14]。

1.2.2　图像语义标注

图像语义标注可看成一种计算机分类系统，该系统的输入是图像，输出是图像对应的语义。即图像语义标注是根据标注了图像语义标签（关键词）的图像数据库，构建一个计算机分类系统，该系统自动将关键词赋予测试图像来描述图像的内容。构建图像语义标注系统时通常需要一个训练图像数据集，在该数据集中，每张图像都附有人工标注的语义标签。

图像语义标注也可以看成一个多示例多标签（multi-instance multi-label，MIML）学习问题，即图像特征包含多个示例（局部特征），图像内容可由多个标签描述。将 MIML 学习问题转化为一般的监督或非监督学习问题有两种方法：将 MIML 学习问题转化为多示例学习（multi-instance learning，MIL）问题，或者多标签学习（multi-label learning，MLL）问题[15]。图像的多示例表示是在提取每个示例的特征之后融合多个示例的特征来表示图像的，因此属于低层图像特征的局部特征。将图像表达为多示例表示时，图像中多个示例的特征合并为一个示例，则图像被表示为单示例多标签的，此时 MIML 学习问题转化为 MLL 问题，如多类别分类问题。而图像的多标签表示是指图像包含多个语义标签，可以将多个语义标签进行组合定义为新标签，则每个图像可以只拥有一个新标签，此时 MIML 学习问题转化为 MIL 问题。图像语义检索或标注可以看成有监

督学习或无监督学习问题[16, 17]。

1.2.3 场景描述与理解

场景描述与理解的主要任务是描述场景中的目标和目标之间的关系。据统计，真实世界中存在几万种目标种类，一般都具有特定的语义。因此真实世界中的场景具有目标繁杂、关系复杂的特点。不同于具有简单背景和单一前景目标的目标识别任务，场景识别具有较高的难度。

图像语义分析实际上是利用先验知识将低层图像特征通过某种函数关系映射到图像语义的过程。在场景描述与理解中，先验知识是对于场景中目标及其关系的描述性知识，如目标的属性知识、不同目标之间的区别和层次关系等。依据场景描述与理解的识别任务也是有层次的：首先识别场景中的目标，然后识别目标之间的关系。在目标之间的关系中，位置关系是场景中特有的关系。低层图像特征的局部特征也常使用位置关系来描述图像特征。不同于局部特征的位置关系，场景中的位置关系更多的是一种递归式的包含关系，如汽车和卡车都有相同的轮胎部分。

1.2.4 图像语义推理描述

图像语义包括使用自然语言或符号语言来表示的方式，因而描述图像语义之间的关系也可以使用语言的句法表示和推理方法。图像语义的推理描述，包括图像语义的概念推理和场景推理。概念推理是一种层次关系的推理，如“交通工具”的概念包括“汽车”和“火车”，根据带有关键词“汽车”的图像可以推理出该图像也应带有“交通工具”的标签；场景推理是一种位置关系的推理，如根据带有“显示器”和“键盘”标签的图像可以推理出该图像有一定的概率带有“计算机”的标签。

基于图像语义推理描述场景目标之间的关系，不仅可以描述低层特征之间的相关性（相似性），也可以描述高层图像语义之间的关系。根据低层图像特征可以推理高层图像语义，可以根据多个低层图像特征的与、或、非关系推导出一个高层图像语义，例如，依据“红色”、“黄色”和“云形”标签的并存关系推导出“晚霞”的标签。将与、或、非关系拓展到函数关系，则多个低层图像特征要满足一定的函数关系，才可以推导出高层图像语义。知识推理的基本形式是根据条件语句推导出结论，多个条件语句可以构建复杂的推理网络，可以采用自底向上和自顶向下两种构建方式。高层图像语义之间也可以进行推理，例如，上述“显示器”和“键盘”两个标签同时出现，可以推导出“计算机”的标签。

1.3　图像语义分析的研究方法

图像语义分析是将低层图像特征通过某种函数关系映射到图像语义的过程，这种映射可以看成一种将图像分类到相应的语义类别的问题，也可以看成建立低层图像特征和图像语义的联合模型的问题。因此图像语义分析的研究方法可以分为模式识别中的两类方法：基于分类的方法（判别模型）和基于概率的方法（生成模型）。

1.3.1　模式识别方法：判别模型

由于图像语义分析问题是 MIML 学习问题，同时也是一个 MLL 问题，判别模型将每个语义标签看成一个类别，则图像标注是一个多类别分类问题，即将待标注的图像分到多个类别，相当于给图像标注多种标签。判别模型常使用贝叶斯分类器或支持向量机分类器[18]。其中，有监督的一对多方法在应用多个两类分类器时，需要学习正图像和负图像，正图像有给定的语义标签，而负图像没有。例如，区分室内和室外图像，区分城市和风景图像，或者从图像包含的目标中检测树、马或者建筑物。基于判别模型的方法可以采用分而治之策略，例如，一种分而治之策略是首先考虑两个类别，即室内图像和室外图像；然后将室外图像细划分为城市和风景图像等。有监督多标签标注则是构建一个多类贝叶斯分类器，对待标注图像的每个示例进行分类，然后统计所有示例的类别作为待标注图像的多个标签[19]。在图像分类模型和优化算法方面，高斯混合模型、支持向量机[14, 20]、K 均值、贝叶斯信念网络[21]、多维隐马尔可夫模型[22]等得到了广泛关注。而对模型优化则采用最大似然学习、基因规划算法[23]等方法。

判别模型还包括人工神经网络。神经网络模型是受人脑内的信息传播和处理方式的启发而产生的，人脑表示信息的方式是：在信号从视网膜传递到大脑皮质再到运动神经的时间里，大脑皮质并没有直接提取数据的特征，而是将信号通过一个复杂的层状网络模型进行处理，从而获得这些信号有什么形式的分布规则[24~26]。

误差反向传播（error back propagation，BP）算法是经典的神经网络训练算法，它的出现，掀起了基于统计模型的机器学习热潮。BP 算法用于训练前馈神经网络模型，该模型是一种凭借人工经验来抽取样本特征的浅层学习模型，这种模型主要用来实现分类或预测，因此特征提取这个至关重要的过程就成为整个系统性能的瓶颈。输入与输出之间通常是非线性映射的，从而使网络误差函数或能量函数空间是一个含多个极小点的非线性空间，而 BP 算法的搜索方向仅是使网络误差

或能量减小的方向，因而经常收敛到局部最小，并随网络层数增加，情况更加严重。理论和实验表明 BP 算法不适于训练具有多隐层单元的深度网络结构[27]。而且由于人工构造样本特征不仅需要使用者投入大量的人力物力，还要求使用者对实际问题具有很好的把握，所以该方法的应用面受到限制。在 20 世纪 90 年代，大量的浅层学习模型被提出，包括支持向量机（support vector machine，SVM）、逻辑回归（logistic regression，LR）方法等。

这些模型之所以被称为浅层学习（shallow learning）模型，是因为它们仅有一层隐层节点甚至没有隐层节点。它们在一些实际应用中表现出较大的局限性，如它们在有限样本和计算单元情况下对复杂函数的表示能力有限，针对复杂分类问题其泛化能力受到一定制约。

深度学习（deep learning）是近年来机器学习研究中最受关注的一个热点，其动机在于模拟、建立人脑进行分析学习的深度神经网络，它模仿人脑的机制来解释图像、声音和文本等数据。它通过将低层的特征组合起来形成更高层的表示，从而发现数据的分布式特征表示[28]。

深度学习与浅层学习明显不同，它不仅强调了模型结构的深度，同时更加突出了特征学习的重要性，即通过逐层特征变换，将样本在原空间的特征变换到一个新特征空间，从而更有利于分类或预测。与人工规则构造特征的方法相比，利用大数据来学习特征，刻画数据所示丰富内在信息的能力更强。而且，深度学习可通过学习一种深层非线性网络结构，实现复杂函数逼近，展现出了强大的从少数样本集中学习数据集本质特征的能力。

1.3.2 模式识别方法：生成模型

生成模型通过建立图像与标签之间的概率相关模型进行图像语义分析。由于图像的语义非常丰富，直接采用原始词汇库构建语义特征表述的学习工作量非常大。借鉴文本分类中词袋（bag of words，BoW）的思想，大多数人采用了视觉词袋（bag of visual-words，BoVW）语义模型[17]，在低层视觉特征的基础上，通过聚类分析来模拟人们对图像的语义分类，然后利用这种语义分类寻找语义和低层视觉特征之间的关系。例如，一种更具普遍性的语义分类方法，可同时处理目标图像中的多个词汇分类[29]。该方法中，用直方图偶（bipartite）表征图像，一半直方图描述适合图像内容的词汇计数，另一半直方图描述相对于适合图像内容的词汇计数的通用词汇计数。基于概率的图像标注算法，例如，概率潜在语义分析（probabilistic latent semantic analysis，PLSA），是一种基于概率的潜在语义分析算法，其基本原理是通过奇异值分解，将文本投影到低维的潜在语义空间中，便可有效地缩小问题的规模[30]。另外，基于相关模型的方法通过构建低层图像特征和图像语义之间的不同