多媒体推理的概念框架*)

A Conceptual Framework for Multimedia Reasoning

庄越挺　潘云鹤　　潘　红
(浙江大学人工智能研究所　杭州 310027)　(杭州师范学院计算机系)

Abstract　Up to now, AI technology is dominated by the Physical Symbolic System (PSS), in which symbolic information is used as the medium for reasoning. In these approaches, information other than symbols, such as image, graphics, and even video should first be represented by symbols, and after reasoning, the symbolic result is again changed into its original media form. In this paper, we will propose a new form of reasoning method called multimedia reasoning (MR), a kind of reasoning that is based on the different media such as text, image, video, audio and so on. By introducing the concept of multimedia transformation theory (MTT), it presents a conceptual framework for multimedia reasoning. In the end, it discusses the importance and potentials in applications.

Keywords　Multimedia, Reasoning, Conceptual framework, Multimedia transformation theory

1. 引言

揭示人类智能是AI、认知科学和心理学的数代学者所致力的目标。许多年来，由 H. A. Simon 提出的物理符号系统(PSS)一直主宰着 AI 系统，其核心是以符号为媒体的知识表达[1]。仔细研究人类的推理，可以发现人类的推理是一种涉及多种感知信息，如听觉、视觉、触觉、味觉等的抽象思维和形象思维共同作用的过程。比如，医生看病，涉及的信息媒体有：

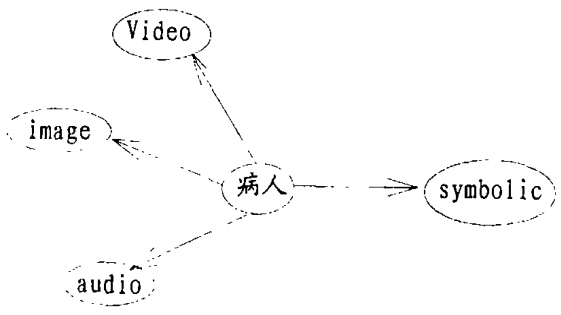


图 1　多媒体推理的一个例子

其中 video：如超声波检查结果；image：如 X 光片；audio：病人的病情陈述；symbolic：病人的性别、血压、血液化验的结果等。医生对病人的诊断便是基于上述多种媒体信息的一种多媒体推理的过程。

多媒体推理(Multimedia Reasoning)指的是一种基于文本、图像、视频、音频等介质之上，为达到某一目标而进行的推理活动，其名称之由来相对于传统的 PSS 之下的符号推理而言。我们认为：人类的推理过程在计算机平台上的降落，便是对应于多媒体推理。在人类活动和计算机平台之间存在着如表 1 所示的对应关系。

表 1　人和计算机之间的对应关系

人(H_r)	计算机平台(C_r)
听觉	音频
视觉	图像、视频、图形
触觉、味觉	(传感器)信号
语言、文字	文本、图形

其次，在推理的形式上，人类的推理(IA)表现为抽象思维(A)和形象思维(I)，二者之间具有如下的类比关系：

$$Hr::Cr = IA::M$$

*)本项研究得到国家自然科学基金和浙江省科技计划项目的资助。

其中的::表示对应关系，Hr 指人，Cr 指计算机平台，M 则为多媒体推理的内容，建立在计算机平台上。称 M 为多媒体推理，是因为有许多不同的媒体参加了推理过程。

从长远来看，智能问题求解系统应该能够访问以不同的媒体形式存储的信息，并按照不同媒体源的本身，而不是把媒体转换成为符号，产生出新的信息。

在本文中，我们将给出媒体转换的理论，然后试图得到关于多媒体推理的概念框架。最后，指出多媒体推理的一些应用。

2. 媒体转换理论

在深入讨论这一问题之前，先察看一下多媒体推理的含义：

1)基于知识的符号化表示以及推理。通过把各媒体源转化为符号的表示，比如图形，表示为点、线、面等实体名称与对应的坐标。典型的有专家系统的方法，如 MYCIN 等。

2)模式识别的方法。研究图像识别居多，即从图像中提取出符号信息，其中图像的范围包含了视觉图像、非视觉图像如超声图像等等。

3)媒体的搜索查找功能。如从 video 中快速定位符合某种要求的画面，模拟医生在思维过程中的心象聚焦过程；从图像数据库中快速查找图像，即多媒体数据库。由于查找操作是如此之重要和频繁，以致许多人认为：搜索查询＝多媒体推理。显然，这是片面的。

我们认为：从媒体转化的角度，推理可看成是一种媒体到另一种媒体的转换过程。对于一个由 m 个不同媒体组成的集合，总共有 m^2 种变换的情形。考虑一个媒体集合{符号，音频，图像，图形，视频}，共有 25 种不同的情形，主要的描述如下：

· 符号→符号：运用了概念和判断，即由传统的符号人工智能系统所解决。符号主义把符号作为人类思维的基本元素，认为思维就是在符号表示上的运算，其思想可以简单地归结为“推理即符号计算”这一基本原理。

· 音频→符号：为语音识别的研究内容，从音频到符号文字到输出的解决，具有重要的意义，如在计算机的输入方面，可以去掉键盘，以及实现复杂的声控。

· 图像→符号：图像对象的识别与理解，以及图标索引(图像的符号化)[5]。

· 图形→符号：图形用输出基元来表示，工程图纸的矢量化便是一例。

· 视频→符号：视频是多幅图像的时序组合，其核心却是图像→符号的推理。如 MIT 的 Weiss 提出的用代数视频的方法实现基于 VIDEO 的查询[6]，其媒体转换过程可以看成为：

$$\text{视频段}\xrightarrow[\text{层次关系}]{\text{逻辑结构或}}\text{符号(称代数视频)}\longrightarrow\text{视频流}$$

· 符号→音频：语音合成的内容，如 PC 机上的各种声卡解决了从数字信号(即符号)到模拟信号(即音频)的转化，又如以文本作为输入，产生出像人的语音。

· 音频→音频：为信号处理技术的范畴，输入的音频信号经过衰减、放大等等得以处理；

· 图像→音频：可视为图像→符号→声音。图形→声音，视频→声音类同。

· 符号→图形与符号→图像。如果每一种符号对应于一种含义，而它又可以被表达为图形的模式，那么这一过程即完成。例如，DXF 格式的文件可以容易地显示在屏幕上。

· 音频→图形：如音频作为一种信号，可以以图形的模式显示出来。

· 图像→图形：这方面的工作称为矢量化或 3D 重建。

· 图形→图形：计算机图形学的内容，如隐藏线和隐藏面的消除，2 维或 3 维的变换；

· 视频→图形：与图像→图形相似但是更复杂；

· 图像→图像。这具有多重含义，首先，它指图像处理，如图像的变形等等；第二、指从众多的图像中，挑选出符合条件的图像，这是图像数据库的内容。第三、源图像指一系列的实例(CASE)，这一过程，称为基于 CASE 的图像推理。

· 图形→图像：这由着色(Shading)完成，为计算机图形学的研究课题。

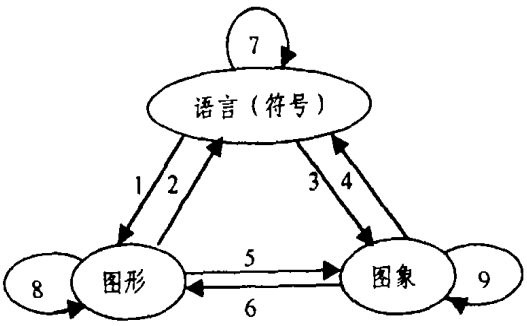
· 视频→图像：视频被看成是图像的集合。以 MPEG 格式的视频，经过解压缩，得到某一图像；

图 2 显示的是三种常用媒体：符号、图形和图像之间的关系。其中的每一种转换代表一种研究领域。

3. 多媒体推理的框架

首先，我们定义多媒体推理是一种基于不同的媒体，如文本、图像、视频、音频等的推理，如此命名

的目的是与仅仅依赖于符号信息的传统推理相区



1 映射对应； 4 图像识别； 7 AI推理；
2 命题提取； 5 真实感显示； 8 空间推理；
3 映射； 6 3D重建； 9 视觉(图像)推理。

图2 三种媒体之间的关系

别。关于多媒体推理一词，国际上学术界尚无定义，但是有类似的研究开始出现。在文[4]中，Brink提出了异质多媒体推理的概念。他实现了一个媒体抽象系统，用来创建和询问媒体的抽象类。用数学术语，他的媒体被抽象为一个七元组：

$M=(ST,fe,\lambda,R,F,V_{ar1},V_{ar2})$

其中：ST：媒体M所具有的各个状态的集合。fe：媒体M中所含的各个特征的集合。λ：从S到功能的映射，即从 $fe\xrightarrow{\lambda}[0,1]$。$V_{ar1}$：状态变量，其值变化范围为ST。$V_{ar2}$：特征变量，其值变化范围为fe。R：ST上的模糊中间状态关系集。F：模糊的特征—状态关系集，F上的每一个关系为：$fe^i\rightarrow[0,1]$（特征间的关系同状态无关）；或者 $fe^i\times ST\rightarrow[0,1]$（i≥1）（特征间的关系同状态相关）。

所有的媒体，无论它是图形，视频还是音频，均可统一成为这样的形式。比如，就图像媒体的数据而言，对应的七元组含义如下：

ST：含有某一图片(Photograph)的文件均表示一个状态，如果是一段N帧的视频，则有N个状态；

fe：包括感兴趣的人（如"克林顿"）和无生命物的特征（如"白宫"、"林肯纪念馆"等），只捕获感兴趣的特征，比如有图片"克林顿在白宫草坪上发表演说"，其中若感兴趣的是可以被清晰识别出的椅子，则它便可作为特征之一。

λ：表示某一特征在给定图像中出现的确信度，例如：$\lambda(S_2)$(Bill Clinton)=0.7，即表示，克林顿出现在状态S_2中的确信度是70%；

R：分二类：一类与给定的状态无关，另一类则与给定的状态有关。考虑：is_wearing有三个参数：(人名、衣服名、颜色)。显然，该关系与状态有关，故在说明时需要给定状态，如：is_wearing('Bill Clinton', 'tie', 'red', file5)：0.99。若是与状态无关的关系，则无需指出状态名。

Brink指出了七元组媒体抽象还可表征其它各种媒体类型：文本数据、视频数据等。Woelk和Kin提出在面向对象的数据库的基础上，稍作增强，便可支持多媒体应用。

一个多媒体系统MMS，是媒体抽象的有限集合。即：$M=\{M_1,\cdots,M_n\}$，其中 $M_i=(ST^i,fe^i,\lambda^i,R^i,F^i,V^i_{ar1},V^i_{ar2})$。

基于媒体的这种抽象表示，马里兰大学提出一种一致的查询语言：一种完全陈述性的逻辑查询语句。其查询格式颇类似于Prolog的逻辑形式，即：

$\leftarrow A_1:\mu_1\&\cdots\&A_n:\mu_n$

其中：A是一个原子，μ是0-1之间的一个数，意为A为真的可信度为μ。

如有这样的一个查询：

Find all pictures of George Bush with the spouse of a person whose taxes have been reported in USA Today.

可表示为如下的形式：

← 'Bush' ∈ flist(Pic) & frametype(Pic, picture) & P∈flist(Pic) & spouse(P,Q) &Q∈flist(Art) &content(taxes, Q, Art) &frametype(Art, usatoday).

可见，在这一表达中，同一般的数据库查询格式相比，它含有多媒体的信息。而对此查询语句的返回结果，则被认为是一种异质多媒体推理。

Brink等所实现的异质推理和协调系统，称为Hermes。从数学术语上讲，Hermes推理依赖的是下面形式的规则：

$A\leftarrow B_1\&\cdots\&B_n\&D_1\&\cdots\&D_m\&E_1\&\cdots\&E_k$

其中，$A,B_1,\cdots B_n$是逻辑原子，$D_1,\cdots,D_m$是形如$(X_i,d_i:f_i((args)))$的原子，这里d_i表示外部软件包，f_i是包d_i中预定义好的过程，(args)是其参数。E_i的形式是relop(V_1,V_2)，relop为比较操作符{=，≥，≤，>，<}，V_1,V_2要么常数，要么形如：X_i.attri。

基于媒体转换的观点，我们给出多媒体推理的框架如下。首先引入若干定义：

定义1 设媒体集$M=\{m_1,m_2,\cdots,m_c\}$。其中$m_i(i=1,\cdots,c)$指出了一种媒体类型。$|M|=c$，"|·|"表示集合的模运算。

如M={符号(m_1)，声音(m_2)，图像(m_3)，图形(m_4)，视频(m_5)，动画(m_6)}，|M|=6，M是目前最

常用的媒体集。

定义 2 同种媒体之间的转换操作集合为 P_x，x 表示媒体类型；如 P_{m_i} 则表示对应于媒体类型 m_i 的操作算子集。

例：m_i＝图像，则 P_{m_i}＝{图像增强、滤波、恢复、色彩校正，…}，即 P_{m_i} 为图像处理领域内的算法集。

定义 3 不同媒体 $m_i, m_j (i \neq j)$ 之间的操作算子集记为 Q_{m_i,m_j}，表示从媒体 m_i 到 m_j 的所有可能操作的集合。

显然有：$Q_{m_i,m_j} \neq Q_{m_j,m_i} (i \neq j)$。例如：$m_i$＝图像、$m_j$＝符号，则 Q_{m_i,m_j}＝{计算机图像识别与理解的各种算法}，而 Q_{m_j,m_i}＝{图像的语义生成或对应关系}。

定义 4 记 $\sum_M PQ$ 为定义于媒体 M 之上的所有操作的集合，则有：

$$\sum_M PQ = P_{m_1} \cup P_{m_2} \cup \cdots \cup P_{m_c} \cup Q_{m_1,m_2} \cup \cdots \cup Q_{m_1,m_c} \cup Q_{m_2,m_1} \cup \cdots \cup Q_{m_2,m_c} \cup \cdots \cup Q_{m_c,m_1} \cup \cdots \cup Q_{m_c,m_c}$$

若令 $A = \begin{pmatrix} P_{m_1} & Q_{m_1,m_2} & Q_{m_1,m_3} & \Lambda & Q_{m_1,m_c} \\ Q_{m_2,m_1} & P_{m_2} & Q_{m_2,m_3} & \Lambda & Q_{m_2,m_c} \\ \Lambda & \Lambda & P_{m_3} & \Lambda & \Lambda \\ \Lambda & \Lambda & \Lambda & O & \Lambda \\ Q_{m_C,m_1} & Q_{m_C,m_2} & \Lambda & \Lambda & P_{m_C} \end{pmatrix}$

则 $\sum_M PQ = \bigcup_{i,j=1}^{n} A(i,j)$。

定义 5 M_x 表示为充分描述客观实体(抽象的或具体的)X 所涉及的媒体类型的集合。显然有：$M_x \subseteq M$。

如上面的医生看病一例，M_x＝{m_1, m_2, m_3, m_5}。若 X 代表一个数学公式，则 M_x＝{m_1}；若 X 代表一个杀人案例，包含了物证、证人录音，则 M_x＝{m_2, m_3}。

定义 6 设 $X_1, X_2, \cdots, X_k$ 表示 K 个客体(或称 K 个源)，Y 表示目标；设 $F: X_1, X_2, \cdots, X_k \rightarrow Y$，则 F 中的所有操作构成了多媒体推理。

在 k＝1 时，F 称为单源的多媒体推理；k＞1 时，F 为多源的多媒体推理。不难证明，k＞1 的情况可以转化为 k＝1 情况。

设 $X = X_1 \cup X_2 \cup \cdots \cup X_k$，则 $M_x = M_{X_1} \cup M_{X_2} \cup \Lambda \cup M_{X_k}$。

现在给出其证明：(采用数学归纳法)已经知道，对于一个客观实体 X，M_x 表示 X 所含有的媒体类型集，设∃ $i, m_i \in M_X$，引入新的符号 $I(x, m_i)$，表示实体 X 之中对应于类型 m_i 的各个体集，所谓个体，此指 m_i 的一个具体例子，如图 3 所示。

设有两个客体 X_1, X_2，若 $m_i \in M_{x1}, m_i \in M_{x2}$，设 $X = X_1 \cup X_2$，则 X 中的类型 m_i 包含的个体为 $I(X_1, m_i) \cup I(X_2, m_i)$，若 $m_i \in M_{x1}, m_i \notin M_{x2}$ 或 $m_i \notin M_{x1}, m_i \in M_{x2}$，则在 X 中包含类型 m_i 及其 X_1 或 X_2 中的某个个体集。

同理，若 $X_1, X_2, \ldots, X_k \rightarrow Y$ 可化为 $X \rightarrow Y$ 的形式，则可推得 $X_1, X_2, \cdots, X_{k+1} \rightarrow Y$ 同样成立。(证毕)

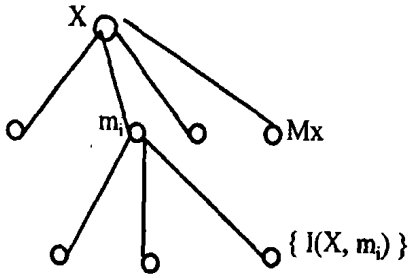


图 3 媒体层次

因此，多媒体推理可以转化为 X→Y 的形式。特别地，对于 F：X→Y，有：

1)当 $|M_X| = |M_Y| = 1$，且 $M_X = M_Y$ 时，则有：$F \subseteq P_{m_1} \cup P_{m_2} \Lambda \cup P_{m_c}$。这是同种媒体间的操作。

2)当 $|M_X| = |M_Y| = 1$，且 $M_X \neq M_Y$ 时，则有：$F \subseteq \sum_M PQ - (P_{m_1} \cup P_{m_2} \Lambda \cup P_{m_c})$。这是不同的二种媒体类型之间的操作。

我们得出如下的结论：

MR 是由两种类型的操作组成的：内操作(inter-operation)和外操作(intra-operation)；

推理结果(R)可被视为一组施加的操作加上时间参数，即 $R = \{(P_i(m_x, t_{i1}, t_{i2}))\}$。在时间 t_{i1} 到 t_{i2}，操作 P_i 施加于媒体 m_x 之上。

引入最小算子集，表示为 P_{MOS}，$P_{MOS} \subseteq \sum_M PQ$，它表示对于任何的操作，如 P_i，下列之一满足：①$P_i \subseteq P_{MOS}$；或②$P_i \notin P_{MOS}$，但是 P_i 可以从 P_{MOS} 中引导出来。如，在一阶谓词逻辑中，{∧，∨，¬}是 P_{MOS}；在计算机图形学上，输出原语是 P_{MOS}，我们把 P_{MOS} 的思想用到了 CAI 软件的实现之中。

MR 中的两个应用领域是多媒体展示(presentation)和设计(design)。多媒体展示的目的是安排下列事情：什么信息在什么时间被显示(如图像、视频)或展示(如音频)。所以在此，多媒体推理的任务是规划：选择媒体，布置媒体。一个基本的事实是，多媒体展示不影响原媒体源的内容。

4. 从图像到图像的推理

现在，让我们察看静态画面的设计过程。所谓的

静态画面，包括了广告、报刊杂志封面等等，范围很广。这里，显然有：

M_x={文字(m_1)，图像(m_2)，图形(m_3)}

M_y={图像(m_2)}

从X→Y包含的操作序列中，核心的操作是m_2→m_2，这包括下列意义：

1)图像的处理：包括空域与频域的操作，如图像的变形(warping，morphing)，这里，输入参数是一个或多个图像，输出参数是另一个图像。

2)如果每一个源图像指向一个CASE(实例)，即如图4：

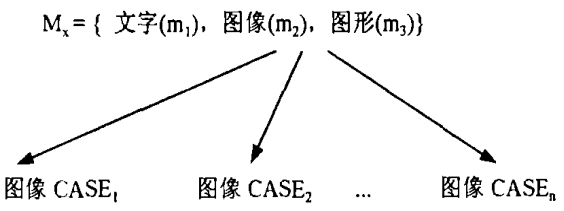


图4 源图指向实例(CASE)

由于源X中的图像均指出了图像的实例(CASE)，因此这时的X→Y是基于CASE的推理(CBR)[8]。

3)源X中的图像构成的是一个图像数据库(IDB)，这时的X→Y是指从图像数据库IDB中得到符合查询条件α的图像，即(α,IDB)→m_2。方法分成二类：传统的数据库方法以及基于图像内容的查询(QBIC)。MARS是针对图像的检索，即从图像到图像的一个快速而高效的实现[7]。

5. 从视频到语义的推理

假设把一段夹杂着爆炸声、汽车燃烧声的视频段呈现给某一个人，他会立即把该段视频描述为"爆炸"。这便是人类从视频到语义的一个推理的例子。

视频是一个非结构化的位流。为了能够智能化地从大量的分布式的视频源中进行索引和搜索，我们需要事先对视频进行内容的摘取。这意味着用户今后的查询可以基于视频的语义进行。从视频中产生出语义信息的需求可以从下列情景中进一步得到说明：

根据用户的兴趣过滤视频内容。例如，给定一个视频，如果系统能自动地从中找出进球的精彩镜头，那么球迷们会发现它是非常有用的。这一过滤器是一种联机的情景。作者在UIUC访问期间，参加了以Thomas S. Huang教授为首的研究小组，引入了Multiject的概念[9]。其基本思想是从一组标了号的训练数据集中估计出模型的参数和结构。该模型用来在未加标记的多媒体段中标识出事件、对象和场地，办法是计算概率，如P(水下 AND 鲨鱼|多媒体数据段)，表示给定某一段多媒体数据，出现鲨鱼以及在水下的概率是P。Multiject的例子有：汽车飞驰而过、爆炸、玻璃粉碎、枪击声、太空等等，每一个都是通过数据训练的。之后，经过良好训练后的Mulitject就可以联机地用来过滤事件、对象和场地。

从多媒体推理的角度，以上可以看成是从视频到语义(以符号的形式)的推理。更进一步的研究，正在伊利诺斯大学、哥伦比亚大学以及Lucent Technologies合作进行之中。

结论 从智能和推理的角度，任何涉及多媒体处理的活动，如多媒体展示、多媒体著作、视觉设计，都可以被看成是多媒体的推理，以区别于传统的AI意义上的推理。本文提出的多媒体推理的框架意义在于：①它为智能多媒体和多媒体智能的研究提供了一个理论框架；②它为多媒体领域中的不同研究提供了一种新的分类方法。

参考文献

1 Flickner M, et al. Query by Image and Video Content: the QBIC System. IEEE Computer, Sept., 1995, 23～32

2 Pan Yunhe. Synthesis Reasoning. Pattern Recognition & Artificial Intelligence, 1996, (4)

3 Horn F, Stefani J B. On programming and supporting multimedia object synchronization. the Computer Journal, 1993, 36(1)

4 Brink A, et al. Heterogeneous Multimedia.

5 Chang S K, Shi Q Y. Iconic indexing by 2D strings. IEEE Trans. Pattern Recognition Mach. Intell, PAMI-9, 1987. 413～428

6 Ron W. Content-based Access to Algebraic Video. IEEE Computer, 1994, 140

7 Huang Thomas S, et al. Multimedia Analysis and Retrieval System (MARS) Project. Proc. of 33rd Annual Clinic on Library Application of Data Processing-Digital Image Access and Retrieval, March, 1996

8 Zhuang Y T, Pan Y H. Case-based Synthesis in Automatic Advertising Creation System. ICIM' 95, SPIE2620, 1995. 6

9 Naphade M R, et al. Probabilistic multimedia objects (Multijects): A novel approach to video indexing and retrieval in multimedia systems. ICIP' 98, Oct. 98, Chicago, USA