文章编号: 1001- 2486(2010) 02- 0079- 06

基于 OWL 本体扩展的视频语义内容分析*

白 亮, 老松杨, 刘海涛, 卜 江, 陈剑赟
(国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 随着可获得视频数据的快速增长, 迫切需要有效的方法在语义层理解和管理视频数据。对 OWL 语言进行扩展, 提出了 V- OWL 本体描述框架, 支持视频内容蕴含的时空关系和不确定性关系的建模, 使用基于贝叶斯网络的 B- 图描述模型, 将 V- OWL 本体概念、关系映射为 B- 图中的节点、边, 利用贝叶斯网络训练推理算法实现视频高层语义的自动推理发现。实验结果显示, V- OWL 本体描述框架对复杂视频内容具有很好的描述能力, 基于 V- OWL 的视频内容分析框架对视频高层语义探测具有较高的查准率和查全率。

关键词: OWL 本体扩展; 视频语义内容分析; 贝叶斯网络
中图分类号: TP391 文献标识码: A

Video Semantic Content Analysis Using Extensions to OWL

BAI Liang, LAO Song- yang, LIU Hai- tao, BU Jiang, CHEN Jian- yun
(College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Due to the rapid increase in the amount of available video data, there has been a growing demand for efficient methods to understand and manage the data at the semantic level. In this paper, the V- OWL is proposed with extensions to OWL, which can describe complex video content including temporal- spatial and uncertain relationships. The B- Graph description model based on Bayesian Net is proposed to map the concepts and relationships in V- OWL ontology into the nodes and edges in B- Graph. Video semantic content can be discovered automatically by using existing training and reasoning methods of Bayesian Net. Results from experiments show that V- OWL has achieved good description of complex video content, and satisfactory precision and recall of high level semantic content detections.

Key words: extensions to OWL; video semantic content analysis; Bayesian net

随着视频资源的快速增长, 迫切需要智能的方法在语义层对视频数据进行理解、存储、索引和检索[1- 3]。

为了共享视频内容的语义层描述, 需要统一定义概念化的语义内容。因此, 使用本体规范化视频语义概念定义成为视频内容分析领域必然的发展趋势[1- 2]。普通的本体已经成功应用在多个多媒体系统中来进行多媒体概念描述[4- 5], 但依然面临两个主要问题: (1) 相同的概念通常表现出某种观察模式, 包括视觉的和听觉的感知特征, 因此概念的本体化定义应该能够建立起概念化模型和基于特征的媒体内容模型, 使之能够相互映射; (2) 视频语义内容的理解需要特定的语境(Context), 这种语境通常反映为特定的背景知识, 而传统的多媒体应用没有对概念、特征和语境提供统一的知识表示和应用模型, 通常只是针对特定问题或领域进行知识表示, 缺乏规范统一的语义描述。

一方面, 已有研究工作主要采用 RDF(s)/OWL 等标准本体建模语言构建描述视频语义内容的领域本体[6- 7], 或是采用本体语言对 MPEG- 7 术语进行重新定义, 来描述视频语义内容[8]。但现有的 RDF(s)/OWL 标准难以支持视频语义内容复杂特性的描述, 因而无法有效完整地构建面向视频语义内容分析的本体。另一方面, 从传统的基于内容的思想出发, 一些研究工作试图根据视频低层特征, 对视频的感知、结构、统计等模式进行建模来描述视频概念[9- 10], 但在这些工作中, 面向视频语义内容分析的低层

* 收稿日期: 2009- 08- 29
基金项目: 国家自然科学基金资助项目(60902094)
作者简介: 白亮(1978—), 男, 博士生。

特征模式都没有与任何领域知识的形式化表示相联系,即没有有效地在语境中(Context)分析视频语义内容;同时,低层特征模式可以较好地描述特定的视频概念,但是难以表示概念的时空关系模式,这为高层语义内容的探测带来了困难。

本文以OWL语言为基础,对其进行扩展,提出了V-OWL本体描述框架,以支持面向视频语义内容分析的本体构建。结合成熟的贝叶斯网络模型,提出了基于V-OWL的视频内容分析方法。

1 OWL扩展框架——V-OWL

OWL是本体建模语言的标准之一,本节提出面向视频语义内容分析任务的OWL本体扩展框架——V-OWL。为了增强V-OWL的标准化和互操作性,对于视频低层特征模式,采用标准的MPEG-7视音特征描述符进行描述,同时可根据具体应用使用MPEG-7 DDL定义新的特征描述符。根据视频语义内容分析的特点和需求,定义新的V-OWL类和属性如下:

定义1 概念与个体

视频内容蕴含的高层语义内容通常可描述为现实世界中的概念、个体及其概念层次结构。使用OWL描述概念与个体是合适的,其分别对应于OWL中的类(Classes)和个体(Individuals)。因此,在V-OWL中采用与OWL相同的构造术语和方法描述视频内容包含的高层语义概念和个体。

定义2 视觉内容与特征

可观察的视频内容包括视听感知内容和低层特征。视听感知内容包括视频本身(例如一个视频片断),以及具有相似感知特征模式的视听内容。为了描述视觉内容和特征,定义V-OWL类:

VContent:表示视觉内容和特征的抽象,是类OWL: Object的一个子类;

VClip:表示一个视频片断,是类V-OWL: VContent的一个子类;

VPerception: 描述视频的感知特征模式,是类V-OWL: VContent的一个子类;

VFeature: 描述视频的特征,是类V-OWL: VContent的一个子类。

基于上述定义,可以进一步通过实例化owl: ObjectProperty定义V-OWL类的属性(Property)关联感知特征模式与特征、概念与感知特征模式、概念与视频片断。

VPerception类可以通过属性hasFeature与VFeature关联,高层概念可以通过hasVClip与一个具体的视频概念关联,同时可通过hasVPerception与感知特征模式关联。

定义3 关系

关系是本体中的重要元素,是视频语义的重要载体和语义推理中的重要依据和线索。本体关系可以关联不同概念间的属性,同时也可以描述构成复杂语义内容的时空模式,例如一个事件由若干个概念的时空组合构成。相对于普通语言概念间的层次结构关系,视频内容具有更为复杂的关系模式,本文声明若干个owl: ObjectProperty的子类来定义V-OWL中的关系,以支持描述视频语义内容蕴含的不同类型关系。OWL定义了多种描述层次关系的属性,V-OWL使用相同的层次关系描述属性构造方法。除层次关系之外,视频语义内容还蕴涵有其他更为复杂的关系模式,这些关系模式是OWL不能支持的,本文主要考虑时序关系、空间关系和不确定性融合关系。

定义3.1 时序关系(Temporal Relation)

视频媒体的一个关键特征是时间性质:一方面视频媒体的表现需要时间;另一方面媒体的时间性质还包含媒体在时间坐标轴上的相互关系,即时序关系。在视频中,时序特性体现了视频拍摄和编辑的规则,时序模式也蕴含了丰富的语义。

本文中,V-OWL定义属性vowl: TemporalProperty描述不同概念和视觉内容间的时序关系,TemporalProperty是owl: ObjectProperty的一个子类。本文中考虑7种时序关系,包括:before(b),meets(m),during(d),overlaps(o),starts(s),finishes(f),equal(e)。

定义3.2 空间关系(Spatial Relation)

与视频媒体时间性质相对的是视频媒体的空间性质:一方面视频媒体的表现需要空间;另一方面空间性质将环境中各种表达信息的部分按照相互的空间关系进行组织,全面整体地反映信息的空间结构,

而不仅仅是零散的信息片段。特定的空间信息结构往往包含特定的语义表达。

本文中, V-OWL 定义属性 vowl: SpatialProperty 描述不同概念和视觉内容间的空间关系, SpatialProperty 是 owl: ObjectProperty 的一个子类。

定义 3.3　不确定性融合关系 (Uncertain Combined Relation)

视频媒体的另一个关键特征是融合性。视频媒体从最初的制作到最终的消费都体现了融合性: 不同的图像序列、对象、伴随音轨、字幕通过多通道获取, 最终对用户统一地合成表现, 表达视频蕴含的语义。

视频中高层语义概念通常由多个概念融合而成。但这种融合关系不是固定不变的, 具有不确定。例如, 进球事件中通常会有球门出现, 但是球门的出现不一定都表示进球事件发生。在这个例子中, 低级概念球门与高级概念进球事件的关系是不确定的。对于特定领域, 不确定性问题通常利用求解相关"分量"的联合概率分布的方法建模解决。视频中组成融合关系的不同"分量"往往表现为"离散状态", 例如"球门", 就只有"出现"和"不出现"两种状态。因此, 可以使用条件概率表(Conditional Probability Table)来描述本体中的不确定性融合关系。在 V-OWL 中声明类 CPTObject 来描述条件概率表建模的不确定性融合关系。一个与概念 A 关联的 CPTObject 实例可描述如下:

```
< vowl: CPTObjects rdf: ID = 'CPT - A' >
    < vowl: RelevantConcept > A < /vowl: relevantConcept >
    < vowl: NumberofStates > 2 < /vowl: NumberofStates >
    < vowl: Conditionalon > b c < /vowl: Conditionalon >
    < vowl: NumberofStatesforChild > 2 2< /vowl: NumberofStatesforChild >
    < vowl: ValueofCPT > p1 p2 p3 p4 p5 p6 p7 p8 < /vowl: ValueofCPT >
< /vowl: CPTObject >
```

上述实例描述了一个针对概念 A 的给定条件概率表对象, 该条件概率表对象对概念 A 和与其关联的因素 b, c 之间的不确定性关系进行了建模。其中, 构造符 *RelevantConcept* 指明相关联的概念实例; *NumberofStates* 描述该概念具有的状态数; *Conditionalon* 指明该条件概率表的两个条件因素 b 和 c; *NumberofStatesforChild* 描述条件因素具有的状态; *ValueofCPT* 指明该条件概率表中的条件概率值。

2　基于 V-OWL 的视频内容分析

视频内容分析的任务是根据观察值(低层特征), 自动地从视频中抽取语义, 即建立视频内容与特定语义概念的关联。基于 V-OWL 的视频内容分析是在给定 V-OWL 本体下, 根据本体定义的概念、概念间关系和观察模式, 自动推理、计算本体概念与特定视频内容的概率关联, 即特定视频内容包含特定语义概念的概率。这里需要解决两个问题: 一是根据本体描述的领域知识, 即概念和概念间关系, 建立可推理计算的视频高层语义描述模型; 二是根据本体定义的关系和观察模式进行推理, 计算高层语义的出现概率。

视频高层语义通常表现为低层特征线索和其时空关系、融合关系的总和。视频媒体在制作过程中具有较强的主观性, 相同的语义内容可能采用不相同的表现手法, 即低层特征线索与高层语义关联具有不确定性, 所以需要采用基于概率统计的方法描述这种不确定性。本文中采用贝叶斯网络来对视频高层语义进行建模。

贝叶斯网络[11] 是图论和概率论相结合的产物, 它是一个带有概率注释的有向无环图, 是一个能够充分利用领域知识和样本信息的模型。贝叶斯网络用弧定性地表示变量之间的依赖关系, 也就是对知识或者规则的建模; 用概率分布定量地表示依赖关系的强弱, 将先验知识与样本信息有机结合起来, 以解决模式分类问题。

对视频高层语义进行建模的贝叶斯网络不同于 V-OWL 中的概念关系图, 它描述的是客观观察到的高层语义事实中概念和概念间的依赖关系, 这里我们称之为 B- 图。在给定 V-OWL 编码下, 本文提出了 B- 图的生成算法, 可自动生成不同高层语义的 B- 图。

B- 图生成算法如下:

步骤 1: 节点映射。

本体中定义的概念映射为 B- 图中的概念节点(Concept Node); 本体中定义的直接可见的观察模式为证据节点(Evidence Node)。

步骤 2: 关系映射。

时空关系映射: 时空关系是特殊的观察模式表现, 同时关联多个概念节点和证据节点, 因此定义时空关系为一种节点, 称为关联节点(Associated Node)。

除去时空关系, 本体中定义的其他关系映射为 B- 图中的边。

步骤 3: 方向映射。

本体中的概念层次关系通过 B- 图中边的方向表示。

步骤 4: 不确定性映射。

根据本体中定义的描述不确定性关系的 CPTObject 将相应的条件概率值指派给相应的节点。

经过上述步骤, 在给定 V-OWL 编码下, 可自动生成描述视频语义内容的贝叶斯网络图。在 B- 图中, 每一个节点都是离散节点, 具有两个状态, 即在相应视频片段中该节点描述的内容是否出现。除过根节点, 每一个节点都具有一组概率参数值, 对应于其父节点的各个状态取值。

这里以体育视频中"进球"高层语义为例, 说明高层语义建模和 B- 图的生成过程。根据对足球领域知识的总结, 可以发现:

ㄋ 进球事件发生时一定会出现球门;

ㄋ 进球后导播一般会使用近景镜头表现球员们的庆祝镜头, 同时伴随有观众的欢呼声;

ㄋ 进球后, 导播一般会使用慢镜头重播上述精彩镜头以飨观众。

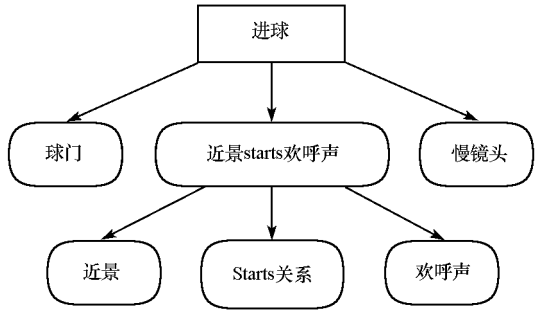

图 1 "进球"语义 B- 图

Fig. 1 B-graph for scored goal semantic event

通过总结上述领域知识, 在 V-OWL 本体中将"进球"语义概念与"球门","慢镜头","庆祝场面"等观察模式相关联;"庆祝场面"通过本体中定义的时序关系属性可以描述为"近景 starts 呼唤声"。采用提出的 B- 图生成算法, 可以得到"进球"语义的 B- 图描述, 如图 1 所示。其中,"进球"节点为概念节点,"近景 starts 欢呼声"(欢呼场景) 为关联节点, 其他为证据节点。

在得到映射完成的 B- 图后, 就得到了一个拓扑结构已知的贝叶斯网络, 下面的问题就是利用给定样本数据确定网络的概率分布, 然后推理高层语义内容的出现概率。基于 B- 图的语义分析就是在得到对高层语义进行建模的 B- 图后, 判断该高层语义在测试视频出现的概率。这包含贝叶斯网络应用中的两个重要问题: 训练和推理。

训练过程是根据训练样本计算对应 B- 图中节点的本体相关 CPTObject 的属性值。贝叶斯网络的训练学习包括网络结构学习(Structure Learning) 和参数学习(Parameter Learing)。通过本体对领域知识和视频拍摄规则的建模, 对视频高层语义建模得到的 B- 图对应的贝叶斯网络的拓扑结构是可知的, 只需要训练学习以求得贝叶斯网络的参数分布。本文采用最大似然估计训练获取贝叶斯拓扑结构中二值节点和连续节点的参数分布。最大似然估计计算给定父节点时, 节点不同取值的出现频度, 并以此作为该节点的条件概率参数。

确定贝叶斯网络的结构和参数之后就需推理, 即计算高层语义出现概率的过程。也就是给定贝叶斯网络中某些节点的值, 求未知节点(也就是变量 x_i) 取值 α 的概率, 即计算 $P(x_i = \alpha \mid E)$。对于非常简单的图模型, 可以根据贝叶斯公式直接求解。但是一般直接求解的时间复杂度过高。因此, 近年来针对图模型的推理, 已经提出了许多算法。本文采用有向图模型的经典推理算法 ——Junction Tree 作为 B- 图的推理算法[12]。该算法首先通过 Moralization 步骤, 连接具有相同子节点的节点, 并且忽略边的方向, 得到 Moral Graph; 然后通过 Triangulation 步骤, 添加无向边, 使得所有长度大于 3 的环都存在连支

(Chord), 即环上非相邻节点之间都有直接连线; 最后, 以图中每个极大完全子图作为节点, 构造一棵树, 即 Junction Tree, 并利用证据传播以及反传等方式, 计算边缘概率求得所需结果。

3　实验结果与分析

为了验证本文提出的 V-OWL 框架和视频高层语义分析方法, 在本文的工作中采集、观察了大量的视频数据, 针对体育视频领域中的足球视频构建了 V-OWL 本体, 并选取了若干足球视频中的高层语义, 构建了其 B- 图模型。采集的视频格式为 4∶2∶2 YUV PAL 制式、MPEG- 1 格式, 出自两个国外转播公司(ITV 和 BBC 体育), 均为 2006 世界杯比赛, 共 10 场比赛, 总时长 15 小时 36 分 52 秒。

如上所述, 足球视频高层语义的 B- 图拓扑结构已知, 只需要训练学习以求得其参数分布。首先, 抽取证据节点的低层特征(MPEG- 7 特征), 计算各个证据节点的概率参数; 然后, 将全部实验数据作为训练学习的样本数据, 采用最大似然估计训练学习获取 B- 图拓扑结构中二值节点(概念节点和关联节点)的参数分布。高层语义推理就是给定 B- 图模型中某些节点的证据值, 求未知节点取某数值的概率。在上述"进球"语义的模型中, 就是给出证据节点和二值节点的取值, 求"进球"节点为真的概率, 当"进球"节点为真的概率大于为假的概率值, 就认为该测试视频片段中含有进球事件。本文采用 Junction Tree 推理算法计算概率值。根据 V-OWL 本体, 构建了"进球"、"黄牌"、"换人"三种高层语义事件 B- 图模型。从实验数据中随机抽取了若干视频片段进行高层语义推理判别, 统计正确判定的高层语义事件数和误判的事件数, 分别计算查准率 Precision 和查全率 Recall 以评估实验结果。实验结果如表 1 所示。

表 1　高层语义事件探测实验结果

Tab. 1　Experiment results for high-level semantic event detection

语义事件	实际事件数	正确探测数	误判的事件数	查准率	查全率
进球	17	17	6	73.9%	100%
黄牌	31	27	7	79.4%	87.1%
换人	24	20	6	76.9%	83.3%
平均值				76.7%	90.1%

从表中可见, 基于 V-OWL 和 B- 图模型的足球视频高层语义事件探测的查准率和查全率的平均值分别为 76.7% 和 90.1%。同时, 本文使用相同测试数据集, 测试了传统的线性加权探测方法, 也就是说如果测试片段同时包含与某个高层语义直接相关的所有概念, 就认为该测试片段包含有该高层语义事件, 这样探测求得查准率和查全率的平均值分别为 57.4% 和 80.1%。本文提出的方法优于基于简单线性加权的融合方法, 并且通用性、鲁棒性大大增强。文献[6]采用 OWL 语言构建视频内容扩展本体, 对视频高层语义内容与感知特征之间的关联进行建模, 并采用本体推理技术自动探测视频高层语义, 采用该方法对足球视频中射门、任意球等精彩事件的平均探测准确率和查全率分别为 85.7% 和 69.0%。可见, 本文提出方法优于文献[6]的方法。

此外, 分析表中的数据可以看出, 查全率的平均值高于查准率的平均值。经分析可知, 在足球视频中, 一些不同的语义内容具有相似的 B- 图模式, 例如: 有一些精彩的射门镜头, 虽然没有破门得分, 但是也出现了球门, 导演也重播了相关镜头; 或者因为越位、犯规导致进球无效, 这些情况下会出现误判为"进球"的情形。类似地, "黄牌"和"换人"语义事件也会由某些相似内容导致误判。但是事实上, 这些相似的语义内容往往也是观众非常感兴趣的语义事件。从某种意义上说, 这种误判反而对实验结果有利。

但实验结果也提示我们, 本文提出方法的性能取决于高层语义建模的准确性, 即 B- 图对高层语义描述的准确性, 而 B- 图对高层语义的描述又基于 V-OWL 本体对体育视频领域知识的抽象总结, 这种抽象总结的概括性降低了描述模型对于语义细节的区分能力, 而往往有些语义细节就决定了不同的语义内容, 例如某些较严重的犯规和黄牌, 从视频内容来看唯一的区别就是裁判会出示黄牌, 所以区分二者的有效因素就是黄牌对象的探测, 但这往往又是比较困难的。因此, 进一步提高语义推理准确率需要从语义内容建模的全面性、准确性和视频内容探测性能两个方面共同努力。

4 结论与展望

使用本体全面、规范的描述视频语义内容、关系及其“语境”中的知识是提升视频语义内容分析性能的有效方法和发展趋势。本文对 OWL 语言进行扩展,提出了 V-OWL 本体描述框架,可以较全面地描述视频语义内容。提出的本体编码到 B- 图的映射方法可以准确地将本体描述自动转化为可推理计算的贝叶斯网络模型,实现视频高层语义的自动推理发现,有效地增强了视频分析性能。实验结果显示,提出的视频内容分析框架对视频高层语义探测具有较高的查准率和查全率,相对于传统的线性加权探测方法,探测性能、通用性、鲁棒性大大提高。

未来的工作需要进一步结合视频内容的特点,扩展现有本体语言的描述能力,以增强视频语义内容描述的全面性和准确性;同时,新的更具区别性的视频低层特征和分类算法地提出也依然是未来具有挑战性的工作。

参 考 文 献:

[1] Rowe L, Jain R. ACM SIGMM Retreat Report on Future Directions in Multimedia Research[J]. ACM Transactions on Multimedia Computing, Communications and Application, 2005, 1(1): 3- 13.

[2] Chang S F, Ma W Y, Smeulders A. Recent Advances and Challenges of Semantic Image/Video Search[C]// IEEE International Conference on Signal Processing, 2007.

[3] Hanjalic A, Xu L Q. Affective Video Content Representation and Modeling[J]. IEEE Transactions on Multimedia, 2005, 7(1): 143- 154.

[4] Snoek C G M, Worring M, Gemert J C, et al. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia[C]// Proc. ACM Multimedia, Santa Barbara, CA, 2006: 421- 430.

[5] Naphade M, Smith J, Tesic J, et al. Large-scale Concept Ontology for Multimedia[J]. IEEE Multimedia, 2006, 13(3): 86- 91.

[6] Bertini M, DelBimbo A, Torniai C. Enhanced Ontoloies for Video Annotation and Retrieval[C]// Proc. ACM MIR' 2005, Singapore, November 10 - 11, 2005.

[7] Bai L, Lao S Y, Smeaton A F. Video Semantic Content Analysis Based on Ontology[C]// International Machine Vision and Image Processing Conference, Maynooth, Ireland, September 5- 7, 2007.

[8] Hunter J. Adding Multimedia to the SemanticWeb: Building and Applying an MPEG- 7 Ontology[C]// Chapter in Book: Multimedia Content and the Semantic Web, Kluwer Press, 2005.

[9] Ekin A, Tekalp A M, Mehrotra R. Automatic Soccer Video Analysis and Summarization[J]. IEEE Transactions on Image Processing, 2003, 12(7): 796- 807.

[10] Sadlier D A, Ó Connor N E. Event Detection in Field Sports Video Using Audio-visual Features and a SVM[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2005, 15(10): 1225- 1233.

[11] 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.

[12] Huang C, Darwiche A. Inference in Belief Networks: A Procedural Guide[J]. International Journal of Approximate Reasoning, 1996, 15(3): 225 - 263.