典型的多通道整合方法之比较*)

The Comparison of Typical Approaches to the Integration in Multimodad Interface

栾尚敏　戴国忠　陈由迪　关志伟
(中国科学院软件研究所　北京100080)

Abstract In this paper, we first introduce several approaches to the integration in multimodal interface, such as the probability-based approach, unification-based approach, task-based approach and knowledge-based approach. Then we discuss the relationships between these approaches and present our view about the integration.

Keywords Multimodal user interface, Multimodal integration, Fusion

多通道用户界面是当前用户界面中研究的热点，人们提出了各种各样的方法。在多通道界面系统中，一个重要的问题是通道整合，本文介绍了人们提出的各种通道整合的方法，对这些方法进行了分类归纳和比较，并介绍了我们在这方面的一些认识和工作。

1　概率统计的方法

Wu，Oviatt 和 Cohen[1]给出了一种基于概率统计方法的整合算法，该方法对于语音和手势进行了整合，并在多通道系统 Quickset[4]中给予了实验研究。假设 $S_i, i=1,2,\cdots,M$ 是语音模式的输出，$G_j(j=1,2,\cdots,N)$是手势模式的输出，该系统用于识别 $C_k(k=1,2,\cdots,K)$多通道类，这里 K 不能超过$(M+1)*(N+1)-1$，并且至少不小于 M 和 N 中的较大者。用关联图表示多通道类到各个多通道输出的影射。假设 X 表示多通道输入特征向量，它是手势输入特征 X^G 和语音输入特征 X^S 的合成，则需要将输入特征空间划分为 K 个不相交的判定区域 $R_k(k=1,2,\cdots,K)$，X 正确识别的概率可以如下定义：

$$P_c=\sum_{k=1}^{K}P(X\in R_k,C_k)=\sum_{k=1}^{K}((P_{X_k})*(P(C_k)))$$

这里 $P_{X_k}=P(X\in R_k|C_k)=\int_{R_k}P(X|C_k)$ 是第 k 个类正确识别的概率，$P(X|C_k)$是它的类条件密度函数，$P(C_k)$是它的先验概率，得到如下的多通道识别概率的界限：

$$\sum_{k=1}^{K}P_{X_j^G}P_{X_i^S}P(C_k)\leqslant P_c\leqslant\sum_{k=1}^{K}\max[P_{X_j^G},P_{X_i^S}]P(C_k)$$

这里 $P_{X_j^G}$和 $P_{X_i^S}$分别是组成第 k 个多通道类的关联手势和语音的正确识别概率。根据他们的经验，类条件概率密度函数 $P(X|C_k)$可以计算如下：

$$P(X|C_k)=\frac{1}{2}[P(X^S|C_k,X^G)P(X^G|C_k)+P(X^G|C_k,X^S)P(X^S|C_k)]$$

若令

$$\alpha_k=\frac{1}{2}P(X^S|C_k,X^G),$$

$$\beta_k=\frac{1}{2}P(X^G|C_k,X^S)$$

则

$$P(X|C_k)=\alpha_k(P(X^G|C_k)|\beta_kP(X^S|C_k)$$

对于 α_k 和 β_k 可以采用如下方法进行估计：

$$\alpha_k=\frac{1}{2}P(x^S\in R_i^G|C_k,X^G\in R_j^G)$$

$$\beta_k=\frac{1}{2}P(x^G\in R_j^G|C_k,X^S\in R_i^G)$$

Wu 和 Oviatt 则提出了优化权值的方法。将不同通道结合后续概率结合起来，采用通道输入条件特征的密度函数来决定通道整合时的优化权值。由于输入特征的信息维度太高，因此很难去评价这些条件密度函数。他们采用了两种模型技术来估计这些条件密度，并为后续识别概率获得交叉相关权值参数。

同时，在多通道整合中，确定影响多通道识别效果的原始因子，从而对整个系统的识别效果进行评价和估计。假设一组单个通道的识别器的识别率已知，那么会对多通道系统的识别效果产生限制作用，他们对在这种情况下的多通道识别效果的上下限进行了分析，从而确定了影响多通道系统识别效果的影响因子。以前的整合方法中，对通道整合过程进行了重新精练。对不同的通道、不同的交互要素附加以不同的权值，这样识别错误能够得到一定程度的避免，从而提高整个系统的稳定性。

通过静态统计过程来定义多通道系统框架。以往的研究认为单通道在多通道系统中都是一个个独立工作的个体，从而多通道命令的后续概率是相关联的各个组成部分的后续概率之间的交集。尽管这种通道独立性的假设提供了一个整合开展的起点，并简化了通道整合的过程，但是这种假设存在着局限，因为语音和唇动或者是语音和手动手势之间已经被公认具有很强的通道联系性。其它的一些基于概率的整合方法，请参考文[2]。

2　基于合一的方法

Johnston 等[5]提出了一种多通道整合方法。他们首先采用带类型的特征结构来表示通道的意义，这样可以很容易地采用带类型的合一方法来实现整合，所以把这种方法称为基于合一的方法。例如，用户若发出“M1A1 PLATOON”的声音，则建立如下结构来描述：

*)本文得到国家自然科学基金(批准号：60033020和60103020)和中国博士后科学基金会资助。栾尚敏　博士后，主要研究领域为人机交互技术、算法设计自动化、信念修正、形式化方法。戴国忠　研究员，博士生导师，主要研究领域为人机交互技术、计算机图形学。陈由迪　研究员，主要研究领域为人机交互技术、计算机图形学。关志伟　博士，助理研究员，主要研究领域为人机交互技术、用户建模、上下文感知。

$$\left[\begin{array}{l}\text{object}:\left[\begin{array}{l}\text{styl}:\text{mlal}\\ \text{echelon}:\text{platoon}\end{array}\right]_{\text{unit}}\\ \text{location}:\left[\quad\right]_{\text{unit}}\end{array}\right]_{\text{create\_unit}}$$

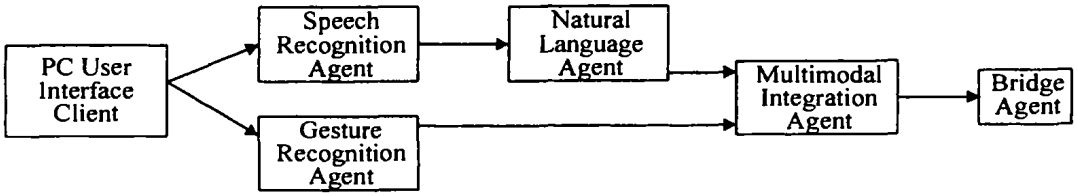


图1

对于手势，需要确定其位置，这就需要点和线的描述，它们的结构分别描述如图1。

$$\left[\text{location}:\left[\begin{array}{l}\text{xcoord}:95305\\ \text{ycoord}:94365\end{array}\right]_{\text{point}}\right]_{\text{command}},$$

$$\left[\text{location}:\left[\begin{array}{l}\text{coordlist}:\\ [(95301,94360),\\ \ (95305,94365),\\ \ (95310.94380)]\end{array}\right]_{\text{line}}\right]_{\text{command}}$$

多通道整合结构如图1所示。

Client 智能体(agent)可以接收来自语音和手势的输入，然后分别传输给语音识别智能体和手势识别智能体，自然语言智能体对语音智能体产生的串进一步进行处理，并且给这些串建立带类型的特征结构。来自手势识别智能体的手势的解释同样用带类型的特征结构表示。多通道整合智能体对将来自语音和手势的输入进行整合，将结果传输给连接(bridge)智能体，连接智能体接收带类型特征结构的命令，然后将这些命令转化为应用能接收的形式，并传递给应用。该方法具有如下特点：带类型的特征结构便于表示来自各通道的不完整信息；避免了手势的二义性，对于一种手势可能有多种解释，只选择与语音通道类型相符的解释，这样就避免了二义性；整合过程可由任意一个通道驱动，可以接收单通道输入的完整命令，也可以接收多通道配合输入。如前面的例子，进行整合以后，结果如下：

$$\left[\begin{array}{l}\text{objct}:\left[\begin{array}{l}\text{type}:\text{mlal}\\ \text{echelon}:\text{platoon}\end{array}\right]_{\text{unit}}\\ \text{location}:\left[\begin{array}{l}\text{xcoord}:95305\\ \text{ycoord}:94365\end{array}\right]_{\text{line}}\end{array}\right]_{\text{create\_unit}}$$

如果上面的发音不是“M1A1”，而是“BARBED WIRE”，则建立如下的结构：

$$\left[\begin{array}{l}\text{object}:\left[\begin{array}{l}\text{style}:\text{barbed\_wire}\\ \text{echelo}:\text{platoon}\end{array}\right]_{\text{unit}}\\ \text{location}:\left[\quad\right]_{\text{line}}\end{array}\right]_{\text{create\_unit}}$$

进行通道整合的结果为：

$$\left[\begin{array}{l}\text{object}:\left[\begin{array}{l}\text{type}:\text{mlal}\\ \text{echelon}:\text{platoon}\end{array}\right]_{\text{unit}}\\ \text{location}:\left[\begin{array}{l}\text{coordlist}:\\ [(95301,94360),\\ \ (95305,94365),\\ \ (95310,94380)]\end{array}\right]_{\text{line}}\end{array}\right]_{\text{create\_unit}}$$

这一方法目前已经用于 QuickSet 系统中。在此系统中，用户可以在一幅地图上用语音和手势进行部署军事力量的模拟，用户可用笔和/或语音来创建和定位大量的实体、点、线和区域。此整合方案也有面向任务的思想，只不过它采用了更精细的数据描述形式。

3 基于任务的方法

Oviatt 等人[6,7]讨论了基于任务的整合方法。他们对基于笔和语音交互中的多通道整合和同步问题进行了详细分析，并给予了模拟实现[8]。该模拟系统支持如下的行为命令：(1)在某个位置、某条线上或某个区域增加对象的命令；(2)将对象移动到新位置的命令；(3)对一些特殊路线或区域进行修改的命令；(4)计算两个位置间距离的命令；(5)查询对象信息的命令；(6)删除对象的命令；(7)进行标记的命令；(8)放缩命令；(9)控制任务的过程，例如，如果你发出“执行下一项任务”的命令，则紧接着就执行下面一项任务；(10)地图卷缩命令；(11)打印命令；(12)对于不在视线中的对象进行自动定位命令；(13)call up overlay；(14)限制条件命令，例如，显示房价在3000元到5000元的房子。该系统将整合分为如下情况：a. 读和写并发重叠进行，b. 读和写顺序进行，c. 指点和语音指点，但只产生一个点，d. 指点和语音整合，并产生图形。并对于各种不同的情况进行了测试，详细的测试结果，请参阅文[6]。

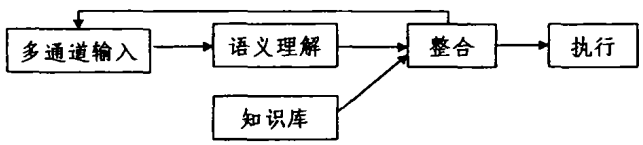
Marsh、Wauchope 和 Gurney[9]也讨论了基于任务的整合方法。他们主要讨论了多通道界面中的话语建模，并且将自然语言集成到了图形对话中。多通道界面对话建模传统的图形用户界面允许直接操纵对象，在菜单、鼠标和其它可操纵方法方面，他们是典型的面向命令和可扩张的。本研究涉及到如何将在多通道和多媒体界面中增加对话功能。1. 在图形系统中，在操作之间或在运算之间的关系很小，各个菜单驱动的命令很简明。在自然语言对话中，在交互中存在或明或暗的关系，反映用户对系统的理解、它的操作符以及对话过程，例如可以使用代词和一般短语。可以理解对话元素和具有对话模型表示方法的自然语言界面系统，允许连续的和连贯的交互。2. 图形界面中，所有的命令都是顺序输入的，但自然语言界面允许使用连接词等，例如，可以使用连接词“AND”将多个断言连接在一起，形成一个句子。3. 具有对话能力的自然语言界面使用对话模型来描述交互的顺序和关系，但图形界面做不到这一点。4. 图形用户界面要求用户理解，并且遵循一个任意操作序列。例如，如果需要改变菜单中的选项，需要很多步操作，这些操作必须以一定的顺序进行，所以这种操作具有重复性、冗余性、操作的前后顺序没有很大关系。总之，通过

自然语言，用户可以有一种自然简单的途径和计算机系统进行交互。各种不同的用户可以非常容易地适应，这大大地缩减了训练时间。为了减少对话模型对多通道界面的影响，他们在对话中的一些特殊交互和应用之间建立了对话管理器。他们将这个对话模型应用到了 Eucalyptus[10] 系统中。Eucalyptus 系统是一个具有有限对话能力的图形用户界面系统，它能处理来自键盘和麦克风的强制性命令和数据库查询。可以通过如下方式进行对话：只用自然语言输入方式、只用自然语言进行输入输出方式、可以只采用图形界面方式、只采用用户初始对话方式。Grasso[11] 给出了一个原型系统，它对语音和鼠标进行了整合，所采用的也是任务整合。

4 其它方法及讨论

在语音和手势的整合中，主要的一个问题就是指代问题，要较好地解决这个问题，需要一个知识库存放指代知识，在整合的过程中使用这些知识，我们把这种方法称为基于知识的方法。很多系统都采用这种方法来实现整合，例如 XTRA[12]，ACCORD 和 MMI²[13]，以及 EDWARD[14]。以 EDWARD 为例，它包含两个单通道的子系统，这两个子系统共享一个对话管理器和知识库，在指称表达式的解释和生成时使用该知识库。

这些方法均存在着很大的不足。首先这些方法都是处于一个比较低的层次上的整合。其次以上各种方法都没有语义层次上的整合，然而没有语义层次上的整合方法并不实用。事实上，人在整合的过程中使用了很多知识，对于计算机来讲，这就需要一个知识库来存放整合中用到的这些知识。其次，在整合之前首先对很多问题进行了理解，从语义层次上进行了整合。当然有些问题只在语法和词法层进行整合就可以了，但大多数的情况还是需要语义层的整合，也只有在语义层上的整合算法才更实用。另外，人在学到很多知识之后，也会对自己的整合方法进行修改，也就是在人人交互的过程中，如果得到的整合结果不正确，则重新进行整合，得到一个新的结果，当然在重新整合之前，可能需要学习一些东西，得到一些新知识和新方法。这个过程可以用下图来表示。



参 考 文 献

1 Wu Lizhong, Oviatt S L, Cohen P R. Multimodal Integration: A Statistical View. IEEE TRANSACTION ON MULTIMEDIA, 1999, 1(4): 334～41

2 方志刚. 多通道用户界面模型、整合方法及可用性测试:[杭州大学博士学位论文]. 1998

3 Lee J, Liu K F R, Chiang W. A Possibilistic-logic-based approach to Integrating Imprecise and Uncertain Information. Fuzzy Sets and Systems, 2000, 113: 309～322

4 Cohen P, et al. Quickset: Multimodal integration for distributed application. In: Proc. of the Fifth ACM Intl. Multimedia Conf. New York, ACM Press, 1997. 31～40

5 Johnston M, et al. Unification-based multimodal integration. In: Proc. of the 35th Annual Meeting of the Assocaition for Computational Linguistics, San Francisco, CA, 1997. 281～288

6 Ovitta S, DeAngeli A, Kuhn K. Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In: Proc. of Conf. on Human Factors in Computing Systems: CHI'97, New York, ACM Press, 415-422. New York: ACM Press

7 Oviatt S, Olsen E. Integration Themes in Multimodal Human-Computer. Interaction. In: Proc. of the Intl. Conf. on Spoken Language Processing, Vol. 2, Acoustical Society of Japan, 1994. 551～554

8 Sharon Oviatt P, Fong M, Frank M. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In: Proc. of Intl. Conf. on spoken Language Processing, 1992, 2: 1351～1354

9 Marsh E, et al. Human-Machine Dialogue for MultiModal Decision Support Systems: [NCARAI Report AIC-94-032]. Navy Center for Applied Research in Artificial Intelligence, Navy Research Laboratory, Washington D. C., 1994

10 Wauchope K, Eucalyptus. Integrating Natural Language Input with a Graphical User Interface: [Naval Research Laboratory Technical Report NRL/FR/5510-94-9711]. 1994

11 Grasso M A, Finin T. Task Integration in Multimodal Speech Recognition Environments. Crossroads, Springer-Verlag, 1997, 3 (3): 19～22

12 Allgayer J, et al. XTRA: a natural-language access system to expert systems. International Journal of Man-Machine Studies, 1989, 32: 161～195

13 Wilson M, Conway A. Enhanced Interaction Styles for User Interfaces. IEEE Tansaction on Computer Graphics \& Applications, 1991, 11: 79～90

14 Bos E, Huls C, Claassen W. EDWARD: Full Integration of Language and Action in a Multimodal User Interface. International Journal of Human-Computer Studies, 1994, 40(3): 473～495

（上接第103页）

2 Workflow Management Coalition. The Workflow Reference Model. http://WWW.wfmc.org/standards/docs.htm. 1995

3 Workflow Management Coalition. Interface 1: Process Definition Interchange Process Model. http://WWW.wfmc.org/standards/docs.htm. 1999

4 缪晓阳，石文俊，吴朝晖. 工作流过程定义规范. 计算机科学，2000，27(11)：53～56

5 Workflow Management Coalition Workflow Standard. Workflow Process Definition Interface - XML Process Definition Language. http://WWW.wfmc.org/standards/docs.htm. 2001

6 卢海鹏，周之英. WWW 应用与标记语言. 计算机科学，1999，26 (1)

7 Martin D, Birbeck M, et al. Professional XML. Wrox Press, Ltd. 2000