媒体计算与内容分析

魏维 魏敏 编著

清华大学出版社

媒体计算与内容分析

Media Computing and Content Analysis

清华大学出版社
北京

内 容 简 介

媒体计算是指对视频、音频、图像、图形、文字等各类媒体信息进行内容分析的计算理论、算法和应用技术，包括各类媒体信息的表示、内容分析与识别算法等内容。本书内容覆盖了媒体内容分析与理解相关的理论和应用技术，共分为 11 章。第 1 章介绍现有媒体内容分析与理解现状；第 2 章主要介绍媒体计算理论及模型；第 3 章介绍了视频内容分析的内容（视频内容分析的前处理）；第 4～6 章主要介绍基于统计学理论的视音频内容分析，包括视觉语义分析与理解、音频语义分析与理解以及视频语义分析两级多模式信息融合；第 7～11 章介绍基于认知机理，从场景整体语义的角度来理解、标注和分析媒体内容和语义概念，包括强依赖关联关系提取、多标记语义标注、媒体场景显著计算、媒体语义相似性计算、媒体显著对象语义本体标注方法等内容。

本书集原理、技术应用为一体，同时有实验分析和原型系统构建，是作者多年来从事图形图像处理与内容分析与理解的相关科研和承担研究生相关课程教学工作的积累。本书主要读者对象为从事图形图像处理的研究人员、大专院校计算机专业及相关专业师生、从事媒体信息处理研究与开发的科研人员和工程技术人员。

图书在版编目（CIP）数据

媒体计算与内容分析/魏维等编著．—北京：清华大学出版社，2012.1
ISBN 978-7-302-26105-6

Ⅰ．①媒…　Ⅱ．①魏…　Ⅲ．①传播媒介—分析方法　Ⅳ．①G206.2

中国版本图书馆 CIP 数据核字(2011)第 131626 号

责任编辑：汪汉友　薛　阳
责任校对：梁　毅
责任印制：杨　艳

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62795954，jsjjc@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185×260　　**印　张**：12　　**字　数**：287 千字
版　　次：2012 年 1 月第 1 版　　**印　次**：2012 年 1 月第 1 次印刷
印　　数：1～2000
定　　价：29.50 元

产品编号：036726-01

前　言

随着计算机技术和网络技术不断发展，视频点播(Video on Demand，VoD)、远程教育、数字交互电视、视频会议、数字图书馆、远程医疗等新的媒体信息交换和应用形式已经完全融入人们的日常工作、生活和娱乐之中。为了对含有丰富时空信息的媒体数据进行高效快捷过滤、浏览和检索，人们提出并发展了内容分析和基于内容的视频检索(Content-Based Video Retrieval，CBVR)技术。目前，实际应用的内容分析与管理主要基于颜色、纹理、形状等低层特征。而人们习惯用概念来描述事物，判断媒体对象是否相似应建立在更高层的、主观的、符合人的感知的相似性上。用户希望根据语义来对媒体进行管理、操纵、检索等应用，让计算机按照人的主观感觉和理解来表示媒体内容。低层视觉特征和高层语义特征之间存在着语义鸿沟(Semantic Gap)，这导致媒体内容抽象、表示与内容分析与理解面临巨大的困难和挑战。如何进行媒体内容分析，跨越低层特征和高层语义概念间的语义鸿沟，以语义概念来管理、访问视频数据，已成为多媒体领域颇具挑战性的研究课题。媒体内容分析已经逐渐成为多媒体领域研究的热点。

媒体计算是指对视频、音频、图像、图形、文字等各类媒体信息进行内容分析的计算理论、算法和应用技术，包括各类媒体信息的表示、内容分析与识别算法。本书围绕媒体内容分析与理解所涉及的理论和技术实际问题，吸取国内外最新媒体计算和媒体内容分析研究成果，系统地介绍了视频、音频、图像、图形等媒体信息的计算理论、算法和应用技术。

本书主要涉及媒体内容分析与理解相关的理论及应用技术。理论主要包括媒体计算中所需要的数学变换、识别理论、数学建模等(第 2 章)。技术部分先介绍视频内容分析的基础知识(第 3 章)，分为两部分内容：一部分内容是基于统计学理论的视音频内容分析(第 4～6 章)，另一部分内容是应用认知机理，从场景整体语义的角度来理解、标注和分析媒体内容和语义概念(第 7～11 章)。

基于统计学理论的视音频内容分析主要提出了一个视频数据多粒度语义分析和提取的通用解决方案。在该方案中，多层次语义分析与多模式信息融合技术在同一模型中得到统一和应用。本部分首先提出了一种基于统计分布的镜头渐变边界检测方法，并用一种具有时间语义语境约束的关键帧选取策略对时域内容进行表示。然后用模式分类方法对注意力模型选择得到的显著区域进行基本视觉语义识别，在此基础上得到一种层次的多粒度视觉语义分析提取框架。随后把时频变换得到的声音频谱作为可观察特征，构建了基本声音语义识别提取的隐马尔可夫模型；通过语义窗口获得基本声音语义组后，按照高层逻辑定义提取音频高层语义。最后仿照人脑多感觉器官信息融合机理，将视频中的多模式特征按不同类别划分为组，得出一种基于仿生的视频语义分析两级多模式融合方法。

基于知识的场景整体语义理解部分首先提出一种提取媒体语义概念间关联依赖关系的方法，获取一个镜头中语义概念间的依赖关系；接着在已获得语义概念间强依赖关系的

基础上得到一种多标记媒体语义概念标注方法；利用显著次序计算得到媒体语义概念对象间的语义相似性值；然后采用动态及静态显著内容对媒体语义内容进行表示；进一步得到媒体场景之间的语义相似性度量准则；最后提出一种媒体显著语义本体标注方法，通过本体实现了媒体语义概念在知识层提供语义（知识）共享和重用，有益于进一步在语义层次对媒体数据进行组织、管理、检索等应用。

目　录

第1章 绪 论

1.1 引 言

随着网络技术和计算机技术的高速发展，网络中的媒体资源日益丰富。媒体中包含了生动丰富的多模式信息，媒体内容的表示、分析是当前多媒体研究领域研究的热门课题之一。

目前，网络上存在着丰富的媒体数据。在搜索引擎中查找需要的媒体数据时，输入一个描述目标的关键词，搜索引擎常会按一定规则进行相关度排序返回成千上万的搜索结果。有时会给用户一种假象：好像计算机自身理解了媒体的“内容”。实际上这种搜索是对人工预先标注描述媒体内容的关键词进行文本的检索，而并非计算机自身理解了媒体“内容”（或自动分析了媒体“内容”）。

对媒体内容进行自动的分析与理解，提取各个语义层次的媒体内容信息，是海量媒体信息进行管理、应用、传播的关键技术。现有的媒体内容分析与理解技术包括机器学习/模式分类、统计学习、基于规则等方法。这些方法涉及不同的计算理论，各有优点和不足之处。

1.2 媒体/视频语义预处理

视频语义内容分析包括时域分割（分镜头）、空域分割（区域分割、视频对象分割）、关键帧选取及低层特征的提取等。

视频镜头（Video Shot）是摄像机在一个拍摄动作（Camera Action）中所录制的由连续视频帧组成的视频序列。在这些帧序列中，彼此间内容有很强的相关性，因此通常将镜头作为视频内容组织的基本单位。目前分镜头技术的研究已经比较成熟，镜头边界的探测算法的效率和准确度也不断提高。镜头边界分为突变和渐变两种，其中突变类边界占大多数（90%以上）。渐变主要是由于渐隐（Dissolving）、淡入/淡出（Fading In/Out）等特殊技术处理形成的。对于突变类的边界检测，在像素域和压缩域中的效果都比较满意。渐变类的边界检测算法在像素域中有一定突破，然而在压缩域中的算法还有待提高[1]。Taskiran 等在 ViBE 系统中直接从压缩域的 DC 序列中提取 GT（Generalized Trace）和衰退树（Regression Trees）进行镜头分割[2]。Ouyang 在 MPEG 压缩域中用 MBs 和运动矢量（Motion Vector，MVs）等信息进行体育比赛中回放镜头的边界探测[3]。以上两种方法都是在压缩域中进行的，而 Wengang 充分利用视频中声音在镜头边界有改变的特征，同时联合音频和可视图像进行边界探测，提高了镜头分割的准确性[4]。

此外，为有效操纵大量的视频数据，有时需要在更高层次分割视频，即场景分割。一个场景由一个或多个语义相关的镜头组成。场景的具体定义与应用有关。媒体数据中包

含的语义形式是多样的，常常相互呼应，以前的方法很多用短时间间隔视听特征变化超过预先定义的门限值来进行分割。这样的分割常会导致一些假的场景变化点出现，而且确切门限值的设定也很困难。而在文献[5]中，作者提出了三种基于 HMM 的联合分类和分割方法，这三种方法中用动态编程技术搜寻最可能的类转换路径。在文献[6]中，作者提出一种视频场景分割和语义表示框架。在这个框架中，首先用由粗到精的算法对镜头边界进行探测，然后用模板匹配选择关键帧，将时空关联的镜头分到同一个场景中，最后对镜头场景进行语义表示。在文献[7]中，作者提出一种基于谱方法和视频结构的场景分割方法。此方法先基于视觉相似性和镜头间的时域相关建立图，然后用谱方法将图中的镜头分为场景。

为了从媒体数据中得到语义概念，往往需要将空域分割的视频内容用特征量的形式表示。这些特征的提取可降低语义模式匹配和识别的难度。语义对象往往与帧中所对应区域的特征有很强的关联，所以对于语义的提取而言，区域分割和视频对象分割具有重要作用[8]。抽取的特征最好是在模式类间具有不变的性质。只有提取显著区分特性的特征才可能用简单模式分类法进行分类和识别。区域分割和视频对象分割对于语义概念抽取具有明显区分意义，是视听信息在时空上分割比较重要的环节。

感兴趣区域(Region-of-interest, ROI)的探测与分割是进一步进行语义分析的基础。ROI 是观众比较感兴趣的区域，会给予比其他普通区域更多的关注。ROI 提供了一种简洁表示视频可视内容的方法[9]。在区域分割方面，现有技术可对均质区域实现自动分割，并且分割准确度较高[10]。Rautiainen 在局部 HSV 颜色直方图的基础上，建立自组织图(Self-organizing Map)，通过对其训练来探测皮肤的区域[11]。Sigal 等提出一种在视频序列中实时分割皮肤区域的新方法[12]。然而，对语义概念提取有较大作用的区域往往是复合颜色并且是非均质的，其实现需要人干预调整参数。

在视频对象分割方面，现阶段可实现用户监督下的半自动视频对象提取[13]。例如，Chen 在视频对象分隔时采用首帧分割，自动对象跟踪和边界精化技术，在用户监督下此方法可实现高效率的半自动视频语义对象分隔[14]。虽然语义提取中急需的快速自动进行视频对象提取当前还不成熟，但此方面的探索性研究也比较多。Kai 提取光流场的相直方图，以此探测视频帧中的语义对象，最终实现非监督的视频语义对象提取[15] Zhou 通过区域提取和运动预测两项技术，解决了视频语义对象提取的速度问题[16]。Lievin 等在联合处理低层颜色和运动量的基础上，进行颜色空间的非线性变化[17]。

视频运动对象也是比较重要的感兴趣区域。大多数自动提取视频对象的算法都在一定程度上利用了空间信息。从时域信息来看，运动对象的运动模式与背景间有明显差异，这是多数视频运动对象分割的基本出发点[18]。较多学者从事视频运动对象分割的研究，并取得了很大的进步和成就。Cucchiara 提出一种基于知识的 Sakbot 随机方法，对图像序列中的运动对象进行探测[19]，解决了背景信息及时更新以及处理背景阴影两个难题。在文献[32][20]中，Wang 通过建立自适应的背景模型，同时用多种算法对运动物体进行识别，可检测出物体的阴影。

低层特征一般用特征描述子表示。描述子(Descriptor)是刻划特征的一个数据结构，一个描述子的维数可以是多维的。常用的可视特征描述子包括颜色描述、纹理描述、形状

描述、运动描述[21]。视频特征大都是高维的(例如,MPEG-7 中颜色结构描述子达一百二十八维)。为准确提取语义概念、减少计算量和避免维数灾难(Curse of Dimensionality),常要进行降维处理。降低维数可以采用特征提取和特征选择两类方法。特征提取通过映射(变换)的方法用低维空间表示样本。而特征选择则从一组特征中挑选最有效的特征,以此达到降低特征空间维数的目的[22]。通常特征子集的产生方法(策略)是穷举法和启发式方法[23]。穷举法把各种可能的特征组合都计算出来,通过比较选择最优的特征组,其典型的算法是 FOCUS[24]。启发式的选择方法依据特定的启发策略来增减搜索空间[25]。

特征提取、选择虽然在理论研究上已经比较成熟,但近年来不断出现新的实现方法[26]。Balaji 选择与分类最相关的联合特征,提出了在高维数据分类中性能优良的联合分类特征最优算法(Joint Classifier and Feature Optimization, JCFO)[27]。然而 JCFO 算法计算复杂,并不适合视频数据应用。其他现有的特征提取、选择算法往往也不能直接用于视频,因此有必要结合视频特征的性质进行降维研究[28]。对于视频中的高维描述子,Ankush 提出了基于最小分类错误的 DABER 算法,同时还提出了减少描述子的 CPDDR 算法[29]。文中的思想对今后视频领域的降维研究有一定启发作用,但 Ankush 在该文中并未进行涉及语义保持的研究。而在文[30]中,Jensen 和 Shen 用基于 rough 集的方法,对数据进行语义保持(Semantics-preserving)的降维处理,为语义视频分析中保持语义降维方法提供了借鉴。

1.3 现有媒体内容分析与理解现状

1.3.1 机器学习/模式分类方法

这种方法将视频语义对象提取看做是待提取视频语义对象(此对象类别未知)的分类问题,利用模式分类方法来尝试跨越语义鸿沟,图 1.1 是机器学习/模式分类方法进行视频内容分析的原理框图。

在这种方法中,语义概念模式分类与识别关注的是模型的概率特性,其核心思想是用随机数学的方法来描述对象的不同特征并在此基础上建立多媒体概念模式分类器。如图 1.2所示,视频语义概念模式的分类器主要包括多媒体语义对象模型和多媒体语义网络模型。建立分类器的过程主要涉及两方面,即给定一般的模型或分类器的形式及利用训练样本去学习或估计模型的未知参数[31]。

1. 多媒体语义对象模型

视频的任意部分(片段)内容都可以理解为在某一地点或场景下存在或发生的事件。依据此理解,提出多媒体对象的语义概念。多媒体对象是多种层次特征(既包括低层次的声音、图像、字幕特征,也包括分割的特征,还包括诸如人脸识别器等高层次的特征探测器)所支持的一种概率模式[32]。多媒体对象利用概率结构模型作为中介,使低层次特征和高层次语义概念间产生联系。通常来讲,多媒体对象不仅在帧内的空域具有空间上的随机性,而且在每一个帧的时间序列及音频时间序列中还具有时间、空间上的随机性质,

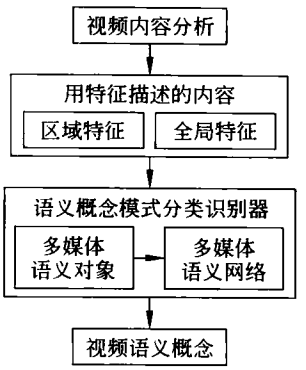

图 1.1 机器学习/模式分类视频内容分析框图

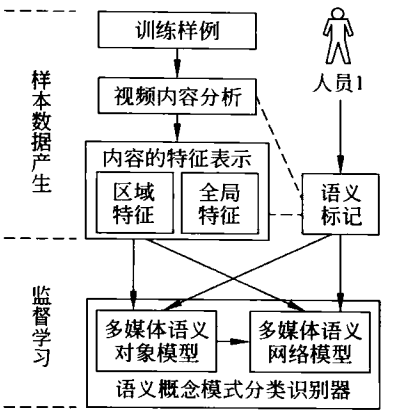

图 1.2 语义概念模式分类器学习过程

所以通常在此模型中将低层的特征作为一个随机变量 X(矢量)。一般可以采用贝叶斯决策理论(Bayesian Decision Theory)建立的贝叶斯分类器来作为语义对象分类模型[33, 34]。

具体来讲,可以把观察到的特征值表示为多维随机变量 X(向量),定义可能的假设 H(较简单的方法是可以采用定义两个假设 H0 和 H1,其中 H0 表示语义概念对象出现,H1 表示语义概念对象未出现)。对每一个假设,定义特征的条件概率密度函数和先验概率。通常用贝叶斯决策理论在可能的假设间做决策时,认为条件概率密度函数是已知的[29]。对静态地点类语义概念,可用高斯混合模型(Gaussian Mixture Models, GMMs)来得到其条件概率密度函数。对于同时具有时空关系特性的事件和对象而言,用隐马尔可夫模型(Hidden Markov Model,HMM)得到每种假设下对应的条件概率密度函数。

由于隐马尔可夫模型在语音识别方面应用效果较好,所以目前主要采用 HMM 建立多媒体声音对象、事件模型。在视频中的声音往往是多个不同声音源的合成(例如背景音乐和前景声音往往同时存在),因此,在混合音源中提取语义概念是音频语义的主要研究内容。

2. 多媒体语义网络模型

用语义对象模型分类得到的语义概念之间并不是相互孤立的。在视频内容的上下文背景中,语义概念间存在彼此的联系。多媒体对象网络就是用来描述对象间这种强关联性的。例如,天空、雪出现在户外的概率较大,人讲话时往往伴随嘴唇的活动等。为描述帧层次上语义概念间的关系,其中一种方法可用加权图(Factor Graphs)来建立模型[35, 36]。加权图包括贝叶斯置信/信念网(Bayesian Belief Network,BN)和马尔可夫随机场(Markov Random Field),其中用得较多的是贝叶斯置信网。BN 是描述联合概率分布的有向无环图(Directed Acyclic Graph,DAG)的拓扑形式。贝叶斯网络是用来表示变量间连接概率的图形模式,它提供了一种自然的表示因果信息的方法,用来发现数据间的潜在关系。在这个网络中,用节点表示语义概念,用有向边表示语义概念间的依赖关系[37,38]。

语义网络可以间接提高模式分类器的语义概念识别能力,一些难以直接探测的语义

概念可以通过其他容易探测的相关对象推理而得[39,40]。如海滩的概念难以直接探测得到，但海滩的景色常伴随水、沙、树、船等容易识别的语义对象。因此可以通过水、沙、树、船间接得到海滩的语义，同时推断出这是一个户外景。

网络模型目前应用较成功。Hoogs 在处理视频中大量的对象、事件和场景时，将语义对象分类与语义数据库相结合。由于此方法结合电子词典数据库 WordNet，所以对象和事件的识别能力得到大大增强[41,42]。Cheng 提出基于语义网络的语义联合模型[37]。此模型利用不同镜头间对象的关系来描述镜头间内容的相互关系，在形式上用六元组定义的联合模型表示。Luo 用动态贝叶斯网(Dynamic Bayesian Network，DBN)和层次隐马尔可夫模型(Hierarchical Hidden Malkov Model，HHMM)建立由粗到精的语义概念模型[43]。Wang 等在文献[44]中以智能代理(Intelligent Agents)确定网球的轨迹和落点，以之作为改进的贝叶斯网络分类特征，分类后得到语义标签。

3. 模式分类器的训练和学习

利用样本数据来确定分类器的过程称为训练分类器。在多媒体语义对象模型方法中需要用到 EM 算法来估计期望，协方差矩阵，GMM 和 HMM 的混合比例(Mixing Proportions)，以及 HMM 中的转移矩阵。在机器学习算法的经验分析方面，可利用 UCI 机器学习知识库中的数据[45,46]。

语义概念模式分类器学习过程如图 1.2 所示。待学习的语义概念或函数称为目标概念(Training Concept)，记做 c。一般来说，c 可以是定义在实例集上的任意布尔函数，即 $c: X \rightarrow \{0, 1\}$。概念定义在一个实例(Instance)集合之上，这个集合表示为 X。在学习目标概念时，必须提供一套训练样例(Training Example)，每一个样例为 X 中的一个观察值 x 以及它的目标概念值 $c(x)$。对于 $c(x)=0$ 的实例称为反例(Negative Example)或目标概念的成员。对于 $c(x)=1$ 的实例称为正例(Positive Example)或非目标概念的成员。训练样本集中每个样本的类别归属是(在人的参与下)“被标记了”的(Labeled)，通常在语义训练中用到的是有监督(Supervised)学习。分类器学习的目标就是寻找一个假设 h，使其对于 X 中的所有 x，有 $h(x)=c(x)$[47,48]。

4. 方法优缺点总结

通过传统的机器学习技术来进行自动媒体内容理解是许多学者的研究重点。传统的语义概念分析、提取采用的是监督学习方法。典型的基于监督学习的媒体语义标注首先应将媒体(视频)分为小的处理单元——镜头(场景)。然后，在每个单元中提取低层物理特征来表示媒体内容。媒体语义标注则分为训练(学习)和分类两个阶段。在训练阶段，针对有限的人工标注好的语义标签的样本进行学习。完成训练后，就可以对大量未知语义类别的媒体进行标注(分类)。其中，最简单的方法就是一个语义概念建立一个 2 分分类器。在这一框架下提出了很多方法，如 Naphade 利用 SVM 作为主动标注和主动学习的分类器[49]。在文献[50]中，作者提出一种基于隐马尔可夫的婚礼视频分析和事件分割系统，通过分类建立婚礼中的关键事件。在文献[51]中，作者提出一种将静态对象(建筑)和动态对象(车辆)进行分类的方法，该方法应用三维动态场景分析方法从 UAV 平台获取的场景中进行分析。Ma 等在文献[52]中提出用隐马尔可夫模型解决图像和视频语义理解和标注中的原因不明和多维问题。该方法把无关联模型分解为多分布，多维的隐马

尔可夫模型。在文献[53]中，作者采用监督多类别、多标记模型对音频音效进行语义标注，最终可以实现基于文字的内容检索。在文献[54]中，作者用适当的关键字对视频进行标注，此方法采用贝叶斯网(Bayesian Network，BN)得到关键词间的关系，并动态、静态地除去无关的关键词。

机器学习方法可以提取一些可靠的中间级语义概念。为提取这些概念，人们提出了许多自动语义概念探测分类方法。其中，有与人相关的人脸识别、主持人探测，讲话、音乐等声音探测器，户外/室内、都市风景等位置场所识别器，以及各种特定对象、物体识别器。多年的研究表明，经过足够的数据训练，这些分类器能取得较好效果。虽然众多不同的方法获得了一些令人鼓舞的效果，但以下几个问题仍困扰着基于机器学习的语义标注性能进一步提升：

① 训练样本不足。要保证监督学习取得较好的学习效果，就必须有一定数量已经标记的样本作为训练集。在很多情况下，人为标注这么多样本很难实现。

② 维数灾难问题。多模式、高维低层物理特征常会导致维数灾难问题。

③ 距离度量准则。许多学习方法的性能与不同语义概念、不同特征间的距离度量准则密切相关。

基于机器学习的媒体标注主要有两大流派，一个是监督学习，另一个是非监督学习。典型的监督学习需要对预先设定的概念进行学习。用统计模型进行语义标注需要大量标记好的样本，而通常这种标记好的样本数目却不多，这导致了最终标记的结果不够准确，成为阻碍这种方法的主要障碍。为解决以上这些问题，大量的方法被提出来，如半监督学习、自主学习等方法。半监督学习技术可以降低监督学习中对标记样本的要求[55]。

监督学习的语义标签类别是预先确定的，非监督学习则可以突破这种限制。在文献[56]中，Moxley E. 等通过搜索和挖掘技术自动进行视频标注，该方法完全是非监督学习，并且不受预先定义词汇的限制。以非监督技术为基础的标注，仅需要少量已标注的数据，每次新数据学习后，分类器参数都会得到相应更新。如文献在[57]中，作者采用非监督主动学习技术进行基于图论语义聚类，使得学习算法可以主动选择未训练过新类别的样本。在文献[58]中，作者提出一种新颖的自主学习框架，从多模式视频数据中提取用于训练的初级语义标签。这完全代替了人工标注语义类别的过程。在此框架中，不需要人参与就可以从多模式特征中得到初级语义标签。系统在通用多例学习(Generalized Multiple-instance Learning)和不确定标签密度(Uncertain Labeling Density)的基础上，推测得到视觉语义概念的相关值。从这些相关度的值中，用支持向量回归(Support Vector Regression)建立通用的视觉语义模型。

在语义概念学习中，或多或少会存在缺少已标注的训练数据，而网络中有大量信息丰富、经过个性定制的，并有一定对应文字描述的媒体数据，如何利用这些网络上的资源来作为语义概念学习中的样本资料是一个值得研究的课题。在文献[59]中，作者通过网络视频资源的挖掘进行语义概念学习。该方法讨论了如何从视频共享站点开始，用 LSCOM，WordNet 或 ConceptNet 本体产生语义概念间的关系。

机器学习提取语义的大多数方法都将语义概念看做是相对独立的，忽略了媒体语义概念间的相互关系。在现实的媒体中，语义概念并不是相对孤立的，在语义层面存在着相

互关系和相互的作用语影响。例如,“道路”这个概念就常和“汽车”一同出现在同一个语义场景中,而“飞鸟”与“飞机”就很少一起出现。若结合语义语境关系进行标注,则可有效降低标注的难度。近期的此类方法开始利用同时出现关系进行语义标注。

1.3.2 统计学习方法

传统的统计模式识别方法研究的是样本数趋向无穷大时的极限特性,是一种渐进理论。其性能在样本数足够多的前提下才能达到理论效果。而视频检索中的样本数目往往有限,因此如何应用有限样本情况下的统计学习理论进行语义概念提取也是研究的重点之一。

支持向量机(Support Vector Machine, SVM)基于统计学习理论,建立在计算学习理论的结构风险最小化原则之上。其目的是在高维空间中寻找一个超平面作为两类的分割,以保证最小的分类错误率。此类模型在只有小训练样例集的情况下,分类效果较好[60]。例如,Naphade 利用 SVM 作为主动标注和主动学习的内在分类器[49]。这种以支持向量机为基础的标注器建立在少量已标注的数据之上,每次新数据学习后,分类器参数都会相应更新。

1.3.3 基于规则推理的方法

以上两种方法的理论基础都是模式分类,实质上是分类器通过学习训练样例由系统内部产生分类标准。而基于规则推理的方法则考虑直接从系统外给定分类标准,即规则。

基于规则推理的方法可以定义为集合 R。

R: $F \rightarrow C$(F 是特征集合,C 是语义概念集合)

若对于 $f \in F$ 和 $c \in C$,c 依赖于 f 则存在一个规则: $f \rightarrow c \in R$。

如图 1.3 所示为基于规则的推理方法的主要组成框图。此类方法主要是根据视频内容特点,结合专业知识(往往由专家参与)定出相关的推理规则。推理规则的实质是给出语义分类的阈值(Threshold),利用一系列的门限值构成语义概念分类器。不同内容的视频流经视频分析后按镜头为单位提取特征,进行推理分类并提取对应的语义概念。语义事件规则的制定有两类,一是依据可视特征和时空关系来定制,另一种是根据对象在现实中的关系(Real-world Relation)定制。同时加上时域上定义的相互位置关系和逻辑运算关系,便可实现预定义的语义对象、事件的检索。确定性事件选择首先以低层的视听线索/特征来表示事件,例如足球比赛中采用场地颜色、摄像机运动和边界等表示。在这些具有特征性质的线索被探测后,再根据针对特定领域所制定的规则进行推理,得出语义概念。

传统的规则方法一般都遵从如图 1.3 所示的方法进行语义提取。Petkovic 定义特征操作符(φ)、空间关系(σ)、时间关系(τ),在此基础上定义对象规则(Object Rule)[61]。Tiecheng 在教学视频中定义局部改变、内容一致和重要改变规则,进行语义内容概要[62]。这些方法通常采用单规则提取语义,其语义概念单一。在此基础上改进提出的多规则语义提取方法,可支持较丰富的语义概念。例如,Li 在对美式足球训练视频分析时,利用帧内绿色像素的比例、镜头时间的长短、是否包含动作等线索制定向前推理规则,设立多个