ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.3, March 2007, pp. 477−489
DOI: 10.1360/jos180477

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

非特定人手语识别进展及关键问题研究思路*

姜 峰[1+], 高 文[1,2,3], 王春立[2], 姚鸿勋[1], 赵德斌[1]

[1](哈尔滨工业大学 计算机学院,黑龙江 哈尔滨 150001)

[2](中国科学院 计算技术研究所,北京 100080)

[3](北京大学 信息科学技术学院,北京 100871)

Development in Signer-Independent Sign Language Recognition and the Ideas of Solving Some Key Problems

JIANG Feng[1+], GAO Wen[1,2,3], WANG Chun-Li[2], YAO Hong-Xun[1], ZHAO De-Bin[1]

[1](School of Computer Science, Harbin Institute of Technology, Harbin 150001, China)

[2](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[3](School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

+ Corresponding author: Phn: +86-451-86416485, Fax: +86-451-86413309, E-mail: fjiang@hit.edu.cn, http://www.hit.edu.cn

Jiang F, Gao W, Wang CL, Yao HX, Zhao DB. Development in signer-independent sign language recognition and the ideas of solving some key problems. *Journal of Software*, 2007,18(3):477−489. http://www.jos.org.cn/1000-9825/18/477.htm

Abstract: Signer-Independent sign language recognition is an unavoidable problem that must be solved in order to promote the practicality of sign language systems. In the signer-independent sign language recognition research, the lack of training data and the signer-independent sign language data variation bring a challenge to the effectivity of the existent research frame. This paper proposes a new research frame for signer-independent sign language recognition, and provides the strategy and ideas to solve the problem. Finding resolution to these problems is significant not only to the research on Chinese sign language recognition but also to other related fields.

Key words: sign language recognition; signer-independent; data synthesis; effort analysis

摘 要: 非特定人手语识别是推动手语系统实用化所必须解决的问题.在非特定人手语识别研究中,训练数据的缺乏和非特定人手语数据的差异性矛盾给原有研究框架的有效性带来了挑战.提出了非特定人手语识别新的研究框架,并给出了解决问题的策略与思路.这些问题的解决将对中国手语识别及其他相关领域具有非常重要的意义.

关键词: 手语识别;非特定人;数据生成;力效分析

中图法分类号: TP391 文献标识码: A

手语是由手形、手臂运动并辅之以表情、唇动以及其他体势表达思想的人体语言,是聋人进行信息交流的最自然的方式和交际工具.手语识别研究的目的在于增进聋人与正常人之间无障碍的交流,提高计算机对人体

* Supported by the National Natural Science Foundation of China under Grant Nos.60332010, 60603023 (国家自然科学基金); the Program for New Century Excellent Talents in University under Grant No.NCET-05-03 34 (新世纪优秀人才支持计划)

Received 2006-06-25; Accepted 2006-11-13

语言的理解能力.手语识别作为多模式人机接口领域的一个重要组成部分,已经吸引了越来越多的专家和学者的注意.从用户角度来看,手势和手语识别的研究可能提供一个从 point-and-click 用户接口到自然的 dialogue-and-spoken 基于命令的用户接口的转换范例.从理论角度看,手语识别的研究不但是各个学科的前沿技术实际运用的结果,反过来,它的研究又推动了各分支领域的发展.手语交流是高度复杂的,具有时间与空间并发的特点,对手语识别要求同时对身体不同部分进行观测,并且精确地对同步信息进行综合.对于手语的理解与机器分析和理解人的运动,行为具有很大的共性.在手语识别中遇到的很多问题在其他研究中也是存在的.手语识别过程既是对计算机的计算能力、知识表示能力的全面考验,也是对人体生理认知机能和模式识别能力工作机理的探索,是一项非常具有挑战性的科学研究.

手语研究具有广泛的应用领域,至少表现在以下几个方面[1]:(1) 特定领域的对话系统;(2) 增加人机交流的方便性,提供功能增强的交流;(3) 从认知科学的角度提高计算机理解人类语言的水平;(4) 在语言学分析需要大量语体的情况下,自动、半自动的手语标注;(5) 为机器人提供示范学习的功能,同时也可以为手语学习者提供识别矫正的能力.手语识别研究不但具有深远的理论价值,而且具有广阔的实际应用前景.

对手语的分类包括手语级和成分级两种策略.在手语级策略中,手语识别使用一个单独的分类过程;在成分级策略中,把同时进行的、组成手语的各部分独立分类,然后整合到一起完成手语词的识别.区分成分级和手语级两种策略的一个优点是,在使用成分级的策略时,与使用手语级策略相比只需处理较少的类别,这样可以使基于成分级的分类器更加简单,所需训练的参数更少.如果在整合阶段使用基于词典的整合方法,在手语识别的过程中仅需要识别各组成成分,而并不是在整个手语词集的范围搜索.如果基于成分的分类已经考虑到了手语中所有可能出现的手语成分情况,那么对词汇集中新出现的手语词识别不需要进行新的训练.再者,使用成分级策略还可以用来处理手语中的语法现象.手语可以通过对其中一个或多个成分变化,形成曲折的手语词,这使得词汇量和所需训练数据量成倍地增加.事实上,手语词发生曲折变化仅发生在有限的成分上,因此如果在成分级上识别,则只需要相对较少的数据量.下面介绍的分类方法,我们同样可以用于在手语级中直接分类手语,或在成分级中对各组成部分进行分类.

人工神经网络是采用大量的简单处理单元广泛地连接起来构成的一种复杂的信息处理网络,因其非线性、自适应地、鲁棒性和学习特性等特点而受到人们的极大关注.尽管统计模型方法在手语识别中占有主导地位,但神经网络的独特优点及其较强的分类能力和输入-输出映射能力在手语识别中具有很强的吸引力.Waldron[2]提出使用自组织特征映射去识别小词汇量的美国手语,使用 Stokoe[3]的定义去分割手语的手形、方向和运动,其系统具有可扩展性.Kim[4,5]对韩国手语进行了研究,利用模糊最大最小神经网络技术进行手指字母及手势词的在线识别,识别 31 个手指字母,识别率为 96.3%;识别 131 个手势词,识别率为 94.3%.Su[6]利用超长方形组合神经网络进行手势识别.这种方法占用的内存少,计算量小,对 51 个手势识别表明这种方法是非常有效的.时序数据,如运动的轨迹,由很多点构成且有不同的单位时间距离长度,而用于对静态数据进行分类的神经网则常常无法利用到所有数据.在进行运动类型分类时,Waldron[2]使用在手语序列开始和中部的位置向量作为 MLP 的输入.Yang 和 Ahuja 通过图像处理的方法提取运动轨迹,使用 TDNN 进行时间上的处理,对运动轨迹类型进行分类[7];并利用 TDNN 在提取的轨迹上去学习运动模式[8],因为仅有一个小的时间窗的手势数据被作为输入,因此仅需要训练少量的权值.最终输入数据的窗口覆盖了所有的数据,但是仍然需要一个标准的时间长度.该方法在识别 40 个词汇的美国手语时,识别率在训练集和测试集分别为 99%和 96%.Murakami 和 Taguchi[9]使用 RNN 考虑时间上下文的因素,识别日本手语中的 10 个手势词,并且不需要固定的时间长度,当输出点的值保持一定时间不变时,则识别为一个手语词.这个时间长度采用启发式决策识别率 80%左右.此外,RNN 在语音识别[10]、手写体识别[11]方面也都取得了一定的效果.尽管时序的神经网络能够处理时序问题,但是它不能作为一个统一的框架用于大词汇量手语识别,即无论是 TDNN 还是 RNN,它们只能利用短距离的上下文,而不能处理长距离的依存关系.这是因为神经网络的结构决定了它缺乏模型化长距离的依存能力,文献[12]中对这个问题进行了理论分析.HMM 可以处理有可变时长的时间序列的数据,通过状态结点之间的跳转和同一状态的跳转来消除时长变化的影响.同时,HMM 可以隐含地对句子进行分割,词或基元的 HMM 被联成树结构,继而使用 Viterbi 找到最

优的路线,恢复词的边界和序列.这种方法被应用到连续手语的识别中,并使用各种不同的方法提高识别的效率,其中一些技术是源于语音识别技术的,包括语言模型,Viterbi-bean 搜索方法和剪枝[13−16],N-best pass[16],快速匹配等方法[15],帧预测[16].语言模型常用到的是 unigram 和 bigram[15−17],或采用严格限制句子成分语法规则的形式[18,19].为了可以扩大识别的词汇量并减少训练所需的数据量,一些研究者定义序列基元,类似于语音中的音素模型,使得每一个手语成为这些基元的联接.Bauer[13]定义了 12 个手语词中的 10 个基元,并使用 boostrap 算法[14]对基元 HMM 参数进行估计.对于由 150 个基元构成的 100 个手语词识别,识别率为 92.5%.对于由此构成的 50 个新词,尽管未进行新的训练,识别率也达到了 81%.Wang[16]和 Vogler[17]没有使用非监督聚类算法,而是从语言学的意义出发,分别定义了手语的基元.Wang 定义了 2 439 个手语基元,并用其识别词汇量为 5 119 的手语,识别率达到 86.2%.其他一些方法应用到手势手语分类的有:决策树[20,21]、最近邻匹配[22]、图像模板匹配[23,24]、相位相关滤波[25].规则的方法基于字典词条或者决策树也被用来对运动的轨迹或手语分类[5],或用来寻找运动类型的特征.由于规则往往基于经验,因此往往不会归纳得很好.Wu[26]提出了半连续动态高斯混合模型,替代 HMM 进行时间序列数的处理,具有较快的训练速度和更少的模型参数,模型用来识别 274 个手语词,但仅仅用到手部关节的数据.

非特定人手语识别可以减少单独一个用户训练所需的样本,是推动手语识别系统实用化所必须解决的问题.类似于语音中的非特定人识别,一个理想的手语识别系统应该可以在非注册集的情况下给出较好的识别结果.目前大部分的非特定人手语识别在训练集和测试集仅仅包括 2 人~10 人的数据.目前,非特定人手语识别中所用最大的训练集规模是 20 人数据[27−29].这远远少于训练一个友好的语音识别系统所需人数.当一个用于训练的数据中只含有较少人数的数据时,非注册集的识别结果会很差.Kadous[30]使用 4 个人训练,非注册集只达到了平均 80%~15%的准确率.Assan[31]使用一个人的数据训练,注册集和非注册集的结果分别为 94%和 51%.当用更多人的数据训练时,效果会有所改善.Vamplew[32]使用 7 个人的数据,得到注册集 94.2%,非注册集 85.3%的识别结果.当仅仅考虑手形数据时,非注册集和注册集的识别率会相对接近,这也许是因为与其他成分相比,不同人手形上的变化相对较小.例如,文献[33,34]报告了使用手形分类器达到了注册集 93%~96%,非注册集 85%~91%的准确率.有趣的是, Kong[35]对非特定人 3D 轨迹进行分类,报告了类似的好结果:注册集 99.7%,非注册集 91.2%的准确率.在我们前期设计的 SOFM/HMM 模型中,将数据转换成一个紧凑的、重要的低维表示形式,使用隐含的方法进行特征提取[36].这种提取是在缺少对手势手语运动理解的情况下进行的,因此不可避免地会造成有意义特征的丧失.在语音识别领域,人们采用说话人自适应算法克服特定人和非特定人系统各自的缺点,利用少量的训练语音调整系统的参数,使得系统性能有所提高.Ong[37]将 MAP 用于基于贝叶斯网络的成分分类器的估计,达到了 88.5%的识别率.MAP 算法仅对适应训练语音中出现的语音模型作更新,而对未出现过的语音模型则无法实现自适应.在大词汇量的识别系统,用户的自适应语音远远无法覆盖所有的语音模型,造成了自适应速度的缓慢.

1 非特定人手语识别面临的主要问题及研究架构

一个理想的手语识别系统应该能够处理大量的、一般既定的词汇;尽最大可能来满足使用者的移动需要;实时准确、在复杂环境下鲁棒地进行识别,并且这种识别应该面向真正的、非特定的操手语者.目前,非特定人大词汇量手语识别与特定人系统相比还有较大的差距,造成这一差距的主要原因在于数据本身的差异性矛盾与训练样本的缺乏.

1.1 目前的主要问题

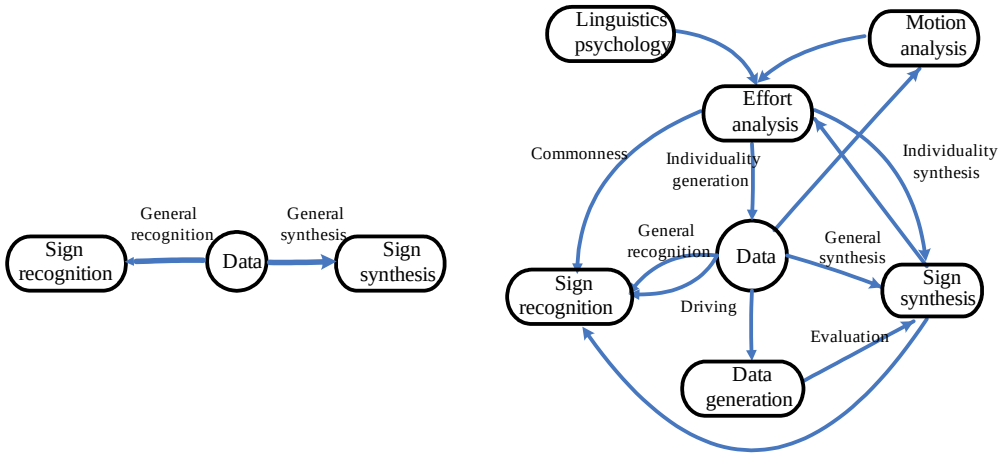
数据差异性矛盾使得非特定人手语识别中提取手语数据有效的共同特征非常困难.一方面,为了使训练的模型参数和操手语者无关,需要采集不同人的数据,但是由于数据中不同人相同词汇的差异甚至比同一个人的不同词汇的差异还大,导致模型训练难以收敛;另一方面,很少有系统使用真正聋人的数据,在使用职业手语者数据的情况下,采集数据规模小,差异性实际上又被忽略,数据的有效性难以得到保证.

对于识别系统来说,识别的效果不仅依赖于识别对象内在的特性以及分类器的设计,也受训练集规模和训

练数据质量的影响.传统统计学所研究的是一种渐进理论,由此提出的各种方法只有当样本数目趋向于无穷大时,其性能才有理论上的保证.在模式识别领域,统计模型需要大量的训练数据才能获得较为满意的识别性能.在实际应用中,模型的表达能力与样本的缺乏之间的矛盾已经成为制约识别系统效果的瓶颈.传统的方式采集训练数据存在以下缺陷:工作量大,这种专业化的数据采集方式增加了手语数据的采集难度;采集设备数量有限且价格昂贵,采集的方式冗长,对于采集的传感器数据,往往无法直观判断其正确性及有效性,这又会对训练效果产生直接的影响.训练数据的缺乏导致训练复杂模型困难,已使手语识别研究工作,特别是大词汇量的非特定人手语识别工作的开展异常艰难.

1.2 基于合成的手语识别架构

在传统的手语识别框架中,数据与统计模型之间为单向联接,从基于传感器或计算机视觉的运动捕获系统传出的数据被送入统计模型训练.大部分工作集中在力图通过统计模型的设计,提高识别的性能.很明显,这种思路很难解决在非特定人手语识别研究中出现的新问题.如上分析,非特定人手语识别研究中,问题的关键在于“数据”.为了有效解决“数据”问题,本文提出了以数据为中心的研究框架,如图 1 所示.



(a) Original framework
(a) 原有框架

(b) New framework
(b) 新架构

Fig.1 Mining interest patterns

图 1 手语识别架构

需要注意的是在手语合成领域已经开展的一些研究工作[38].在以往的工作中,手语识别与手语合成是独立存在的.在新框架中,手语识别与手语合成不再彼此独立,手语合成方法被有机地整合在系统中,二者通过数据及其分析相互联系,其最重要的一点就是可以直观地表现数据,从而验证数据的有效性.新框架有效辅助非特定人手语识别研究中关键问题的解决,又为手语合成研究,特别是个性化的手语合成提供了很好的研究思路.在新的框架中,利用不同的生成数据的方法,一方面可以从更深刻地认识手语数据本身具有的特性,一方面又可以利用合成的新的个性化手语数据,对统计模型进行有效的驱动.从数据出发,解决目前非特定手语识别中的问题,可以有如下的研究思路:(1) 针对训练数据的缺乏,利用有效的数据生成算法,生成新数据.利用手语合成模块验证数据的有效性,继而驱动非特定人手语识别模型,以期得到较好的识别效果;(2) 针对数据差异性矛盾,利用手语语言学、人体运动分析等领域的知识,研究手语数据本身携带的共性与个性的表现方法与提取方法.这具有以下两方面的作用:提出手语数据的归整方法;从人体运动角度出发,进行手语数据的生成,以生成自然度更高的手语数据.

2 手语数据合成策略及合成数据驱动方法

手语数据生成的主要目的就是在小训练样本集情况下,有效解决大词汇量非特定人手语识别问题.手语数据生成过程的目的是衍生出有广泛代表性与充分有效性的样本.在我们的前期工作中[39]提出了一种基于静态

手势量化与离散余弦变换(DCT)相结合的手语生成方法.该方法能够根据现有手语者的数据生成新的手语者的手语数据.但是,这种方法只能有条件地将打手语者的特征参数化,且训练过程初始化对生成数据有逆转的影响,因此,借助于参数化的手段解决更加复杂的样本生成问题存在其固有的局限性,必须研究新的方法以解决这一问题.

本文介绍了两种从已有的自然手语数据合成新数据的方法,以期通过将合成数据加入到自然数据中来扩大训练集规模,用于分类器的训练,即采用合成数据驱动的方法提高分类器的性能.合成数据不会被识别系统初始化过程逆转.我们已对该方法在多种实验环境下进行评估,在某些例子中,效果提高得很明显,说明利用合成数据驱动方法,可以有效地提高手语识别系统的性能.采用合成数据驱动方法,向原始训练集中加入合成数据,与分类器识别效率提升没有必然关系.可以预期,有两种相反的结果.其一,训练集规模的扩大有可能潜在地提高识别器的识别效果;其二,合成数据固有的不自然性又会导致识别效果的下降,特别是在测试集只采用自然数据的情况下.为使识别器的效果有显著的提高,关键在于设计有效的合成数据的方法.

2.1 基于演化计算的静止手势词数据的生成

由于手语数据采集很困难,每个特定人的数据不可能采集很多遍,利用这些数据训练出来的模型不可能具备很好的泛化性能,因此,我们可以在已经采集的数据基础上生成更多的数据.对于属于基本手势词集合的每个静止手势词,由于手语数据中的左手手形、右手手形、位置和方向来自不同的硬件设备,彼此之间不存在关联性,因此我们可以借鉴基因遗传算法中的交叉和变异思想,把这几部分的数据重新进行组合,生成新的数据.具体思路如图 2 所示.

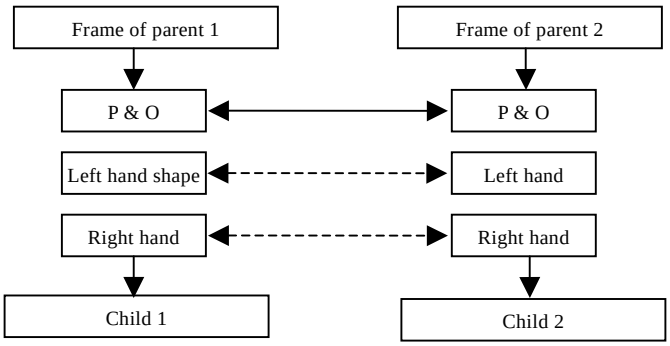


Fig.2 Crossover operator during the data generateon

图 2 样本生成中的交叉操作

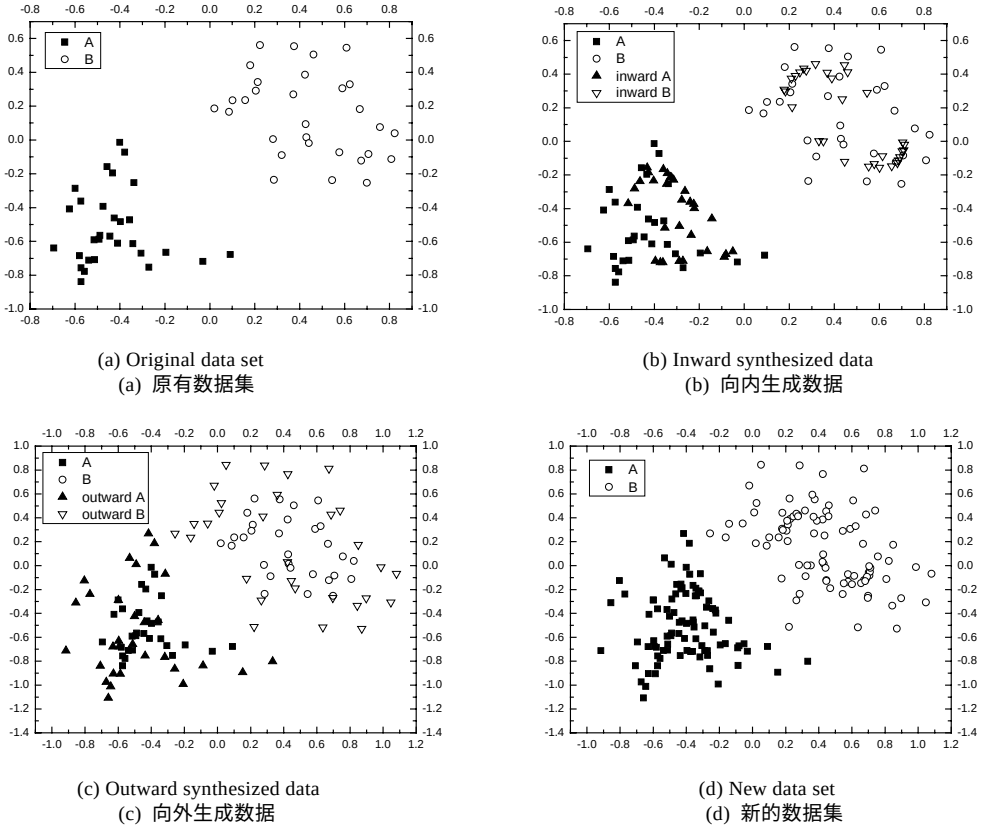
首先任意取出同一手势词的两个样本,将这两个样本进行规整,在时序上对齐,因为是静态手势,则直接进行线性变换即可.其次,对应的帧被分为左手手形、右手手形、位置和方向 3 部分,然后依次在每一部分进行交叉,每次只在一个部分交叉,每次交叉就得到两个新的训练数据.对于不在基本手势词集合中的静止手势,根据手势描述,从基本手势词集合中找到构成该手势词的基元的数据,每个基元可能不只一个数据,然后将这些基元数据进行组合,就可得到任意静止手势词的训练数据.

2.2 基于**mean-shift**样本生成

mean-shift 过程是一种基于模式识别中 Parzen 窗方法的核密度估计方法[40].仿照 Parzen 核密度估计的方法,我们可以根据样本估计概率密度的梯度.基本思路是,先用无参估计获取样本空间概率密度分布,然后对其进行微分以得到梯度信息.由于它是一种没有嵌入假设的非参数方法,因此能够适用于任意结构的特征空间分析.而 mean-shift 算法的非参性,很恰当地满足手语数据合成过程的自然性要求,相反,使用参数化的方法解决更为复杂的数据合成问题是行不通的.

如图 3 所示,基于 mean-shift 方法的样本生成可以从类内生成和类间生成两种方式去考虑.类内生成是利用类内已有的样本生成新的样本;类内生成又包括内向和外向两部分.内向部分使生成的样本朝向类内样本分

布密度增大的方向,外向部分朝向类内样本分布密度稀疏的方向;类间生成是利用类内样本与其易混类的样本生成新的样本.类内生成的样本有较高的有效性,但广泛的代表性较弱;类间生成的数据有广泛的代表性,但有效性难以保证.类间生成的情况比类内生成要复杂,一方面要使生成的样本涵盖非特定人的数据,另一方面还要规定扩张的边缘即界,以保证扩张在界内进行.



(a) Original data set
(a) 原有数据集

(b) Inward synthesized data
(b) 向内生成数据

(c) Outward synthesized data
(c) 向外生成数据

(d) New data set
(d) 新的数据集

Fig.3 Process of synthesized data generating

图 3 二维特征空间样本的生成过程

2.3 使用合成数据驱动的识别系统框架

采用数据生成技术的手语识别系统结构如图 4 所示.由输入设备采集进来的数据被送入特征提取模块,然后被归一化,数据库中包含少量样本,其中部分样本作为测试数据,其余的作为用于合成样本的初始样本,这些初始样本和合成样本用于训练 HMMs.由于训练样本数增多了,训练时间会增加,但在一般情况下,训练时间并不作为识别系统性能评价的重要因素.识别时间和识别率更为重要,本系统中采用 Viterbi 搜索算法进行译码,识别过程与一般的系统相同,计算复杂度和识别速度并不受影响.

3 手语数据力效分析及非特定人手语识别

手语作为一种结构化手势,是聋哑人进行信息交流的最常用形式.目前,对于手势手语的研究有两条线索.一条线索是基于语言学、心理学、神经学、人体动作分析等方面的工作.基本上,这些领域的研究不关心计算模型的建立,它们关注的是理论上手势手语及其功能的理解.尽管这些工作包括较深层次上的分析,但大部分的模型是理论的,这样很难说明它们的正确性和通用性.另外一条线索是基于机器学习、计算机视觉、计算机图形学等方面的工作,这些领域的研究关注的是对手势手语的操作,尤其是对手势手语的词义标注,大部分工作都

面向手势手语识别系统,这些工作虽然研究了问题的不同层面,但却没有回答一些基础性问题.

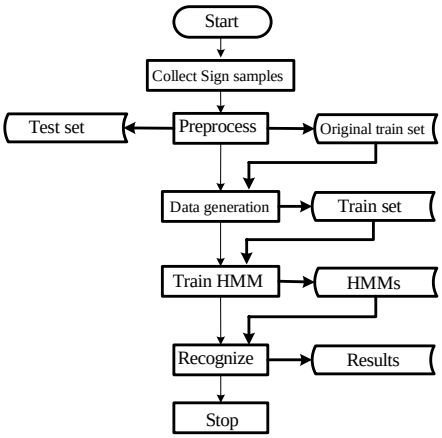


Fig.4 Overview of the recognition system

图 4 识别系统框架

3.1 力效理论-人体运动的内在本质

LMA[41,42]有效地描述、解释了人体运动,从而提高了对运动理解的效率及容易度.Laban 主张人体运动分作内容和形式两个方面,力效表示了运动质量的形式及执行情况,赋予动作以含义和表现性,并进一步综合成动作的力度变化和节奏现象.Laban 认为,每个动作都包含了运动的四大元素——空间、时间、重量和流动,每个元素对应一个闭连集.在 LMA 中,这些力效要素被认为是运动基础的、不可约简的属性,意味着它们是在描述和观测运动时的最小单元,为我们对手语运动理解提供了全面、紧凑的参数集合.

3.2 手语数据力效分析及非特定人手语识别

在数据采集过程中,从传感器中传出的数据携带着超过手语本身词义的更多信息.这些内容反映在手语词典的规定以外,但是与手语者个人情况紧密相关.因此,对手语信号可以在保证其结构性的前提下,消除其由各种因素造成的差异性,得到泛化的手语数据.将泛化的手语数据用于训练,对待识别的手语词,同样可以通过预处理消除其一定的个性,然后进行识别.手势手语的这些特点是语音所不具备的.与手语数据相比,语音缺少这种结构性,所以对语音的规范更加困难.对于手语中个性的消除,关键在于对手语信号的分析与理解.目前,对手语进行数据分析缺少现存的方法.手语在语言学上缺少类似语音中对音素的定义,这类似于在字母表诞生之前对语音进行研究.从数学上分析,发掘隐藏在模式之后的公式及其参数是复杂的,甚至是不可解决的.我们的手语力效分析开始于“原始”的测量,通过估计运动特性和力效要素之间的关系,对力效的每一维进行解释分析与处理,得到规范化的数据.在如图 5 所示的框架中,训练集中的多个手语者的数据构成力效分析统一的目标参考集,作为对待规整数据进行规整的参照,其力效分析反映了手语共性;而测试集和训练集中的不同手语者的数据被作为多个本体参考集,其力效分析反映了待规整数据的手语个性.本体参考集与目标参考集所对应的词汇要求相同.

3.2.1 空间维

在力效的要素中,“空间”体现了对周围环境的关注.对手语来说,空间的概念与保持阶段手的位置、手臂的伸展或收缩,以及手形与朝向相关.在日常交流的手语中,这些范畴具有不确定性.在手语数据中,有时手语者手的位置可能上下浮动,这种位置的变化并不影响手语词义的表达.虽然手形在手语定义的范围内具有相对稳定的特点,但由于手语者习惯或生理方面的原因,不确定性也很明显.同样,这种不确定性也是手语定义所允许的.一般说来,这些范畴的不确定性有一定的规律,即同一手语者在操手语时,会保持类似的变化趋势.可以把不同人的手语中手的位置、朝向、手形进行聚类,聚类中心作为预测值,实际观测值通过滤波处理,减少由于习惯等

因素对手语数据的影响.

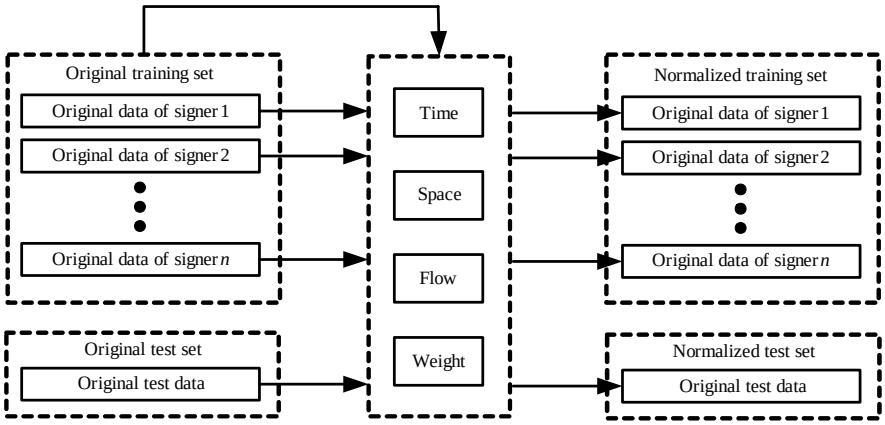


Fig.5 Analysis of effort and normalization of signer independence sign data

图 5 力效分析及非特定人手语数据规整

3.2.2 时间维

“时间”维体现手语者对时间的感知.聋人交流简单、直接,速度快慢取决于掌握程度和对方辨认程度,没有一定的限制.不同手语词经历的时间是不同的,因此很难根据当前手语的时间长度,使用参数化的方法对规范的手语长度进行预测.本文使用直方图对不同手语词的经历时间进行统计.图 6 分别为目标参考集的平均帧数与手语者 A 的本体参考集帧数直方图.定义函数:

$$P(k)=\frac{n_k}{N},k=0,1,\ldots,L \tag{1}$$

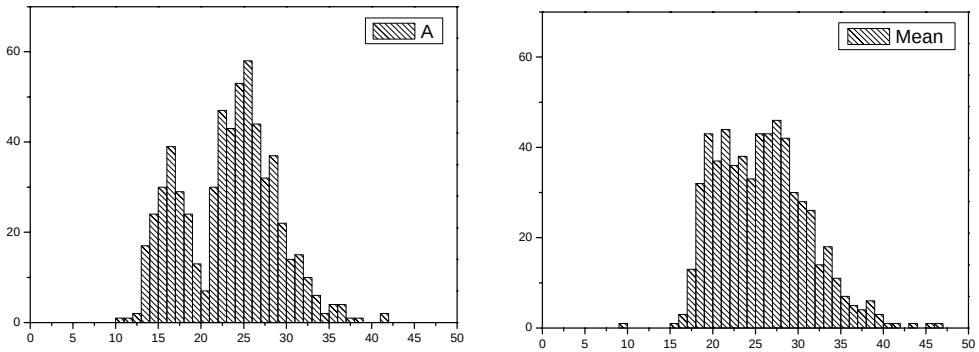
L 为孤立手语词序列时间长度可能出现的最大值;N 为目标参考集规模.这里的 $P(r_k)$给出了目标参考集中手语词序列平均帧长为 r_k 的概率估计值.$P(z_k)$对应本体参考集手语词序列时间长度为 z_k 概率密度.令分布函数

$$\Gamma_k=T(r_k)=\sum_{j=0}^{k}P(r_j)=\sum_{j=0}^{k}\frac{n_j}{N} \tag{2}$$

$$\Psi_k=G(z_k)=\sum_{i=0}^{k}P(z_i)=S_k \tag{3}$$

则

$$z_k=G^{-1}\left[T(r_k)\right] \tag{4}$$



(a) “Time” histogram of subject reference (b) “Time” histogram of object reference set

(a) 本体参考集“时间”直方图 (b) 目标参考集“时间”直方图

Fig.6 Effort element “time”

图 6 力效要素“时间”

3.2.3 流动维

力效要素中“流动”源于身体控制的紧张程度,表现为运动序列的光滑性和连贯性.根据手语信号的特殊性,对流动的分析可以从过渡段经历的时间与保持段的比例分析着手.定义流动度量

$$\Phi = \left\lfloor Con \times \frac{T_{Tran}}{T_{Tran} + T_{Hodl}} \right\rfloor \tag{5}$$

常数 Con 反映统计及操作的精度.对于每个手语词,我们统计目标参考集中同一样本的平均流动度量作为流动量的共性体现.定义函数 $Q(k) = \frac{m_k}{N}$,m_k 为目标参考集中平均流动度量 Φ=m_k 的个数,N 为目标参考集规模.令 $Q(r_k)$为目标参考集中流动度量为 r_k 发生的概率估计值.目标参考集中流动度量分布函数表示为

$$\gamma_k = G(z_k) = \sum_{i=1}^{k} Q(z_i) \tag{6}$$

同样,令 $Q(z_k)$对应本体参考集中流动度量为 z_k 概率密度.本体参考集中流动度量分布函数表示为

$$B_k = T(r_k) = \sum_{i=0}^{k} Q(r_i) \tag{7}$$

其中,$k \in [0 \ldots Con]$,同理可得:

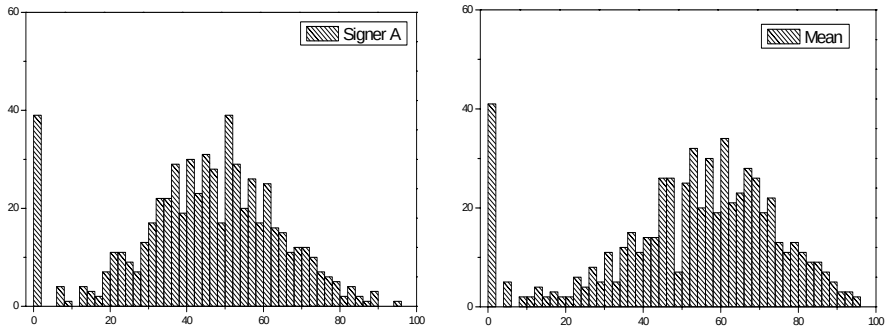
$$z_k = G^{-1}[T(r_k)] \tag{8}$$

图 7 中分别显示了目标参考集的平均流动度量与手语者 A 本体参考集的流动度量直方图.规整后保持阶段和过渡阶段经历的时间调整为:

$$T'_{Tran} = \frac{z_k \times (T_{Tran} + T_{Hold})}{Con} \tag{9}$$

$$T'_{Hold} = \frac{(Con - z_k) \times (T_{Tran} + T_{Hold})}{Con} \tag{10}$$

经过对“时间”与“流动”的处理,可以得到规范情况下整个手语词及保持阶段和过渡阶段经历的时间,规整的数据可由插值得到.



(a) “Fluidity” histogram of subject reference (b) “Fluidity” histogram of object reference set

(a) 本体参考集“流动”直方图 (b) 目标参考集“流动”直方图

Fig.7 Effort element “Fluidity”

图 7 力效要素“流动”

3.2.4 重量维

“重量”维表现为手语序列中的运动总的压力和动态性.在手语中包含多个过渡段或者一个复杂的过渡段包含一系列分段成分的情况下,我们希望这些分段保持规律的重量使用情况.这里,每一个运动的片断经历的时间满足 Fitt’s 准则的规定[43],即

$$T = b \cdot \log_2 \left(\| \vec{x} - \vec{x}_s \| + 1 \right) \tag{11}$$

对于一个包含 N 段基本运动片段的手语词,每一个基本片断的经历时间可以近似估计为

$$T = b\ln\left(\sum_t QOM\right) \tag{12}$$

其中 QOM 为运动性度量,定义为运动捕获系统在相邻时间间隔内检测到的手语各数据流相对运动总量.为了使重量状态均衡,对每一基本运动段占手语词过渡段比例作以下规整:

$$T_i'' = \frac{\log\left(\sum_t QOM + 1\right)}{\sum_N \log\left(\sum_t QOM + 1\right)} \sum T'_{Tran} \tag{13}$$

4 实 验

在没有约束数据质量的前提下,评价实验结果是困难的.影响手语数据质量的因素包括词汇集的选取和手语者的选择.在词汇集选取方面,不仅存在词汇集规模、词汇上的限制,减少最小词汇对的数量也会对识别的结果造成影响.在数据的选取方面,一些系统所使用的训练数据来自操手语很规范的专业手语教师,这样一个负面影响使得非特定人不同的操手语习惯就更加难以捕捉.本文的实验采集数据的对象是聋校随机抽取的学生,有各自不同的节奏感和操手语的习惯方式,他们的数据更具代表性.

4.1 小词汇集力效分析下的数据规整效果

实验 1 的数据由 6 130 个样本组成,它是以聋校手语小学教材中的 613 个词汇作为对象,由 5 位不同的手语学生将每个手语词分别采集 2 遍.在进行力效分析时,参考集包含了所有的 613 个词汇数据.统计模型分别采用 HMM 与 SOFM/HMM 模型.HMM 采用连续有跳转的拓扑结构,3 个状态节点、5 个混合项.SOFM/HMM 采用 3 个状态节点,每个状态下的 SOFM 初始神经元节点取 5 个.实验结果见表 1.

Table 1 The effect of effort analysis and normalization in the case of small vocabulary

表 **1** 小词汇集数据力效分析及归整实验结果

Signer	HMM			SOFM/HMM		
	Reg. (%)	Effort (%)	Unreg. (%)	Reg. (%)	Effort (%)	Unreg. (%)
A	85.1	84.4	76.2	89.1	88.4	80.2
B	83.3	81.3	75.3	87.4	85.6	78.6
C	81.6	82.5	74.2	86.3	85.8	79.3
D	85.1	84.7	79.6	89.8	88.8	83.1
E	84.2	85.1	78.3	89.3	90.1	80.2
Average	83.9	83.6	76.7	88.4	87.7	80.2

其中,Reg 是测试集作为已注册测试集的结果,即训练集使用测试者 1 遍和其他 4 个人的 8 遍数据;Effort 是使用力效分析对训练集规整后用于训练,并使用规整后的测试集进行识别的结果;UnReg 是测试集作为未注册测试集的结果,即训练集使用测试者之外 4 个人的 8 遍数据.在两种统计模型下,使用力效分析对手语数据规整用于识别训练的效果均高于非注册集测试的概况,识别率获得了明显的提高,可以说明方法是有效的.与注册集相比,10 种情况下有两种方法比注册集的效果要好,整体识别效果逼近注册集的实验效果.可见,小训练集情况下,所有词汇参与了力效分析取得了良好的效果,有效地剥离了手语数据个性.

4.2 大词汇集合成数据驱动方法效果

实验 2 的目的在于检测在大词汇集、较多打手语者情况下合成数据驱动方法对识别效果的影响.实验数据包含 61 356 个样本,分别由 6 位手语者将中国手语词典中 5 113 个手语词每个采集 2 遍组成.实验采用交叉验证的方式,使用 5 个人的数据进行训练,剩下一个人的自然数据作为未注册测试集.

实验考虑到影响实验结果的两个实验环境配置参数:模型的容量和合成强度.模型的容量表现为模型中参数的个数,它反映了模型可以表达的信息量.模型的容量过大,会导致模型训练不充分;反之,容量较小的模型有可能因为相对简单而无法对数据进行有效的描述.对于合成数据驱动方法来说,由于合成数据往往带有更强的不自然性,一般来说,它的最优模型容量要高于只使用原始数据集训练的最优模型容量.这是因为,如果模型的

容量在训练集扩张之后没有提高,有模型容量过小的危险,这样,模型就会偏向于非自然数据,即对合成的不自然的数据的描述,从而导致识别效果的下降.在本实验中,模型的容量是通过调整 HMM 发射概率中高斯混合中心的个数来实现的.增加混合中心个数,可以使特征提取更加精细.另外,还要考虑合成强度的影响,在实验中分别考虑了弱外、强内、弱内 3 种情况.

在使用原始训练集时,HMM 采用连续有跳转的拓扑结构,3 个状态、5 个混合项,识别效果达到最好.经过对状态节点和混合项数目的反复实验,在使用添加了合成数据的训练集时,适当增加观测概率混合项的数目,会提高系统的识别效果,这里,HMM 采用连续有跳转的拓扑结构,3 个状态 8 个混合项.在未注册测试集,平均识别结果为 78.40%;在使用合成数据驱动的情况下,测试结果获得了明显的提高,较强向内合成数据的情况下,平均识别结果为 80.66%,在同一个识别框架下,通过有效地对原始训练数据进行扩展,系统的识别率提高了 2.26%.

5 结束语

在非特定人手语识别中,训练的样本缺乏和数据的差异性矛盾给样本的有效性及识别系统性能带来了挑战.鉴于原有的识别架构已经无法有效解决这些问题,本文提出了一种新的基于合成的手语识别架构,用于指导对于非特定人手语识别研究中上述问题的解决.针对训练样本缺乏这一问题,提出了两种合成手语数据驱动方法.本文提出的手语数据生成方法还可以应用于其他方面,例如,用来检测识别系统可能存在的缺陷.针对手语数据的差异性矛盾这一问题,在保证手语词典规定的共性前提下去除手语数据中的个性,是解决这种差异性矛盾的有效途径.本文通过引入 LMA 中的“力效”分析,对差异因素进行了归纳,总结了关于手势手语的假设和理论,提出了“力效”各要素的计算模型及手语数据的规整方法,规整后的数据用于训练与识别.本文提出的手语力效要素分析方法还有其他方面的应用:对手语合成中感情色彩的合成有很好的借鉴作用;反映情感变化的动态结构特征,对情感识别起重要作用;此外,对建立标准手语数据库也有一定的指导意义.

对手语识别研究本身,是需要建立在更为深刻的认识之上的,这需要机器学习研究者和人体运动、语言学、心理学研究者的进一步合作.

References:

[1] Gao W, Chen XL, Ma JY, Wang ZQ. Building language communication between deaf people and hearing society through multimodal human computer interface. Chinese Journal of Computers, 2000,23(12):1253−1260 (in Chinese with English abstract).

[2] Waldron MB, Kim S. Isolated ASL sign recognition system for deaf persons. IEEE Trans. on Rehabilitation Engineering, 1995, 3(3):261−271.

[3] Stokoe WC. Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. University of Buffalo: Linstok Press, 1960.

[4] Lee CS, Park GT, Kim JS, Bien Z, Jang W. Real time recognition system of Korean sign language based on elementary components. In: Proc. of the IEEE Int’l Conf. on Fuzzy Systems. IEEE Press, 1997. 1463−1468. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=619759

[5] Kim JS, Jang W, Bien Z. Dynamic gesture recognition system for the Korean sign language (KSL). IEEE Trans. on Systems, Man, and Cybernetics, 1996,26(2):354−359.

[6] Su MC, Huang H, Lin CH, Huang CL, Lin CD. Application of neural networks in spatio-temporal hand gesture recognition. In: Simpson PK ed. Proc. of the IEEE World Congress on Computational Intelligence. New York: IEEE Press, 1998. 2116−2121.

[7] Yang MH, Ahuja N. Extraction and classification of visual motion patterns for hand gesture recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Santa Barbara: IEEE Press, 1998. 892−897. http://citeseer.ist.psu.edu/yang98extraction.html

[8] Yang MH, Ahuja N. Recognizing hand gestures using motion trajectories. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR’99). Fort Collins: IEEE Press, 1999. 466−472. http://vision.ai.uiuc.edu/mhyang/papers/cvpr99.pdf

[9] Murakami K, Taguchi H. Gesture recognition using recurrent neural networks. In: Proc. of the CHI’91 Human Factors in Computing Systems. New York: ACM Press, 1991. 237−242. http://portal.acm.org/ft_gateway.cfm?id=108900&type=pdf

[10] Robinson T. An application of recurrent nets to phone probability estimation. IEEE Trans. on Neural Networks, 1994,5(2): 298−305.

[11] Senior A, Robinson AJ. Forward-Backward retraining of recurrent neural networks. Advances in Neural Information Processing Systems, 1996,8:743−749.

[12] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. on Neural Networks, 1994,5(2):157−166.

[13] Bauer B, Kraiss KF. Towards an automatic sign language recognition system using subunits. In: Proc. of the Gesture Workshop. London: ACM Press, 2001. 64−75. http://www.springerlink.com/index/07DPUY0RPPMNE9M0.pdf

[14] Bauer B, Kraiss KF. Video-Based sign recognition using self-organizing subunits. In: Kasturi R., Laurendeau D., Suen C. eds. Proc. of the Int'l Conf. Pattern Recognition. Quebec: IEEE Computer Society, 2002. 434−437.

[15] Gao W, Ma J, Wu J, Wang C. Sign language recognition based on HMM/ANN/DP. Int'l J. Pattern Recognition Artificial Intelligence, 2000,14(5):587−602.

[16] Wang C, Gao W, Shan S. An approach based on phonemes to large vocabulary Chinese sign language recognition. In: Proc. of the Int'l Conf. Automatic Face and Gesture Recognition. New York: ACM Press, 2002. 393−398.

[17] Vogler C. American sign language recognition: Reducing the complexity of the task with phoneme-based modeling and parallel hidden Markov models [Ph.D. Thesis]. University of Pennsylvania, 2003.

[18] McGuire RM, Hernandez-Rebollar J, Starner T, Henderson V, Brashear H, Ross DS. Towards a one-way American sign language translator. In: Proc. of the Int'l Conf. Automatic Face and Gesture Recognition. Seoul: ACM Press, 2004. 620−625. http://ieeexplore.ieee.org/iel5/9123/28919/01301602.pdf

[19] Starner T, Weaver J, Pentland A. Real-Time American sign language recognition using desk and wearable computer based video. IEEE Trans. on Pattern Analysis Machine Intelligence, 1998,20(12):1371−1375.

[20] Hernandez-Rebollar JL, Lindeman RW, Kyriakopoulos N. A multi-class pattern recognition system for practical finger spelling translation. In: Proc. of the Int'l Conf. Multimodal Interfaces. 2002. 185−190. http:ieeexplore.ieee.org/iel5/8346/26309/01166990.pdf

[21] Hernandez-Rebollar JL, Kyriakopoulos N, Lindeman RW. A new instrumented approach for translating American sign language into sound and text. In: Proc. of the Int'l Conf. Automatic Face and Gesture Recognition. Seoul: ACM Press, 2004. 547−552. http://ieeexplore.ieee.org/iel5/9123/28919/01301590.pdf

[22] Kramer J, Leifer L. The talking glove: An expressive and receptive verbal communication aid for the deaf, deaf-blind, and nonvocal. In: Brown. C. ed. Proc. of the 3rd Annual Conf. on Computer Technology, Special Education, Rehabilitation. Northridge：California State University Press,1987. 335−340.

[23] Gupta L, Ma S. Gesture-Based interaction and communication: Automated classification of hand gesture contours. IEEE Trans. on Systems, Man, and Cybernetics. Part C: Applications and Reviews, 2001,31(1):114−120.

[24] Sutherland A. Real-Time video-based recognition of sign language gestures using guided template matching. In: Harling PA, ed. Proc. of the Gesture Workshop. London: Springer-Verlag, 1996. 31−38.

[25] Terrillon JC, Piplr A, Niwa Y, Yamamoto K. Robust face detection and Japanese sign language hand posture recognition for human-computer interaction in an "Intelligent" Room. In: Proc. of the Int'l Conf. Vision Interface. Banff: ACM Press, 2002. 369−376. http://www.cipprs.org/vi2002/pdf/s7-4.pdf

[26] Wu J, Gao W. A fast sign word recognition method for Chinese sign language. In: Proc. of the Int'l Conf. Advances in Multimodal Interfaces. San Diego: Academic Press, 2000. 599−606. http://www.springerlink.com/index/8H84T698MF0AU23V.pdf

[27] Chen FS, Fu CM, Huang CL. Hand gesture recognition using a real-time tracking method and hidden Markov models. Image and Vision Computing, 2003,21(8):745−758.

[28] Huang CL, Jeng SH. A model-based hand gesture recognition system. Machine Vision and Application, 2001,12(5):243−258.

[29] Kobayashi T, Haruyama S. Partly-Hidden Markov model and its application to gesture recognition. In: Taylor FJ, ed. Proc. of the Int'l Conf. on Acoustics, Speech and Signal Processing. New York: Academic Press, 1997. 3081−3084.

[30] Kadous MW. Machine recognition of auslan signs using PowerGloves: Towards large-lexicon recognition of sign language. In: Messing L ed. Proc. of the Workshop Integration of Gestures in Language and Speech. Delaware: IEEE Computer Society Press, 1996. 165−174.

[31] Assan M, Grobel K. Video-Based sign language recognition using hidden Markov models. In: Proc. of the Gesture Workshop. Bielefeld: Springer-Verlag, 1997. 97−109. http://www.springerlink.com/index/T34LRHBYXFQAL4CD.pdf

[32] Vamplew P, Adams A. Recognition of sign language gestures using neural networks. Australian Journal of Intelligent Information Processing Systems, 1998,5(2):94−102.

[33] Handouyahia M, Ziou D, Wang S. Sign Language recognition using moment-based size functions. In: Proc. of the Int'l Conf. on vision interface. Kerkyra: CRC Press, 1999. 210−216. http://www.gel.ulaval.ca/~parizeau/vi99/PDF-files/S7/paper76.pdf

[34] Su MC. A fuzzy rule-based approach to spatio-temporal hand gesture recognition. IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2000,30(2):276−281.

[35] Kong WW, Ranganath S. 3-D Hand trajectory recognition for signing exact english. In: Proc. of the Int'l Conf. Automatic Face and Gesture Recognition. Seoul: ACM Press, 2004. 535−540. http://ieeexplore.ieee.org/iel5/9123/28919/01301588.pdf

[36] Gao W, Fang GL, Zhao DB, Chen YQ. A Chinese sign language recognition system based on SOFM/SRN/HMM. Pattern Recognition, 2004,37(12):2389−2402.

[37] Ong S, Ranganath S. Deciphering gestures with layered meanings and signer adaptation. In: Proc. of the Int'l Conf. Automatic Face and Gesture Recognition. Seoul: ACM Press, 2004. 559−564. http://ieeexplore.ieee.org/iel5/9123/28919/01301592.pdf

[38] Wang ZQ, Gao W. A method to synthesize Chinese sign language based on virtual human technologies. Journal of Software, 2002,13(10):2051−2056 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/2051.pdf

[39] Zhang CX, Jiang F, Yao HX. Static gesture quantization and DCT based sign language generation. In: Tao J, Tan T, Picard RW, eds. Proc. of the 1st Int'l Conf. on Affective Computing and Intelligent Interaction (ACII 2005). Beijing: Springer-Verlag, 2005. 168−178.

[40] Comaniciu D, Ramesh V, Meer P. Kernel-Based object tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003, 25(5):564−575.

[41] Laban R. The Mastery of Movement. Boston: Plays, Inc., 1971.

[42] Laban R, Lawrence FC. Effort: Economy in Body Movement. Boston: Plays, Inc., 1974.

[43] Kopp S, Wachsmuth I. Synthesizing multimodal utterances for conversational Agents. The Journal of Computer Animation and Virtual Worlds, 2004,15(1):39−52.

附中文参考文献:

[1] 高文,陈熙霖,马继勇,王兆其.基于多模态接口技术-聋人与正常人交流系统.计算机学报,2000,23(12):1253−1260.

[38] 王兆其,高文.基于虚拟人合成技术的中国手语合成方法.软件学报,2002,13(10):2051−2056. http://www.jos.org.cn/1000-9825/13/2051.pdf

姜峰(1978－),男,黑龙江呼兰人,博士生,讲师,主要研究领域为模式识别,机器学习,图像处理,神经网络.

高文(1956－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为多媒体数据压缩,图像处理,计算机视觉,多模式接口,人工智能.

王春立(1972－),女,博士,副教授,主要研究领域为模式识别,人机交互.

姚鸿勋(1965－),女,博士,教授,博士生导师,主要研究领域为视觉语言,多媒体技术,数字水印.

赵德斌(1963－),男,博士,教授,博士生导师,主要研究领域为多媒体数据压缩,图像处理,计算机视觉,多模式接口,人工智能.