# 多媒体语义模型研究进展

栾悉道[1,2] 谢毓湘[2] 谭义红[1] 陈治平[1] 赵碧海[1] 胡 赛[1]
(长沙大学信息与计算科学系 长沙 410003)[1] (国防科学技术大学信息系统与管理学院 长沙 410073)[2]

**摘 要** 多媒体语义研究是多媒体数据处理与多媒体信息服务领域的核心和关键问题。多媒体数据的语义问题源于多媒体的数据获取方式,在多媒体数据的应用阶段,这一问题成为制约多媒体数据使用和创作的重要瓶颈。语义模型研究是多媒体语义研究的重点,是多媒体数据处理过程的总结和抽象,其实质就是研究多媒体数据整个生命周期的语义问题。介绍了近几年多媒体语义模型在内容描述、语义表示、数据检索三个方面的研究进展情况。

**关键词** 多媒体语义模型,内容描述,语义表示,检索模型

## Research Development of Multimedia Semantic Model

LUAN Xi-dao[1,2] XIE Yu-xiang[2] TAN Yi-hong[1] CHEN Zhi-ping[1] ZHAO Bi-hai[1] HU Sai[1]
(Department of Information and Computing Science, Changsha University, Changsha 410003, China)[1]
(College of Information System and Management, National University of Defense Technology, Changsha 410073, China)[2]

**Abstract** Multimedia semantic research is one of the most important issues in multimedia analyzing and multimedia information service field. This issue originated from multimedia data's capture becomes an important bottleneck in multimedia applications. Multimedia semantic model is a summary and abstraction of multimedia processing, and focuses on all semantic problems existing in the lifecycle of multimedia data. The theme introduced the research developments of multimedia semantic models in content description, semantic representation and retrieval models.

**Keywords** Multimedia semantic model, Content description, Semantic representation, Retrieval model

## 1 引言

通常,多媒体是指表示媒体的多样化,常见的有文字、图形、图像、声音、动画、视频等多种数据形式。目前多媒体数据的研究与应用,都是围绕着如何为用户提供更好的多媒体信息服务,也就是广义的视听觉服务来展开的。其中,广义的视听觉服务是包含文本、图像、视频等多种媒体数据所提供的服务。

从多媒体分析应用的角度来说,可将视觉特征对应地分为间接视觉特征和直接视觉特征两部分。间接视觉特征指的是基于文本的特征(如关键字、注释等),其语义内容的理解需要知识的密切参与;直接视觉特征,如色彩、纹理、形状、对象表面等,可以从多媒体数据本身分析得到。间接视觉特征在信息系统等领域研究得比较深入。直接视觉特征又可以分为通用的视觉特征和领域相关的视觉特征。前者用于描述所有图像、视频共有的特征,与素材的类型或者内容无关,主要包括色彩、纹理和形状;后者则建立在对所描述内容的某些先验知识(或者假设)的基础上,与具体的应用紧密相关,例如人的面部特征或指纹特征等。

对于某个特定的特征,通常有许多不同的表达方法。由于人们主观认识上的差别,对于特定特征并不存在一个所谓的最佳表达方式。事实上,特征的不同表达方式从各个不同角度刻画了该特征的某些性质。另外,目标特征的提取与刻画一般来说并不能做到完全和完备,因而无法将特征与目标对象做到严格的对应。

多媒体数据特征的研究并不是最终目的,在多媒体研究与应用领域,多媒体语义问题的研究才是关键。

## 2 多媒体语义问题研究的困境

目前多媒体语义研究的关键,还集中在如何克服语义鸿沟问题。随着计算视觉与图像理解的研究深入到多媒体索引与检索领域,产生了基于模型的方法。该方法就是把内容中所探测到的概念,包括对象、事件等的低层特征通过监督学习的方法进行训练。如果应用领域已知,这些方法可以取得较好的效果,例如在体育视频中基于内容的描述。但是,总的说来,结果极度依赖于与概念相结合的低层特征。从TRECVID[1]所介绍的研究结果来看,抽象语义概念的探测率是非常低的。

目前,描述和检索媒体的内容,如图像、视频或者声音片段,大都使用文本技术以及元数据。文本与元数据可以手工添入,但这是一个费时费力的过程;也可以使用文字识别技术(OCR)得到,或者使用自动语音识别技术(ASR)得到。直接

到稿日期:2009-12-10 返修日期:2010-02-25 本文受国家自然科学基金(60802080),863高技术项目(2009AA01Z335),长沙市科技计划项目(k0902210-11)资助。

栾悉道(1976-),男,博士,讲师,主要研究方向为多媒体信息系统,E-mail:xidaoluan@sina.com;谢毓湘(1976-),女,副教授,主要研究方向为多媒体信息系统。

从媒体数据中抽取对象并以自动语义标注技术为之命名，要比媒体的获取困难得多，更不用说揭示各个对象在一起所要表达的各种抽象语义。在这一点上，不得不借助于知识的辅助。

另外，用户对多媒体数据的使用大多还停留在浏览阶段。人们被动地得到在某地拍摄的图像或者视频，了解在记录中所发生的事件。对于这种经过采集、压缩等处理后的数据，如果想对其进行进一步的应用，存在着诸多的困难。

首先，这些数据是经过融合了的。这包括两个方面：一种是物理层次的融合，是指通过数据采集设备，如照相机、摄像机，得到某一时刻、某一地点所发生的某一事件的片段。这些数据采集设备的设计与实现原理决定了它们无法得到所采集对象的全部数据。也就是说，这个数据采集过程并不完全，丢失了或者无法采集到一些对象的固有属性和关系，如对象间的空间层次关系。对于普通照相机这种采集设备来说，还丢失了对象间的时间关联关系。因此，对于一张图片和一段视频，用户能够对其再利用的空间非常有限。

第二，用户在浏览多媒体数据时，对于数据内容的理解与判断，是基于自身所积累的知识和经验。在信息不完全或者扭曲的情况下（如图像中的物体被遮挡），甚至是变形的情况下，用户能够根据周围场景、上下文等其他信息特征，对数据所要表达的对象和事件做出正确的识别和判断。另外，人具有感知及情感基础，可以从多媒体数据中得到关于（空间和时间）组织结构、喜怒伤悲等高级认知。这是人全面参与多媒体数据内容理解与分析的情况。在这种情况下，人是作为处理的主体出现的，以人的知识和经验为基础，对多媒体数据的内容、高低语义，都能很好地识别和判断。这种情形大多适用于多媒体数据量不是很大或者要精确了解数据内容的情况。对于数据量非常庞大，而且无需全面、细致了解各种语义的情况，这种以人为主体的处理方式，就不是十分适宜。

以计算机作为多媒体数据理解与处理的主体，虽然取得了很多的进展，但也面临着巨大的困难。这种困难的关键在于，如何让计算机像人一样，可以根据多媒体数据所包含的各个特征，来了解其所要表达的信息。目前对这个问题的研究，有两个最为关键的过程和阶段，即针对所要提取的内容定义并选取特征，以及从多媒体数据中提取特征。在这两个阶段中，人对表达特征的选择，以及用程序来自动分析多媒体数据特征，都存在特征的选取与提取是否完全和完备的问题。即便这两个过程能够得到充分的保证，也要对多媒体数据中存在的每种（或者每类）对象进行特征的选取与提取，才能保证计算机能够对所有的对象进行识别。而大千世界对象的种类繁多，形态各异，要全部进行特征的选取与提取，显然是不可能的，更不用说以心理、感知等知识为基础的高级语义的识别和判断。经过十多年的发展，TRECVID 在 2006 年发布的高级特征任务研究（High-Level Feature Task）中也只包括了 39 种特征，如 sports，weather，office，meeting，animal 等。因此，在现有情况下，目前的计算机还不能胜任多媒体数据内容理解的重任。

第三，声音、图像与视频等多媒体数据的制作和使用问题。现今，在多种软件的辅助下，用户已经可以制作一些动画、图像和视频。制作这些媒体所需的素材一部分由手工绘制或采集，另一部分则来自过往的图像和视频。对已有素材的使用主要存在两个问题：一是素材的检索与选取问题；二是素材的使用问题。多媒体素材中对象的存储与表现大多不是独立的。通常，关注的对象都处于某些背景、环境之中，或者与其他对象发生遮挡。因而这些素材使用起来相当不便。因为缺少所需对象的光照、空间等数据，从图像或者视频中截取出来的部分还需要经过相当多的处理。这使得图像与视频等多媒体数据无法像文本一样，能够被绝大多数人所轻易使用与制作。视听觉感受优良的图像与视频，只有专业人士才能完成，无法为普通用户提供更加有效的控制和使用数据的手段，限制了多媒体的应用。

另外，多媒体数据的应用还面临几个比较棘手、普遍并且重要的问题：

· 适应性与一致性

因为同一张图像、视频可能会面对许多种不同的重用需求，可能要根据许多不同的、无法预料甚至可能是矛盾的需求对同一种内容进行解释。这就需要媒体数据内容以某种更有条理的语义领域模型而不是关键词来描述，可以根据“概念相似”来进行选择、分类或者聚类。

语义数据涵盖媒体所要表达的内容和对象，将它们联接成领域概念集，在媒体数据本身与“概念世界”中架起桥梁。多媒体数据及其内容一定要和这些概念联系起来，才能实现一个理想的多媒体信息系统。

目前通常使用关联的关键词进行语义标注，通过索引来组织，并使用 SQL 查询语句进行检索。如果不将关键字词赋予媒体数据，没有一个通用的词汇表，就不能实现精确检索。但是，关键词系统通常被认为是孤立于任何领域模型，使用关键词时也很少使用结构信息，例如关键词的同义词。因此，关键词的标注经常是独立的，不能成为一个结构良好、具有一致性的模型。

· 动态分类及重用性

还有一种方法，是预先定义关键字词的标注。这虽然可以保证一致性，但是规定了在多媒体数据中“应该是什么”，而不是确实描述“那里真正有什么”。事先规定好关键词使得这种标注是静态的，不能扩展，而且内容的解释不能支持推理，也无法支持其他应用。因此，对于同样的一些数据，要想能够根据不同的应用需求进行重用，需要对数据的描述进行具有适应性的扩展，并能够对同一内容的解释进行动态分类，虽然这些解释不可预测，甚至可能彼此矛盾。例如，一幅“安南秘书长出访中国，他走下飞机并向人群挥手致意”的图像，可以根据其内容分为秘书长、男人、加纳人、飞机、人和飞机等许多种类。如果要找一幅包含飞机的图像，这幅图像虽然主要是为了表现一个人，但也应该被检索到。要使得数据能够根据其内容被正确地分类，需要将这些数据的内容以某种有条理的语义领域模型描述出来，而不是依赖于对关键字词的使用。

· 框架扩展与重分类以及不完全性

对于绝大多数的多媒体数据库，如图像数据库、视频数据库、新闻数据库等，这些数据库在描述多媒体数据时，和传统数据库一样，预先定义并不能适应说明性的数据库框架。只有通过不断扩展或者进化，才能将数据实例与框架联系起来，或者从描述实例开始扩展框架。实际应用决定了要有相应的基础与之相匹配，才能将框架扩展。这些框架在将这些实例联系起来之前，就已经存在了。例如，“安南秘书长”这一概念

在图像将“秘书长”与“安南”这两个概念联系之前就已经存在了。实例开始时，没有标注，将他们与框架联系起来时，会揭示更多的内容，也会需要检索更多的内容。因此，实例一直是不完全的，或者描述的完整性一直在变。而描述在被扩展之后，需要重新归类。因此这种不断增添、不断扩展的框架支持是非常关键的。

·基于标注的相似性检索

不准确或者不完全的多媒体数据内容描述将导致不准确、不精确的匹配和查询。不过，如果检索出来的数据是那种“概念相似的”，这在许多应用中也非常有用。例如，“检索有关安南秘书长的视频”。但是，实际上用户真正感兴趣的是所有有关领导人物的数据，或者，如果没有安南的视频，那么一个相关领导人物的视频或者一幅安南的图像，也完全可以替代。

## 3 多媒体语义模型研究进展

多媒体语义模型也是比较宽泛的概念和研究领域，它涉及多媒体的内容和数据模型、表现模型和检索模型等多个领域，它是每个多媒体数据处理过程的总结和抽象，其实质就是研究多媒体数据整个生命周期的语义问题。

多媒体语义模型应具有一定的基本属性[2]：

(1)模型的提炼性：语义模型的设计与内容应以多媒体数据为基础，但要高于多媒体数据本身。建立多媒体数据语义模型的目的，是要能够表达多媒体数据的语义内容结构、指导相关的各种语义处理过程。

(2)模型的继承性：语义模型的继承性有多种含义。一方面，语义模型本身具有一定的继承性，是一种模型在演变与进化过程中的继承，是模型发展与多样性的前提条件；另一方面，语义模型的内容，即语义概念和关系存在着继承。各种概念与关系并非独立存在，在相关的语义对象之间存在着继承上层对象对应属性等内容的现象。

(3)模型的自适应性和可扩充性：语义模型的设计，要能够适应不同语义领域的应用需求。语义模型的自适应性与可扩展能力，对于语义模型的生存能力具有至关重要的意义。一个灵活、高适应性的多媒体数据语义模型将具有更大的通用性。

### 3.1 多媒体内容描述模型

目前可用于多媒体数据内容描述的模型有很多，如 Dublin Core[3]和资源描述框架 RDF(Resource Description Framework)[4]等。它们对多媒体数据内容的描述比较粗糙，缺乏对视频中镜头、场景等结构单元信息以及运动对象等高层语义信息的描述。RDF 是一个用于元数据的框架，其目标是提供一种适合描述任何领域信息的机制，强调工具可以自动处理 Web 资源。以 XML 为基础，W3C 提出了多媒体同步综合语言 SMIL(Synchronized Multimedia Integration Language)，采用多媒体表现的 Web 开放式设计，使用户可以像看电视一样通过浏览器享受 Web 漫游[5]。

目前从视听角度来综合探讨多媒体数据语义的内容描述模型，主要采用 MPEG-7。它的目标是作为一种描述多媒体内容数据的标准，满足实时、非实时以及推-拉应用需求。自 MPEG-7 成为 ISO/IEC 的标准以来，一直受到众多研究者的关注，基于 MPEG-7 的多媒体数据语义相关研究有很多。

1)多媒体内容描述项目和框架

哥伦比亚大学的 A. B. Benitez 教授发起的 MPEG-7 Project[6]项目，集中研究用于图像、视频等多媒体集合的符合 MPEG-7 标准的描述模式的开发。OpenDrama[7]项目将 MPEG-7 用到戏剧的创作中，采用 MPEG-7 中标准的描述模式对多媒体进行描述，在时间和空间上描述多媒体内容的结构。在这些描述的基础上开发了一个叫做 Medefi 的多媒体著作工具，但是该项目仅仅集中在多媒体结构信息的描述研究上，对结构信息之上的语义、抽象概念考虑得不多。Panorama[8]是一个用于完成视听信息及其附属信息的数字化存储和检索系统。它开发设计了一系列的算法来抽取视觉特征，例如颜色和纹理信息、移动的物体、人脸以及文字区域。另外，这个基于 Web 的系统使用了 MPEG-7 的 XML 框架以实现子系统之间信息的交互。虽然，MPEG-7 对于建立一个通用系统来整合现有的多媒体检索技术具有重要意义，但是，MPEG-7 对语义信息的支持非常少，并没有提供相应的检索手段。

2)图像的语义标注与检索

奥地利格拉茨技术大学的 Caliph & Emir 项目[9]采用 MPEG-7 标准对图像进行标注，除了可以实现 Who，What，When 等语义标注以及底层特征的处理，最大的特色就是用图形化的方式与用户进行交互，实现高层语义的标注。

3)视频建模与处理

文献[10]介绍了 COSMOS-7，它是一个针对数字视频的 MPEG-7 元数据建模与过滤框架。它引入了一个基于内容的过滤策略来提取用户的偏好。过滤操作的结果可能是元数据，也可能是满足过滤条件的帧序列。文献[11]将视频语义信息分为三层，并提出了一个基于 MPEG-7 的视频数据模型，运用 XML 实例阐述视频内容的视频对象、事件和元数据构造和描述方法，能够支持不同层次的复杂语义关系描述。IBM 研究机构开发了 VideoAnnEx[12]标注工具，以辅助用户用 MPEG-7 标准的元数据对视频中的帧序列进行标注，并将各种标注以 XML 形式存储。此外，Yavuz[13]使用 MPEG-7 描述工具建立了一个基于 ST-AVIS 的视频数据库管理系统。文献[14]提出了一个新颖的基于 MPEG-7 描述的视频自动摘要系统以及一种基于肤色过滤进行视频摘要的方法。

4)结合本体理论的 MPEG-7 研究

文献[15]运用本体理论，结合 MPEG-7 的视频特征描述接口以及视频语义信息的层次结构，建立了基于 MPEG-7 的视频语义信息模型。Tsinaraki 等人描述了一个与 MPEG-7 相结合的框架[16]，用于领域本体和视听内容的标注，并提出使用 XML 框架作为本体设计语言。他们还试图将领域词汇对应到 MPEG-7。例如，将音乐本体中的 Artist 概念映射到 MPEG-7 中的概念 CreatorType。在本体领域的很多情况下，在这些具有丰富语义的概念之间建立起相等或者包含关系并不合适。

Troncy[17]认为，多媒体文档的结构判断与其内容判断同等重要，将 MPEG-7/XML Schema 用于表示多媒体数据的结构语义。Bloehdorn 等人[18]描述了一个本体框架以及一个软件环境，用以基于原型方法在 MPEG-7 的低层视觉描述子与领域本体概念之间建立映射关系。文献[19]与 Tsinaraki[16]，Troncy[17]以及 Hunter 等人[18]的工作存在相似之处，使用语

义网本体语言来表示 MPEG-7 标准，以解决互用问题，并使用领域词汇表对该标准进行补充。

像其他的 MPEG 家族成员一样，MPEG-7 是满足一般需求的视听信息的标准表示。MPEG-7 强调的是提供新的视听内容描述方案，视听内容可以包含文本。MPEG-7 正考虑现有的其他标准组织开发的文本处理方案，并适当地支持这些标准。

作为多媒体数据中的一种常见数据类型，视频数据的内容处理经常会使用到文本、图像和声音的相关模型和技术。因此，这里将视频作为一个具有代表性的多媒体数据，介绍视频的内容模型相关研究情况。视频中的“内容”是指基于内容分析技术中的“内容”概念，主要包括视频结构模型和视频语义模型。

a)视频结构模型

视频结构模型把视频表示成一些连贯的基本结构单元的组合体，这些基本结构单元一般通过时序或者时空分割产生得到。基于镜头的结构模型以时序分割的镜头作为基本结构单元。镜头是摄像机一次连续拍摄动作所记录的图像序列及其伴随音轨。在镜头之上还存在场景、故事单元等视频结构单元，镜头是基本结构单元，镜头分割是视频结构层次化的基础。

早期的视频模型很多是用于表示视频的结构的。其中，每个镜头都用一个具有代表性的帧来表示，称为关键帧，再对每个场景用一个或数个关键帧来表示，从而形成镜头的目录。所有的镜头组合用代表帧的“线性”排列方式来表示，代表帧可用于快速浏览视频，代表帧的物理特征可用来建立索引；在检索时，根据关键帧的物理特征(颜色、纹理、形状等)来匹配用户所提供的示例查询。

b)视频语义模型

区别于结构建模，视频语义模型致力于开发高层概念化的模型，用于表示和管理视频中包含的对象、事件及关系等语义信息，并提供实现语义查询的基础。

在文献[21]中，王煜等人通过对 VIMSYS，AVIS，Videx，VideoGraph，Extended ExIFO2，BilVideo，THVDM 等系统的研究，对视频的语义模型进行了很好的比较和总结，认为视频语义模型可以分为 16 种，其中有 5 种是基于标注的模型，还有 11 种被称为“丰富语义模型”。丰富语义模型可以描述现实世界中的实体(如概念、对象、事件)以及它们之间各种复杂的关系。实体可以是出现在视频中的具体事物，也可以是抽象的概念，甚至可以是不在视频中出现但可以作为背景知识的信息。丰富语义模型通常采用分层的结构。最底层对应原始的视频流，最高层表示语义信息，这两层之间还可能存在一些中间层，如逻辑视频段层、特性层、媒体对象层等。这些模型采取的策略分为两类：将现有的模型扩展应用到多媒体领域表示视频内容，或针对视频应用设计全新的模型组件和表示方法。

### 3.2 多媒体数据语义表示模型

多媒体数据内容的语义表示，实质上是一种知识表示。由于多媒体数据的内容非常丰富，对于不同的用户，对内容的理解可能存在一定的差异。即便是同一用户，随着年龄、阅历、经验的增长，对于同一数据的理解也会有所不同。这种语义表示与理解的模糊、不确定性问题，一直是研究的热点。

首先，最为常用，也是最简单的语义表示方法，就是使用文本对图像、视频、音频数据进行解释和标注，相当于在这些数据上叠加一个文本标注层数据。对于视音频来说，每个标注信息要与数据的逻辑段相关联。这种标注式的语义表示方法，被用于针对视频语义的 OVID，VideoStar，CCM，VideoText，Smart VideoText 以及图像的 IRIS 等系统中。当然，这些系统在方法的实现与侧重上，还有一些不同。例如，Smart VideoText 是对 VideoText 的扩展，引入概念图表示自由文本标注中的知识[21]。文献[22]利用词典 WordNet 将文本表示的语义概念联系起来，以大大增强文本表示语义的效果，并提高检索的准确度。

使用文本进行多媒体数据的语义表示，其优点是可以描述一些高层、抽象的语义，比较直观且易于处理。但是文本描述不容易自动获取且存在内在的主观性，以及对于复杂的概念缺乏足够的表达能力，因此使用文本表示语义，虽然普遍，但并不是一个十分有效的手段。

其次，是使用人工智能领域中的知识表示方法，比如语义网络、数理逻辑、框架等方法，它们具有表达复杂关系的能力。最近的一些研究者使用了一些不同的语义表示模型，例如 Zhuang Y 等使用模糊布尔模型、概率布尔模型[23]；Bertin 等使用形式语言理论表示[24]；Fang 使用模糊逻辑语言[25]；Yoon 等人则使用符号语言学方法[26]。

也有学者使用 Petri 网进行多媒体数据的表示与检索研究[27,28]。文献[27]提出了一种基于基本语义单元合成 Petri 网的足球视频查询描述模型。该模型首先定义了一种类似文本字词集合的足球视频基本语义单元集合，在此基础上采用基本语义单元合成 Petri 网模型建立了一种足球查询语义的描述模型，并分别构建了进球、进攻、角球、犯规、换人等足球语义。该模型具有一定的有效性和推广性。

描述逻辑(Description Logic)是知识表示方法中的一种形式化方法，它起源于 20 世纪 70 年代的知识表示方法——语义网络(Semantic Network)和框架(Frame)系统。在众多知识表示的形式化方法中，描述逻辑受到人们特别关注的主要原因在于：具有严格、清晰的模型-理论语义；对概念性知识的处理，特别是对概念分层的处理非常有效；提供有效的推理机制，支持可判定的推理服务。

虽然，描述逻辑在自然语言处理中主要用于表现术语，但是如今越来越多地使用描述逻辑来建立数据模型，它在多媒体系统中也能够起到更大的作用，可以广泛用于标准化服务、模式分类、理解图像的逻辑、与知识库的接口、空间和时间信息的处理等多媒体数据分析与应用领域。

多媒体数据所包含的对象以及语义内容的判断，是多媒体数据分析中较难的部分。利用描述逻辑在分类、知识表示等方面的优势，有助于多媒体数据内容的分析与语义推理。文献[29]将模糊描述逻辑与多媒体数据的知识表示、推理和检索从形式上结合到一起；文献[30]则将描述逻辑与空间推理结合到一起。文献[31]使用描述逻辑来描述复杂的形状。但是，因为没有考虑形状的位置，无法对位置信息进行推理。文献[32]在描述逻辑结构方法的帮助下，建立起图像底层特征与对象所具有的复杂结构之间的映射关系，使用对象的结构描述来识别图像，并进行图像的分类和推理。文献[33]使用描述逻辑来描述 BCIO(Breast Cancer Imaging Ontology)，

用于辅助X光照片中癌变位置、概念的标注、诊断和推理。Antonio采用基于本体和描述逻辑的方法来融合各种数据集的地理信息[34]。该方法从语义的角度，以本体为核心定义了一个语义框架，该框架表示了各主题概念及各概念之间的关系，用于解决多媒体数据类型的主题地理数据（如图像、视频）的上下文索引和检索。

上述模型具有一定的语义描述功能，但很难在普遍性和面向领域的特定性方面取得一致。另外，这些模型具有主要的目标领域，如视频领域或者文本领域，并没有对多媒体数据从语义这个角度建立统一、一致的模型。

此外，使用MPEG-7来表示多媒体数据语义的研究，目前也是比较热门的领域。MPEG-7标准致力于制定一个标准化的框架来描述多媒体内容，以便有效表示和方便地检索多媒体内容。

### 3.3 多媒体数据检索模型

要实现多媒体数据的检索，先要对数据进行组织。图像、视频、音频等多媒体数据，并不能直接进行检索，需要先从这些原始数据中抽取逻辑视图。用户使用查询来表示其信息需求。检索系统根据查询的表示，搜索数据集，获取与该查询相关的数据，并根据相似性匹配的程度，按序返回检索结果。

在响应用户检索需求，对相关数据进行相关度排序时，不同的检索模型对相关度的衡量有着不同的方法。布尔模型、矢量模型和概率模型是信息检索中的三个传统模型。在布尔模型中，文档和查询被表示成索引项的集合，因此这个模型是基于集合理论的；在矢量模型中，文档和查询被表示成为维空间中的一个矢量，因此这个模型是基于代数理论的；在概率模型中，文档和查询的建模框架是基于概率理论的。近些年来，传统的模型都有了不同的改进和演化。在基于集合论的检索模型中，又提出了模糊布尔模型和扩展布尔模型。在代数型模型中，衍生出广义矢量模型、隐含语义索引模型、神经网络模型等三种。在概率型检索模型中，发展出推理网络模型等。

基于内容检索CBR（Content-Based Retrieval），是20世纪90年代发展起来的一种有效的多媒体信息检索方法。它是根据媒体的内容，包括视觉和听觉特性、时间和空间结构等，进行信息检索的一种方法。基于内容检索，就是根据其内容而不是外部属性，从多媒体数据集中找到与给定的查询请求相关的媒体子集。著名的图像检索系统包括IBM的QBIC系统，哥伦比亚大学的VisualSEEK、VideoQ，MIT多媒体实验室开发的PhotoBook，UC Berkeley开发的Chabot系统等。这些系统的共同特点就是利用多媒体内容的底层特征（如颜色、纹理、形状等）进行检索，但是底层特征相似的多媒体在语义上却未必相似。

有关多媒体数据语义的研究在传统的CBR中并没有占据主流的地位，这也与语义所涉及到的高层含义及其所面临的巨大挑战有关。用户在产生检索需求时，往往只存在所要描述的对象、事件以及表达情感等含义上的概念，用户需要返回的是多媒体数据的含义，而不是颜色、纹理、形状等特征。这些含义，往往指的是高层语义，它与人对多媒体数据的内容的理解息息相关，而这种理解无法直接从多媒体数据的视听觉特征获得，需要根据人的主观知识来判断。为了克服简单视觉特征的多媒体检索方法的不足，人们提出了基于语义的多媒体检索方法[35,37,38]。与基于底层物理特征查询不同，语义特征查询是基于文字的查询，包含了自然语言处理和传统的信息检索技术，其目标是最大限度地减少媒体底层特征与丰富语义之间的鸿沟。

Yang认为，语义鸿沟问题产生的根源在于无法从图像、视频等多媒体数据和用户的检索请求中准确地提取语义[36]。可以通过提取图像周围文本的语义来克服这种语义鸿沟，并通过自组织映射SOM来建立图像隐含语义与文本之间的相关性。该方法对数据来源与类型有一定的要求，例如对采自互联网的图像数据集能得到较好效果。Mats则提出了一种根据文本来推断语义信息的方法[37]，该方法是一种自学习方法。它通过对图像和视频进行训练，将它们的底层视觉和听觉描述子与文本特征相联系。通过底层描述子训练所得到的自组织映射将训练集中的一个语义类映射到检测集中的一类相似对象。实验结果证明可以进一步用于解决底层视听觉数据与语义概念之间的语义鸿沟问题。Urban等人[38]为了提高多媒体数据的检索效率，提出一种检索模型和学习框架的方法，以把上下文相关的特征融合到一起用于交互式信息检索。该方法将视觉与文本特征融合到一个统一的框架之中，并使用random walks理论来对这个模型进行查询。

国内越来越多的学者开始关注多媒体检索领域所面临的语义鸿沟问题，并开展了各自的研究。有代表性的研究单位包括微软亚洲研究院、清华大学、国防科技大学、浙江大学等。国防科技大学侧重于从多模态融合的角度建立分析和检索模型[39-41]；浙江大学[42]提出了一个多媒体交叉参照图检索模型，在该模型的辅助下，通过支持向量聚类引发各个单模态进行内容检索。鲍永生利用语义网络建立底层视觉特征与高层语义特征之间的关联，然后利用相关反馈技术提高检索的准确率[43]。文献[44]将PCA降维技术应用到视频语义标注中，并提出视频镜头语义匹配策略，同时采用改进的语义网络对标注的视频语义进行动态更新与扩展。

一个完整的检索系统应该既包括传统的基于内容检索，也提供语义级的检索。这就需要从以下三方面开展工作：(1)提供高层语义的描述和表达方式；(2)提供将低层的视听觉特征映射到高层语义的语义提取方法；(3)语义检索系统的语义处理方法，尤其是自然语言的处理。

**结束语** 从上面的叙述可以看出，一个理想的多媒体信息系统，它应该至少具备以下几个条件，才能满足多媒体数据内容与语义的应用要求：(1)具有推理功能，能够对概念的等价、包含及一致性等进行推理与验证；(2)具有灵活、可扩展的系统框架，能够实现内在的概念的扩展，以及外在的服务与应用上的扩展；(3)知识库的支持，不但包括概念间关联关系的知识，也能对多媒体信息服务与应用提供应有的支持。但是，这种提高也只是现有多媒体数据获取、处理模式下的一个补充，或许会缩小语义问题在实际需求与服务水平之间目前存在的巨大鸿沟，但并不见得能够真正解决多媒体数据领域的全部语义问题。

多媒体语义问题的出现，根源于多媒体数据的获取方式。在这种方式中，数据创作者与使用者分离。前期便利的数据“获取”方式，是以牺牲后期用户便利地“使用”多媒体数据为代价的。多媒体数据在获取过程中，损失了大量的数据，导致这一过程不可逆，多媒体数据中的对象、场景、事件等语义无法与现实世界对应。

要想从根本上解决这一问题，就要从多媒体数据的生成、获取方式上入手，提出新的多媒体数据获取方案，进而真正解决多媒体数据的语义问题。这也将是下一步研究、探索记录、分析现实世界的新思路。

## 参 考 文 献


[1] http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

[2] 余卫宇，余英林. 视频语义信息的研究[J]. 计算机工程与应用，2004(6)：27-29

[3] Ward J. Unqualified Dublin Core Usage in OAI-PMD data providers[J]. OCLC systems & services，2004，20(1)：40-47

[4] Stuckenschmidt H，Vdovjak R，Broekstra J. Index Structures and Algorithms for Querying Distributed RDF Repositories[C]// Proc. 13th Int'l World Wide Web Conf.（WWW'04）. ACM Press，2004：631-639

[5] http://www.w3.org/TR/REC-smil/

[6] Benitez A B，Zhong D，Chang S-F. Perspectives on MPEG-7：Metadata for Multimedia Enabling MPEG-7 structural and semantic descriptions in retrieval applications[J]. Journal of the American Society for Information Science and Technology，2007，58(9)：1377-1380

[7] Mieza C E. An opera information system based on MPEG-7C[C]// Proc. AES 25th International Conference. London，UK，2004

[8] Wallace M，Mylonas P，Akrivas G，et al. Automatic Thematic Categorization of Multimedia Documents using Ontological Information and Fuzzy Algebra[J]. Studies in Fuzziness and Soft Computing，Springer Berlin / Heidelberg，2006，204

[9] Lux M，Granitzer M. A Fast and Simple Path Index Based Retrieval Approach for Graph Based Semantic Descriptions[C]// Proceedings of Workshop on Text-Based Information Retrieval TIR'05. In Context of the 28th German Conference on Artificial Intelligence. Koblenz，Germany，Sep. 2005

[10] Aguis H，Angelides M C. Modeling and Filtering of MPEG-7-Compliant Meta-Data for Digital Video[C]// ACM Symposium on Applied Computing. 2004

[11] 朱华宇，孙正兴，王箭，等. 基于 MPEG-7 的视频语义描述方法[J]. 南京大学学报：自然科学，2002，38(1)：74-82

[12] VideoAnnEx. http://www.research.ibm.com/VideoAnnEx

[13] Yazici A，Yavuz O，George R. An MPEG-7 Based Video Database management System[M]. Flexible Querying and Reasoning in Spatio-Temporal Databases：Theory and Applications，In Springer's Geo-sciences/Geoinformation series by Springer Verlag，2004：181-210

[14] Fonseca P M，Pereira F. Automatic video summarization based on MPEG-7 descriptions[J]. Signal Processing：Image Communication，2004，19(8)：685-699

[15] 陈贤明，王小铭. 基于本体与 MPEG-7 视频语义描述模型[J]. 华南师范大学学报：自然科学版，2007(2)：51-56

[16] Tsinaraki C，Polydoros P，Kazasis F，et al. Ontology-based semantic indexing for MPEG-7 and TVAnytime audiovisual content[J]. Special issue of Multimedia Tools and Applications Journal on Video Segmentation for Semantic Annotation and Transcoding，2005(26)：299-325

[17] Troncy R，Bailer W，Hausenblas M，et al. Enabling Multimedia Metadata Interoperability by Defining Formal Semantics of MPEG-7Profiles[J]. Lecture Notes in Computer Science，2006，4306

[18] Bloehdorn S，Petridis K，Saatho C，et al. Semantic annotation of images and videos for multimedia analysis[C]// Proceedings of the Second European Semantic Web Conference. Heraklion，Crete，Greece，2005

[19] Vembu S，Kiesel M，Sintek M，et al. Towards bridging the semantic gap in multimedia annotation and retrieval[C]// Proc. of First International Workshop on Semantic Web Annotations for Multimedia（SWAMM）. Edinburgh（Scotland），May 2006

[20] Hunter J，Drennan J，Little S. Realizing the hydrogen economy through semantic web technologies[J]. IEEE Intelligent Systems-special eScience issue，2004

[21] 王煜，周立柱，邢春晓. 视频语义模型及评价准则[J]. 计算机学报，2007，30(3)：337-351

[22] Pinto F J，Martinez A F，Perez-Sanjulian C F. Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet[J]. International Journal of Computer Applications in Technology，2008，33(4)：271-279

[23] Zhuang Y，Mehrotra S，Huang T S. A multimedia information retrieval model based on semantic and visual content [EB/OL]. http://citeseer.nj.nec.com/zhuang99multimedia.html

[24] Bertini M，D'Amico G，Bimbo A D，et al. Using knowledge representation languages for video annotation and retrieval，Flexible Query Answering Systems[C]// Proceedings Lecture Notes in Computer Science. vol. 4027，2006：634-646

[25] Fang Hui，Jiang Jianmin，Feng Yue. A fuzzy logic approach for detection of video shot boundaries[J]. Pattern Recognition，2006，39(11)：2092-2100

[26] Yoon，JungWon. Improving recall of browsing sets in image retrieval from a semiotics perspective[D]. University of North Texas，2006：184

[27] 老松杨，黄广连，Smeaton A F，et al. 一种基于基本语义单元合成 Petri 网的足球视频查询描述模型[J]. 计算机研究与发展，2006，43(1)：159-168

[28] Bai Liang，Lao Song-yang，Liu Hai-tao，et al. Video shot boundary detection using Petri-Net[C]// International Conference on Machine Learning and Cybernetics. Kunming，July 2008：3047-3051

[29] Stoilos G，Stamou G，Tzouvaras V，et al. A Fuzzy Description Logic for Multimedia Knowledge Representation[C]// Proc. of ESWC2005 Workshop on Multimedia and the Semantic Web. Heraklion，Greece. June 2005：12-19

[30] Na Kwan-Sang，Kong Hyunjang，Cho Miyoung，et al. Multimedia information retrieval based on spatiotemporal relationships using description logics for the semantic Web[J]. International Journal of Intelligent Systems，2006，21(7)：679-692

[31] Georgieva L，Maier P. Description logics for shape analysis[C]// ThirdIEEE International Conference on Software Engineering and Formal Methods. Sept. 2005：321-330

[32] Schober J-P，Herzog H T. Picturefinder：Description Logics for Semantic Image Retrieval[C]// IEEE International Conference on Multimedia and Expo，ICME 2005. July 2005：1571-1574

[33] Sun Shanghua，Taylor P，Wilkinson L，et al. An Ontology to Support Adaptive Training for Breast Radiologists[J]. Lecture Notes in Computer Science，2008，5116

[34] Terrasa A N. Semantic integration of thematic geographic information in a multimedia context[D]. Universitat Pompeu Fabra，2006

[35] Shyu C-R，Klaric M，Scott G J，et al. GeoIRIS：Geospatial Information Retrieval and Indexing System-Content Mining，Semantics Modeling，and Complex Queries[J]. IEEE Transactions on Geoscience and Remote Sensing，2007，45(4)：839-852

[36] Yang H-C，Lee C-H. Image semantics discovery from web pages for semantic-based image retrieval using self-organizing maps [J]. Expert Systems with Applications，2008，34(1)：266-279


（下转第 37 页）

```
21:        If (m.type=EN) state←ENDN;
     //若回复消息类型为 EN 则置协商结束状态
22:        SendMsg(m);} //发送回复消息
23:    Else                    //收到终止协商消息
24:        state←ENDN;}    //置协商结束状态
25:While (state=NEGO){     //在协商状态
26:    m←ReceiveMsg();     //接收消息
27:    If (m.type=ID){     //收到信息披露消息
28:        m←LocalStrategy(M,L,R,m,A);
29:        SendMsg(m);
30:        If (m.type=EN) state←ENDN;}
31:    Else state←ENDN;}
32:If (m.content=∅) //终止协商消息内容为空
33:    return FAIL;//返回协商失败
34:Else
35:    return m.ticket;//返回授权票据
END of NegotiateAgent.
```

以上算法时间复杂度为 $O(M+N)$，其中 $M$ 和 $N$ 分别为双方最大信认证个数和访问控制策略条数。

## 4 分析与讨论

本协商协议的优越性主要体现在以下几个方面：

(1) 适用于不同应用场景的协商

现有研究中应用场景大部分是用户为请求访问某个单一的、离散的和具体的资源而展开协商。而实际应用中，服务器提供连续的、相互关联的和整体资源服务的情形并不少见。比如，开放系统在用户登录时需要先经过协商来确定用户能够访问哪些资源，然后将用户有权访问的资源组合起来，由用户选择访问，而不是在用户要求访问一个个具体资源时一次次地进行协商。本协议 RR 消息中用 URI 和一系列可选的属性可以描述广泛多样的不同规模和粒度的资源，可适应不同应用场景的协商需求。

(2) 支持多种协商策略和访问控制策略描述语言

本协议 ID 消息中封装的 4 种格式披露信息，能够描述现有协商策略所能处理的各种披露信息。比如，文献[5,7,8]中分别提出的契约、隐藏证书和 DL-TNL 语义身份断言等这些特定协商策略披露的信息都可以按 4 种格式分类描述并封装到 ID 消息中。同时，第三种格式披露的信息能够表达不同种类访问控制策略语言描述的访问策略。

(3) 允许在协商过程中同时使用多种协商策略

本协议对多种协商策略的支持不仅使不同协商策略之间的协商成为可能，而且允许协商策略模块在一次协商过程中灵活切换多种协商策略，可满足协商中随信任级别提升而调整协商性能的需求。

此外，由于协商成功后授权票据中包含了有效期，用户可以“一次协商多次使用”。而授权票据与用户主体标志进行绑定并提供服务器对授权的数字签名，一方面能防止授权票据被滥用，另一方面可保护用户免遭抵赖和欺骗。

**结束语** 本文提出的信任协商协议将消息分为 3 类，采用极具灵活性的 URI 命名机制描述广泛多样的资源，4 种披露信息格式能完全区分和表示现有协商策略所能处理的各种信息，因而使协议独立于具体的协商策略而具有明显的通用性。

## 参 考 文 献


[1] Lee A J, Winslett M, Basney J, et al. The Traust Authorization Service [J]. ACM Transactions on Information and System Security, 2008, 11(2): 1-33

[2] Smith B, Seamons K E, Jones M D. Responding to policies at runtime in TrustBuilder[C]// Proc. of the 5th Int'l Workshop on Policies for Distributed Systems and Networks. Washington: IEEE Computer Society Press, 2004: 149-158

[3] Bertino E, Ferrari E, Squicciarini A C. Trust-X: A Peer-to-Peer Framework for Trust Establishment[J]. IEEE Trans. Knowledge and Data Eng, 2004, 16(7): 827-842

[4] Nejdl W, Olmedilla D, Winslett M. PeerTrust: Automated trust negotiation for peers on the semantic Web[C]// Proc. of the Workshop on Secure Data Management in a Connected World (SDM 2004). LNCS 3178. Springer-Verlag, 2004: 118-132

[5] 李建欣，怀进鹏. COTN：基于契约的信任协商系统[J]. 计算机学报，2006，29(8)：1290-1300

[6] Winsborough W H, Seamons K E, Jones V E. Automated trust negotiation[C]// DARPA Information Survivability Conf. and Exposition. New York: IEEE Press, 2000: 88-102

[7] Holt J, Bradshaw R, Seamons K E, et al. Hidden credentials[C]// Jajodia S, Samarati P, Syverson PF, eds. Proc. of the 2003 ACM Workshop on Privacy in the Electronic Society. New York: ACM Press, 2003: 1-8

[8] 张妍，冯登国. 吝啬语义信任协商[J]. 计算机学报，2009，32(10)：1989-2003


(上接第 6 页)


[37] Sjöberg M, Laaksonen J, Honkela T, et al. Inferring semantics from textual information in multimedia retrieval[J]. Neurocomputing, 2008, 71(13-15): 2576-2586

[38] Urban J, Jose J M. Adaptive image retrieval using a graph model for semantic feature integration[C]// 8th ACM International Workshop on Multimedia Information Retrieval MIR '06. Santa Barbara, CA, USA, October 2006: 117-126

[39] Luan Xi-dao, Xie Yu-xiang, Ying Long, et al. SATS: A News Story Detection Method Based on Multi-feature Fusion[J]. Journal of Information and Computational Science, 2008, 5(1): 267-274

[40] Liu Yuchi, Luan Xidao, Wu Lingda, et al. Narrative structure analysis of lecture videos with hierarchical hidden markov model for e-learning[C]// Technologies for E-Learning and Digital Entertainment. First Internation Conference, Edutainment 2006. Hangzhou, China, April 2006: 429-437

[41] 栾悉道，谢毓湘，应龙，等. 基于 EDU 模型的新闻视频摘要技术研究[J]. 系统仿真学报，2007，19(16)：3770-3774

[42] 庄越挺，吴聪苗，吴飞，等. 多媒体交叉参照检索系统研究[J]. 计算机辅助设计与图形学学报，2005，17(4)：834-839

[43] 鲍永生，任建峰，郭雷. 支持语义的图像检索[J]. 南京航空航天大学学报，2005，37(1)：75-78

[44] 丁国祥，吴仁炳，张振亚，等. 一种用于 MAM 的语义可扩展视频编目与检索方法[J]. 中国图象图形学报，2005，10(8)：1036-1041