ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.9, September 2008, pp.2302−2312
DOI: 10.3724/SP.J.1001.2008.02302

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

基于 HMM-FNN 模型的复杂动态手势识别*

王西颖[1,2+], 戴国忠[1], 张习文[1], 张凤军[1]

[1](中国科学院 软件研究所 人机交互技术与智能信息处理实验室,北京 100190)

[2](中国科学院 研究生院,北京 100049)

Recognition of Complex Dynamic Gesture Based on HMM-FNN Model

WANG Xi-Ying[1,2+], DAI Guo-Zhong[1], ZHANG Xi-Wen[1], ZHANG Feng-Jun[1]

[1](Laboratory of Human-Computer Interaction and Intelligent Information Processing, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: xiying04@ios.cn

Wang XY, Dai GZ, Zhang XW, Zhang FJ. Recognition of complex dynamic gesture based on HMM-FNN model. *Journal of Software*, 2008,19(9):2302−2312. http://www.jos.org.cn/1000-9825/19/2302.htm

Abstract: Recognition of complex dynamic gesture is a key issue for visual gesture-based human-computer interaction. In this paper, an HMM-FNN model is proposed for gesture recognition, which combines ability of HMM model for temporal data modeling with that of fuzzy neural network for fuzzy rule modeling and fuzzy inference. Complex dynamic gesture has two important properties: Its motion can be decomposed and usually being defined in a fuzzy way. By HMM-FNN, complex gesture is firstly decomposed into three components: Posture changing, movement in 2D plane and movement in Z-axis direction, each of which is modeled by HMM. The likelihood of each HMM to observation sequence is considered as membership value of FNN, and gesture is classified through fuzzy inference of FNN. In this proposed method, high-dimensional gesture feature is transformed into several low-dimensional features, as a result, computational complexity is reduced. Furthermore, human's experience or prior knowledge can be used to build and optimize model structure. Experimental results show that the proposed method is an effective method for recognition of complex dynamic gesture, and is superior to conventional HMM method.

Key words: gesture recognition; HMM-FNN model; complex dynamic gesture; human-computer interaction

摘 要: 复杂动态手势识别是利用视频手势进行人机交互的关键问题.提出一种 HMM-FNN 模型结构.它整合了隐马尔可夫模型对时序数据的建模能力与模糊神经网络的模糊规则构建与推理能力,并将其应用到复杂动态手势的识别中.复杂动态手势具备两大特点:运动特征的可分解性与定义描述的模糊性.针对这两种特性,复杂手势被分解为手形变化、2D 平面运动与 Z 轴方向运动 3 个子部分,分别利用 HMM 进行建模,HMM 模型对观察子序列的似

* Supported by the National Basic Research Program of China under Grant No.2002CB312103 (国家重点基础研究发展计划(973)); the National Natural Science Foundation of China under Grant Nos.60673188, 60605018 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z328 (国家高技术研究发展计划(863))

Received 2006-11-14; Accepted 2007-04-18

然概率被作为 FNN 的模糊隶属度,通过模糊规则推理,最终得到手势的分类类别.HMM-FNN 方法将高维手势特征分解为低维子特征序列,降低了模型的复杂度.此外,它还可以充分利用人的经验辅助模型结构的创建与优化.实验表明,该方法是一种有效的复杂动态手势识别方法,并且优于传统的 HMM 模型方法.

关键词: 手势识别;HMM-FNN 模型;复杂动态手势;人机交互

中图法分类号: TP391 **文献标识码:** A

近年来,基于视觉手势的人机交互界面正在成为计算机视觉与人机交互领域研究的热点[1−4].通过摄像机捕获用户的手势动作,并由计算机进行分析理解,然后完成交互任务,这种方式使人与计算机之间的信息传递方式更加符合人自身的习惯,摆脱了传统交互方式对人的束缚,从而使交互的舒适性与效率得到提高.特别是在 3D 虚拟现实环境下,传统上利用数据手套的方式给用户带来很大的约束,并且设备价格昂贵,很难为一般用户所接受.徒手手势作为一种最自然的 3D 输入方式,可以极大地改善交互的效果并且易于推广.视频手势的分析与理解是基于视频手势交互的关键问题,特别是复杂的动态手势识别问题仍然是亟需解决的难点之一.

根据不同的角度,手势的种类可以有多种不同的分类方法,例如,从手势动作的目的性出发,可以分为操作类(manipulative)与通信类(communicative)手势[1];从手势的运动特点出发,又可以分为静态手势(posture,也称手形)与动态手势.静态手势是仅依靠手的外部形状与轮廓传递信息的方式,可以被视为动态手势的特例;动态手势是指手的形状与位置都随时间发生变化的手势,它可以表达更加丰富和准确的信息,也是人们日常生活中最为常用的交流方式之一.动态手势识别的早期研究工作主要集中在低分辨率的简单手势动作上,随着交互应用的发展和对交互手势精度要求的提高,高分辨率的复杂手势识别成为亟需解决的问题.高分辨率的复杂手势与简单手势的明显区别在于:简单手势将手作为一个具有面积属性的点进行处理,它的识别只需要考虑在某个时间段内这个点所经过的轨迹以及它与身体其他部位(例如人脸)的相对位置关系;复杂手势除了具备简单手势的全部特点外,还必须考虑到各个手指的变化情况(例如手指的弯曲与伸展)、手指与手指之间的相对关系(例如手指之间的距离变化)、手指与手掌的相对关系以及手掌的变化情况等.

本文的研究对象是空间位置不断变化、手形外观不断变化而且在 3D 空间进行运动的复杂动态手势.复杂动态手势在运动规律上具备如下 3 个明显特点:(1) 时间上的可变性.动态手势的运动速度是不定的,对同一个手势,不同的人可能会用不同的速度来完成;即使是同一个人,每次的完成速度也不尽相同;(2) 空间上的可变性.手势的运动空间与活动范围是不同的,不同的人完成同样手势的幅度总是存在差异;(3) 手势的完整性是可变的.很多情况下,操作者的手势动作与系统预定义的手势相比是不完整或存在冗余的.此外,经过对复杂动态手势进行细致地分析我们发现,它还具备如下两种特性:(1) 动态手势可以看作是由手形(posture)变化、整体手的 2D 平面运动与 Z 轴方向运动 3 部分组合构成,而且这 3 部分可以看作是相互独立的运动过程;(2) 动态手势在定义和描述上存在模糊性,而且它的 3 个构成部分都具有模糊的语义特性.通常情况下,人们在对动态手势进行描述时,很难用确定性的语言描述一个手势的运动变化情况,而往往采用模糊性的语言进行手势的定义.例如,一个将手掌从完全伸开逐渐变化到握紧,同时,手的运动轨迹像一个圆形的手势,是一个表示全部选中的操作手势,这就启发我们考虑用模糊逻辑与模糊规则对手势进行描述与识别.

本文在对手掌与指尖视频跟踪的工作基础上,提出一种基于 HMM-FNN 模型的单目视觉条件下动态视频手势建模和识别的方法.它充分利用了复杂动态手势的自身特点,即复杂手势具备了运动特征可分解性以及类别定义的模糊性特点.HMM-FNN 模型充分整合了隐马尔可夫模型(hidden Markov model)对时序数据建模的能力,以及模糊神经网络(fuzzy neural network,简称 FNN)的模糊规则建模与模糊推理能力.动态手势的 3 个独立运动分量可以通过 HMM 模型分别进行描述,然后通过 FNN 的模糊规则网络将这 3 个独立部分组合成一个完整的动态手势模型,利用模糊规则和推理实现手势的分类与识别.本文方法将高维的动态手势特征分解为 3 个低维的特征向量序列分别进行识别,从而避免了高维特征的计算和推理;同时,还可以充分利用人的先验知识和经验,辅助手势的建模与识别.

本文首先对手势识别的相关工作进行介绍,分析现有方法的特点.第 2 节介绍在单目视觉条件下,动态视频

手势特征的提取方法.第 3 节给出一种新的 HMM-FNN 动态手势模型,并介绍该模型的训练过程与分类方法.在实验及分析部分给出实验结果,并与传统的基于 HMM 模型进行手势识别的方法进行比较.最后进行总结.

1 相关工作

20 世纪 90 年代,Starner[5]首先利用 HMM 模型进行美国手语(ASL)的识别研究,他利用 4 个特征值构成单手手势的特征向量,只考虑了手的 *XY* 坐标、最小惯量的角度以及外包椭圆的离心率.Rigoll[6]的方法是首先提取图像帧间的差值图像(difference image),然后用差值图像的七维向量作为动态手势的特征表示,通过 HMM 模型对一组手势进行了识别.由于手势特征的提取只是利用了运动差值图像,所以 Rigoll 的方法无法准确获得手部形态的完整信息,对于高分辨率的手势识别无法完全适用.Bobick 和 Wilson[7]提出一种基于状态的手势表示与识别方法,该方法将手势描述为在手势特征空间的一条轨迹,然后将轨迹曲线划分为不同的状态,从而将一种手势表示成为一组连续状态的序列.与 HMM 不同,这种基于状态的方法并不是通过训练学习来而得到手势模型的参数,它只是通过对一组原型数据建模后,将待识别的手势图像序列与原型进行匹配,得到它们的匹配度(match score)作为识别的依据.任海兵与祝远新等人[8,9]利用手势运动的表观信息建立动态手势的时空表观模型,提出综合利用颜色、运动与形状等信息的融合策略抽取模型参数,并利用动态时空归整(DTW)的方法对手势进行识别.这种方法是一种完全基于图像的底层特征信息对动态手势建模与识别的方法,无法应用高层语义规则进行辅助识别.HyeSun Park[10]利用双手区域与人脸之间的位置关系为手势建立六维的特征向量,通过一个整合的 HMM 模型实现了 13 种双手手势的识别,而不是传统方法中为每一种手势分别建立 HMM 模型.这种方法的优点是有可能利用手势与手势之间的相关关系改善识别的结果,但同时,由于其结构过于复杂,训练和识别的效率势必受到影响.

神经网络的方法通常被用来对静态手势进行识别[11,12],特征的采集是通过数据手套的方式而非视觉的方式进行.也有研究者尝试将神经网络用于动态手势识别[13],但在时序建模方面,其能力有限,而且训练计算量非常大,难以成为动态手势识别的主流方法.然而,由于神经网络作为估计器的性能比传统的统计方法要强,研究者常常通过训练神经网络产生 HMM 模型中的输出后验概率,也就是将 HMM 模型与神经网络进行结合,利用神经网络或多层感知器(multi-layer perceptron,简称 MLP)进行后验概率的估计.Cohen[14]等人提出将 MLP 与 HMM 结合进行概率估计,从而可以减少关于观察向量中特征之间相互独立的假设条件的局限性.

简单的动态手势通过低层次(low-level)的动态模型就可以完成,例如 HMM 模型等;而对于复杂的动态手势,仅依靠低层次模型则难以清楚表达它的含义,必须依靠某些高层次(high-level)模型,基于规则的方法[15]就是其中的一种.通过定义规则,可以表达手势的高层次语义信息.模糊神经网络也是定义规则的一种方式,它定义的规则是一系列的模糊规则,通过规则条件在网络中的传播可以自动进行模糊规则的推理.

2 动态手势特征定义与提取

手势建模建立在手势特征提取的基础上.复杂动态手势特征是一种高维度特征,为了避免计算量过大的问题,我们首先根据其特点,对复杂手势进行分解.3D 动态手势在时间-空间上的运动变化可分解为投影平面内的图像特征变化与 *Z* 轴方向的运动变化,进一步可细分为手形的变化、整体手的 2D 平面位置变化以及 *Z* 轴方向的运动变化 3 大部分.为降低问题的复杂度,本文中的 3D 手势动作在垂直于投影平面 *Z* 轴方向的运动分量由手掌面积 S_p 的大小来表示.当 S_p 逐渐增大时,表示手在向摄像机靠近运动;反之,则表示远离摄像机的运动.

特征的定义与分类器的选择密切相关.HMM 模型具备较强的时序建模能力,但高维的观察值向量对模型参数学习和模型评估,都会带来较大的计算负担.为了提高系统的实时性能,需要对观察值的特征维度进行降维处理.此外,HMM 模型又分为离散与连续两种:离散模型针对较为简单的离散观察值序列,但对于连续型的时间序列,则需要进行矢量量化,这就会导致部分信息的丢失;连续 HMM 模型可以直接应用到连续的观察值序列,从而保留了更多的原始信息.本文中对于动态手势的手形序列和 *Z* 轴方向运动采用一维连续型 HMM 进行建模,手势的平面轨迹模型则采用一维离散型 HMM.

手形序列中的手形特征是一个一维变量,它需要既能够表示出当前手形的基本类别,又能够提供手形的基本参数特征.本文的手形特征值计算建立在我们前期工作的基础上.在前期工作中,我们利用了一种模糊集运算的方法实现了手区域的图像分割,它综合利用了人手的肤色信息与运动信息,实现了整个手区域的完整分割[16].此外,还实现了对变形手势的连续跟踪[17].针对连续运动并且外形不断变化的手势,应用 Camshift 算法对手的整体进行跟踪,利用改进的粒子滤波(particle filter)算法对多个手指的指尖进行连续跟踪,同时,利用跟踪检测技术可以实时地发现手形的变化,动态地调整跟踪模板.手势跟踪的效果如图 1 所示,其中,大圆标识出手掌的位置,两个小圆标识出手指指尖的位置.本文在前面工作的基础上,提出了一种基于跟踪结果的一维表示方法,作为手形状态的特征矢量.

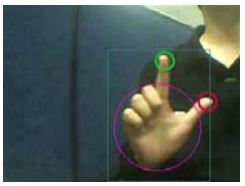


Fig.1 Gesture tracking

图 1 手势跟踪效果图

二维图像中,手形的变化可以简化为手指长度的改变以及手指之间距离的改变,所以,手形的特征就可以通过各个手指的长度以及手指之间的距离进行表示.如图 2(a)所示,设大拇指、食指、中指、无名指和小指与手区域重心连线的长度分别为 $Len(0)$~$Len(4)$,设相邻手指之间的距离分别为 $Dis(0)$~$Dis(4)$,其中,$Dis(0)$表示大拇指与食指指尖的距离,$Dis(1)$表示食指与中指指尖的距离,依此类推,$Dis(4)$表示小指与大拇指的距离.当手势图像中存在手指遮挡情况时,被遮挡手指的长度设定为 0,相邻指尖的距离则成为可见的手指指尖间的距离.通过 $Len(0)$~$Len(4)$,$Dis(0)$~$Dis(4)$可以大致表示出手形的基本特征.目前,这种手形特征已经具备了平移与旋转不变性,为了使其具备尺度缩放的不变性,我们对其进行如下的归一化处理:跟踪过程中得到当前帧中手区域的外包圆直径 D_h,将上面的 10 个特征数值除以外包圆直径 D_h,得到的新特征数值即为归一化的特征数值,从而具备了尺度缩放不变性.为了能够较为明显地区分不同类别手形的特征值(手形的类别是通过跟踪阶段的静态手势识别得到的),我们为不同类别手形的特征值 V 设定了不同的基数数值 V_{base},从而可以使不同手形的 V 值分布在不同的数值区间段.

此外,我们引入了偏心率(eccentricity)作为 V 计算公式的组成部分.偏心率也称为伸长度,在一定程度上,描述了区域的紧凑性.这样,V 既体现了手形中手指位置的形状特征,又体现了整个手区域的统计特征.根据上面的描述,手形特征值 V 可以按照公式(1)的方法计算得到:

$$V=\alpha[V_{base}+100\times\sum_{i=0}^{4}(Len(i)/D_h)+100\times\sum_{j=0}^{n}(Dis(j)/D_h)]+(100-\alpha)E \tag{1}$$

其中:$\alpha(0\le\alpha\le100)$为比例常数;n 为手势图像中可见的手指个数;E 为当前手形区域的偏心率,其计算由公式(2)给出:

$$E=\sqrt{\frac{(A+B)-\sqrt{(A-B)^2+4H^2}}{(A+B)+\sqrt{(A-B)^2+4H^2}}},\ A=\sum m_i(y_i^2+z_i^2),B=\sum m_i(z_i^2+x_i^2),H=\sum m_i x_i y_i \tag{2}$$

其中,A,B 分别为刚体绕 X,Y 轴的转动惯量,H 称为惯性积,E 的值不受平移、旋转和尺度变换的影响[18].

对于不同的手形,由于 V_{base} 的不同,V 值被明显划分到不同值域区间;对于相同的手形,由于手指伸出长度以及指间距离的不同,V 值亦会有明显区别.因此,V 值基本可以明确表示手形的状态特征.

手势的 2D 平面运动轨迹是由一系列的位置点组成,其特征值可以由一组离散的运动方向值进行表示.对于二维运动空间,我们将其平均划分为图 2(b)所示的 8 种运动方向.对于某一时间段内的手势运动轨迹点,提取任意两个相邻点 p_1,p_2 的位置坐标(相邻点的间距应大于设定的阈值),计算两点间连线 $\overrightarrow{p_1p_2}$ 的方向角$\beta(0\le\beta<2\pi)$,然后按照图 2(b)中对角度β所属方向的划分,得到 p_1 到 p_2 运动方向的离散值.类似地,可以得到全体运动轨迹的方向序列 R:$r_1,r_2,\ldots,r_i,\ldots,r_T$ $(0\le i\le7)$.

为了获取手势运动在垂直于成像平面的 Z 轴方向的运动信息,我们采用了通过手掌面积变化获得 Z 轴运动方向的方法.手掌面积的特征值也需要进行归一化处理,归一化的方法是将各个时刻的手掌面积与初始时刻(t=0)的手掌面积 S_0 进行比较,最终得到一组连续的归一化面积数值 S:$1,s_1,\ldots,s_i$.

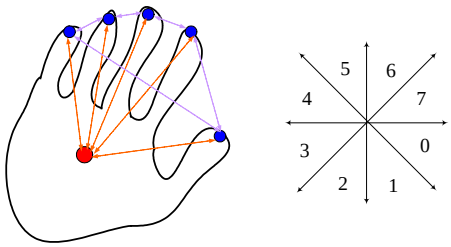


Fig.2 Calculation of gesture feature value

图 2 手势特征的计算

3 手势建模及分类器设计

本节中,我们提出一种新的动态手势建模方法——基于 HMM-FNN 的动态手势建模方法,并对 HMM-FNN 模型的训练与分类过程进行了详细描述.普通的 FNN 模型是进行模糊规则建模与模糊推理的工具,但它只能处理离散的特征向量,对带有时序关系的连续输入则不适用.HMM 模型是常用的时序数据建模方法,将 HMM 模型引入 FNN 中来,则可以解决传统 FNN 无法处理时序数据的问题.本文提出的 HMM-FNN 模型是一种复合模型结构,主要针对具有结构性特征的复杂时序数据进行建模,它同时具备了 HMM 模型时序数据建模的能力以及 FNN 模型模糊规则建模与推理能力,并与复杂动态手势的运动特点相吻合.

3.1 HMM-FNN模型

HMM-FNN 模型是将 HMM 模型与 FNN 模型进行整合的结果.由于 HMM 模型具备时序建模与推理能力,它被用来对手势的 3 个时序分量进行建模.模糊神经网络是近年来的研究热点,它将模糊逻辑推理与人工神经网络结合为一体,充分利用了模糊逻辑的知识表达与推理能力以及神经网络强大的自学习能力.普通的模糊神经网络不具备时序建模的能力,我们将 HMM 模型的时序建模能力与模糊神经网络相整合,为动态手势建立模型,其结构图如图 3 所示.

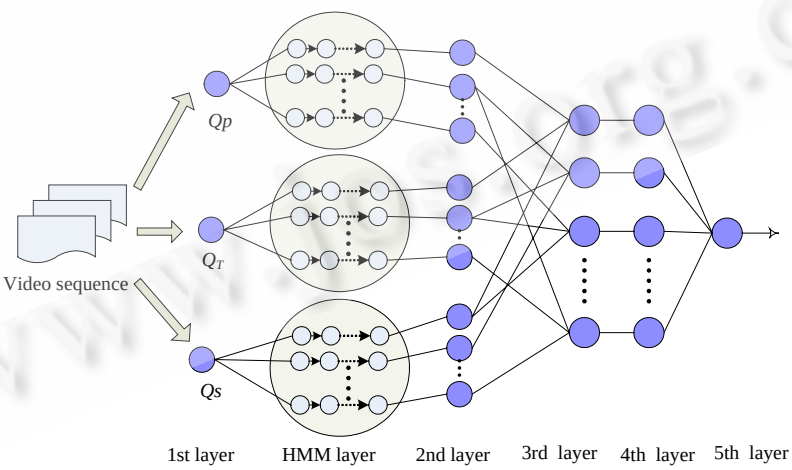


Fig.3 HMM-FNN model of dynamic gesture

图 3 HMM-FNN 动态手势模型

在模糊系统中,模糊化→模糊推理→模糊判决是构成模糊系统的最基本模块,将模糊系统表达成联结主义形式的网络结构,就得到了模糊神经网络.普通的模糊神经网络包括 3 个主要部分:模糊预处理部分、模糊推理部分与输出部分,其中的模糊预处理部分往往是通过隶属度函数得到的输入变量对应于各个模糊集的模糊隶

属度.与普通 FNN 不同,本文中利用 HMM 模型关于观察值序列的似然概率作为模糊隶属度,从而完成模糊预处理.如图 3 所示,HMM-FNN 模型包括 6 层结构,其中第 1 层、第 2 层以及中间的 HMM 层(HMM layer)属于模糊预处理部分;第 3 层与第 4 层属于模糊推理部分;第 5 层属于去模糊化的输出部分.下面对各个层次分别进行说明.

第 1 层(The 1st layer)是输入层.动态手势被分解为 3 个部分:手形变化序列 Q_P、手的 2D 轨迹序列 Q_T 和手掌面积变化序列 Q_S,作为 3 个独立的观察值序列输入 HMM-FNN 网络,与输入层的 3 个节点相对应.

第 2 层(The 2nd layer)以及 HMM 层构成了模糊化层.第 1 层的 Q_P,Q_T 与 Q_S 观察值序列分别与 HMM 层的若干 HMM 模型相关联,表示它们可能的模糊子类别.例如,Q_T 序列与 3 个 HMM 模型关联,表示可能的 3 种运动轨迹类别:类似圆形、类似三角形与类似方形的运动轨迹.根据输入的观察值序列,可以对各个 HMM 模型进行可能性评估,并将评估结果输入第 2 层节点.第 2 层的每个节点与一个 HMM 模型相对应,代表了一个具有模糊语义的子类别分量.输入的观察值序列 Q 关于各个 HMM 模型具备似然概率 $p(Q/\lambda)$,它同时也代表了输入观察值属于该语义变量的模糊隶属度.第 2 层节点构成了模糊规则的前件部分,节点的总数为 $m_1+m_2+m_3$,其中,m_1 为 Q_P 的类别个数, m_2 为 Q_T 的类别个数,m_3 为 Q_S 的类别个数.

第 3 层(The 3rd layer)为模糊规则层或模糊推理层,每个节点代表一条模糊规则.模糊规则的前件由该节点的输入决定,模糊推理方法采用了 Sum-Product 推理方法,得到节点的输出值.一个模糊规则的例子如下:若手形的变化是一个由五指伸开到握拳的过程(规则前件 C_1),手的运动轨迹是一个近似圆形(规则前件 C_2),而且手掌的面积基本保持不变(规则前件 C_3),则这个运动手势的类型为 A,即 $C_1\wedge C_2\wedge C_3\rightarrow A$.第 3 层节点个数即模糊规则的个数.第 3 层与第 2 层之间的连接是带有权重(weight)的连接,由于第 2 层节点代表着模糊规则的前件,连接权重值则代表了各个规则前件对推理结果的贡献度.第 3 层模糊规则节点 j 的输出计算如公式(3)所示,其中,m 为节点 j 的输入值个数,也就是该规则的前件个数:

$$O_j^{(3)}=\sum_{i=0}^{m}\omega_{ij}I_{ij}^{(3)}=\sum_{i=0}^{m}\omega_{ij}p(Q/\lambda_i),\ \sum\nolimits_i\omega_{ij}=1 \tag{3}$$

第 4 层(The 4th layer)为归一化层,其节点个数与第 3 层节点个数一致,目的是计算规则的归一化激励强度,避免在学习过程中由于各修正量过大而产生震荡,影响收敛速度.通过归一化计算,全部规则的输出值之和为 1,其计算如公式(4)所示:

$$O_j^{(4)}=I_j^{(4)}\Big/\sum_{j=0}^{N}I_j^{(4)}=O_j^{(3)}\Big/\sum_{j=0}^{N}O_j^{(3)} \tag{4}$$

第 5 层(The 5th layer)为输出层或去模糊化层,实现的是清晰化计算,并给出了系统的最终判别结果.第 5 层的输出结果是各条规则输出的加权代数和,如公式(5)所示:

$$O^{(5)}=\sum_{j=1}^{N}\omega_jO_j^{(4)},\ \sum_{j=1}^{N}\omega_j=1 \tag{5}$$

其中,ω_j 相当于第 j 条规则对分类结果的贡献率权重,N 为规则个数.

HMM-FNN 模型中的 HMM 层所采用的隐马尔可夫模型如图 4 所示.HMM 模型的类型选择并没有统一的规则方法,通常是根据各个应用系统的特点决定的.考虑到手势运动的特点,这里的 HMM 模型采用直观的 left-right 形式的 Bakis 模型[19],Bakis 模型的拓扑结构简单、参数较少,是目前研究较多且成熟的类型.

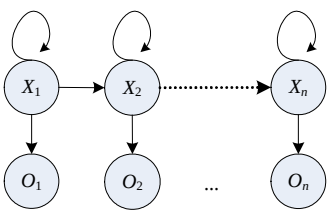


Fig.4 HMM model

图 4 HMM 模型

可以看到,HMM-FNN 中的 HMM 模型主要针对 3 种动态时序过程建模,它们分别描述了动态手势运动过程中手形的变化、运动轨迹与手掌面积的变化过程.上一节中对这 3 种变化的特征定义方法进行了说明.与特征类型相对应,手形变化是通过一维连续 HMM 模型进行建模,平面运动轨迹是通过一维离散 HMM 模型进行建模,而 Z 轴方向的运动则是通过手掌面积变化的一维连续 HMM 模型进行建模.

对于连续 HMM 模型,观察特征的输出概率密度采用高斯混合模型(Gaussian mixture model,简称 GMM)的形式.高斯混合模型是一种具备多个高斯模型分量的混合模型,设观察矢量为 O,则 O 在 GMM 模型下的似然度可表示为

$$p(O/\lambda)=\sum_{i=1}^{M}\omega_i g_i(x) \tag{6}$$

其中,ω_i 为第 i 个分量的权重,$g_i(x)$为一维的高斯密度函数 $g_i(x)\sim N(\mu,\sigma^2)$.

HMM 模型参数的选择是关系到 HMM 模型工作效果的关键问题.由于各种手势的运动特点不同,HMM 模型中的状态个数也不同.手掌面积变化情况较为简单,我们设定其 HMM 模型状态个数统一为 3;手形变化与运动轨迹变化情况较为复杂,其各个类别对应的 HMM 模型的状态个数则是通过关键帧提取与人为经验共同确定的.关键帧提取就是对作为训练数据的视频片断选取其中最不相关的若干帧,这个过程主要涉及手势视频中帧间特征差异度的计算.在提取手形变化关键帧时,与传统方法不同,我们关心的只是手的形态变化而不是整个场景的变化情况.首先,第 1 帧被定为关键帧和当前参考帧,对其后的各帧手势进行跟踪,并计算各帧中手势形状与参考帧手势的差值.当差值大于设定阈值时,当前帧被选定为关键帧,并作为参考帧继续对后续帧进行比较,直至结束.轨迹 HMM 模型的状态数是根据轨迹的曲率变化得到的,从训练轨迹的第 1 个位置点开始按固定长度,例如轨迹总长的 1/5,扫描整个轨迹路线,曲率变化大于设定阈值的点的个数即为该轨迹 HMM 模型的状态数.利用经验调整阈值的大小,可以将最终状态数控制在合适的范围之内.

3.2 模型的训练

HMM-FNN 模型的训练分为两个过程:首先是 HMM 模型的参数训练,也就是调整各个 HMM 模型的状态转移概率矩阵、离散观察值的输出概率矩阵或连续观察值的高斯混合模型参数等.然后是模糊神经网络的参数学习,即连接权重的调整以及对多余连接的剪枝.HMM 模型的训练采用的是 Baum-Welch 算法[19],也称为 Forward-Backward 算法.它是一种基于期望最大化(expectation maximum)的算法,根据最大似然准则(maximum likelihood,简称 ML)对模型参数进行迭代的重新估计,直至得到最大似然意义下的最优模型参数值.

由于与传统 FNN 不同的结构特点,本文学习算法中关于模糊隶属度函数参数的训练部分被省略掉,HMM 模型对输入特征序列的似然概率 $p(Q/\lambda)$被作为输入样本对该类别的隶属度.模型中,FNN 参数的学习主要涉及两部分:公式(3)中各规则前件的权重以及公式(5)中各规则的贡献率.FNN 的学习采用的是反向误差传播算法,即 BP(back-propagation)算法.它是一种典型的无反馈的多层前向神经网络的学习算法,其实质是求误差函数最小值的问题,通过采用最快梯度下降法,按误差函数的负梯度方向修改权系数.设整个网络模型的学习误差函数 J_e 的定义为 $J_e=\frac{1}{2}(t-z)^T(t-z)$,其中,矢量 t 表示网络的期望输出,矢量 z 表示网络的实际输出.由 BP 算法可得到第 5 层和第 3 层的权值修正量 $\Delta\omega_k^{(5)}$,$\Delta\omega_{ji}^{(3)}$ 如公式(7)所示:

$$\begin{cases}\Delta\omega_k^{(5)}=-\eta\delta^{(5)}\bullet O^{(5)}=-\eta(t-z)I_k^{(5)}\\ \Delta\omega_{ij}^{(3)}=-\eta I_i^{(3)}\delta_j^{(3)}=-\eta(\omega_j^{(5)}\delta^{(5)})I_i^{(3)}=-\eta(\omega_j^{(5)}\bullet(t-z))I_i^{(3)}\end{cases} \tag{7}$$

其中,t 为第 5 层节点的期望输出值,z 为它的实际输出,η为学习率. $O_i^{(k)}$ 表示第 k 层第 i 个节点的输出值, $I_i^{(k)}$ 表示第 k 层第 i 个节点的输入值.更新后的权值按照 $w(T+1)=w(T)+\Delta w$ 的方法进行迭代,当网络的参数学习实现收敛或迭代次数达到最大时,训练过程即可结束.

3.3 手势识别

识别过程是建立在对手和手指跟踪的基础上,首先将一组视频序列 Q 分解为相应的平面轨迹点序列 Q_T,手形变化序列 Q_P 和手掌面积变化序列 Q_S,其中,Q_P 序列与 Q_S 序列为连续值序列,直接输入对应的连续型 HMM 模型中进行分类评估,Q_T 序列则需要根据上一节中的方法进行运动方向的量化,将离散化后的方向特征序列输入对应的离散 HMM 模型进行分类评估.在将上述 3 组特征序列分别输入到 HMM-FNN 模型中对应的 HMM 组后,按前向递推算法计算各 HMM 模型的 $p(Q/\lambda)$,步骤如下:

设转移概率$\alpha_{ij}=p(s_j/s_i)$,观察概率 $b_j(k)=p(o_k/s_j)$,设 $a_t(i)=p(o_1,o_2,\ldots,o_t;q_t=s_i/\lambda)$

1) 初始化 $a_1(i)=\pi_i\times b_i(o_1)$,$\pi_i$ 为第 i 状态的初始概率.

2) 递推 $a_{t+1}(j)=\left[\sum_{i=1}^{N}a_t(i)a_{ij}\right]b_j(o_{t+1});\ t=1\sim T-1,\ j=1\sim N$,N 为总的状态个数.

3) 当 $t=T$ 时终止,计算得到 $p(O/\lambda)=\sum_{j=1}^{N}a_T(j)$.

按照第 3.1 节中神经网络各层的输入输出计算公式,最后得到网络的输出值 Y,根据 Y 值所属的手势类别范围即可得到最终的分类结果.例如,A 类手势的 Y 值范围为(0,2],B 类手势的 Y 值范围为(2,4],等等.

4 实 验

我们用 VC++实现了上述算法程序,并在一台 CPU 为 Pentium IV 1.7G,内存为 256M 的普通 PC 机上进行了手势识别实验.首先是为跟踪过程定义了 10 种静态手势类型,如图 5 所示,其中包括 1 种握拳手形、3 种单手指手形、2 种两手指手形、2 种 3 手指手形、1 种 4 手指手形和 1 种手指全部伸出的手形,分别标记为 posture A~posture J.

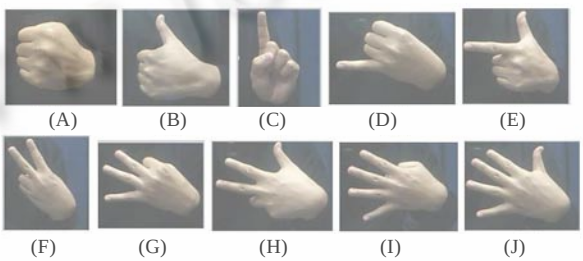


Fig.5 Ten predefined static postures

图 5 10 种预定义的静态手形

训练过程分两步进行:首先是 HMM 模型的参数训练过程,然后是 FNN 的连接权重训练过程.HMM 模型分为 3 种:手形变化 HMM 模型——posHMM、平面轨迹 HMM 模型——traHMM 和 Z 方向运动 HMM 模型——zHMM.

动态手势的手形变化类型共设计了 8 种,例如:由 Posture E 变为 Posture A、由 Posture A 变为 Posture F、Posture B 保持不变等等,这些变化序列同时也是可逆的过程.实际上,手形的变化序列并不局限于两种手形之间的变化,它可以包括若干种不同的手形,只是出于简化的目的,我们在实验中只考虑了两种手形之间的变化.PosHMM 的结构中包含了 3~6 个状态节点,混合高斯模型中的高斯模型分量的个数设置与状态节点个数相同.

实验中,动态手势的平面运动轨迹共有 10 种,除了最简单的向上、向下、向左与向右运动外,还定义了 6 种较为复杂的类型,例如圆形运动、三角形运动、N 形运动和 Z 运动等.如图 6 所示,带圆点一端为运动的起始端,箭头一端为运动轨迹的终端.TraHMM 模型的状态个数为 2~5 个,观察值类型为 8 种.

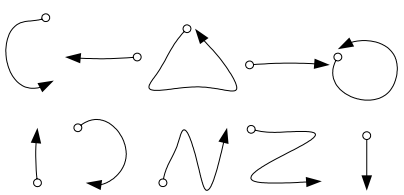


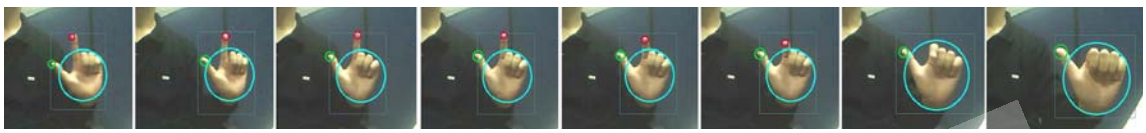
Fig.6 Ten predefined trajectories

图 6 10 种预定义轨迹

Z 轴方向运动类型分为 4 种:向摄像头的靠近运动、远离运动、先靠近再远离以及先远离再靠近的反复运动.zHMM 模型的状态个数为 3 个,混合高斯模型中,高斯模型分量的个数为 3.

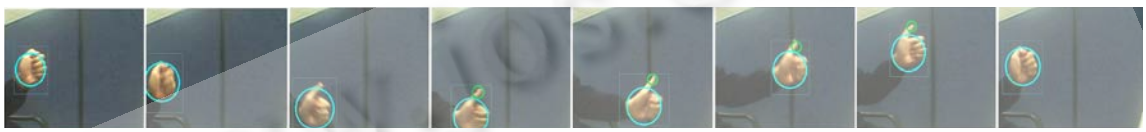
根据手势交互中的任务需要,实验设计了 14 种不同的动态手势,图 7 给出了其中 3 组动态手势的视频截图,各帧截图中用圆形标识出了手掌以及各个手指的跟踪结果:图 7(a)所示的动态手势,是由两手指的手形 Posture E 变化为单手指的手形 Posture B,并同时向摄像头靠近的手势过程;图 7(b)所示动态手势的手形变化过程是一

个由握拳变为大拇指伸出、再到大拇指与食指同时伸出的过程.在手形变化的同时,整个手在进行向右侧的水平运动;图 7(c)中所示动态手势是一个整体手部作划圆运动,同时,手形由握拳变化到大拇指伸出,最后又变化回握拳的过程.



(a) Dynamic gesture A
(a) 动态手势 A

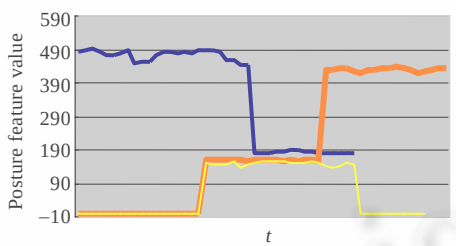
(b) Dynamic gesture B
(b) 动态手势 B



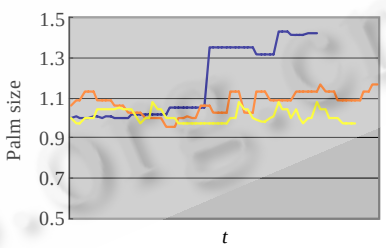
(c) Dynamic gesture C
(c) 动态手势 C

Fig.7 Three dynamic gestures
图 7 3 种动态手势示意图

应用式(1),可以得到图 7(a)~图 7(c)所示动态手势的手形特征值序列,如图 8(a)中实线曲线、虚线曲线以及点划线曲线所示.图 8(b)中实线曲线、虚线曲线以及点划线曲线分别表示图 7(a)~图 7(c)所示动态手势的手掌面积变化序列.



(a) Curves of posture feature
(a) 特征值序列曲线



(b) Curves of palm area
(b) 手掌面积曲线

Fig.8
图 8

实验中,HMM-FNN 模型的模糊规则是通过人的经验确定的,共包含 24 条模糊规则.此外,通过对网络结构进行剪枝,去除一些冗余或不存在的网络连接,可以提高训练的效率以及识别过程的速度.例如,对于图 7(a)所示的手势,由于我们知道它与手势的运动轨迹关系不大,而与手形变化关系较为密切,在确定网络结构时,去除与平面轨迹 HMM 节点的连接;而在设定规则条件的初始权重时,手形条件的权重比 Z 轴运动条件的权重稍大.

每种手势由不同的 5 个实验者各做 3 次,共得到 15 个手势视频片断,对于 14 种手势则产生 210 组训练数据.训练时,首先将手形变化、轨迹变化与手掌面积变化数据分别提取出来,进行 HMM 模型的训练.在 HMM 模型训练完成后再进行 HMM-FNN 模型的连接权重训练.在训练过程中,随机抽取其中的 180 组作为训练数据,剩余的 30 组作为测试数据.当系统的实际输出结果与预期结果误差低于阈值或迭代次数达到最大限制时,训练结束.训练结束后,我们对训练完成的 HMM-FNN 模型进行了测试,并与传统的离散型 HMM 进行了比较.结果见

表 1.

Table 1 Comparison of recognition rate of proposed method and conventional HMM method

表 1 本文方法与传统 HMM 识别率的比较

Gesture type	A	B	C	D	E	F	G
Conventional HMM (%)	73.3	70.0	73.3	80.0	70.0	83.3	70.0
Proposed HMM-FNN (%)	86.6	86.6	83.3	90.0	86.6	90.0	80.0
Gesture type	H	I	J	K	L	M	N
Conventional HMM (%)	66.6	80.0	73.3	70.0	70.0	76.6	70.0
Proposed HMM-FNN (%)	83.3	80.0	83.3	90.0	83.3	83.3	86.6

实验中我们发现,HMM-FNN 模型中的 3 组 HMM 模型分别对测试手势动作的手形变化序列、运动轨迹以及 *Z* 轴方向运动进行分类,它们的平均分类正确率分别为 76.6%,80.0%和 73.3%,而动态手势的最终平均分类正确率为 88.3%.也就是说,HMM 分类中间结果的某些错误,经过模糊规则的推理被纠正而得到最终正确的分类结果.此外,本文方法对手势运动的节奏快慢是不敏感的,只要手势的运动过程类似,就会得到类似的分类结果.

5 总 结

本文提出了一种基于 HMM-FNN 模型的动态视频手势建模与识别方法.首先,它充分利用了动态手势本身的特点,即动态手势运动特征的可分解性与语义描述上的模糊性,将其分解为手形变化、2D 平面运动与 *Z* 轴方向运动 3 个组成部分,通过对手及手指指尖的位置跟踪,获得 3 组特征值序列作为 HMM-FNN 模型的输入数据.HMM-FNN 模型是一种结合了隐马尔可夫模型(HMM)时序建模能力与模糊神经网络(FNN)的模糊逻辑表达与分类能力的模型,HMM 对观察值序列的似然概率作为对各子类的模糊隶属度,通过 FNN 的模糊规则推理得到最终手势类别.

与普通 HMM 模型相比,本文提出的方法在对复杂动态手势识别时,通过利用手势本身的特点将复杂问题进行分解,避免了用高维度特征对手势进行描述,从而降低了运算复杂度,提高了系统性能.此外,本方法充分考虑到手势的模糊特性,并通过 FNN 的形式进行模糊规则的建模与模糊推理.较之简单的确定性推理,系统的鲁棒性更好.而且,HMM-FNN 模型可以充分利用人的先验知识,在模糊规则的构造与网络连接结构上进行优化处理,使系统的训练与识别效率更高.实验表明,本文提出的方法是一种高效、可靠的复杂动态手势识别方法.

在下一步工作中,我们将考虑利用多个摄像机捕捉更加准确的手势运动的 3D 特征,从而进一步改善手势识别的效果.此外,引进空间、时间上下文约束关系对手势视频进行自动时序分割也是我们进一步研究的方向.

References:

[1] Pavlovic VI, Sharma R, Huang TS. Visual interpretation of hand gesture for human-computer interaction: A review. IEEE PAMI, 1997,19(7):677−695.

[2] Wu Y, Huang TS. Vision-Based gesture recognition: A review. In: LNCS 1739, London: Springer-Verlag. 1999. 103−105.

[3] Ye GQ, Corso JJ, Hager GD. Gesture recognition using 3D appearance and motion feature. In: Proc. of the IEEE CVPR 2004. 2004. 160−170.

[4] Alpern M, Minardo K. Developing a car gesture interface for use as a secondary task. In: Proc. of the CHI 2003. 2003. 932−933.

[5] Starner T, Pentland A. Real-Time american sign language recognition from video using hidden Markov models. Technical Report, 375, MIT Media Lab., Perceptual Computing Group, 1995.

[6] Rigoll G, Kosmala A, Eickeler S. High performance real-time gesture recognition using hidden markov models. In: Proc. of the Gesture Workshop. 1997. 69−80.

[7] Bobick AF, Willson AD. A state-based approach to the representation and recognition of gesture. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997,19(12):1325−1337.

[8] Ren HB, Zhu YX, Xu GY, Lin XY, Zhang XP. Spatio-Temporal appearance modeling and recognition of continuous dynamic hand gestures. Chinese Journal of Computers, 2000,23(8):824−828 (in Chinese with English abstract).

[9] Zhu YX, Xu GY, Kriegman DJ. A real-time approach to the spotting, representation, and recognition of hand gesture for human-computer interaction. Computer Vision and Image Understanding, 2002,85:189−208.

[10] Park HS, Kim EY, Jang SS, Kim HJ. An HMM-based gesture recognition for perceptual user inerface. In: Proc. of the 5th Pacific-Rim Conf. on Multimedia (PCM 2004). LNCS 3332, 2004. 1027−1034.

[11] Murakami K, Taguchi H. Gesture recognition using recurrent neural network. In: Proc. of the CHI'91. 1991. 237−241.

[12] Fels S, Hinton G. Glove-TalkII: An adaptive gesture-to-formant interface. In: Proc. of the CHI'95. 1995. 456−463.

[13] Vamplew P, Adams A. Recognition and anticipation of hand motions using a recurrent neural network. In: Proc. of the IEEE Int'l Conf. on Neural Networks. 1995. 2904−2907.

[14] Cohen M, Franco H. Hybrid neural networks hidden Markov model continuous speech recognition. In: Proc. of the Int'l Conf. on Spoken Language Processing. 1992. 915−918.

[15] Quek F. Unencumbered gesture interaction. IEEE Multimedia, 1996,3(3):36−47.

[16] Zhu JY, Wang XY, Wang WX, Dai GZ. Gesture recognition based on structural analysis. Chinese Journal of Computers, 2006, 29(12):2130−2137 (in Chinese with English abstract).

[17] Wang XY, Zhang XW, Dai GZ. A novel approach to tracking deformable hand gesture for real-time interaction. Journal of Software, 2007,18(10):2423−2433 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/2423.htm

[18] Zhang YJ. Context-Based Visual Information Retrieval. Beijing: Science Press, 2003 (in Chinese).

[19] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 1989,77(2): 257−286.

附中文参考文献:

[8] 任海兵,祝远新,徐光祐,林学闾,张晓平.连续动态手势的时空表观建模及识别.计算机学报,2000,23(8):824−828.

[16] 朱继玉,王西颖,王威信,戴国忠.基于结构分析的手势识别.计算机学报,2006,29(12):2130−2137.

[17] 王西颖,张习文,戴国忠.一种面向实时交互的变形手势跟踪方法.软件学报, 2007,18(10):2423−2433.http://www.jos.org.cn/1000-9825/18/2423.htm

[18] 章毓晋.基于内容的视觉信息检索.北京:科学出版社,2003.

王西颖(1974－),男,江苏睢宁人,博士,主要研究领域为计算机视觉,模式识别,人机交互.

张习文(1971－),男,博士,副研究员,主要研究领域为模式识别,图像理解.

戴国忠(1944－),男,研究员,博士生导师,CCF 高级会员,主要研究领域为计算机图形学,人机交互.

张凤军(1971－),男,博士,副研究员,CCF 高级会员,主要研究领域为虚拟现实,计算机图形学.