基于关键帧的多级分类手语识别研究 *

姜华强[1,2], 潘 红[2]
(1. 上海大学 机电工程与自动化学院, 上海 200072; 2. 杭州师范大学 信息科学与工程学院, 杭州 310012)

摘 要: 提出了一种基于关键帧识别的多级分类的手语识别方法,该方法采用 HDR(多层判别回归)/DTW(动态时间规正)模板匹配多级分类方法。根据手语表达由多帧构成的特点,采用 SIFT(尺度不变特征变换)算法定位获取手语词汇的关键帧,并提取其特征向量;根据手语词汇的关键帧采用 HDR 方法缩小搜索范围,然后采用 DTW 比较待识别的手语词特征与该范围内每一个手语词进行匹配比较,计算概率最大的为识别结果。这种方法在相同识别率的情况下比 HMM 识别方法速度提高近 8.2%,解决了模板匹配法在大词汇量面前识别率快速下降的问题。

关键词: 手语识别;多层判别回归方法;模板匹配
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)02-0491-03
doi:10.3969/j.issn.1001-3695.2010.02.023

Key frame based multi-level classification of sign language recognition

JIANG Hua-qiang[1,2], PAN Hong[2]
(1. *College of Mechatronics Engineering & Automation, Shanghai University, Shanghai* 200072, *China*; 2. *School of Information Science & Engineering, Hangzhou Normal University, Hangzhou* 310012, *China*)

Abstract: This paper presented a sign language recognition method based on the multi-level classification of key frame recognition. This method adopted hierarchical discriminant regression (HDR) and dynamic time warping (DTW) template to match multi-level classification. According to the multi-frame characteristic of sign language, adopted the scale-invariant feature transform (SIFT) algorithm to orient and obtain the key frames of sign language vocabularies, and extracted the feature vectors. Based on these key frames of sign language vocabularies, the adopted HDR method could narrow the search scope. Then used the DTW compare the irrecognition features of sign language vocabularies with every sign language word inside this scope, and the maximal calculate probability was the recognition result. With the same recognition rate, this method could be 8.2% faster than the HMM recognition method, and solved the problem that the template matching was suddenly slow down in the face of a large vocabulary.

Key words: sign language recognition; hierarchical discriminant regression; template matching

0 引言

手语是使用手的指势、动作、位置和朝向,配合面部表情、按照一定的语法规则来表达特定词意的交际工具[1]。手语的物质载体是手,通过手的形状、位置、运动来传递信息。手语识别是利用计算机对手语进行识别从而获得手语相应的文本、语音等的技术,进而达到帮助聋人与正常人进行正常交流的目的。手语识别的最终目标就是使计算机能像人那样通过对手语视觉信息[2]的处理来观察和理解[3]。

手语识别的尝试始于 20 世纪 80 年代末,根据手语输入设备的不同,分为数据手套和视觉两种方法。a)通过数据手套获取的手势空间运动轨迹和时序信息来识别手语。此类比较有代表性的是 Takahashi 和 Kishino 使用 VPL 数据手套识别 46 个日本手指字母[4],可正确识别出其中的 30 个[5];Wang Chunli 等人[6]开发出了大词汇量的中国手语识别系统,1 064 个孤立词的识别率为 90% 左右[7]。此方法的优点是采集到的数据可直接应用在训练和识别中,在小词汇量和大词汇量都取得了很好的效果。b)通过计算机视觉分析获取图像来识别手语。这一方法主要有 Charaphayan 和 Marble 使用图像处理方法来识别 31 个美国手语词,可以正确识别出 27 个;Starne 等人对 40 个词进行识别,识别率为 99.2%;香港中文大学 Deng 和 Tsui 识别 192 个美国手语词,识别率为 93.3%。此方法的优点是输入设备比较便宜,但识别率相对较低,实时性较差。但是,此方法是手语识别的发展趋势[8]。

目前基于视觉的手语识别方法有:a)神经网络方法(neural network)。人工神经网络的方法具有很强的分类效果和抗噪声能力,在静态的手势识别中被广泛应用。但该方法不具备描述信号时空变化的能力,所以在动态识别领域内一直没有成为主流的方法。b)统计识别方法(如隐马尔可夫模型(HMM))。该方法已经成功地应用在语音识别中,HMM 是众所周知并广泛使用的统计方法,它具有很强的描述动态时空变化的能力,在动态识别领域中一直占有主导地位。c)模板匹配方法(template matching)。该方法的抗噪声能力差,以及当词汇量增加时会造成模板在空间上的重叠,使得识别率快速下降。

收稿日期: 2009-03-26; **修回日期**: 2009-08-03 **基金项目**: 国家自然科学基金面上资助项目(60773051);杭州师范大学科研重点资助项目(2007XNZ10)

作者简介:姜华强(1978-),男,浙江淳安人,讲师,博士研究生,主要研究方向为模式识别、人机交互等(jhq@hznu.edu.cn);潘红(1967-),女,浙江杭州人,高级工程师,学士,主要研究方向为计算机教育、图形图像等.

近年来所开发的手语识别系统中,主要采用统计识别方法识别图像的本征特征匹配识别。这些方法识别率较好,但识别效率方面相对较差。基于提高识别效率的考虑,本文提出了一种具有基于关键帧的多级分类的手语识别方法,并利用该方法设计实现了一种快速的手语识别器。实验表明,这一方法在识别速度和精度上得到了较大的提高。

1　基于关键帧的多级分类手语识别方法

1.1　快速手语识别思想的提出

目前手语识别系统几乎都是在人工配合下完成手语词典的构建,并以全局检索的方法实现手语的识别。最典型的就是中科院自动化所研究的中国手语识别系统,它采用美国 Virtual Technologies 公司的 CyberGlove 型号数据手套获取手语信息,并通过 HMM 方法进行统计分析,产生手语特征数据,然后通过全局检索实现手语的识别[9, 10]。

手语识别最重要的依靠就是手语词典。在手语词典中,每个手语都有一个对应的手语序列,这个序列可以通过 HMM 提取模型。一个手语词的帧数在 20 ~ 70 帧不等,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。庞大的 HMM 手语数据在手语的快速识别上带来了很大的问题。根据对《中国手语》教材的分析可知,中国标准手语词汇的描述一般由 1 ~ 3 帧标准手语图像构成,如图 1 所示。因此,本文提出了大量剔除过渡帧,依靠关键帧的识别达到手语的快速识别。

为了快速识别手语,通过识别视频中的关键帧信息,并提取手语的骨骼结构本征特征,能够大量地减少手语匹配过程中的计算量。对手语词通过关键帧进行多级特征选择可以大幅地提高识别效率。

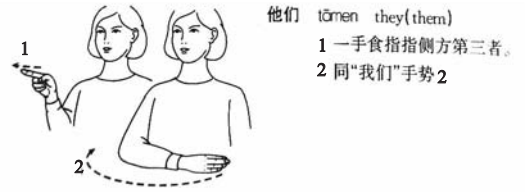

图1　中国标准手语词汇“他们”的描述图像

1.2　使用 SIFT 算法提取关键帧

SIFT 算法[11, 12]由 D. G. Lowe 提出,这是一种提取局部特征的算法,能够在尺度空间寻找极值点,提取位置、尺度、旋转不变量。这一算法提取的图像局部特征,其旋转、尺度缩放、亮度变化均具有保持不变性,对视角变化、仿射变换、噪声也保持一定程度的稳定性。该算法主要包含以下步骤:

a)建立尺度空间,寻找候选点;

b)精确确定关键点,剔除不稳定点;

c)确定关键点的方向;

d)提取特征描述符。

通过 SIFT 算法对每个关键点产生 128 个数据,即最终形成去除了尺度变化、旋转等几何变形因素影响的 128 维的 SIFT 特征向量 x_i。由手及手臂的抽象模型可知,一只手及手臂共有 18 个运动单元、18 个关节。每个关节有 1 个或多个自由度,一只手及手臂的运动由 27 个参数控制,所以一个手势的手形需要 54 个参数。因此,为每一幅手势图像(手语词根关键帧图像)提取 108 个关键点,形成该手语词根的向量空间 $X_t = [x_1, x_2, \cdots, x_{108}]$。在提取关键点的过程中,采用关键点特征向量的欧式距离来作为两幅图像中关键点的相似性判定度量。取手语图像前一帧中的某个关键点,并找出其与后一帧中欧式距离最近的前两个关键点。在这两个关键点中,如果最近的距离除以次近的距离少于某个比例阈值,则认为是过渡帧,予以排除;否则记录为关键帧。

1.3　应用 HDR 算法缩小搜索范围

HDR(hierarchical discriminant regression)方法[13, 14]描述了一种由机器人感知(sensor)系统到行为(action)系统的映射建立过程。其主要优点是能够更好地处理较高维数的输入向量,并且可以从每一个实例中学习新的知识。HDR 的主要思想是建立从输入空间到输出空间的映射,如图 2 所示。在学习的初级阶段,由于输入空间和输出空间的学习素材不是特别多,通过粗糙的聚类就可以构建起一些输入向量类到输出向量类的对应。随着学习的深入,越来越多的细节和输出情况被引入,这样就需要对原先的类分解,进行更精确的分类来构建对应关系。经过一段时间的学习,输入空间就会像树的结构一样,有些节点会分裂产生分支。而识别的过程就是一个在树中找到最相似的节点的检索过程,并输出对应的输出向量。HDR 方法在找到某一个节点时,就能够找到最为接近的输出响应。

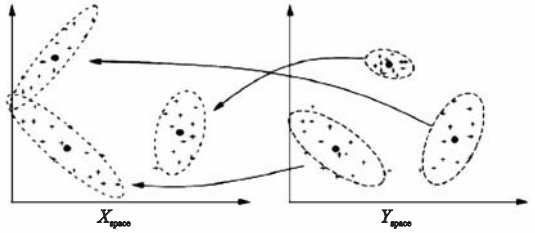

图2　输入空间(*X*)中的聚类和相对应的输出空间(*Y*)的聚类示意图

例如给定一幅手语图像,就能找到这幅图像与哪些以往学习过的图像最接近,每一个节点只需记录所代表类的概率分布信息,如类中心、方差等。学习则是通过更新树的结构和节点记录的信息来完成的。其算法如下:

输入:多层判别回归树 T 和样本输入空间向量 X_t($t = 1, 2, 3$),系数 k,检索敏感系数 ε。

输出:相应的输出向量 Y。

a)从根节点开始,计算节点中每个聚类到样本输入空间向量 X_t 的距离,挑选出到样本输入空间向量 X_t 距离最小的前 k 个输入空间聚类,把它们记为活动的输入空间聚类。设样本 X_t 到聚类 c 的距离最小,将其距离值与检索敏感系数 ε 比较,如果小于 ε,则检索结束,返回聚类 c 输出空间聚类平均向量 ym_c 作为样本 X_t 的输出向量 Y;否则,继续执行下一步。

b)对每个活动的输入空间聚类,按距离递增排序,且依次处理每个活动的输入空间聚类,如果它有子节点,记为非活动,并且搜索它的子节点。对于子节点,递归调用此过程,直到所有最终活动的输入空间聚类都没有子节点。在所有最终的活动聚类中,设聚类 c 到样本 X_t 的距离最短,输出聚类 c 的输出空间聚类的平均向量 ym_c 作为样本 X_t 的输出向量 Y。

1.4　通过 *DTW* 进行模板比较

DTW 算法[15, 16]的目的是在标准手语特征向量 Y 和非特定人手语特征向量 O 的特征之间找到一条优化的时间校准匹配路径。设 Y 是一个将测试信号的样点映射到参考信号的弯曲函数式中:

$$Y = y(1), y(2), \cdots, y(k), \cdots, y(K)$$
$$y(k) = (i(k), j(k))$$

其中:i 和 j 分别代表参考信号 Y(总长 I 点)和测试信号 O(总长 J 点)的能量特征点,表示在作 k 次特征匹配时,Y 第 i 点与 O 第 j 点比较。弯曲函数的限制条件为

a)单调性:$i(k-1) \leqslant i(k), j(k-1) \leqslant j(k)$

b)连续性:$i(k) - i(k-1) \leqslant 1, j(k) - j(k-1) \leqslant 1$

c)边界:$i(1) = 1, j(1) = 1, i(K) = I, j(K) = J$

d)窗:$|i(k) - j(k)| \leqslant r$。其中 r 是允许窗的长度。

限于篇幅,这里不作详细的讨论,定义 $D(p_k) = d((i(k), j(k))) = \| Y_i - O_j \|$。DTW 算法的实质是寻找匹配路径 P 使 Y 和 O 总距离最小,即 $D(P) = \min_P \sum_{k=1}^{K} d(p_k)$。从 DTW 角度说,匹配路径 P 是 Y 和 O 的最优时间匹配。

2　基于关键帧的多级分类手语识别器的设计与实现

本文的手语识别系统基于关键帧的多级分类手语识别系统。其中手语词典的数据输入是本文在 1.1 节中提到《中国手语》的标准手语图像,非特定人的手语数据是通过一个正面的摄像头采集到的手语视频。整个手语识别系统工作分为两个子模块,即手语词典生成过程、非特定人手语识别过程。

手语识别器的结构图如图 3 所示。

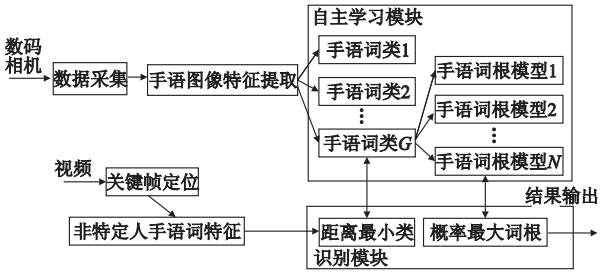

图3　手语识别器结构图

2.1　手语词典构造过程

手语词典生成过程可以分成两部分:a)针对一幅标准手语图像,运用 SIFT 算法提取出 128 维特征向量,用 HDR 分类方法(具体步骤见 1.3 节)对手语词样本集合进行分类,并建立每个类的词汇表;b)根据对各个类中的每一个词根建立一个手语骨骼模型。模型训练时使用 SIFT 算法对标准人体骨骼框架和标准手语轮廓图像进行匹配,取得手势的每个关节点,最后形成 108 维的特征空间向量。

2.2　识别过程

训练过程可以分成两部分:a)采用 SIFT 算法识别手语的关键帧,提取特征向量;b)用 HDR 方法(具体步骤见 1.3 节),对手语词样本集合进行分类,并建立每个类的词汇表; c)对各个手语词类根据关键帧的特征向量进行 DTW 算法匹配。

3　实验结果与结论

实验中使用了 20 个汉语手语词汇,并加入阿拉伯数字 10 个手势和 26 个字母手势,共 56 个手语;共采集两套标准手语图像库、两套手语视频,采用标准手语图片库作为学习样本。两套手语视频作为测试,识别率达到 85%,单个词汇的识别速度在 1.3 s 左右,比 HMM 识别方法提高了 0.1 s,速率提高了约 8.2%。如表 1 所示。

表 1　手语识别结果

项目	识别正确	识别错误
数量	48	8
百分率/%	85.71	14.29

经分析,个别手势不能识别的原因在于手势较为雷同以及关键帧的捕获存在着误差。本文实验在 PC(DELL-Pentium 4,256 MB 内存)上进行。

4　结束语

本文的创新点在于:提出了基于关键帧的多级分类的手语识别方法,在识别率基本不变的情况下较好地提高了识别器的识别速度。在今后的研究工作中,要继续探索如何把具有多级分类的识别方法应用到连续语句识别当中,这样才能够让此方法在手语识别中产生更加重要的作用。

参考文献:

[1] VALLI C, LUCAS C, MULROONEY K J. Linguistics of american sign language:an introduction[M]. 4th ed. Washington DC:Gallaudet University Press, 2005.

[2] ULRYCH J, KOPECKY M. Visual similarity in sign language[C]// Proc of the 24th International Conference on Data Engineering. 2008:53-60.

[3] 胡友树. 手势识别技术综述[J]. 中国科技信息, 2005(2):42.

[4] AOKI Y, TANAHASHI S, SUGIYAMA H. Tracing of arm motion by matching video images with 3D arm model for intelligent communication of sign language[C]//Proc of the 3rd IEEE International Conference on Electronics, Circuits, and Systems. 1996:53-56.

[5] SUGIYAMA H, TANAHASHI S, AOKI Y. Recovering three dimensional hand motions of sign language from monocular image sequence [C]//Proc of the 1st International Conference on Information, Commuincations, and Signal Processing. 1997:1098-1101.

[6] WANG Chun-li, CHEN Xi-lin, GAO Wen. A comparison between etymon- and word-based chinese sign language recognition systems [C]//Proc of the 6th International Gesture Workshop. 2006:84-87.

[7] ZHOU Yu, GAO Wen, CHEN Xi-lin, *et al*. Signer adaptation based on etyma for large vocabulary Chinese sign language recognition [C]//Proc of the 8th Pacific-Rim Conference on Multimedia. 2007:458-461.

[8] Von AGRIS U, ZIEREN J, CANZLER U, *et al*. Recent developments in visual sign language recognition[J]. Universal Access in the Information Society, 2008, 6(4): 323-362.

[9] 吴江琴, 高文. 基于 DGMM 的中国手语识别系统[J]. 计算机研究与发展, 2000,37(5):551-557.

[10] 张良国, 高文, 陈熙霖, 等. 面向中等词汇量的中国手语视觉识别系统[J]. 计算机研究与发展, 2006,43(3):476-482.

[11] STRAKER D. The SIFT model[J]. Quality World, 2003, 29(5): 45-46.

[12] LOWE D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[13] WENG Ju-yang, HWANG W S. Incremental hierarchical discriminant regression[J]. IEEE Trans on Neural Networks, 2007, 18(2): 397-415.

[14] 王增进,危辉. 改进的多层判别回归树算法及其在遥感图像分析中的应用[J]. 计算机学报, 2004, 27(1): 92-98.

[15] KAR B, DUTTA P K, BASU T K, *et al*. DTW based verification scheme of biometric signatures[C]//Proc of IEEE International Conference on Industrial Technology. 2006:381-386.

[16] FANG P, WU Z C, SHEN F, *et al*. Improved DTW algorithm for online signature verification based on writing forces[C]//Proc of International Conference on Intelligent Computing. 2005:631-640.