视觉语言 ——唇读综述

姚鸿勋[1],高 文[1,2],王 瑞[1],郎咸波[3]
(1. 哈尔滨工业大学计算机科学与工程系,哈尔滨 150001;2. 中国科学院计算技术研究所,北京 100080;3. 黑龙江省外国专家局,哈尔滨 150001)

摘 要: 本文介绍了目前唇读研究的现状与发展水平,详细阐述了唇读研究的内容和方法,以及唇读研究的意义,旨在引起大家对此新兴研究方向的关注与兴趣,从而积极参与对唇读问题的研究,并推动与此相关问题的进展.

关键词: 唇读; 唇动; 自动语音识别; 手语识别; 情感计算
中图分类号: TN391 **文献标识码**: A **文章编号**: 0372-2112 (2001) 02-0239-08

A Survey of Lipreading ——One of Visual Languages

YAO Hong-xun[1], GAO Wen[1,2], WANG Rui[1], LANG Xian-bo[3]
(1. *Dept. Computer Science and Engineering, Harbin Institute of Technology, Harbin* 150001, *China*; 2. *Chinese Academy of Sciences, Beijing* 100080, *China*; 3. *Foreign Experts Bureau of Heilongjiang Province, Harbin* 150001, *China*)

Abstract: The actuality and the developing level of lipreading at the present time is introduced in this paper, and the contents, approaches and the significance on the research of lipreading are paid particular attention to. This paper aims to arouse people's attention and interests into this new field, to participate in the study of the lipreading problems actively, and to achieve more successes correlated to this problem.

Key words: lipreading; lip movement; automatic speech recognition; gesture recognition; expression computing

1 引言

所谓唇读(lip-reading/ speech-reading),是指通过观察说话者的口型变化,"读出"或"部分读出"其所说的内容. 研究唇读目的是利用视觉信道信息补充听觉信道的信息,以提高计算机系统的理解力. 唇读技术源于听力弱者或者听力障碍者学习、了解正常人的表达的一种技巧,它亦可用于特定场合的信息获取(如情报等). 人类的语言认知过程本身就是一个多通道的感知过程. 生活经验告诉我们,在人与人交流的过程中,人们在理解他人讲话内容时不仅通过声音来感受信息,而且还用眼睛观察对方口型、表情等的变化,以期更准确地理解对方所讲的内容[1]. 听力弱者或听力障碍者会从视觉中得到更多的暗示,有的甚至已经达到了完美听觉的程度[2]. 即便是正常人也利用视觉信息来提高语音感知,尤其在噪音环境下[3,4]. 一些音素在语音上难于区别而在视觉上却易于分辨;反之依然. 因此,视觉信号通常对语音噪声敏感的音素提供更多可区分的信息. 而唇动信息即使在没有噪音的情况下,也是有用的,它能提高对说话人语义的正确理解.

近年来,语音识别技术得到迅速发展,自动语音识别系统有了长足的进步,已有许多听写机之类的实用产品,较好的有 IBM 开发的 Viavoice 语音系统,但尚未有较强的抗干扰能力,原因是它只单纯从语音信道获取信息. 一旦这些系统用到真实环境中,有背景噪声或交叉的说话者,它们的性能大大下降[5]. 而这样的应用环境却很常见,如:办公室、汽车、工厂或机场等等. 唇动(lip movement,一个与唇读密切相关的概念,或者说是唇读的另一种提法,但它更强调口型的可视变化过程,意在跟踪和识别,而唇读侧重于理解)作为视觉信息通道的信息源,用作语音的理解源,最早是由 Sumby 在 54 年提出来的[3]. 唇读是对唇动的解释,更一般的意义可解释为人脸面部情感的表达,有着人的情感状况和所要表达的意义[6]. 以往的语音识别系统忽略了语言感知的视觉特性,仅仅利用了听觉特性,使得现有的语音识别系统在噪声环境或多话者条件下,其识别率都大大下降,限制了它的应用领域. 近年来,唇读作为语音识别的辅助手段引起了越来越多的研究人员的关注,初步的研究结果表明,将唇读与语音进行融合能有效地改善识别率,特别在噪声环境下,效果更为明显[7,8].

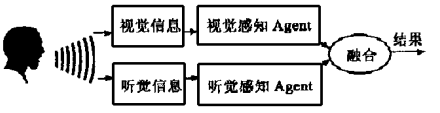

图 1 视觉信息与听觉信息的协同

2 研究内容综观

唇读的研究内容包括基本口型、口型的视觉特征,口型特征的提取、描述与表示方法,以及口型识别与理解,包括对来

收稿日期:2000-05-30;修回日期:2000-07-30
基金项目:国家 863 计划项目(863-306-QN99-4);国家 863 计划项目(863-306-ZT03-01-2);国家自然科学基金重点项目(69789301)

自多个通道的相关信息的综合理解. 研究视觉信道与听觉信道的互补性关系,研究视觉信道中人说话时口型运动规律,分离出属于人个性的运动特点和发音时共同的运动规律,将个性特点用于生物个性特征的识别与合成,而将发音时的共同运动规律用模型和参数来描述,用于辅助语音识别和运用到基于 *MPEG-4/7* 的视频编码与解码中. 通过对口型运动规律及其识别技术的研究,作为语音识别、手语识别、身份特征识别等的其他通道的辅助识别手段和必要的补充形式;同时指导解决 *MPEG-4/7* 编码、电话会议、动画唇动合成、网络自动代理中的虚拟人唇动和配音等方面的内容表示、描述等关键技术问题. 因此,唇读研究的内容大约可分为两个层面:一方面着重于口型变化序列的识别与理解,另一方面强调对口型进行编码和描述,以配合可视化输出.

图 2 是以唇读为中心的面部感知系统概貌.

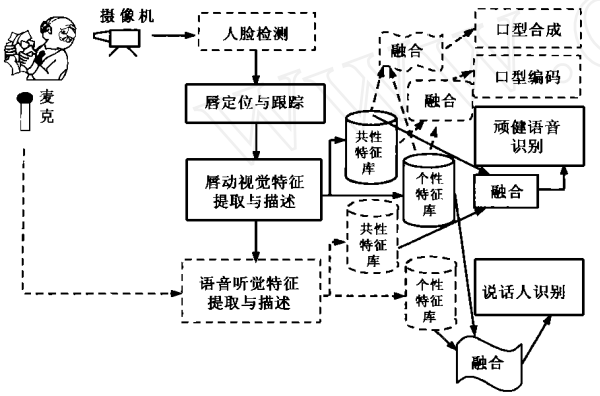

图 2 唇读与面部感知概貌

目前,国内外对唇读的研究只限于与语音识别相结合的视觉特征提取和识别,以及在此过程中可同时考虑到的关于说话者个性的特征提取和识别工作,而对其它方面的研究尚未展开. 唇读的思想起源很早,但唇读的第一个系统于 1984 年才由 *Petajan* 在文献[9]中详细给出. 80 年代后,唇读的研究才在世界范围内兴起,伴随着语音识别技术的不断成熟而逐渐显露出它的重要性的. 它是新一代人机交互技术的重要组成部分. 它代表着一个年轻的研究方向,蕴涵着许多人类亟待解决的问题,亟待有更多的有志者前来加盟. 而对于先驱者们已经做过的工作及使用的方法值得后来者学习和鉴借.

3 唇读识别研究方法比较

近十几年来,开展唇读研究的主要有美国加利福尼亚大学,加利福尼亚 Ricoh 研究中心,卡耐基梅隆大学(CMU),麻省理工大学(MIT)人机交互研究所,斯坦福大学,伊利诺伊斯大学,福罗里达大学,明尼苏达州立大学,德克萨斯州立大学,加州圣地亚哥大学认知学系,华盛顿大学电子工程系信息处理实验室,乔治亚工学院电子工程院 AT&T 实验室,德国斯图加特大学,卡尔斯鲁厄大学,瑞士 IDIAP 人工智能感知研究所、英国舍菲尔德大学电子电气工程系、东安格列大学、日本大板大学、雄本大学、NTT(日本电报电话公司)人机交互研究所、意大利科学技术研究所、国立研究会声学研究所,法国国际高级通信中心、EURECOM 欧洲通讯研究所多媒体通信部、芬兰的赫尔辛基大学,国内的哈尔滨工业大学、中国科学院声学研究所、杭州大学、大连理工大学、北京理工大学、北京大学视觉与听觉信息实验室等已经有人做了大量工作. 但总的来说,对唇读的研究目前还处于研究和探索的初级阶段. 加利福尼亚大学的唇读网页 http://mambo.ucsc.edu/psl/lipr.html,提供了世界范围内的唇读研究的情况和最新进展.

通常,唇读问题的解决需要经过唇的检测定位、特征提取、识别和融合理解的过程,将按着这样的顺序,分别来比较现有系统所采用的方法及各自优缺点.

3.1 检测与定位

嘴是唇读过程中最关心的区域. 准确地将人脸和嘴唇实时检测和定位,是一切唇读系统的首要任务. 早先的系统为了能够得到嘴唇或得到轮廓,采用手动的办法框出唇的区域,或将唇涂上深色的口红或者贴反光片[10],且在特定的光照条件下摄取. 也有的人在唇的周围布上发光二极管作为特定的标志来跟踪唇动. 少数的唇读系统将摄像头固定在头盔上,与人脸成固定位置[9],不允许人脸自由移动. 而唇读的目标是在无任何交互和限制条件下,能够自动地将不同光照、不同皮肤颜色、不同话者准确定位、跟踪、识别和理解. 因此智能型的检测手段在不断研究出新.

随着人脸检测技术日益成熟,人脸定位变得容易了. 有关人脸检测技术,请参阅文献[11～13].

一旦检测到人脸以后,就可以有针对性的定位嘴了. 文献[14]介绍的最佳阈值二值化算法以唇的边缘是平滑的和左右形状对称的条件作为二值化阈值选定的约束条件,得到平滑而对称的唇图像.

3.2 特征提取

可视语音信息的特征提取方法可分为两大类:基于像素的方法和基于模型的方法.

所谓基于像素的方法,就是直接利用包含嘴的灰度级图像或直接利用经过若干预处理后得到的特征向量(feature vector)的一类方法. 这类方法的优点是所有的数据都起作用;其缺点是分类器的学习过程中对一般的二维或三维的平移、旋转、放缩,或光照变化或说话人的变化都很敏感;另一个缺点是,其特征向量是高维度的和高冗余度的.

所谓基于模型的方法,就是对可见的发音部位 ——主要是唇的轮廓建立一个模型,而外形的描述用一个小的参数集合. 这类方法的优点是重要特征被表示成低维量且通常不因平移、旋转、放缩或光照而改变. 缺点是独特的模型有可能没能包括所有相关的语音信息.

3.2.1 直接像素的方法

® 直接像素法

Yuhas 等人提出将包含嘴的整个区域灰度图像作为特征向量的方法[15].

加利福尼亚 Ricoh 研究中心的 G.J. Wolff[16] 提出对关键点直接在图上进行标注,对标注后的图以嘴为中心的水平和垂直扫描线作为特征向量,直接用神经网络进行唇读识别和信道融合.

该类方法前期预处理过程简单,但后期需计算数据量大,

复杂度高,需要很大的软硬件开销.它需要收集近于不可能达到的所有模式种类及模式样本去进行训练,且需要很长训练时间.否则,由于其样本的覆盖小,对于变化的情况无法得出正确的结论.其致命弱点就是对光照变化没有鲁棒性.

Movellan[17]提出以图像的垂直中线为对称轴对图像进行对称规范化处理的视觉特征抽取方法,只有对称图像的一半用于进一步的处理,图像的另一半为相邻帧间对应像素的差代替,得到的图像再经过滤波、抽样、比例变换,最后得到每帧300维的特征向量.该方法能有效地减少数据量,提高了计算速度和识别精度.

为了缩小该类方法的盲目性,减少数据冗余,可以选用有较高针对性的矢量量化方法.

® 矢量量化法

Silsbee[18,19]采用了矢量量化的方法,其使用的17个码本矢量是在训练中手工选择建立的,以反映嘴部的不同变化结构.其特征提取过程为:先对原始图像进行直方图平滑以减小反射点以及总的照明不均匀的影响,再进行从左到右的直方图均衡化以减小照明的不对称性,然后最小化图像平移与码本矢量间的失真测度,以得到图像的量化表示[20].由于该方法对数据量进行了有效的控制,对两个小规模的音素数据库和一个中规模的孤立词数据库进行了实验,证明了双通道语音识别对系统性能的提高.该方法对开唇的宽度和高度的差值极为敏感,对牙齿的露与否亦极其敏感.

® 特征唇(主成分分析)法

特征唇方法通常是把整个唇部区域作为一个向量,通过大量的样本进行主成分分析(PCA).这种方法的优点是保留了唇部的大部分信息,不要求有明显的边缘信息.缺点是对唇的变形、旋转非常敏感,而且没有直观的中间处理结果,即定位、跟踪的结果很难检验,一旦识别结果不理想,很难找到问题所在.这种方法也称为主成分分析法或KL变换法.采用主成分分析的目的,就是要求在大规模口型序列图像数据中,迅速将能够代表唇读的信息提取出来,对唇读的主要特征进行认识.它属于非模型化的研究方法.采用主成分分析保证在数据信息损失最小的前提下,对高维数据进行降维处理,迅速揭示系统中的主要因素.

德国的卡尔斯鲁厄大学的Christoph Bregler在加利福尼亚大学计算机科学系进修期间提出了"eigenlips"概念[21].他将特征脸10个主要特征序列和语音信号用TDNN(时间延迟的神经网络)方法进行融合识别,对不同程度多话者环境下进行了实验比较,识别词汇2955个.主成分分析方法的优点是不要求图像有非常清楚的轮廓信息,而且信息丢失量最小,缺点是它对嘴的姿态、大小、旋转、比例变化非常敏感.

福罗里达中心大学计算机系视觉实验室实现了一个特征序列的唇读系统[22],他们将整个序列当作一个矢量进行PCA训练,提出了能量识别率的算法,识别十个英文字母{*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J*}.由于人们在不同时间说话的速度是不一样的,所以同一个词或句子的序列长度会不同.该系统采用了动态时间归正(DTW)的方法,将序列归正成等长的序列.

® 基于光流的分析方法

这种方法是一种基于运动的方法.它的思想是从二维图像序列检测唇运动,提取运动参数,分析运动规律,主要研究的是唇动的瞬时位置速度场,也称为光流场.一般认为瞬时图像灰度没有变化.

采用这种方法的有日本NTT(日本电报电话公司)人机交互研究所的间濑健二[23]、雄本大学电气情报专业的清田等人,他们尝试了用每帧与固定帧的差值和相邻两帧的差值来作为唇读识别的特征,初步证明了用光流的可行性.该算法简单,只识别五个日本人姓名单词.日本大板大学用可变模板技术求口型参数,用光流法跟踪唇动,对5个单元音{[a], [i], [u], [e], [o]}进行了尝试.

使用光流有以下优点:首先,因为人类视觉对运动是敏感的即使在变化的照明条件下,因此光流特征是最有效的;其次,因为词间间隔唇动瞬时速度为零,以此作为切分连续词或句子序列的特征变得容易;再次,因为发同样的音,从生理学角度看其肌肉运动是相同的,这样可以消除个人差别的影响.

MIT媒体实验室Mase和Pentland等人采用了一种基于光流的方法[24],他们采用时空梯度方法,并将嘴唇分为上下左右四个窗,通过计算嘴部4个窗口中的光流,得到语音产生过程中的嘴部肌肉运动,用四个窗的平均速度作为识别的变量,对十个英文数字进行识别.但是由于运动光流估计的约束条件:运动中径向不变和运动物体为刚体,在唇动过程中难以保持,该方法存在着局限性.

福罗里达中央大学计算系G. A. Martin[25]采用视觉流相关技术来解决唇读问题,由于不同人发音速度不一样,所以必须对光流序列进行时间和空间的归正,提出了相关匹配算法,并对三个单音{[d], [h], [t]}进行了实验.

光流方法是一种很有潜力的方法,它对于描述清唇动变化规律有独特的长处.这种方法之所以受到限制没有发展起来,是因为它对预处理的准确定位要求很高,这是目前计算机视觉中的难题.

3.2.2　基于模型的参数描述方法

基于模型的参数描述方法,是对与语音有密切关联的唇的轮廓,用若干参数表示之,并将部分参数及参数的线性组合作为特征送入识别器.这类方法最简单的可以是符合某种模型的特定的点,通过测量这些点的位置、距离等得到的参数送入识别器.将这些点用某种外力和内力去作用它,使它可以活动,就称为"Snake"模型,或主动轮廓模型(active contour model).也可用曲线去逼近轮廓,去描述口型,就称为可变模板(deformable template).可变模板是通过几条曲线来定义轮廓,然后通过一定的限制,用最优化方法将曲线贴近最合适的唇的位置.可变模板不受嘴唇的变形、旋转和缩放的影响,很好地刻画了唇的形状.但可变模板算法的前提是能够有效地提取唇的轮廓边界信息.这种前提常常受不同光照条件的影响,威胁到内轮廓因受牙齿、舌头和口内阴影的影响而不能准确定位.

® 模型点方法

Petajan在系统[9]中采用了几何特征,如嘴张开的高度、宽度、面积等,用阈值技术获得口腔位置,并采用了简单距离测

度,没有进行时间规正. Petajan 的系统后被 Goldschen 发展为用 HMMs 的连续唇读系统,该系统针对的是特定词汇的小集合. 该系统的瓶颈问题即是特征抽取,它没有很好的鲁棒性,需要手工干预. Cosi 的系统[26]中用反光标志标定 8 个点,包括 3 个参考点(左右耳垂与鼻尖)与 5 个目标点(4 个唇点与一个下颌点),由此可以得出 14 个视觉参数:上唇垂直运动 UL、下唇垂直运动 LL、上唇突出 ULP、下唇突出 LLP、唇开高度 LOH、唇开宽度 LOW、下颌开度 JO、上唇垂直运动速率 ΔUL、下唇垂直运动速率 ΔLL、上唇突出速率 ΔULP、下唇突出速率 ΔLLP、唇开高度速率 ΔLOH、唇开宽度速率 ΔLOW、下颌开度速率 ΔJO. Lavagetto 从语音视觉合成的角度给出了 6 个视觉参数:下颌位置、唇外缘高度、唇外缘宽度、唇内缘宽度增量、上唇外缘位置和舌,其中的位置均为相对于鼻尖测量.

这类方法对视觉行为是有效的但却难于实用. 自动地提取特征,不加人工干预和特殊的限定条件,是唇读系统不断追求的目标.

® Snake 方法或主动轮廓模型

主动轮廓模型(Active Contour Model,Active Shape Model)或叫 Snake 算法,采用对若干个唇上特征点的描绘,用一定的限制条件检测到这些点,用内部能量和外部作用力共同作用 Snake 线,得到一个稳定态. 对"Snake"精辟的解释由 Horbelt 在文献[27]给出:"Snake,从数学角度讲,是一条能量最小曲线;从物理角度讲,是一条被内外力驱动的链."

Kass 在文献[28]中给出了原始的"Snake 模型". Snake 是一条变形曲线,它的形状由内部齿条能量(约束曲线的光滑性)和外部特征能量(定义清晰特征)共同作用的:

$$E_{snake(V)} = \sum_{i=1}^{N} E_{snake}(i) = \sum_{i=1}^{N} (E_{int}(i) + E_{ext}(i)) \quad (1)$$

$$E_{int(i)} = \alpha_i \| V_i - V_{i-1} \|^2 + \beta_i \| V_{i-1} - 2V_i + V_{i+1} \|^2 \quad (2)$$

其中 N 为 Snake 的点数,$V_i = (x_i, y_i)$ 为 Snake 点第 i 维坐标,E_{int} 为内力以约束 Snake 的光滑性,α_i 为调节相邻 Snake 点间内部的力量(张力)的常数,β_i 为压制相邻 Snake 点的外部力量的常数. 外部能量

$$E_{ext}(i) = E_{image}(i) + E_{con}(i) \quad (3)$$

其中 $E_{image}(i)$ 可以是任意图像函数(如强度、梯度或两者的综合权值),其作用使 Snake 朝特定的图像演变. 显然,如果边缘检测是目标的话,则 就是一些图像梯度函数.

$E_{con}(i)$ 为任意的约束函数. 例如,它可以用来吸引 Snake 朝用户指定的点运动.

这类 Snake 方法与下面介绍的可变模板方法本质是相同的具有相似的特点:不受嘴唇的变形、旋转和缩放的影响,能得到直观的唇型参数. 但这类方法对自然条件下直接获取的口型照片区分边界并不容易. 因此有人考虑用彩色信息来增强目标和背景的区别. Chiou[29]和姚[13]的方法结合了彩色信息,是对这一目标分离问题的更好解决方案.

® 可变模板方法

可变模板是一个参数化描述的物体模型,其实质是一特定的主动轮廓模型. 在唇读识别中,它就是一个用多条曲线来描述的口型模型.

亚特兰大的乔治亚工学院电子工程院的 R. R. Rao 使用嘴唇的形状进行唇读识别[30]. 他使用一个 3 ×3 线性滤波算子将唇的各个边缘区域分开,分别用四条抛物线拟合,然后用所得到的参数序列进行识别.

而四次曲线比抛物线能更精确反映外唇活动形状[31].

斯坦福大学电子工程系的 Hennecke 和加利福尼亚 Ricoh 研究中心 Stork 等对可变模板方法进行了改进,通过增加一些启发式限制条件[31,32],得到了很好的效果.

德国斯图加特大学的 M. Vogt 利用彩色信息加强了特征模板提取的鲁棒性,但是匹配速度更慢.

文献[13]提出了一种基于色度分析的彩色坐标变换方法,能有效将唇色从相近色系的肤色中区分出来. 其思路是根据不同性别、年龄及肤色的人脸图像肤色和唇色分布的统计规律,得出肤色和唇色各自具有相对稳定聚类特性. 但肤色和唇色处于相近的色系,相互之间有部分交叠. 为把唇色和肤色有效的区分开来,采用对 *YUV* 坐标进行旋转和平移变换,新坐标轴 $\tilde{U}$ 为唇色和肤色聚类中心的连线,$\tilde{V}$ 则是此连线的垂直平分线,这样唇色和肤色的聚类中心到 $\tilde{U}$ 的投影具有最大距离,变换后坐标的 $\tilde{U}$ 值对唇色和肤色具有很好的可分性.

如图 3 所示为将 *RGB* 值的真彩色图转换成旋转后的 $\tilde{U}$ 值图结果对比. 可以看出,增强后的图像,再获得唇形的精确定位就不难了.

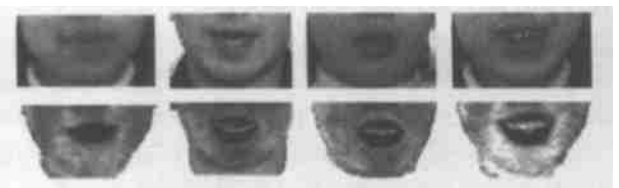

图 3 唇色增强变换前后图像对比
(上行是 *RGB* 真彩色图,下行为 $\tilde{U}$ 值图)

这类方法都直观地再现了特征提取的准确性,降低了特征维数,提高了特征的有效性,但其缺点是预处理过程要求高,需要有鲁棒性和有效性都很强的算法(如文献[13])支持.

3.2.3 混合特征提取法的唇读系统

华盛顿大学电子工程系信息处理实验室 G. I. Chiou 和 J. N. Hwang 使用彩色口型序列图像的主成分分析和 Snakes 算法进行结合[29],并将他们分别作为两组特征用 HMM 识别 10 个模拟在驾驶室里发出命令的词.

瑞士 IDIAP 人工智能感知研究所的 Luettin 提出一种基于训练模型的方法[10,33],这种方法实质上也是主动轮廓模型和主成分分析相结合的方法,取得了较好的跟踪和识别结果. 采用多态 HMM 模型,完成动态的孤立数字集合{one,two,three,four}的识别. 这种方法既保留了轮廓信息,又保留了灰度信息.

美国 AT&T 实验室的 Potamianos 运用了基于轮廓的特征和基于像素的图像变换[50](如离散小波变换、离散余弦变换、Walsh 变换、KL 变换-主成分分析方法)的特征,用 HMM(隐马尔可夫模型)结合高斯 GMM 高斯混合模型的方法,开创了特征提取的新思路,并对图像序列的处理取得了较好的结果.

主动轮廓模型(ASM ——Active Shape Model)的扩展形式为主动面模型(AAM ——Active Appearance Model),由英国曼彻斯特大学的 Cootes 在美国麻省理工大学(MIT)人机交互研究所读博士后期间,和来自英国东安格列大学 Matthews 等人共同提出的[34~36].这种方法的本质与 Luettin 的方法(轮廓模型和主成分分析)很相似.他们还提出了多尺度空间分析法[37](MSA —Multiscale Spatial Analysis).这是特征提取的主流方向.它既含有轮廓模型的信息,又含有纹理和强度信息,是一种特征包含全面数据量又有效降低的表达形式.

3.3　识别方法

® 模板匹配方法

早期唇读识别最简单的方法是从静态图像提取的特征和存入的模板进行比较[15,24,38],忽略特征是随时间变化的事实.这种方法简单,但是只能对简单的元素进行分类,对词一级的识别就没有办法了,所以对语音识别贡献不大.后来人们逐渐采用动态特征.早期唇读识别系统大多是以模板匹配原理为基础建立的,其基本思想是:在训练阶段,将词汇表中的每一个词的特征矢量序列作为模板存入模板库中,在识别阶段,将输入唇读的特征矢量序列依次与模板库中的每一个模板进行相似度比较,将相似度最高的词作为识别结果输出.

Mase 和 Pentland 采用线性时间归正技术进行模板匹配[24],效果得到了改进.采用 DTW 是为解决孤立词识别时说话速度不均匀造成时间伸缩变化的难题而提出来的.DTW (Dynamic Time Warping)其实质是动态规划(DP)概念的扩展.DTW 显著提高了系统性能,但是当词表扩大、孤立词语音转为连续语音或者是非特定人的唇读识别时,这种简单的唇读识别系统就不能胜任了.这是由于 DTW 算法必须有一个精确的起点定位且匹配速度慢等固有缺点所造成的.所以 HMM、TDNN 方法就相继替代模板匹配方法.

® HMM 模型在唇读中的应用

HMM 隐马尔可夫模型(Hidden Markov Model)的基本思想是:认为唇读信号在极短时间内是线性的,用线性模型参数表示,再将许多线性模型在时间上串接起来组成一条马尔可夫链.HMM 用马尔可夫链来模拟信号的统计特性的变化,而这种变化是间接地通过观察序列来描述的.因此,HMM 过程是一个双重的随机过程,这与人的语言唇动过程是相吻合的.唇动信息本身就是一个可观察的序列,它是头脑里(不可观察的)根据言语需要和语法知识(状态选择)所发出的音素(词句)的参数流.因此,唇动信息精确模型化必须用 HMM 来描述才行.HMM 方法一般采用一个半连续的 HMM 模型,它是吸收型的,并且一个状态只能转移到当前状态或下一个状态,既无跨越从左向右模型,并且此模型限定状态 1 为起始状态,如图 4 所示.由此看出,转移矩阵 A 中只有主对角线或右副对角线上的元素允许非零,因而 A 比较稀疏,大大减少了模型参数估计的计算量,因此在唇读识别中采用了无跨越的自左向右模型.

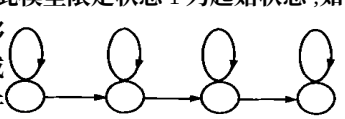

图 4　无跨越的自左向右模型

诺克福老多米尼亚大学电子与计算机工程系目前正在开发一个大型的自然语音系统,借助唇读来改善在噪声环境下单纯的语音识别系统.该系统试图增加唇读识别率,研究唇读和语音的融合策略,开发一个适合唇读研究自然语音的数据库,示范唇读和语音融合系统能比单纯的语音系统更顽健.开发的数据库至少包括 100 人,并且允许扩充.采用可变模板来描绘唇的物理运动,并且作为参数进行识别.在模板提取时采用多模板方法,第一步采用分类的模板进行初始化,然后再采用通用的办法提取更精确的唇的轮廓.识别模型采用 SCHMM 模型,融合采用早期融合模型.针对高斯分布中在尾部出现小值引起大比率概率变化,改进了高斯分布.采用的融合模型能够随噪声水平手动改变权值大小来适应噪声环境.目前能够识别简单的孤立词汇,减少误识率 20 %～50 %.

® TDNN 模型在唇读中的应用

TDNN(Time-Delay Neural Network)是一种延迟神经网络,采用多层结构,输入层是一个随时间变化的时序窗,它同 HMM 一样非常适合于识别序列特征.

德国斯图加特大学与美国 CMU 合作的交互系统实验室 NLIP 小组 1992 年开始进行唇读研究,使用神经网络对连续拼写的德语字母进行识别.该系统在不同的光照或变换环境的条件下不需要人工干预就能够达到实时跟踪和识别的要求.系统的主要研究者 Bregler 到 Berkley 市国际计算机科学研究中心工作,继续从事交互式视觉语言系统的视觉语言模型的研究,并开发一个 BeRP 系统,这是一个交互式连续语言理解媒体字典系统,其中采用唇读改善语音识别率.Bregler 等人使用的 TDNN 包含一个输入层、一个隐层和一个音素状态层,网络训练采用 BP 算法.最后在状态层使用 DTW 发现词模型的最佳音素组合路径.

® BZ 链在唇读中的应用

Boltzmann 机是 Hopfield 网的推广形式,它属于神经网络的一种.时间序列模型可以被看成大量点的外表现,即观察序列的联合概率分布.Boltzmann 链是对上述模型的一种综合.文献[32]采用了该方法.

3.4　融合

文献[39]对三种唇读融合模型:前融合、中间融合和后融合模型进行了详细的探讨,并对后融合模型采用各信道可信度评价的算法机制;而对语音特征力图进行前融合,因为实验表明语音和口型运动的生理过程应该是一个前融合过程(故意用其他声音干扰一段口型序列,人会得到一种假想的意念);对中间融合,则探讨了如何自适应外界噪声对融合的权值进行有效的调整,探讨自适应机制.

表 1 给出了各唇读系统所采用的预处理方法、特征提取方法、识别方法、融合方法以及所处理的对象集合比较情况.

4　唇读识别应用

(1)辅助语音识别

唇读首当其冲的应用是与语音识别进行融合,辅助提高语音识别率.在噪音环境和多话者条件下,将唇读和语音进行融合能明显改善语音识别率[15,40].

(2)辅助手语识别

手语和唇读是分不开的,唇动是手语的一个组成部分.对

于聋哑人来说,看手语的时候,不仅看手的动作和形状,也要看人脸表情和唇动.所以,对口型序列识别的研究是手语识别的不可缺少的组成部分,它有助于对手语的正确理解.其融合方法与和语音融合的方法相类似.

表 1 各种唇读系统比较

研究者与文献	研究单位	检测定位	特征提取	识别算法	融合	任务集
Petajan 84[9]	伊利诺伊斯大学(现在贝尔实验室)	画面固定并利用鼻孔位置	阈值	距离	后融合	字母
Finn 86[37]	乔治州大学	画面固定	反射点	距离	无	单音
Yuhas 89[15]	乔治华盛顿大学	画面固定	直接基于像素	神经网络	先融合	元音
Mase 91[24]	MIT 媒体实验室	画面固定	光流法	LTW(线性时间归正)	无	数字
Stork 92[7]	CRC(加利福尼亚 Ricoh 研究中心)	画面固定	反射点	TDNN(时间延迟神经网络)	先中后融合	单音
Goldschen 93	乔治华盛顿大学	画面固定	阈值	HMM(隐马尔可夫模型)	无	句子
Silsbee 93[18]	德克萨斯州立大学	画面固定	矢量量化	HMM(隐马尔可夫模型)	后融合	单音、单词
Rao 94[30]	乔治亚工学院	边缘检测	可变模板	DTW(动态时间归正)	无	2 个单词
Bregler 94[48]	卡耐基梅隆大学-(德)卡尔斯鲁厄大学	画面固定	轮廓	NN(神经网络) + HMM(隐马尔可夫模型)	先融合	字母、单词
Hennecke95[40]	CRC(加利福尼亚 Ricoh 研究中心)-斯坦福大学	肤色模型 + 对称性	可变模板	HMM(隐马尔可夫模型)	先融合	单词
Movellan95[17]	加州圣地亚哥大学认知学系	只有嘴部	直接基于像素	HMM(隐马尔可夫模型)	先后融合	数字
Waibel 95	卡耐基梅隆大学-(德)卡尔斯鲁厄大学	肤色模型 + 神经网络	PCA(主成分分析)	TDNN(时间延迟神经网络)/DTW(动态时间归正)	中融合	字母
Silsbee 96[19]	德克萨斯州立大学	轮廓、差分	可变模板	HMM(隐马尔可夫模型)	后融合	单音
Stork 96[32]	CRC(加利福尼亚 Ricoh 研究中心)	轮廓、肤色	可变模板	BZ(玻耳兹曼链)	中融合	单音
Cosi 96[26]	(意)国立研究会声学研究所	画面固定	反射点	TDNN(时间延迟神经网络)	中融合	单音
Vogt 97	(德)斯图加特大学	彩色	可变模板	时间延迟神经网络	无	单词
Luettin 97[46]	瑞士 IDIAP-(英)舍菲尔德大学	画面固定	主动轮廓 + 主成分分析	HMM(隐马尔可夫模型)	无	4 个数字
Chiou 97[29]	华盛顿大学	画面固定 + 彩色单体值之比	Snake 模型 + 主成分分析	HMM(隐马尔可夫模型)	单纯唇读(无语音信息)	10 个孤立词
Li 97[22]	福罗里达中心大学	画面固定	直接基于像素	DTW(动态时间归正)	无	10 个字母
Matthews 98[43]	(英)东安格列大学	固定光照	AMM(主动面模型)	HMM(隐马尔可夫模型)	后融合	{A…Z}
Cootes 98[34]	MIT 人机交互研究所-(英)曼彻斯特大学	标志点	AAM(主动面模型)	模板匹配	后融合	{A…Z}
Pentland 98	MIT 媒体实验室	标志点	3D 形状模型	TDNN(时间延迟神经网络)	无	数字
Potamianos98[50]	AT&T 实验室	色彩 + 轮廓	Fourier 傅立叶系数和图像变换	HMM(隐马尔可夫模型) + GMM(动态高斯混合模型)	无	数字字符串
徐 98[20]	(中)中科院语音所	——	——	——	——	建双语料库
Cienkowski 99	明尼苏达州立大学	固定	直接基于像素	HMM(隐马尔可夫模型)	后融合	单字
Edwards 99[44]	MIT 人机交互研究所	固定	AAM(主动面模型)	模板匹配	后融合	字母
Cootes 2000[45]	MIT 人机交互研究所	标志点	AAM(主动面模型)	模板匹配	后融合	字母
Grant 2000[37]	MIT 人机交互研究所	标志点	MSA(多尺度分析)	距离	后融合	字母
姚 2000[13]	(中)哈尔滨工业大学	肤色-唇色增强模型	可变模板	HMM(隐马尔可夫模型) + GMM(动态高斯混合模型)	无	单元音

(3) 聋人辅助教育

聋人通过后天练习可以恢复其说话能力.开发一个口型和语音相对应的系统能够帮助聋人学校教师对其学生进行发音练习.

(4) 口型合成

目前人们在口型合成方面的研究,主要集中在文本驱动的口型合成.其方法就是将音素(元音和辅音)的口型进行分类,定义几个基本口型(主要是单韵母的发音口型),并根据基本口型集,衍生出一个音素口型库,然后将切分出的文本转化为拼音,通过查询韵母口型库,得到文本发音时对应的口型.但是目前还没有文献提及用声音驱动口型合成,这方面的工作在可视电话中有着非常的意义.

(5) 口型编码

将发音的口型特征如果能按照某种标准进行描述和编码的话,无疑对数据压缩和通讯有极其重要的意义.MPEG-4 给出了这种标准.在 MPEG-4 视频编码标准中定义了人的脸部的合成编码,通过定义人脸的模型及运动参数,在编码过程中只传输模型和运动参数,这样能极大地提高编码效率.例如在虚拟电视会议系统和视频电话中,人们最感兴趣的是人的脸部.根据 MPEG4 中定义的人脸模型参数和人脸的运动参数,对于面部视频对象定义了面部动画参数 FAPs(Face Animation Parameters)和面部定义参数 FDPs(Face Definition Parameters),以及缺省值.采用 SNHC 面部和身体运动编码可以获得 1kbps 的超低码率.对于像可视电话等低带宽应用是非常有价值的.

另一方面,纹理、形状和运动的独立描述为充分利用视频对像素材,制作新的面部和身体对像提供了方便.

(6) 说话人基于唇动特点的生物特征的识别

人说话都有各自的特点,包括说话时的口型运动变化都有其自身的规律.利用人说话时口型变化的每个人不同的特点,可以进行人的身份鉴别,并已有研究者做了这方面的尝试.瑞士 IDIAP 人工智能感知研究所于 1995～1998 开发一个多模型身份认证系统[41,42],以唇的张开速度、唇的关闭速度和时空面积作为识别特征,采用孤立词依赖来进行身份认证,只对几个人进行识别.目前该系统在特定数据库视觉认证能达到 85 %～90 %,合并视觉和语音认证能达到 99 %精确度.

5　结论

唇读研究的困难在于口型和语音是一个一对多的对应关系,单从口型来映射语音是不确定的.如果没有其它相关知识,要识别绝对是不可能的.而这些相关知识,需要很多.比如,特定语言学、音韵学的规律,相关领域的专业知识背景,这些知识库的建立和支持并不是一件容易的事.另一方面,唇读研究不可回避的计算机视觉问题,如光照变化、深度信息的缺乏,给唇的描述和识别带来巨大的障碍.随着计算机视觉、知识表达、知识推理技术的不断发展,我们有理由相信:视觉语言一定会被人们完全掌握和运用.

就目前的唇读技术而言,它尚属于初级阶段,相当于 70 年代末语音识别的发展阶段.笔者以为,在唇读过程中,预处理不宜太简单,否则把问题遗留到了后期的计算和处理中;在特征选择方案中,可考虑选择带一定基于模型的特征与一些重要像素的强度和纹理特征相结合的方法,既减少数据的相关性和冗余度,又包含具有隐藏性和非模型化的充分的信息量,促使后期识别处理的算法不过于复杂,以保证实时性,同时又不失准确性;在识别阶段的模型中,则一定要采用具有描述动态变化能力的模型,如 HMM,或其改进形式;最后融合其它通道的信息,得出正确的理解结果.这是一个可行的模式.

参考文献:

[1] B. Dodd and R. Campbell, editors. Hearing by Eye: The Psychology of Lip-Reading [M]. Lawrence Erlbaum Associates Ltd., London, 1987.

[2] A. Q. Summerfield. Lipreading and audio-visual speech perception [J]. Philosophical Transactions of the Royal Society of London, Series B, 355, 1992: 71 - 78.

[3] W. H. Sumby and I. Pollak. Visual contributions to speech intelligibility in noise [J]. Journal of the Acoustical Society of America, 1954, 26: 212 - 215.

[4] K. W. Grant and L. D. Braida. Evaluating the articulation index for auditory-visual input [J]. Journal of the Acoustical Society of America, 1991, 89(6): 2952 - 2960.

[5] Y. Gong. Speech recognition in noisy environments: a survey [J]. Speech Communication, 1995, 16: 261 - 291.

[6] 王瑞. 连续语音唇读识别的研究 [D]. 哈尔滨工业大学计算机系博士论文开题报告, 哈尔滨工业大学档案馆, 1998.

[7] D. G. Stork, G. J. Wolff and E. P. Levine. Neural network lipreading system for improved speech recognition [A]. Proceedings International Joint Conference on Neural Networks [C], 1992, Volume 2, 289 - 295.

[8] M. E. Hennecke, D. G. Strok, and K. V. Prasad. Visionary Speech: Looking Ahead to Practical Speechreading Systems [M]. In David G. Stork and Marcus E. Hennecke, editors, Speechreading by humans and machines, Springer Verlag, Berlin, volume 150 of NATO ASI Series, Series F: Computer and Systems Sciences. 1996: 331 - 350.

[9] E. D. Petajan. Automatic lipreading to enhance speech recognition [D]. Ph. D. thesis, University of Illinois at Urbana - Champain, 1984.

[10] J. Luettin, N. A. Thacker. Speechreading using probabilistic models [J]. Computer Vision and Image Understanding, 1997, l65(2): 163 - 178.

[11] Y. Dai and Y. Nakano. Face-texture model based on SCLD and its application in face detection in a color scene [J]. Pattern Recognition, 1996, 29(6): 1007 - 1017.

[12] M. B. Liu, H. X. Yao, W. Gao. Real-time human face tracking in color images [J]. Chinese Journal of Computer, 1998, 21(6): 527 - 532.

[13] 姚鸿勋, 刘明宝, 高文等. 基于彩色图像的色系坐标变换的面部定位与跟踪法 [J]. 计算机学报, 2000, 23(2): 158 - 165.

[14] H. Yao, R. Wang, W. Gao. Method of deformable optimum threshold for lip-reading [A]. IEEE Fourth International Conference on Signal Processing [C], October 12 - 16, 1998, Beijing, China, ICSP '98-II: 912 - 915.

[15] B. P. Yuhas, M. H. Goldstein and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural nets [J]. IEEE Communication Magazine, November 1989: 65 - 71.

[16] G. J. Wolff, K. V. Prasad, D. G. Stork & M. Hennecke. Lipreading by neural networks: visual preprocessing, learning and sensory integration [A]. Proceedings of the Neural Information Processing Systems-6 [C], Morgan Kaufmann, 1994: 1027 - 1034.

[17] J. R. Movellan. Visual Speech Recognition with Stochastic Networks [M]. In G. Tesauro, D. Touretzky, and T. leen, editors, Advances in Neural Information Processing Systems, volume 7. MIT press, Cambrige, 1995.

[18] P. L. Silsbee. Computer lipreading for improved accuracy in automatic speech recognition [D]. Ph. D. Thesis, University of Texas at Austin. 1993.

[19] P. L. Silsbee and A. C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition [J]. IEEE Transactions on Speech and Audio Processing, 1996, 4(5): 337 - 351.

[20] 徐彦君. 中文双语料语音识别关键技术研究 [D]. 博士论文. 北京: 中科院语音所, 1998.

[21] C. Bregler, Y. Konig. "Eigenlips" for robust speech recognition [A]. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE_ICASSP) [C], Adelaide, Australia, 1994.

[22] N. Li, S. Dettmer and M. Shah. Visually recognizing using eigensequences [DB/OL]. http://www.cs.ucf.edu/～vision/papers/shah/97/NDS97.pdf, 1997.

[23] 间濑健二. 读唇 [J]. 电子情报通信学会论文志, 1990, J73-D-II (6): 796 - 803.

[24] K. Mase and A. Pentland. Automatic lipreading by optical flow analysis [R]. Technical Report 117, MIT Media lab, 1991.

[25] G. A. Martin and M. Shah, Lipreading using optical flow [A]. Proc.

Nat. Conf. Undergraduate Research [C], March 1992.

[26] P. Cosi, E. M. Caldognetto, F. Ferrere, M. Dugatto and K. Vagges. Specker indepent bimodal phonetic recognition experiments [A]. Proc. ICSLP '96 [C]. October 3-6, 1996, Philadelpha, Pennsylvania, USA.

[27] S. Horbelt, J. Dugelay. Active contours for lipreading combining snakes with templates [A]. 15th GRETSI Symposium on Signal and Image Processing [C], Juan Les Pins, France, 18 - 22 Sept., 1995. URL: http://www.cica.fr/~image.

[28] M. Kass, A. Witkin & D. Terzopoulus. Snakes: active contour models [J]. International Journal of Computer Vision, 1988: 321 - 331.

[29] G. I. Chiou and J. N. Hwang. Lipreading by using snakes, principal component analysis and hidden Markov models to recognize color motion video [J]. IEEE Trans. on Image Processing, 1997, 6(8): 1192 - 1195.

[30] R. R. Rao, Russell M. Mersereau. Lip modeling for visual speech recognition [A]. Proceeding of 28th Annual Asilomar Conference on Signals [C], Systems, and Computers, Pacific Grove, CA, 1994.

[31] M. E. Hennecke, K. V. Prasad and D. G. Stork. Using deformable templates to infer visual speech dynamics [A]. 28th Annual Asilomar Conference on Signals, Systems and Computers [C], Pacific Grove, CA. IEEE, IEEE Computer Society Press, 1994, Volume 1, 578 - 582.

[32] David G. Stork and Marcus E. Hennecke, editors. Speechreading by Humans and Machines: Models, Methods, and Applications [M]. volume 150 of NATO-ASI Series, Series F: Computer and Systems Sciences, Berlin, Springer-Verlag, 1996.

[33] Juergen Luettin, Neil A. Thacker, Steve W. Beet. Visual speech recognition using active shape models and hidden markov models [R]. University of Sheffield, Electronic Systems Group Report No. 95/47, 1996.

[34] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models [A]. Proc. European Conference on Computer Vision [C], June 1998: 484 - 498.

[35] T. F. Cootes and C. J. Taylor. A mixture model for representing shape variation [J]. Image and Vision Computing, 1999, 17(8): 567 - 574.

[36] T. F. Cootes, C. Beeston, G. J. Edwards and C. J. Taylor. A unified framework for Atlas matching using active appearance models [A]. Proc. Int. Conf. on Image Processing in Medical Imaging [C], (Springer, LNCS 1613) 1999: 322 - 333.

[37] K. W. Grant, P. F. Seitz. The recognition of isolated words and words in sentences: individual variability in the use of sentence context [J]. Journal of the Acoustical Society of America, 2000, 107: 1000 - 1011.

[38] K. Finn. An investigation of visible lip information to be used in automatic speech recognition [D]. Ph. D. thesis, Georgetown University, Washington, D. C, 1986.

[39] E. D. Petajan, B. J. Bischoff, D. A. Bodoff and N. M. Brooke. An improved automatic lipreading system to enhance speech recognition [R]. Bell Labs Tech. Report TM 11251 - 871012 - 11. 1987.

[40] M. E. Hennecke, K. Venkatesh Prasad and David G. Stork. Automatic speech recognition system using acoustic and visual signals [A]. The 29th Asilomar Conference on Signals, Systems and Computers [C], Pacific Grove, CA, IEEE Computer Society Press, November 1995.

[41] J. Luettin, N. A. Thacker and S. W. Beet. Speaker identification by lipreading [A]. Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96) [C], 1996.

[42] P. Jourlin, J. Luettin, D. Genoud and H. Wassner. Acoustic-labial speaker verification [A]. Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA '97) [C], 1997.

[43] I. Matthews. Features for Audio-Visual speech recognition [D]. Ph. D. thesis, University of East Anglia, 1998.

[44] G. J. Edwards, C. J. Taylor and T. F Cootes. Improving identification performance by integrating evidence from sequences [A]. Proc. IEEE CVPR 1999, I: 486 - 491.

[45] T. F. Cootes, K. N. Walker and C. J. Taylor. View-based active appearance models [A]. Proc. Int. Conf. on Face and Gesture Recognition [C], 2000, 227 - 232.

[46] J. Luettin. Visual speech and speaker recognition [D]. Ph. D. Dissertation of Department of Computer Science University of Sheffield, UK, 1997.

[47] K. N. Walker, T. F. Cootes, and C. J. Taylor. Automatically building appearance models [A]. Proc. Int. Conf. on Face and Gesture Recognition [C], 2000, 271 - 276.

[48] C. Bregler, S. M. Omohundro. Surface Learning with Applications to Lipreading [M]. in J. D. Cowan, G. Tesauro and J. Alspector, editors, Advances in Neural Information Processing Systems, volume 6. Morgan Kaufmann, 1994.

[49] M. Vogt. Interpreted multi-state lip models for audio-visual speech recognition [A]. Proceeding of the AVSP '97 work shop [C]. Rhodes (Greece), Sept. 1997: 26 - 27.

[50] G. Potamianos, H. P. Graf and E. Cosatto. An Image Transform Approach for HMM Based Aotumatic Lipreading [A]. Proceeding of the International Conference on Image Processing [C]. Chicagao, 1998, III: 173 - 177.

作者简介:

姚鸿勋 哈尔滨工业大学博士生,副教授,1987 年 7 月、1990 年 3 月先后获得哈尔滨船舶工程学院计算机信息与工程系计算机应用专业学士和硕士学位. 主要研究方向:图像处理、模式识别、多媒体技术及自然人机接口. 已发表论文 20 余篇.

高 文 教授,博士生导师,国家八六三计划智能计算机主题专家组组长,1988 年获哈尔滨工业大学计算机应用博士学位,1991 年获日本东京大学电子学博士学位. 主要研究领域: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、虚拟现实等. 已发表论文 200 余篇.