中国科学: 信息科学 2018 年 第 48 卷 第 4 期: 433–448
SCIENTIA SINICA Informationis
智能时代的人机交互专刊 · 论文

多通道人机交互信息融合的智能方法

杨明浩[1*], 陶建华[1,2,3*]

1. 中国科学院自动化研究所模式识别国家重点实验室, 北京 100190
2. 中国科学院大学人工智能技术学院, 北京 100049
3. 中国科学院脑科学与智能技术研究中心, 上海 200031
* 通信作者. E-mail: mhyang@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn

收稿日期: 2017–10–30; 接受日期: 2018–03–01; 网络出版日期: 2018–04–13
国家重点研发计划 (批准号: 2017YFB1002804) 和国家自然科学基金 (批准号: 61332017, 61425017) 资助项目

摘要 本文首先简要回顾了认知科学在单通道信息加工及多通道信息融合方面的假定; 其次, 介绍了计算机科学在多通道信息融合方面相对于单通道信息处理增强的理论模型及实验验证. 在各通道特征能够同时获得并统一表示的前提下, 多通道人机交互信息的融合可以转化为分类或者回归问题求解. 对于实际的交互系统, 目前的多通道信息融合技术除了依赖单通道信息识别的准确性外, 还依赖于交互系统设计的合理性. 最后通过一个多通道信息融合的人机交互的实例, 讨论了目前多通道交互系统的缺陷, 并给出多通道人机交互信息融合智能方法未来的一个突破方向.

关键词 多通道信息融合, 人机交互, 机器学习, 模式识别, 认知科学

1 引言

因为符合人的交互模式, 多通道交互 (multi-modal human-computer interaction, MMHCI) 被认为是更为自然的人机交互方式 [1,2]. 相对于传统的单一通道交互方式, 多通道人机交互方式在移动交互和自然交互存在着更为广泛的应用潜力, 如智能家居 [3]、智能人机对话 [4,5]、体感交互 [6~8]、教育 [9] 等. 近年来, 人工智能技术使得单一通道认知感知技术, 如语音识别 [10,11]、人脸识别 [12,13]、情感理解 [14~18]、手势理解 [19~22]、姿态分析 [7,23,24]、笔 [25~27]、眼动 [28~30]、触觉 [31~34] 等性能得到快速提升, 计算机能够比较准确理解用户单通道行为. 同时, 高速发展的便携式硬件技术, 催生了一些价格低廉却便于随身穿戴的小巧便捷的传感器. 这些传感技术和设备为准确判断用户行为提供了更多数据.

传统的单一通道人机交互方式, 如较为广泛使用的鼠标键盘, 或者基于笔触的图形界面交互方式, 因为输入设备信息精确和直观, 计算机不用关注用户行为. 然而在多通道移动交互和自然交互条件下,

引用格式: 杨明浩, 陶建华. 多通道人机交互信息融合的智能方法. 中国科学: 信息科学, 2018, 48: 433–448, doi: 10.1360/N112017-00211
Yang M H, Tao J H. Intelligence methods of multi-modal information fusion in human-computer interaction (in Chinese). Sci Sin Inform, 2018, 48: 433–448, doi: 10.1360/N112017-00211

系统需要准确地判断用户 “在做什么” 和 “要做什么”, 才可能对用户行为进行准确反馈. 如在家庭服务机器人领域, 用户指着桌上的苹果对机器人说 “请帮我把苹果拿过来”, 则机器人首先需要根据言语内容和手势所指目标, 准确理解用户的意图是拿取桌上的苹果, 然后机器人凭借目标识别与路径规划技术完成相应任务. 再譬如在人机对话中, 如果用户对计算机的回答表示满意, 然后带着高兴的表情说 “不错, 还以为你不会回答呢”, 计算机如果没能正确理解用户意图, 着力于解释其回答问题的能力, 反而带给用户不好的体验. 因此, 多通道人机交互中用户意图的准确理解是交互自然与否的关键, 而如何根据不同通道信号进行有效融合并计算是意图准确理解的重要手段.

多通道人机交互是一个非常动态和广泛的研究领域, 本文目标不是为了呈现一个该领域完整的概括, 而是从面向用户意图理解的多通道信息融合角度出发, 分析单通道信息处理的特点, 然后介绍多通道信息融合对单通道信息理解增强的分析和实验验证; 通过分析具有一定特色的多通道用户行为融合模型、方法及应用实例, 讨论未来人机交互在多通道信息融合的智能计算方面可能存在的突破.

2 多通道信息处理模型

2.1 单一通道信息处理

交互中的通道有很多类别, 尽管这些通道数据获取以及存储方式存在很大区别, 但从认知科学的角度看, 它们在信息处理流程上具有一些共同特点.

2.1.1 不同通道信息处理的认知假定

认知心理学认为人类处理信息时遵从 3 个假定: 通道加工、容量有限及主动加工假定. 前两个假定主要与单通道信息处理密切相关, 后一个假定更多与多通道信息处理相关. 单通道加工假定指人类首先对各通道信息进行单独加工, 并形成各自通道的认知特征 [35,36], 根据通道关注的外界情况, 选取学习对象的表征. 这些外界对象的表征可以视为外界事物的抽象, 甚至是已有的抽象的组合 [37].

容量有限假定指各通道一次性加工的数据容量是有限的, 即在没有对信息进行加工的前提下, 学习者能够保持的通道信息容量具有一定限制 [38,39]. 实验发现, 人类的记忆广度大约为 7 个单元, 比如被试者以每秒 1 字的顺序读一串数字, 紧接着, 被试者能够准确复述的数字广度大约为 7 个数字左右 [40]. 还有学者认为容量更少 [41]. 这说明, 人类在没有利用联想、推理、记忆等认知功能的前提下, 其所能关注的信息是非常有限的 [42].

近年来, 更多的心理学及认知科学实验和发现进一步验证了通道加工和容量有限假定 [37,43], 认知科学认为人类的记忆结构分为 3 级: 感觉记忆、工作记忆和长时记忆 [37]. 容量有限假定大部分发生在感觉阶段以及工作记忆阶段, 而通道加工则从感觉记忆、工作记忆持续到长时记忆阶段 [44].

2.1.2 不同通道信息处理的一般计算模式

交互系统中, 研究者们习惯将不同通道信号转化为界面操作, 如触觉交互中的位置和触碰判断、笔式交互中的触点和点击判断、眼动交互中的视点和视线判断等. 另外还有一些不便于转化为界面操作的单通道用户交互行为, 如情感交互中语音对话、手势跟踪等. 这类通道信息首先被转变为行为识别, 然后系统根据行为识别结果给出反馈. 这些单一通道信号转化为意图理解的过程可以采用式 (1) 简化描述:

$$y^t = f(x^t, x^{t-1}, \ldots, x^{t-l}), \tag{1}$$

其中, x^t 表示时刻 t 的某通道输入信号, l 表示通道加工的信息长度, y^t 表示时刻 t 的通道信号的认知结果, 比如笔触位置、手势姿态、情感类别、语音对应的文本等. 单一通道信息的处理过程在计算机中的处理就变为了函数拟合或者分类的问题.

认知科学中的通道加工和容量有限假定对应了式 (1) 中的两个关键变量: x 和 l. x 对应于通道加工中的特征表示, 如语音信号处理中, 针对不同的应用, 不同的声学特征会带来不同的效果提升. 在深度学习兴起之前, 传统的单通道信息计算中, 人工设计的通道加工特征成为意图理解准确性的一个关键因素. 深度学习兴起后, 深度网络结构可以自动抽取通道特征, 其自动获得的特征在用于识别和分类方面, 甚至表现出比人工精心设计的特征更好的性能 [45,46]. 总的说来, 针对不同交互任务, 如何在通道加工过程中发现更好的特征表示是交互意图理解的一个重要因素.

l 对应于容量有限假定. 在传统的采用机器学习求解过程中, l 作为输入信号所观测的时间窗, 其值的大小会影响到 y 的精确程度. 如在脸部的连续维度情感识别中, 连续 11 帧支持向量机得到的情感识别结果送入到新的级联支持向量机中, 得到的情感判断准确度相对更高 [47]; 同样在语音识别 [11,48]、手势识别应用 [22,49] 中, 研究者们发现, 合适的 l 值将会获得更好的识别效果. 即便是本身就考虑了记忆因素的长短记忆型递归神经网络, 在中间层有目的的加入以时间片为单位的池化特征, 也会不同程度地提高模型精度, 这样的处理方法在情感识别 [16]、手势识别 [19,20]、语音识别 [11,48] 中也被证实是有效的.

2.2 多通道信息融合

2.2.1 多通道信息融合认知假定

随着计算机计算能力的提高, 在某些领域, 计算机对单一通道信号的分辨能力接近甚至超过一般用户. 即便如此, 单一通道信息处理的认知处理流程从输入信息和输出信息的对比看, 其过程仍旧类似样本学习后的一个再辨识的过程.

认知科学认为认知的结果分成两类: 记忆和理解. 记忆是指人能够回忆、辨认和识别过去呈现的学习材料或者信息的能力. 理解是对呈现的教学内容建构连贯的心理表征的能力, 其发生在记忆之后, 表现为可以在一个新的情境中应用所学到的知识. 这种能力可以使学习者在一个新的情境中应用所学到的知识 [50], 其效果可以采用迁移性测验来测量, 即, 学习者将面对一些新的情境, 他们需要将学过的知识进行迁移, 才能解决问题. 多通道信息融合有利于通过记忆形成理解, 反过来继续促进记忆 [51,52].

2.2.2 多通道信息融合认知模型

认知增强建立在认知科学学习过程的第 3 个假定上: 主动加工假定. 主动加工指在通道加工与容量有限假定之后, 人们会为了对学习到的不同通道知识和经验建立一致的心理学表征, 会主动调用注意机制, 对信息进行编码和组织并将新知识与旧知识融合等策略参与认知加工 [44]. 美国教育心理学家 Mayer 在其专著《Multimedia Learning》中介绍了一个人类多通道信息认知加工过程的概念图, 其中的一些观点和实验与多通道交互中人类的认知过程非常类似 [53]. 多通道认知加工过程如图 1 所示.

图 1 中, 从左到右的模块为多通道信息呈现、感觉记忆、工作记忆以及长时记忆模块, 分别对应着多通道信号获取之后的人类认知的 3 个假定. 感觉记忆和工作记忆的部分对应于通道加工. 此时, 人类的感知通道获取得到的信息, 包括图像、语音、触感等信号, 完成特征表示和编码, 有选择地进入工作记忆区. 因为容量有限原则, 被选择的特征在大脑对应区域保留很短的一段时间, 并进行多通道

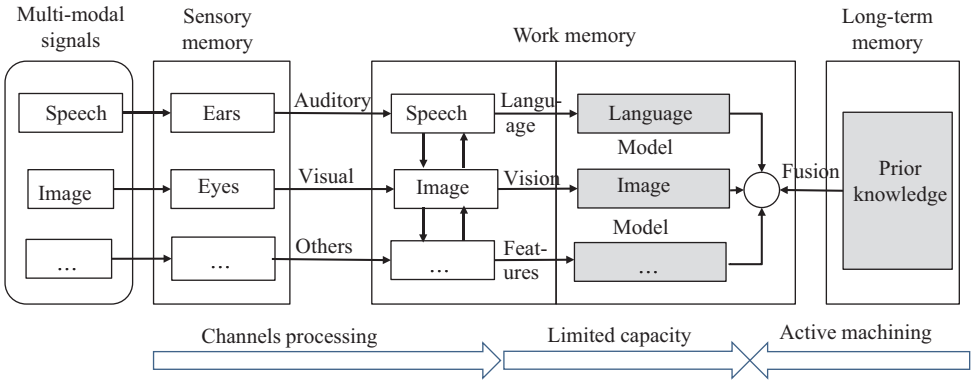

图 1　多通道信息加工及认知过程

Figure 1　cognitive process of multi-modal information in brain

特征的关联和融合. 针对具体任务, 主动加工调用包括注意、联想和推理的机制, 触发长时记忆学到的历史知识, 综合工作记忆中的多通道融合信息得到最后的判断结果.

多通道学习的认知过程表明多通道信息的融合发生在工作记忆中, 同时融合过程会触发长时记忆形成的知识. 虽然至今为止, 知识在大脑中的形成过程和存储形式究竟没有明确的结论, 但认知科学家采用实验的方法探索了多通道信息融合相对于单通道信息处理的增强效应, 首先需要验证的是: 基于已经学到的单通道知识, 多通道信息融合是否会为信息理解带来帮助? 如果有, 相对于单通道信息而言, 后者会在精确度上带来多大的提升?

2.2.3　多通道信息融合增强验证

简化各单一通道信息的表达形式, 有助于定量的探索多通道信息对于单通道信息的增强程度. 在式 (1) 的基础上, 式 (2) 给出了多通道信息融合的描述形式:

$$y^t = f\left[\oplus_{k=1}^{K}(x_k^t), \oplus_{i=1}^{k}(x_k^{t-1}), \cdots, \oplus_{i=1}^{k}(x_k^{t-1})\right], \tag{2}$$

其中, k $(2 \leqslant k \leqslant K)$ 为参与信息融合的通道数, x_k^t 表示时刻 t 的第 k 通道的输入信号, 符号 $\oplus_{k=1}^{K}(x_k^t)$ 表示时刻 t 多通道信号的融合, y^t 为时刻 t 的多通道信号融合的标注结果, 尽管式 (2) 可以简化为一个函数拟合问题, 但其真正难点在于各通道信号的表示差别迥异, 这使得 $\oplus_{k=1}^{K}(x_k^t)$ 难以得到准确的表达, 多通道信息难以统一描述.

文献 [54] 介绍了一个定量分析多通道信息融合性能的实验. 假定第 k 个通道信号服从均值为 u_k, δ_k 的 Gauss 分布, 记为 $s_k \sim N(u_k, \delta_k)$, 记每个通道信号的置信度为 $w_k = \frac{1/\delta_k^2}{\sum_k (1/\delta_k^2)}$, 则某一时刻, 多通道信号融合后表示为

$$\tilde{\mathrm{S}} = \oplus_{k=1}^{K}(s_k) \sim N\left(\sum_{k=1}^{K} w_k u_k, \sum_{k=1}^{K} w_k \delta_k\right). \tag{3}$$

式 (3) 实际上为多个 Gauss 信号的最大似然估计, 文献 [54] 进一步借助虚拟现实技术构建了视觉和触觉双通道感知场景深度信息的信息融合感知实验. 多人实验后的统计结果表明: 在视觉和触觉双通道感知场景深度信息中, 场景深度估计实验的结果服从式 (3) 的描述, 即, 在单通道信号服从 Gauss 分布的情况下, 多通道信号的融合可以表示为多个信号的最大似然估计. 同时, 多个通道中, 置信度越高 (或者方差更小) 的通道在融合后占有更加明显的主导地位.

3 多通道信息融合计算

多通道信息融合按照发生的时间顺序, 可以分为前期融合和后期融合; 按照信息融合的层次来分, 融合可以分别发生在数据 (特征) 层、模型层及决策层; 如果按照处理方法来分, 可分为基于规则的融合, 或者基于统计 (机器学习方法) 的融合 [55,56]. 也有文献根据多通道信息的相关性, 把它们的关系分为信息互补、信息互斥、信息冗余的特点, 然后根据其信息特点进行分别融合 [57]. 以上这些融合策略中, 基于统计方法或者基于机器学习方法的融合在计算层面发生. 其他的融合方法更偏重于设计, 因此后面介绍几种比较广泛实用的基于统计和机器学习的信息融合计算模型.

(1) Bayes 决策模型. Bayes 决策的特点在于其能够根据不完全情报, 对部分未知的状态采用主观概率估计, 然后用 Bayes 公式对发生概率进行修正, 最后利用期望值和修正概率做出最优决策 [58]. 在多种通道信号联合分布概率部分已知的情况下, Bayes 决策可以根据历史经验反演得到某些缺失的信号, 从而得到整个多通道信号融合整体最优评估. 设不同通道信号在某时刻的联合分布概率记为 $p_s(S)=p_s(s_1,s_2,\ldots,s_D)$, D 表示通道数, 已知各通道联合建模的概率联合分布, 则某通道观测信号的边缘分布概率 $p_D(d_{\rm obs})$ 为 $p_D(d_{\rm obs}|S)\times p_s(S)$, 其中 $d_{\rm obs}$ 表示第 d 通道信号的观测值. 根据 Bayes 公式, 假定先验知识的情况下, 已知某通道信号的边缘概率, 联合分布概率为

$$p_D(S|d_{\rm obs})=p_D(d_{\rm obs}|S)\times p_s(S)/p_D(d_{\rm obs}), \tag{4}$$

$$p_D(d_{\rm obs})=\int_S p_D(d_{\rm obs}|S)\times p_s(S)\times {\rm d}V_s. \tag{5}$$

式 (4) 和 (5) 中, $p_D(d_{\rm obs})$ 为对某通道的边缘分布观测, 对应于某通道信号的实际观测. 根据 $p_D(d_{\rm obs})$, $p_D(d_{\rm obs}|S)$ 以及部分已知的 $p_s(S)$ 先验信息, 联合式 (4) 和 (5), 迭代获得精确的 $p_s(S)$ 以及补全缺失的通道信息 $p_D(d_{\rm obs})$.

因为 Bayes 决策方法根据不完全情报反演出部分观测条件下的最优决策的优点, 其在多通道信息整合分析、多通道观测手段联合建模分析上体现出一定优势, 其在鲁棒人脸跟踪 [59]、用户行为感知 [60]、机器人姿态估计和避障 [61]、情感理解 [62]、多源传感器信息对齐及观测数据分析 [63] 等方面得到非常好的应用.

(2) 神经网络模型. 传统的神经网络模型在非线性函数拟合方面表现出很好的性能, 结构更深的神经网络模型在语音识别、人机对话、机器翻译、语义理解、目标识别、手势检测与跟踪、人体检测与跟踪等领域广泛应用. 如在情感识别领域, 采用相似度评估, 目前采用深度长短时记忆神经网络模型 (long short-term memory neural networks, LSTM), 计算机得到的最好结果与专业人士识别相差 10% 左右 [16,18]; 在语音识别领域, 目前针对方言口音的语音识别, 深度递归神经网络 (recurrent neural networks, RNN) 在字识别准确度可以达到 95% [64], 接近人类水平; 在图像目标识别领域, 超大规模深度卷积神经网络 (convolution neural network, CNN) 已经超过普通人类辨识水平 [65,66]. 在单一通道的深度神经网络模型技术上, 很多研究者综合上述 LSTM, RNN, CNN 结构, 构建面向多通道信息融合的大规模深度神经网络模型, 力图在融合阶段无差别的处理多通道信息. 通常而言, 多模态信息融合应用于深度神经网络, 其和普通的多通道信息融合一样, 在结构上发生在 3 个层次, 分别是早期数据级融合、中期模型级融合及后期规则级融合 [55]. 图 2(a)~(c) 分别对应了上述 3 种多通道信息融合的抽象表示.

基于上述结构, 多种多样的多通道信息融合可以进一步组合上述结构, 构建更为复杂的大规模结构, 实现多任务学习 [67~69]、跨模态学习 [70] 等功能, 同时在基于多模态数据的联合训练情况下, 这类

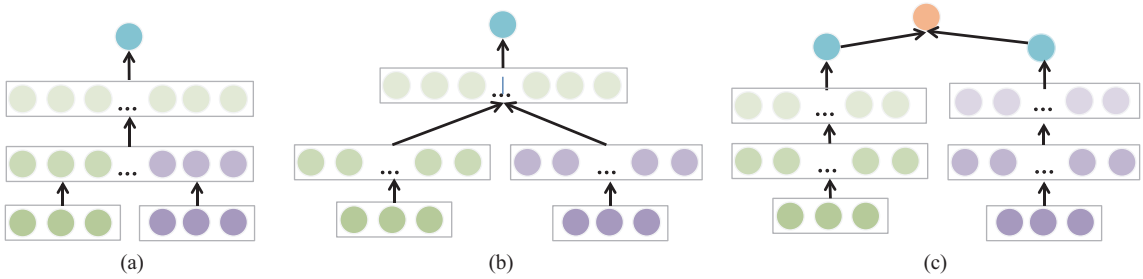

图 2 (网络版彩图) 神经网络信息融合结构抽象

Figure 2 (Color online) Structure of classical neural networks for information fusion. (a) Fusion at data or feature level; (b) fusion at model level; (c) rules based fusion

结构在运行时可以做到即使某一个模态信息缺失整个网络也能取得不错的效果, 在多通道情感识别、语义理解、目标学习等领域取得很好的效果. 尽管如此, 这类网络相对于任务来说还是相对 "具体", 如果要换一个任务, 用户就需要修改网络结构包括重新调整参数, 这使得深度神经网络结构的设计是一个耗时耗力的过程. 因此研究者们希望一个混合的神经网络结构可以同时胜任多个任务, 以减少其在结构设计和训练方面的工作量. 鉴于此, 研究者开始致力于首先采用大数据联合训练构建出多通道联合特征分享层, 然后在识别阶段可以同时进行多任务处理的深度多模态融合结构. 如 Google 的学者尝试建议一个统一的深度学习模型来自适应地适配解决不同领域、不同数据模态下的多个不同类型的任务, 且在特定任务上的性能没有明显损失的模型 [71]. 该模型构架请见文献 [71] 的图 2, 由处理输入的编码器、编码输入与输出混合的混合器、混合输出的解码器 3 个部分构成, 文献 [71] 的图 3 给出了这 3 个部分的详细描述. 每一个部分的主体结构类似, 均包含多个卷积层、注意力机制和稀疏门控专家混合层. 其中, 不同模块中的卷积层的作用是发现局部模式, 然后将它泛化到整个空间; 注意力模块和传统的注意力机制的主要区别是定时信号, 定时信号的加入能让基于内容的注意力基于所处的位置来进行归纳和集中; 最后的稀疏阵列混合专家层, 由前馈神经网络 (专家) 和可训练的门控网络组成, 其选择稀疏专家组合处理和鉴别每个输入.

(3) 基于图模型的信息融合. 图模型将概率计算和图论结合在一起, 提供较好的不确定性计算工具, 其构成上的节点以及节点之间的连线, 使得其在计算变量与周围相连变量的关系上具有一定优势. 根据节点之间的连线是否有方向, 图模型主要可分为无向图模型和有向图模型. 进一步借助多尺度分析, 无向图在场景分割、视频内容分析、文本语义理解等方面应用广泛, 如基于 Markov 场模型的视频中人体运动分割、检测与跟踪 [72], 无向图模型联合多文档摘要抽取 [73], 基于分组的无向图估计, 用于找回多通道信息中丢失的特征 [74], 基于无向图模型的文本、视频及音频信息情感分 [62], 基于 Gauss 图模型的多模态脑分区中大脑欢愉活动分布情况检测 [75].

相对于无向图模型, 有向图模型节点之间的连线不仅记忆了数据流向, 还记录有学习过程中的状态跳转概率, 有向图模型除了可以用于不确定性计算外, 还可用于面向时序问题的决策推理, 如基于动态 Bayes 模型模仿人类对文字的书写过程 [76], 基于 Markov 决策过程理解手势与姿态 [77], 基于有向图模型的多用户行为冲突最优决策 [57,78], 基于加权有限状态自动机的多通道人机对话模态冲突对话决策策略 [57] 等.

除了以上多通道信息融合计算模型外, 还有很多其他的模型也用于多通道信息融合, 如多层支持向量机、决策回归树、随机森林等方法, 由于篇幅这里不一一叙述.

4 多通道交互信息融合管理

多通道信息提供了界面鲁棒性、纠正错误或复原方式, 因而其对不同的状况和环境增加了交互可选方式. 然而就其本质而言, 多通道用户行为不属于界面操作, 因此如果仅仅把多通道信号简单转化为界面操作或者事件可能是无效的, 甚至是无益的 [1]. 有研究者提出多通道交互会产生一种多通道困扰问题 (common misconceptions or myths), 原因在于多通道人机交互给予了用户较大的表达自由度, 如交互中用语音、姿态、情感表达具有不确定和随意的特点. 这种表达的丰富性和模糊性难以准确映射为传统人机交互的界面操作, 使得系统反馈不够准确 [79]. 系统需要对多通道交互信息的融合进行管理. 本节首先介绍多通道交互信息管理的概念, 然后介绍一个多通道信息融合和交互管理的实例.

4.1 多通道交互信息管理

管理实际上就是把交互过程规划起来, 本质上是多通道信息在决策层上的语义融合过程. 早期的多通道信息管理方法多采用基于规则的方法, 为了减少各分析模块的不确定性和歧义, SmartKom 对各通道单独解析, 通过自适应的置信度测量, 逐步进行扩展达到各通道的融合, 形成用户意图网格加以后续处理 [80]. 文献 [81] 利用有限状态机将来自手势和语音的通道信息进行整合和理解, 让机械臂完成抓取物体的交互任务. 在人机对话中, 语音信息是主要交互通道, 文献 [56] 据不同通道对语音交互的影响, 根据它们与语音信号的关系把它们的处理方式分为 3 种模式: 信息互补模式、信息融合模式和信息独立模式. 然后根据语音识别、表情跟踪和识别、身体姿态跟踪和识别、情感识别的结果, 进行用户的语义理解和对话管理. 文献 [82] 在构建面向小孩朗读陪伴机器人交互系统中, 将交互过程分为 8 个状态, 根据小孩不同的交互需求和情况, 系统在 8 个状态之间跳转, 保证交互的鲁棒和连贯. 可以看到, 交互的管理的目的是对多通道信息的计算在时间上和空间上进行规划, 使用户交互更加鲁棒和自然, 因此面向任务的人机交互的管理重在设计, 规划在很大程度上涉及推理和决策等关于时序的不确定性计算, 因此有向图模型是人机交互任务管理的重要工具之一 [57,83,84].

4.2 一个实例: 具有交互学习能力的智能机械臂写字系统

文字的学习过程是实现其形音义的同时记忆, 这是一个典型的音视频通道联合学习的过程, 其中形和音的学习记忆是汉字学习的基础. 在家庭教育方面, 构建具有交互学习能力的机器人用于家庭写字陪伴, 在增加教育产品的趣味性的同时 [9], 也有利于部分缓解家长在陪伴小孩方面的压力 [82], 在小孩书写教育方面具有广阔的应用前景.

对小孩学习而言, 语音和视觉呈现是自然的交互通道. 深度学习技术使得语音识别和语音合成技术可以非常准确地理解人类语音及模拟人类说话. 图像处理方面, 机械臂借助图像处理技术已经可以非常准确地提取汉字笔画的起点、交叉点和结束点 [85], 根据一定规则, 机械臂可以根据这些得到的汉字特征点规划出汉字的书写轨迹, 但是基于规则书写的笔画不一定正确. 此时, 如果儿童采用语音和视觉信息融合的方式教机械臂写字, 可以达到进一步巩固书写汉字的目的.

实际系统中, 交互中的语音和图像通道采集具有在空间上隔离, 在时间上分散的特点. 同时, 机械臂从用户实时书写的过程中学习书写汉字是一个少量样本甚至是单样本的过程, 因此采用仅仅依靠目前流行的深度学习方法并不适用于用户交互学习的智能机械臂写字系统. 在语音识别、视频序列分析、对话意图理解的基础上, 有向图模型是本系统交互流程的核心. 图 3 给出了具有智能交互学习能力的机械臂写字系统流程图.

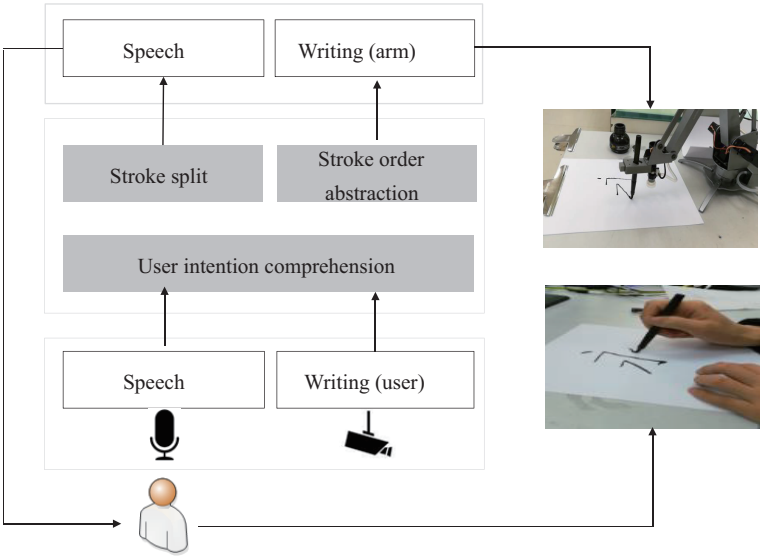

图 3 (网络版彩图) 具有智能交互学习能力的机械臂写字系统

Figure 3 (Color online) Intelligent robot arm writing system with interactive learning ability

(1) 系统结构. 系统分为 3 个主要部分, 信息输入、信息处理以及输出反馈. 系统的输入信息模块包含用户的语音信息以及摄像头观察到的文字的图像信息. 图中灰色部分为本系统关键技术模块, 主要包含两大部分, 一是对用户的语音信息进行分析, 获得用户想要写的关键字及用户意图, 并进行对话管理; 二是通过对摄像头看到的图像信息进行分析, 对检测到的汉字进行自动笔画拆分和笔顺提取, 对于正在教授的字, 跟踪笔迹顺序, 学习新写法. 最后通过对话管理, 机械臂会以对话的形式进行反馈与用户交互, 并调用机械臂写字程序, 书写需要写的字. 根据任务, 系统中机械臂的功能主要包括如下几方面.

语义分析: 分析用户对话语音, 如用户对机械臂说 “请写一个天气的天字”、“写的不好” 等, 系统理解用户语音的意图.

图像分析: 摄像头实时检测场景中是否有要写的字, 如果有, 会记录该字并自动提取笔画; 同时用户教机械臂写字的过程中, 摄像头会实时记录用户教的笔画顺序, 并进行学习和记忆.

对话管理: 控制整个系统的对话流程, 是本系统的核心, 用于管理整个系统的交互逻辑.

状态分类: 系统中的状态分类在一定程度对应到用户意图理解, 系统共设置 Write, Teach, Positive, Negative, Others 等意图, 其中 Write 表示机械臂写字, Teach 表示教机械臂写字, Positive 表示用户的肯定或正面评价, Negative 表示用户的否定或负面评价, Others 表示其他意图, 是闲聊等与书写无关的意图.

笔画自动拆分: 机械臂对看到的字提取特征点, 然后进行自动笔画拆分, 并基于规则进行笔画顺序规划.

笔顺提取: 对于用户正在教授的字, 提取用户笔画顺序, 学习记忆新的写字方式.

语音反馈: 通过对话管理, 机械臂将需要反馈的信息以语音形式输出, 如机械臂对用户提问或者反馈 “请问您要写什么字”、“那您教我吧” 等.

机械臂书写: 机械臂根据自动拆分笔画的笔画顺序, 或者机械臂根据用户教授的笔画顺序写字.

(2) 多通道信息融合及交互过程管理. 系统的交互过程管理设计为有限状态自动机. 该有限自动

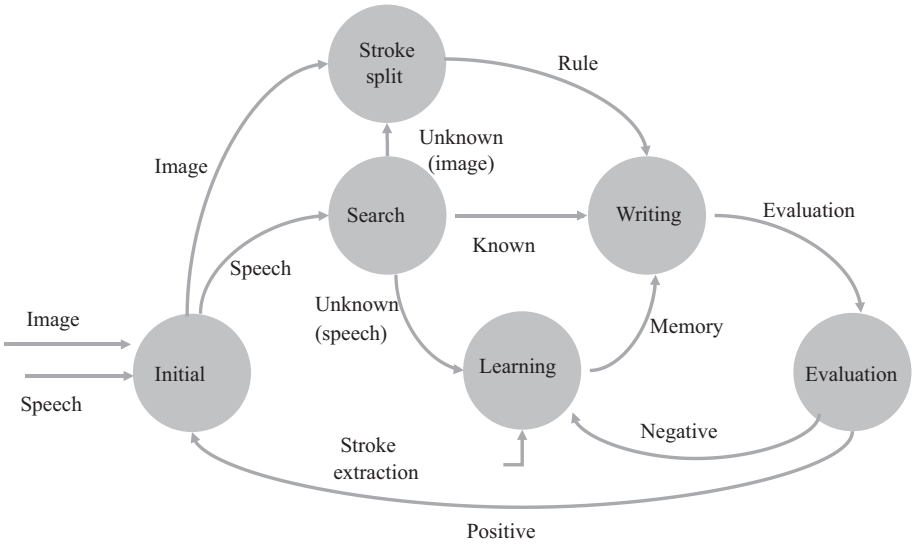

图 4 具有交互学习能力的智能机械臂意图管理及策略

Figure 4 Intention management and strategy for intelligent robot arm writing system

状态机如图 4 所示. 其中, “图像输入”、“语音输入” 及 “笔顺提取” 为音视频输入信息, 圆圈内表示系统状态, 在当前状态条件下, 系统根据不同的用户操作进入不同的系统状态. 如系统在写完一个字后, 系统主动询问用户字写的是否正确. 如果用户评价是负向的, 则系统会提示用户教授正确的写法, 然后学习用户的正确写法, 并实现记忆; 如果用户评价是正向的, 则系统巩固记忆, 返回到初始接收图像和语音输入状态.

(3) 实验分析. 对用户而言, 交互体验十分重要. 实验主要采用用户评估的方式, 验证该机械臂的写字效果及用户体验是否达到设计最初需要.

首先, 测试了用户个性化交互对实验的影响. 本系统中, 在用户采用书写教授机械臂书写的过程中, 不同用户的书写具有各自的特点, 只有将不同用户书写的个性化的字体对齐, 才能使得系统可以识别并理解不同用户书写同一字体的书写顺序. 针对用户个性化书写的问题, 系统采用了基于中文字特征点匹配的方法消除用户个性化书写时可能导致的字体匹配的影响 [86]. 图 5 给出了部分不同用户书写汉字的对齐效果, 其中第 1 行和第 2 行为不同用户书写的同一个汉字, 第 3 行为两个不同用户书写的汉字的对齐结果. 实际测试, 在笔画小于 10 以内的字体, 匹配正确率达到 90%, 使得用户在书写时可以采用个性化字体书写方式教授机械臂, 系统可以很好地对这些不同个性化的字进行较好的匹配. 这也是本系统中笔画匹配的多通道信息融合及对话管理的基础.

其次, 实验分析了交互过程中影响用户体验的主要因素. 系统给用户带来不好体验主要来源于语音识别或者语义理解的错误, 这些错误对用户意图识别的准确性判断造成一定的影响. 本实验对 150 轮中的 624 句有效对话的用户意图识别进行测试, 并分析在语音识别正确与不正确情况下对交互意图分类的影响. 实验结果如表 1 所示.

从实验结果可以看出, 语音识别的正确率为 81.9%, 在语音识别正确的情况下, 对话意图的正确率为 86.7%. 同时, 在 14.3% 的语音识别错误率, 约 113 句语音识别错误的情况下, 大约只有 62 次的意图可以被机械臂正确理解. 在这种情况下, 对话意图准确率仅为 54.9%. 因为语音识别错误, 导致意图理解的准确率下降了 32% 左右. 比如用户对机器臂说 “请写一个用户的用字”, 在噪声环境下, 如果识

图 5 (网络版彩图) 不同用户个性化书写对齐效果

Figure 5 (Color online) Points matching automatically between different users' writing

表 1 用户对话意图识别准确率

Table 1 Accuracy of users' dialogue intention [a)]

	Correct ASR	Correct ASR and intentions	Inaccurate ASR with correct intentions
Number of correct feedback	511	443	62
Accuracy	0.819	0.867	0.549

a) ASR: Automatic speech recognition.

别为 "请写一个拥护的拥字", 则这种错误在后续的语义理解阶段会大幅度降低用户交互体验. 可以看到, 语音识别的准确性对用户交互意图的正确理解有很大影响.

最后, 在语音识别正确的情况下, 实验评估了不同通道信息对交互体验的影响. 本实验中, 用户可采用语音的方式告诉系统 "请写一个永远的永字", 同时给系统展示一张 "永" 字的卡片给系统, 系统根据已经学到的笔顺写 "永" 字. 如果系统没有该字信息, 则系统会主动询问字的信息. 整个过程中, 这两种交互方式分别为用户主动交互和系统主动交互. 在系统主动交互时, 用户根据系统的提问分别采用语音和图像单通道给与交互信息. 而在用户主动交互时, 用户可采用两通道混合的方式教授系统写字. 实验邀请了 20 位参与者进行机械臂系统的不同通道体验和评测. 在体验之前, 系统设计人员详细描述了单通道和多通道交互模式的区别, 并要求他们每人至少分别进行 2 次单通道和多通道交互模式的写字测试. 图 6 给出了 20 位参与者的主观评分的平均值、最高值、最低值、上下均方差值的统计情况, 其中 5 分为最高分.

从图 6 可以看到, 在平均分、最低值、上下均方差值方面, 多通道交互模式比单通道交互模式取得更好的分数. 总体看来, 采用较为灵活的用户主动的多通道交互模式比单一通道交互模式能取得更好的交互体验. 另外, 在个别情况下, 部分用户也认为系统主动的有步骤的单一交互模式也能得到很好的交互体验, 比如基于单通道的被动交互模式的最高值为 0.405, 比多通道交互模式的最高值 0.395 更高一些. 给出该分数的用户认为在单通道交互模式下, 系统询问用户书写要求, 并指示用户有步骤的教授自己写字, 显得比较有章法.

为了评估系统总体体验, 实验另外邀请了 67 位参与者参与机械臂系统的体验和评测, 每人至少进行 3 次写字测试. 与上次实验不同, 本次实验中, 用户没有被详细告知单通道和多通道交互模式, 只是要求他们按照自己的方式与系统交互, 然后要求每个体验者对系统进行满意度投票, 总共 5 个选项, 分别是很满意、满意、一般、不太满意和很不满意, 其统计分布如图 7 所示. 由图 7 评测结果可知, 70.1% 的

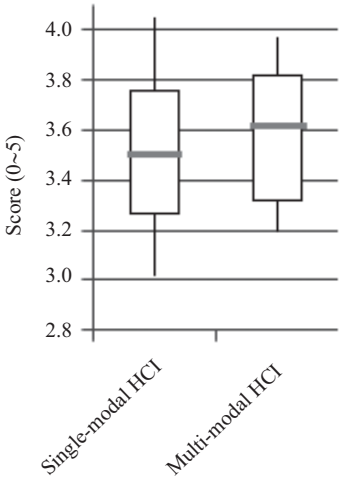

图 6 多通道体验主观评测

Figure 6 Evaluation of user experience on multi-modal interaction

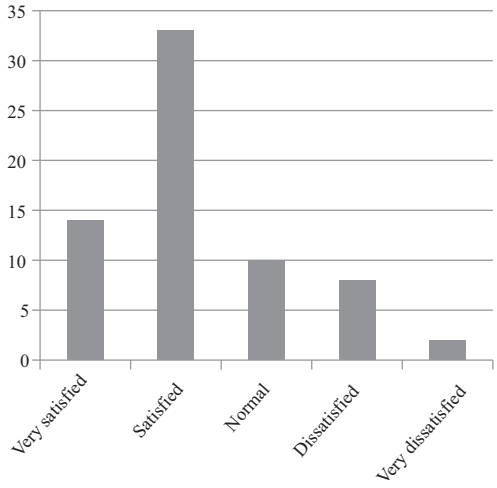

图 7 用户体验满意度评测

Figure 7 Evaluation of user experience

人对交互体验感觉满意或者很满意, 有 14.9% 人对系统体验不太满意或者很不满意. 从用户的主观评测可以看出, 该系统在多通道信息融合的设计方面, 获得了不错的交互体验, 得到大多数体验者认可.

5 总结与展望

本文简要回顾了认知科学和计算机科学在多通道信息融合方面的假定和实验验证. 然后介绍了常见的多通道信息融合模型, 以及面向交互学习的智能机械臂写字系统实例. 在认知科学对人类多通道信息的整合能力并不完全了解的情况下, 可以看到目前的融合方法及交互系统依旧依赖于经验设计. 同时, 模型的迁移能力在支持个性化用户行为理解方面需要提高, 比如语音识别和视频分析发生微小错误情况下, 系统不能很好适应, 如智能机械臂中 “用户” 识别为 “拥护” 的情况, 会导致系统后续操作偏离用户意图. 另外, 现存的多通道融合模型和交互系统缺乏自我增长的能力. 未来的人机交互应用, 如家庭服务机器人、智能教育等, 由于用户的行为比较开放自由, 使得交互过程中会不间断地出现新环境和新实物, 目前的交互模型, 不管是基于统计的模型, 还是基于模式设计的面向任务的交互模型, 因为其模型固定, 新知识如何和旧知识有效融合成为一个难点.

要适应人机交互中用户行为自由, 交互环境变化的特点, 多通道信息融合技术至少需要具有通过简单预设定, 就能够伴随人机交互任务在新的环境里和用户一起学习成长的能力, 这些能力为 (1) 准确识别用户已知交互意图; (2) 准确归纳已知意图的新个性化行为到已知意图; (3) 准确判断用户行为中出现的未知意图类别, 在交互中对未知意图类别进行标识, 并添加到已知学习模型; (4) 模型中加入已知意图的新个性化行为和新意图类别, 不影响原有意图理解准确度, 同时模型得到增长. 目前, 已有的人工智能方法对上述前 3 点研究内容进行了部分探索, 在一些数据集上取得一定的成果, 然而在多通道人机交互应用方面, 缺乏一个同时满足上述 4 个特点的多通道人机交互模型. 构建出具有智能增长的多通道信息融合和理解模型, 使得计算机系统具有在与用户的交互中学习、理解并整合新知识到已有知识的能力, 将是人机交互在多通道信息融合方面的一个重要的突破方向.

参考文献

1 Cohen P R, McGee D. Tangible multimodal interfaces for safety-critical applications. Commun ACM, 2004, 47: 1–46
2 Jaimes A, Sebe N. Multimodal human-computer interaction: a survey. Comput Vis Image Und, 2007, 108: 116–134
3 Meyer S, Rakotonirainy A. A survey of research on context-aware homes. In: Proceedings of Australasian Information Security Workshop Conference on ACSW Frontiers, Adelaide, 2003. 159–168
4 Yang M H, Tao J H, Li H, et al. A nature multimodal human-computer-interaction dialog system. In: Proceedings of the 9th Joint Conference on Harmonious Human Machine Environment (HMME), Nanchang, 2013
5 Yang M H, Gao T L, Tao J H, et al. The error analysis of intention classification and speech recognition in speech man-machine conversation. In: Proceedings of the 11th Joint Conference on Harmonious Human Machine Environment, Huludao, 2015 [杨明浩, 高廷丽, 陶建华, 等. 人机对话中的意图及语音识别错误对交互体验的影响分析. 见: 第十一届全国人机交互学术会议 (CHCI 2015), 葫芦岛, 2015]
6 Duric Z, Gray W D, Heishman R, et al. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. Proc IEEE, 2002. 1272–1289
7 Wang L, Hu W, Tan T. Recent developments in human motion analysis. Pattern Recogn, 2003, 36: 585–601
8 Seely R D, Goffredo M, Carter J N, et al. View invariant gait recognition. Adv Pattern Recogn, 2009, 23: 1410–1413
9 Chin K Y, Hong Z W, Chen Y L. Impact of using an educational robot-based learning system on students' motivation in elementary education. IEEE Trans Learning Technol, 2014, 7: 333–345
10 Oudeyer P Y. The production and recognition of emotions in speech: features and algorithms. Int J Human-Comput Studies, 2003, 59: 157–183
11 Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition. Comput Sci, 2015, 10: 429–439
12 Yang M H, Kriegman D, Ahuja N. Detecting faces in images: a survey. IEEE Trans Pattern Anal Mach Intell, 2002, 24: 34–58
13 Zhao W, Chellappa R, Phillips P J, et al. Face recognition: a literature survey. ACM Comput Surv, 2003, 12: 399–458
14 Pantic M, Rothkrantz L J M. Automatic analysis of facial expressions: the state of the art. IEEE Trans Pattern Anal Mach Intell, 2000, 22: 1424–1445
15 Tao J H, Tan T. Affective computing: a review. In: Proceedings of Affective Computing & Intelligent Interaction. Berlin: Springer, 2005. 981–995
16 Chao L L, Tao J H, Yang M H, et al. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, Brisbane, 2015. 65–72
17 Wang S J, Yan W J, Li X, et al. Micro-expression recognition using color spaces. IEEE Trans Image Process, 2015, 24: 6034–6047
18 He L, Jiang D, Yang L, et al. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proceedings of ACM International Workshop on Audio/visual Emotion Challenge, Brisbane, 2015. 73–80
19 Ge L H, Liang H, Yuan J S, et al. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In: Proceedings of Computer Vision and Pattern Recognition, Honolulu, 2017. 1991–2000
20 Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. In: Proceedings of International Conference on Computer Vision, Venice, 2017. 4903–4911
21 Ruffieux S, Lalanne D, Mugellini E, et al. A survey of datasets for human gesture recognition. In: Proceedings of International Conference on Human-Computer Interaction. Berlin: Springer, 2014. 337–348
22 Hasan H S, Kareem A. Human computer interaction for vision based hand gesture recognition: a survey. Artif Intell Rev, 2015, 43: 1–54
23 Hu W, Tan T, Wang L, et al. A survey on visual surveillance of object motion and behaviors. IEEE Trans Syst Man Cybernet, 2004, 34: 334–352
24 Fagiani C, Betke M, Gips J. Evaluation of tracking methods for human-computer interaction. In: Proceedings of the 6th IEEE Workshop on Applications of Computer Vision. Washington: IEEE Computer Society, 2002. 121–126
25 Oviatt S L, Wu L, Vergo J, et al. Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. Human-Comput Interact, 2000, 15: 263–322

26 Tian F, Xu L S, Wang H A, et al. Tilt menu: using the 3D orientation information of pen devices to extend the selection capability of pen-based user interfaces. In: Proceedings of Conference on Human Factors in Computing Systems, Florence, 2008. 1371–1380

27 Tian F, Lu F, Jiang Y. An exploration of pen tail gestures for interactions. Int J Human-Comput Studies, 2013, 71: 551–569

28 Pelz J B. Portable eye-tracking in natural behavior. J Vis, 2004, 4: 14

29 Santella A, de Carlo D. Robust clustering of eye movement recordings for quantification of visual interest. Eye Tracking Res Appl, 2004, 23: 27–34

30 Cheng S, Sun Z, Sun L, et al. Gaze-based annotations for reading comprehension. In: Proceedings of ACM Conference on Human Factors in Computing Systems, Seoul, 2015. 1569–1572

31 Yu C, Sun K, Zhong M Y, et al. One-dimensional handwriting: inputting letters and words on smart glasses. In: Proceedings of CHI Conference on Human Factors in Computing Systems, San Jose, 2016. 71–82

32 Yu C, Wen H Y, Xiong W, et al. Investigating effects of post-selection feedback for acquiring ultra-small targets on touchscreen. In: Proceedings of CHI Conference on Human Factors in Computing Systems, San Jose, 2016. 4699–4710

33 Wang D, Zhao X, Shi Y. Six degree-of-freedom haptic simulation of probing dental caries within a narrow oral cavity. IEEE Trans Haptics, 2016, 9: 279

34 Yang W, Jiang Z, Huang X. Tactile perception of digital images. In: Proceedings of International AsiaHaptics Conference, Singapore, 2016. 445–447

35 Paivio A. Mental Representation: A Dual Coding Approach. New York: Oxford University Press, 1986

36 Baddeley A D. Working Memory. Oxford: Clarendon Press, 1986

37 Nelson C. What are the differences between long-term, short-term, and working memory. Prog Brain Res, 2008, 169: 323–338

38 Baddeley A. Working memory: looking back and looking forward. Nature Rev Neurosci, 2003, 4: 829–839

39 Service E. The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. Quarterly J Exp Psychol Sec A, 1998, 51: 283–304

40 Carpenter P A. A capacity theory of comprehension: individual differences in working memory. Psychol Rev, 1992, 99: 122–149

41 Nelson C. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci, 2001, 24: 87–185

42 Chooi W T, Thompson L A. Working memory training does not improve intelligence in healthy young adults. Intelligence, 2012, 40: 531–542

43 Barrouillet P, Bernardin S, Camos V. Time constraints and resource sharing in adults' working memory spans. J erimental Psychol, 2004, 133: 83–100

44 Yukio M, Satoru S. The relationship between processing and storage in working memory span: not two sides of the same coin. J Memory Lang, 2007, 56: 212–228

45 Hinton G E, Salakhutdinov R R. Salakhutdinov reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507

46 Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res, 2010, 11: 3371–3408

47 Chao L, Tao J, Yang M, et al. Bayesian fusion based temporal modeling for naturalistic audio affective expression classification. In: Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction, Geneva, 2013

48 Miao Y J, Gowayyed M, Metze F. Eesen: end-to-end speech recognition using deep rnn models and wfst-based decoding. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, 2015. 167–174

49 Caramiaux B, Montecchio N, Tanaka A. Adaptive gesture recognition with variation estimation for interactive systems. ACM Trans Interact Intell Syst, 2015, 4: 18

50 Mayer R E. Multimedia learning. Psychol Learn Motivation, 2002, 41: 85–139

51 Revlin R. Cognition: Theory and Practice. New York: Worth Publishers, 2012

52 Fournet N, Roulin J, Vallet F. Evaluating short-term and working memory in order adults: French normative data.

Aging Mental Health, 2012, 16: 922–930
53 Mayer R E. Multimedia Learning. 2nd ed. New York: Cambridge University Press, 2009
54 Ernst M O, Banks M S. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 2002, 415: 429–433
55 Gunes H, Piccardi M. Affect recognition from face and body: early fusion vs. late fusion. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, 2006. 3437–3443
56 Yang M H, Tao J H, Li H, et al. A nature multimodal human-computer-interaction dialog system. In: Proceedings of the 9th Joint Conference on Harmonious Human Machine Environment (CHCI2013), Nanchang, 2013 [杨明浩, 陶建华, 李昊, 等. 面向自然交互的多通道人机对话系统. 见: 第九届全国和谐人机环境联合学术会议 (CHCI 2013 年), 南昌, 2013]
57 Yang M H, Tao J H, Chao L. User behavior fusion in dialog management with multi-modal history cues. Multimedia Tools Appl, 2015, 74: 10025–10051
58 Li X, Gao F, Wang J. A priori knowledge accumulation and its application to linear BRDF model inversion. J Geophys Res-Atmo, 2001, 106: 11925–11935
59 Liu F, Lin X, Li S Z. Multi-modal face tracking using Bayesian network. In: Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures, Nice, 2003. 135
60 Town C. Multi-sensory and multi-modal fusion for sentient computing. Int J Comput Vis, 2007, 71: 235–253
61 Pradalier C, Colas F, Bessiere P. Expressing bayesian fusion as a product of distributions: applications in robotics. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, 2015. 1851–1856
62 Savran A, Cao H, Nenkova A. Temporal bayesian fusion for affect sensing: combining video, audio, and lexical modalities. IEEE Trans Cybern, 2015, 45: 1927
63 Li W, Lin G. An adaptive importance sampling algorithm for Bayesian inversion with multimodal distributions. J Comput Phys, 2015, 294: 173–190
64 Yu D, Deng L, He X, et al. Large-margin minimum classification error training for large-scale speech recognition tasks. In: Proceedings of IEEE International Conference on Acoustics, Honolulu, 2016
65 He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on imageNet classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 1026–1034
66 Yang F, Choi W, Lin Y. Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 2129–2137
67 Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning. In: Proceedings of International Conference on Machine Learning, Bellevue, 2011. 689–696
68 Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine learning, Helsinki, 2008. 160–170
69 Seltzer M L, Droppo J. Multi-task learning in deep neural networks for improved phoneme recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, 2013. 6965–6969
70 Tzeng E, Hoffman J, Darrell T. Simultaneous deep transfer across domains and tasks. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 4068–4076
71 Lukasz K, Gomez A N, Shazeer N, et al. One model to learn them all. 2017, arXiv:1706.05137
72 Wang C, Gorce M D L, Paragios N. Segmentation, ordering and multi-object tracking using graphical models. In: Proceedings of IEEE International Conference on Computer Vision, Kyoto, 2010. 747–754
73 Wei F, Li W, Lu Q. A document-sensitive graph model for multi-document summarization. Knowl Inform Syst, 2010, 22: 245–259
74 Myunghwan K, Jure L. Latent multi-group membership graph model. Comput Sci, 2012, 80
75 Honorio J, Samaras D. Multi-task learning of gaussian graphical models. In: Proceedings of International Conference on Machine Learning, Haifa, 2010. 447–454
76 Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. Science, 2015, 350: 1332–1338
77 Wu J X, Cheng J, Zhao C Y, et al. Fusing multi-modal features for gesture recognition. In: Proceedings of the 15th ACM on International Conference On Multimodal Interaction, Sydney, 2013. 453–460

78 Hamoud L, Kilgour D M, Hipel K W. Strength of preference in graph models for multiple-decision-maker conflicts. Appl Math Comput, 2006, 179: 314–332
79 Oviatt S L. Mutual disambiguation of recognition errors in a multimodal architecture. In: Proceedings of ACM Conference Human Factors in Computing Systems, Pittsburgh, 1999. 576–583
80 Wahlster W. SmartKom: symmetric multimodality in an adaptive and reusable dialogue shell. In: Proceedings of the Human Computer Interaction Status Conference, Berlin, 2003. 47–62
81 McGuire P, Fritsch J, Steil J J. Multi-modal human-machine communication for instructing robot grasping tasks. In: Proceedings of LEEE/RSJ International Conference on Intelligent Robots and Systems, Lausanne, 2002. 1082–1088
82 Michaelis J E, Mutlu B. Someone to read with design of and experiences with an in-home learning companion robot for reading. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, 2017. 301–312
83 Cheng A, Yang L, Andersen E. Teaching language and culture with a virtual reality game. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, 2017. 541–549
84 Sun M, Zhao Z, Ma X. Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, 2017. 556–567
85 Pham T A, Delalandre M, Barrat S. Accurate junction detection and characterization in line-drawing images. Pattern Recogn, 2013, 47: 282–295
86 Zhao B, Yang M, Pan H, et al. Nonrigid point matching of Chinese characters for robot writing. In: Proceedings of the IEEE International Conference on Robotics and Biomimetics, Macau, 2017. 762–767

Intelligence methods of multi-modal information fusion in human-computer interaction

Minghao YANG[1*] & Jianhua TAO[1,2,3*]

1. *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190;
2. *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing* 100049;
3. *Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai* 200031
* Corresponding author. E-mail: mhyang@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn

Abstract We first introduce the concepts of single-modal information processing and multi-modal information fusion in cognitive science. Some classical multi-modal information fusion models and their computer implementations in history are also explained. Under the conditions that each channel's information can be obtained, and their features could be unified representation synchronously, the fusion of multi-modal information can be transformed into classification or regression problems. For practical human-computer interaction systems, the performance of multi-modal information fusion largely relies on the accuracy of the single-modal information identification and the design of the interactive system. We present a practical example of multi-modal information fusion system, and discuss its performances on human computer interaction. Finally, the possible and important development trends for multi-modal human-computer interaction techniques and systems are discussed.

Keywords multi-modal information fusion, human-computer interaction, machine learning, pattern recognition, cognitive science

Minghao YANG was born in 1977. He received his Ph.D. degree in 2000 from the Institute of Automation, Chinese Academy of Sciences (IACAS). Currently, he is an associate professor at the National Laboratory of Pattern Recognition Institute of Automation in IACAS. His research interests include multi-modal human-computer interaction, human-machine interaction learning, pronunciation production and visualization, and augmented reality technology.

Jianhua TAO was born in 1972. He received his Ph.D. degree in 2001 from Tsinghua University. Currently, he is a professor and deputy director of NLPR, Institute of Automation, Chinese Academy of Sciences. His research interests include speech synthesis and recognition, speech coding, human-computer interaction, multimedia information processing, and pattern recognition.