基于多分辨率的多层分类器的手语识别方法*)

张晨曦　姚鸿勋　姜　峰
(哈尔滨工业大学计算机科学与技术学院　哈尔滨150001)

摘　要　本文提出了一种基于多分辨率的多层分类器的手语识别方法，该方法对来自数据手套的手语输入，先用多分辨率选择特征，然后根据这些特征数据先进行低分辨率识别，再使用全部数据进行高分辨率识别。实验结果表明，该方法比传统HMM(隐马尔可夫模型)识别过程识别速度平均提高了约0.6秒，识别率提高了6.73%。

关键词　手语识别，特征选择，多分辨率分析，隐马尔可夫模型

A Method of Sign Language Recognition Based on Multiresolution Multilayer-Classifier

ZHANG Chen-Xi　YAO Hong-Xun　JIANG Feng
(Department of Computer Science, Harbin Institute of Technology, Harbin 150001)

Abstract　A method of sign language recognition based on multiresolution multilayer-classifier is proposed in this paper. For the input from data gloves, firstly features are selected using multiresolution theory, then low-resolution recognition with those selected features is made, at last high-resolution recognition with all the data is executed. The result of experiments shows that the recognizing average speed is 0.6 second faster than the traditional HMM recognition process, and the recognizing ratio has been enhanced by 6.73%.

Keywords　Sign language recognition, Feature selection, Multiresolution analysis, Hidden markov model

1　引言

中国目前约有1800万聋哑人，虽然他们懂得手语，彼此之间可以进行交流，但是正常人并不懂得手语，与正常人之间的交流困扰着他们。这就使对手语的研究成为必要。随着计算机技术的发展，人工智能在计算机领域内，得到了愈加广泛的发展。通过自然的人机交互(HCI)方式，数据手套作为人机交互的输入设备，使计算机能够对手语进行识别与合成，再加上语音识别与合成技术的帮助，这就使聋哑人与正常人之间顺利交流成为现实。

国际上对手语识别的研究最初是从静态手势识别开始的，20世纪80年代末，J.Kramer 和 L.Leifei 提出了 Talking Glove 系统[1]，它能够将手势输入翻译成语音输出。随后，S.S Fels 和 G.E. Hinton 提出的 GloveTalk 系统[2]，是手语识别的原型系统。90年代初，Schlenzig 等人提出基于视觉手势的识别[3]。随后的研究工作从静态的手势识别逐渐发展到动态手语识别，如 CMU 的美国手语识别系统[4]和台湾大学的台湾手语识别系统[5]。目前国内的手语识别发展水平在国际上是最高的，主要的方法都是基于 HMM[6]模型下的研究。由于人工神经网络研究高涨，出现了 ANN/HMM[7]的方法。随后又提出了 DGMM/HMM[8]和 SOFM/HMM[9]的方法，这两种方法都是对手语词序列在 HMM 每个状态下观察概率进行估计，都得到了很好的识别率。但是以上这些方法都存在识别速度的问题，即当随着手语库中的词汇量增多时，识别速度会随之下降。针对既要保证识别率又要提高识别速度的问题，本文提出了多分辨率手语识别的方法。本文主要分成三个部分：1.多分辨率思想的提出及特征选择方法；2.手语识别系统的设计；3.实验结果及结论。

2　基于多分辨率的多层分类器设计

2.1　多分辨率思想的提出

我国现在所做的手语识别系统都是基于 HMM 下的系统。HMM 是一个双随机过程，其中的状态转移分布为一个马尔可夫分布，状态的概率密度函数描述了该状态观测到特征向量的概率。手语识别的训练过程是为每个手语词建立一个 HMM，如果有100个词语，那么构造100个 HMM 与之对应。在识别的过程中，每个手语词在 HMM 模型下看成是一个观察序列，该观察序列与所有 HMM 进行匹配，用向前向后算法分别计算出该观察序列在每个 HMM 下产生的概率。概率值最大的手语词作为识别结果。如果随着手语识别系统中词库数量的增多，每个词语对应的 HMM 也会随之增多，那么识别一个词所用的时间也会随之增多。这就有必要降低手语识别的时间，而且提高识别速度，在以后的工作中对句子识别有重要意义。

首先分析所要识别手语词的数据结构。目前手语识别系统从数据采集上分为两种，基于视觉的手语识别系统和基于数据手套的手语识别系统。本文所要研究的是后者，输入设备采用美国 Virtual Technologies 公司的具有18个传感器的 CyberGlove 型号数据手套及其配套的3个 Polhemus FASTRAK 3-D 位置跟踪器，数据传输频率均取38400波特。以身体颈部的三维方向为参照系，位置跟踪器是用于采集手的运动方向、位置(各3维数据)，左右手与坐标原点距离(2维数据)和左右手之间距离(1维数据)。最后可计算出，每个时刻从输入设备上得到51维的向量作为输入数据。一个手语词的帧数为20帧到70帧之间不等。要想提高识别速度最直接的方法是减少每个手语词的数据量，为达到该目标本文提出了多分辨率识别。

*)本文得到国家863项目资助。

对手语词进行“二进”特征选择，如51维数据先特征选择其一半的数据(25维)进行低分辨率识别，然后再用51维数据进行高分辨率识别。当然25维数据可以再进行“二进”特征选择进行识别。本文的方法只进行一次“二进”特征选择。

2.2 特征选择

在进行第一步的低分辨率识别时，欲从51维数据中选择25维数据使其对识别率效果最好，如果对每一种可能的情况都进行测试，那么它的解空间是$\binom{51}{25}$，如果对5000个手语词进行训练和测试一遍的时间是4个小时，那么得到此问题的解需要约200亿年。所以无法找到最优的解，本文提出一种数据特征选择的方法，该方法是选择所有中国手语词(按51维计算)平均变化最剧烈的25维数据，即取出51维数据中方差最大的25维。具体算法如下：

定义1 一个手语词的格式$O=[A_1,A_2,\cdots,A_i,\cdots,A_T]$，$T$代表一个手语词的帧数(每个手语词$T$是不同的)，其中$A_i$表示第$i$帧的一个51维向量：$A_i=\langle a_1,a_2,a_3,\cdots,a_{51}\rangle^T$。用$O(j)$表示第$j$个手语词，$O_j(A_i)$表示第$j$个手语词的第$i$帧向量。

(1)计算每一维数据的数学期望$E[k]$，其中$k=1,2,\cdots,51$，假设各个手语词每帧之间以及51维数据之间都是相互独立的等概率分布。

N=训练手语词总数

$T(j)$=第j个手语词帧数，$j=1,2,\cdots,N$

$$sum=\sum_{j=1}^{N}T(j),\ j=1,2,\cdots,N$$

$$E[k]=\frac{1}{sum}\sum_{j=1}^{N}\sum_{i=1}^{T(j)}O_j(A_i),\ i=1,2,\cdots,51$$

(2) 计算每一维数据的方差$cov[k]$。

$$cov[k]=\frac{1}{sum}\sum_{j=1}^{N}\sum_{i=1}^{T(j)}(O_j(A_i)-E[k])^2,\ k=1,2,\cdots,51$$

(3)在$cov[k]$中选择最大的25维，返回该25个k值作为特征选择结果。

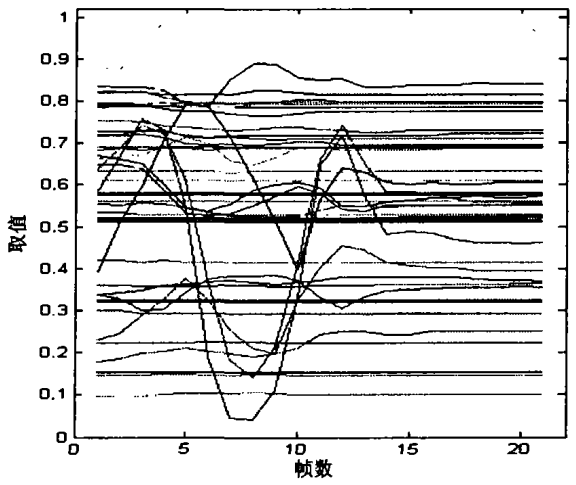

图1 手语词“地球”通过传感器传出的各维数据随时间变化状况图

图1是手语词“地球”图像表示，横轴是一个该手语词的帧数(该词为21帧)，可看成时间变化，纵轴是传感器的数据表示。图中画出了51条曲线，每一条曲线表示其对应某维数据随时间变化的情况。由上述算法可得出所选择的25维数据就是图1中所示曲线变化最剧烈的25条曲线，它们对识别的结果明显比随机选择特征的结果好。本文在第4节给出了随机选择25维特征数据和用此算法选择特征数据时其识别结果的比较。

2.3 建立易混词表

如果只依赖一半的特征数据进行识别显然识别率不高，但是可以先用一半特征数据进行测试，得到每个手语词对应的错误识别词，然后为每一个手语词建立一个易混词表。算法描述如下：

(1)N=训练手语词总数。

(2)为每个手语词O_i建立一个易混词表$Table(i)$，初始化$Table(i)$为空，其中$i=1,2,\cdots,N$。

(3)如果手语词O_i被错误识别成手语词O_j，那么将O_i放入O_j的易混词表中。

(4)如果没有任何手语词被识别成手语词O_j，那么O_j的易混词表为空。如果O_j不为空，将其本身加入易混词表中。

(5)返回易混词表$Table(i)$，$i=1,2,\cdots,N$。

对每个手语词建立一个易混词表的意义是：在对一个未知的手语词进行识别的时候，先用一半的数据进行低分辨率识别，假设识别结果是O_i这个手语词，说明这个未知的手语词不一定是O_i，而有可能是它对应的易混词表中的某个手语词。所以需要继续在它的易混词表中，用全部数据进行高分辨率识别，最后得到识别结果。

3 多分辨率手语识别系统

本文的手语识别系统是基于DGMM/HMM的多分辨率手语识别系统。其中数据输入是本文在2.1节中提到的CyberGlove型号数据手套及其配套的3个Polhemus FASTRAK 3-D位置跟踪器所采集的51维向量序列。整个手语识别系统工作分为三大步骤：训练过程、生成易混词表过程和识别过程。

训练过程：首先，为所有手语词建立一个词汇表作为索引。以每个手语词为单位，训练出一个HMM和对应该HMM下各个状态的DGMM，其中DGMM用于估计每个手语词各个帧在其HMM下每个状态的发生概率值。其中DGMM的训练公式参考文[8]。但是本文要对一个手语词建立两个HMM和DGMM，分别对应25维的手语词和51维的手语词。

生成易混词表过程：用大量测试数据对25维的手语词进行测试，得到测试的结果后，再用本文2.3节提出的算法生成易混词表。

识别过程：以上两个过程是对手语识别系统的准备工作，有了这两个过程后，系统可以完成识别工作。识别过程也是对系统测试和评价的过程，其具体工作流程如下：

图2为本系统流程图，工作流程分以下4步进行。

(1)对于输入系统51维向量序列的手语词，进行特征选择得到25维向量。

(2)特征选择后向量进入HMM识别器，在所有25维HMM库中找到最大的$P(O|\lambda)$。

(3)找到(2)中最大$P(O|\lambda)$所对应的手语词的易混词表，查看该词在易混词表中是否有易混词。如果没有则此识别过程结束，该手语词为输出结果，否则执行(4)。

(4)将开始输入的51维向量带入HMM识别器，在(3)中找到的易混词中找到最大的$P(O|\lambda)$。将最大$P(O|\lambda)$所对应的手语词作为识别结果输出。

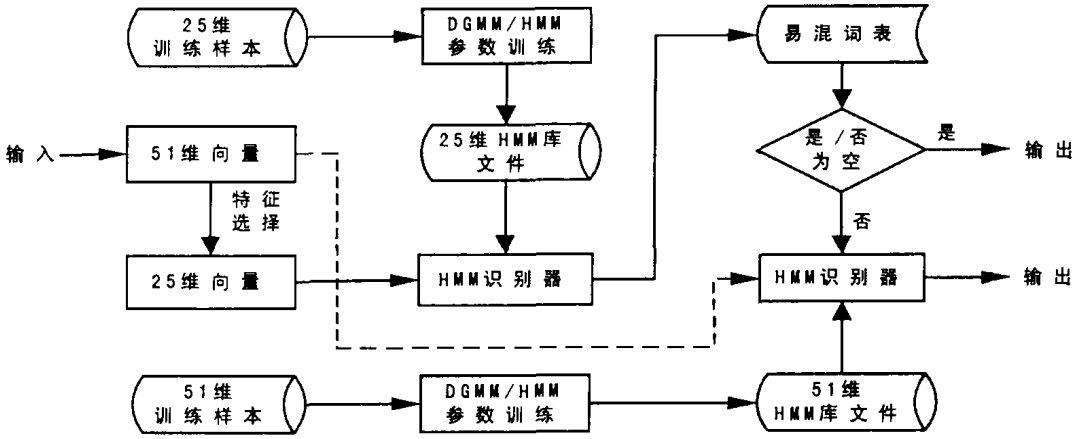

图2 系统流程图

4 实验结果

本文实验所选用的手语数据是由5位手语老师分别打2遍的手语词，把这一共10遍的手语词分成5组，每组中测试数据是每位老师打的一遍数据，训练数据是该老师第二遍数据加上其他4位老师的2遍数据。手语词选择的是《中国手语词典》中常用的4942个手语词。

表1 随机选择特征与取方差最大选择特征对识别率的影响

人名	ljh	llq	lwr	pfz	ygy	平均值
随机选择(%)	49.41	39.84	43.52	54.43	50.71	47.58
使方差最大选择(%)	67.62	64.77	66.06	75.61	70.13	68.84

进行手语识别测试时，先对51维手语数据进行特征选择后，实现低分辨率手语识别，表1是随机选择25维数据与用本文2.2节中所述特征选择方法其识别率的比较。由此可见，该方法比随机选择特征对识别率的贡献提高了21.26%，从而能保证下一步进行高分辨率识别后，提高最后的识别率，也减少了每个手语词的易混词表中易混词数，从而降低了时间复杂性。

表2 多分辨率识别系统与传统HMM识别系统比较

人名	识别率(%)		识别每个词时间(秒)	
	传统HMM系统	本文系统	传统HMM系统	本文系统
ljh	84.42	91.52	2.369	1.839
llq	83.97	91.76	2.366	1.689
lwr	86.30	93.97	2.358	1.794
pfz	91.87	96.03	2.368	1.699
ygy	85.90	92.82	2.412	1.829
平均值	86.49	93.22	2.375	1.770

表2是本文所使用的多分辨率手语识别系统与传统的单一HMM系统识别率的比较。实验结果表明，本文提出的多分辨率识别系统比传统单一HMM识别系统在平均识别一个手语词时间上缩短了0.605秒，识别率提高了6.73%。

由此可见，本文提出的多分辨率识别系统在识别速度和识别率都有了提高。首先，在进行低分辨率识别及对手语数据进行特征选择的过程中，把对识别没有贡献的那些维(相当于噪声)滤掉，从而提高了识别率。其次，从系统处理的数据量上看，本系统处理的数据量约是单一HMM识别系统的一半，所以可以达到提高识别速度的目的。

结论 本文提出了基于多分辨率思想的多层分类器的手语识别方法，给出了手语数据特征选择的算法，并实现了多分辨率手语识别系统。试验结果表明，本文提出的多分辨率识别系统比传统单一HMM识别系统在识别速度和识别率上都有明显的提高。在本系统中若进行多步"二进"特征选择、多层识别将会提高系统识别性能，此工作是本文今后研究的方向。

参考文献

1 Kramer J, Leifer L. The Talking Glove: A speaking aid for non-vocal deaf and deaf-blind individuals. In: Proc. of RESNA 12th Annual Conf. 1989. 471～472

2 Fels S S, Hinton G E. GloveTalk: A neural network interface between a DataGlove and a speech synthesizer. IEEE Transactions on Neural Networks, 1993. 2～8

3 Schlenzig J, Hunter E, Jain R. Recursive Identification of Gesture Inputers Using Hidden Markov Models. In: Proc. 2nd Annual Conf. on Applications of Computer Vision, 1994. 187～194

4 Starner T, Pentland A. Real-time American sign language recognition form video using hidden Markov models. MIT media Lab Perceptual Computing Section: [TR-375]. 1996

5 Liang R, Ouhyoung M. A sign language recognition system using hidden Markov model and context sensitive search, In: Proc. of the ACM Symposium on VR software and Technology, Hongkong, 1996. 59～66

6 Rabiner L R, Juang H. An Introduction to Hidden Markov Models. IEEE ASSP Magazine, 1986. 4～16

7 Wu Jiangqin, Gao Wen. Sign Language Recognition Method on ANN/HMM. Computer science and application, 1999(9): 1～5

8 Wu Jiang-Qin, Gao Wen. A Hierarchical DGMM Recognizer for Chinese Sign Language Recognition. Journal of Software, 2000, 11(11): 551～552