连续动态手势的时空表观建模及识别

任海兵 祝远新 徐光祐 林学訚 张晓平
(清华大学计算机科学与技术系媒体研究所 北京 100084)

摘 要 论述了复杂背景下连续动态手势的分割、建模及识别;融合手势运动信息和皮肤颜色信息,进行复杂背景下的手势分割;通过结合手势的时序信息、运动表观以及形状表观,提出动态手势的时空表观模型,并提出基于颜色、运动以及形状等多模式信息的分层次融合策略抽取时空表观模型的参数.最后,提出动态时空规整算法用于手势识别.实验表明,利用上述提出的手势分割、建模、特征参数抽取及识别方法识别 12 种手势,平均识别率高达 97%.

关键词 手势分割;手势识别;时空表观模型;多模式信息的分层次融合;动态时空规整
中图法分类号: TP391

Spatio-Temporal Appearance Modeling and Recognition of Continuous Dynamic Hand Gestures

REN Hai-Bing ZHU Yuan-Xin XU Guang-You LIN Xue-Yin ZHANG Xiao-Ping
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract This paper presents the segmentation, spatio-temporal appearance modeling and recognition of continuous dynamic hand gestures with complex background. Fusing the coarse motion information and skin color chrominance, the segmentation of hand gesture can be got robustly. By integrating temporal information, motion appearance, and shape appearance, a spatio-temporal appearance model is proposed to represent dynamic hand gestures, and a novel scheme is proposed for recovering the model parameters by hierarchical fusion of multi-modal information such as skin chrominance, motion, and shape. In addition, Dynamic Space-Time Warping is suggested for recognition of hand gestures. Experiment results indicate that the average recognition rate of the 12 hand gestures is 97% based on the proposed segmentation, modeling, feature extracting, and recognition techniques.

Keywords hand gesture segmentation, hand gesture recognition, spatio-temporal appearance, hierarchical fusion of multi-modal information, dynamic space-time warping

1 引 言

最近几年,基于计算机视觉的手势识别逐渐成为计算机视觉领域的一个研究热点.将手势识别系统用于人机接口,从而实现直接的、自然的人机交互方式,是手势识别的发展趋势和基本目标.其中,最关键的 3 个问题是动态手势的分割、建模以及识别.

在基于计算机视觉的手势识别技术中,复杂背景下的手势分割非常困难,特别是在单目视觉情况

收稿日期: 1999-06-18;修改稿收到日期: 2000-06-12.本课题得到国家自然科学基金(69873022)资助.任海兵,男,1975 年生,博士研究生,主要研究方向为计算机视觉、模式识别和人机交互.祝远新,获博士学位,现在密苏里哥伦比亚大学从事博士后研究工作,研究方向为计算机视觉、模式识别和人工智能.徐光祐,教授,博士生导师,主要研究领域为计算机视觉、人机交互技术和多媒体技术.林学訚,教授,博士生导师,主要研究方向为计算机视觉和人机交互技术.张晓平,硕士研究生,研究方向为计算机视觉和人机交互技术

下. 这主要是由于背景各种各样, 环境因素也不可预见. 不仅没有成熟的理论作为指导, 而且现有的方法实现困难, 计算复杂度很高, 效果也不是很理想. 目前, 主要有以下 3 种解决方法:

(1) 增加限制的方法. 这是最常用的方法, 研究人员对手势图像加上种种限制, 如使用黑色或白色的墙壁、深色的服装等简化背景, 或要求人手戴特殊颜色的手套等强调前景, 来简化手区域与背景区域的划分. 如美国MIT 的 Thad Starner[5]以深色的背景、左右手不同颜色的手套来简化人手的分割, 利用双手形状和运动轨迹实现 40 个美国手语的识别. 由于人为增加了诸多的限制, 不利于自然的人机交互.

(2) 大容量手势形状数据库方法. 密西根州立大学计算机系的Cui Yuntao[9]建立了一个数据库, 其中有各种手势类(m 种)在各个时刻(n 个时刻)不同位置(l 个位置)不同比例(k 个比例因子)的手形图像, 故手势形状数据库是($m \times n \times l \times k$)量级. 分割时以 Prediction-and-verification 方法, 从中找出最接近的形状作为模板, 然后对形状求精, 计算复杂度为$O(m \times n \times l \times k)$. 即使将数据库按分层的 quasi-Voronoi 图进行组织分类, 搜索复杂度也达到$O(\lg(m \times n \times l \times k))$, 在 SGI INDIGO2 工作站上, 平均分割一副图像需要 58.3s.

(3) 以上两种都是单目视觉的方法, 还有立体视觉方法. 如纽约哥伦比亚大学计算机系的 Gluckman[8]利用两个不在同一平面镜子的反射图像, 计算物体与摄像机之间的距离, 根据距离信息分割出人手.

另外, 有些研究人员, 如 Micorsoft Korea 的 Lee[10]仅仅根据手运动的轨迹识别 10 种手势, 完全避开了手形分割, 丢失了手势的重要信息, 很难推广应用. 为了使基于单目视觉的手势识别方法能够付诸实用, 我们结合人手颜色信息和手势运动信息提出了一种复杂背景下手势分割的方法, 以消除以上几种方法的缺点.

在手势建模方面, 现有的手势模型可以分归两大类: 基于三维人手模型的手势模型和基于表观的手势模型. 原理上, 基于三维人手模型的手势模型非常精细, 适合于给所有手势建模. 然而, 模型参数多, 计算复杂度高, 而且为抽取模型参数而使用的许多近似过程导致模型参数的估计很不可靠. 而基于表观的手势模型的计算复杂度低, 易于达到实时. 因此, 在目前条件下绝大部分手势识别研究都采用了基于表观的手势模型. 本文提出了动态手势的时空表观模型, 融合了手势的时序信息、运动表观及形状表观信息. 其中的运动表观描述由于人手运动所引起的图像表观变化, 而不是静止手姿态特征在时序上的简单排列; 形状表观则描述了手形的拟合椭圆特征.

目前的手势识别方法有模板匹配、动态规划以及隐马尔可夫模型等. 模板匹配的方法多用于静态手形识别中, 如Cui Yuntao[9]文中计算模板的相关系数进行预测匹配. 文献[11]利用动态规划匹配运动模式. 而文献[5, 10]都采用隐马尔可夫模型. 本文根据浏览三维全景图所需手势的特点提出了动态时空规整技术, 度量两个时空模式之间的距离.

2　复杂背景下的手势分割

在基于计算机视觉的手势识别技术中, 把图像中的人手区域与其它区域(背景区域)划分开来始终是一个难点. 我们根据手势图像的特点, 综合运用手势的运动信息和人手的皮肤颜色信息, 提出一种全新的手势分割方法, 具体实现如图 1 所示.

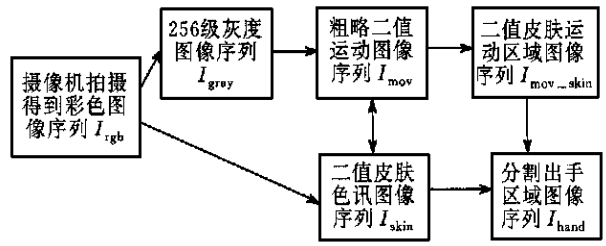

图 1　手势分割原理图

首先, 由拍摄得到彩色图像序列 I_{rgb}, 一方面将其转换为 256 级灰度图像序列 I_{gray}, 用于运动参数的分析; 另一方面根据 RGB 颜色在 HSI 空间的分布, 得到二值皮肤色讯图像序列 I_{skin}, 将其划分为皮肤颜色区域和非皮肤颜色区域. 对灰度图像序列 I_{gray}, 处理得到粗略的二值运动图像序列 I_{mov}. 同时, I_{mov}和 I_{gray}对应帧图像之间的与操作即得到二值皮肤运动区域图像序列 I_{mov_skin}, 我们认为序列 I_{mov_skin} 中区域就是运动的皮肤区域(基本上属于手区域).

值得注意的是, 在检测人手运动的过程中, 并不是手的每一部分都有明显的图像表观变化(特别是手心部分的图像灰度相差不是很大、纹理不清晰, 手心小的运动不能产生灰度的明显变化), 所以得到的粗略运动区域图像并不一定包含完整的手形. 因此, I_{mov_skin}也并不一定包含完整的手形, 把它作为手形计算形状特征, 会造成很大的偏差. 本文回溯到二值皮肤色讯图像序列 I_{skin}, 利用种子算法寻找完整的手形, 具体方法如下:

1、假定手的运动区域是 I_{mov_skin} 中的主要部分，所以在 I_{mov_skin} 的二值图像中应用种子算法，找到最大的块 B，把这个块 B 作为人手的一部分. 由于手势者是坐在摄像机前面，面向摄像机镜头做手势的，这个假定是合理的. 同时，这也排除了背景中与皮肤颜色相似部分的微小运动.

2、把块 B 映射到 I_{skin} 上的相同位置 B'，应用种子算法以块 B' 为中心，在 I_{skin} 中扩展得到手区域图像序列 I_{hand}.

手区域图像序列 I_{hand} 就是手势分割得到的只包含手区域的图像序列，用来抽取运动和形状特征参数. 本文利用合理的假设进行人手形状的分割，计算复杂度是常数量级.

3 动态手势的时空表观模型

给定包含动态手势的灰度图像序列 I_{gray} 以及对应的手区域图像序列 I_{hand}，我们利用参数化的图像运动模型、鲁棒回归算法以及基于图像矩的形状分析技术估计手势的时空表观模型的参数. 手势的时空表观模型由一序列帧间特征向量组成，而帧间特征向量由运动表观和形状表观两部分组成.

3.1 帧间运动表观

本文利用参数化的图像运动模型(平移模型、仿射模型或者平面模型)近似表示物体(即人手)运动所引起的图像表观变化. 把 I_{hand} 图像对应的区域作为分析区域，通过鲁棒回归估计图像运动模型的参数. 既然绝大多数数据点都属于运动物体，所以回归结果就是物体的帧间运动参数. 另外，基于回归结果分析残差(即检测内点，摒弃外点)就可以得到运动物体的精细分割.

图像运动的平面模型可以写为式(1)的形式，其中 a_i (对整个运动区域而言)是常数，$\mathrm{u}(\mathrm{x})$ 是像素点 $\mathrm{x}=(x,y)$ 的帧间位移向量，$u(x,y)$ 和 $v(x,y)$ 分别是它的水平和垂直分量.

$$\mathrm{u}(\mathrm{x})=\begin{bmatrix}u(x,y)\\ v(x,y)\end{bmatrix}=\begin{bmatrix}a_0\\ a_3\end{bmatrix}+\begin{bmatrix}a_1 & a_2\\ a_4 & a_5\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}+\begin{bmatrix}x^2 & xy\\ xy & y^2\end{bmatrix}\begin{bmatrix}a_6\\ a_7\end{bmatrix} \tag{1}$$

式中的参数 $[a_0,\cdots,a_7]$ 表示区域的运动和形变信息，可以分解为一些独立的、有几何意义的分量 $[m_1,\cdots,m_7]$：X 轴方向的纯平移 $m_1=a_0$，Y 轴方向的纯平移 $m_2=a_3$，各向同性的膨胀 $m_3=a_1+a_5$，错切 $m_4=\sqrt{(a_1-a_5)^2+(a_2+a_4)^2}$，旋度 $m_5=-a_2+a_4$，沿观察方向的左右偏转 $m_6=a_6$，沿观察方向的上下翻转 $m_7=a_7$.

这里的旋度、散度以及错切量具有尺度不变性，并且不依赖于图像坐标系统的选择. 我们定义一个 7 维向量 $\mathrm{m}[t]=[m_1,m_2,m_3,m_4,m_5,m_6,m_7]$ 来描述相邻两帧图像之间的图像运动表观特征.

3.2 帧内形状表观

给定分割出的人手图像区域，本文通过形状分析抽取手势的形状表观. 即首先用一个椭圆去拟合人手图像区域，然后利用拟合椭圆的几何特征描述形状信息. 图像坐标为 (x,y) 的区域 R 的二维 p,q 阶中心矩 $\tilde{m}_{p,q}$ 定义如下：

$$\tilde{m}_{p,q}=\frac{1}{\tilde{R}}\sum_{(x_i,y_i)\in R}(x_i-\bar{x})^p(y_i-\bar{y})^q \tag{2}$$

其中，$(\bar{x},\bar{y})$ 是区域 R 的质心坐标；$\tilde{R}$ 是区域 R 的面积，即像素数目.

有了中心矩，就可以得到区域空间分布的协方差矩阵，即式(3). 通过计算空间协方差矩阵的特征值以及相应的特征向量，可以计算出该椭圆长轴的长度 (a)、短轴的长度 (b) 以及长轴与图像平面的水平 X 轴之间的倾斜角 (θ). 这三个形状参数描述了区域的形状特征，因此，可以定义一个三维向量 $\mathrm{s}[t](=[a/2,a/b,\theta])$ 来描述第 t 帧图像中的静止手姿态信息.

$$\begin{bmatrix}\tilde{m}_{2,0} & \tilde{m}_{1,1}\\ \tilde{m}_{1,1} & \tilde{m}_{0,2}\end{bmatrix} \tag{3}$$

3.3 时空表观模型

给定分割出的动态手势样本，令 L 表示该样本的时间长度(帧数)，令第 $t(=0,1,\cdots,L-1)$ 帧与第 $t+1$ 帧之间的运动表观是 $\mathrm{m}[t]$，$\mathrm{s}[t]$ 是第 t 帧内的形状表观. 本文通过把帧间运动表观跟帧内形状表观集成起来构成时刻手势的图像表观，即定义一个 10 维特征向量，$\mathrm{f}_t=[\mathrm{m}[t],\mathrm{s}[t]]^{\mathrm{T}}$. 图像表观的时间序列就构成该手势样本的时空表观特征 g，即 $\mathrm{g}=[\mathrm{f}_0,\mathrm{f}_1,\cdots,\mathrm{f}_{L-2}]$.

4 基于动态时空规整的手势识别

不同用户(甚至同一个用户)做同一种手势时，有人做得快，有人做得慢，因此手势速率的变化会在时空表观模型的时间轴上引起非线性波动，并且由于手势的采样率很低(如本系统的采样率为 10Hz)，因此时间轴上的波动比较强烈. 如何消除这种波动一直是动态手势识别研究的一个重要问题. 本文针对这个问题提出动态时空规整算法，把时空表观模式动态地规整到一个固定的时间长度 K.

4.1 动态时空规整算法

由于时空模式中的运动表观描述了图像序列中

的运动变化, 是动态信息, 而形状表观描述的是静态手姿态信息, 因此对它们二者的规整要区别对待. 具体地说: 对于一个时间长度为L 的手势, 规整后时刻l 的动态表观信息对应于原手势$(l-1)L/K$ 时刻到lL/K 时刻的运动量(包括平移、旋转、错切等); 而静态表观信息则对应于原手势lL/K 时刻的手姿态信息.

手势的时空表观模式经过规整后, 就变成了时间长度相等的表观模式. 本文基于规一化相关计算规整后两个表观模式之间的距离, 并称之为时空规整距离. 令$^{l}\hat{A}=(^{l}\hat{a}_{ij})_{K\times 10}$, $^{l}\hat{B}=(^{l}\hat{b}_{ij})_{K\times 10}$分别表示动态时空规整后得到两种手势A, B 的表观模式, K 是规整长度, 则A, B 之间的时空规整距离$D(A,B)$的定义如下:

$$D(A,B)=1-\frac{\sum_{i=0}^{K-1}\sum_{j=0}^{9}(w_j\,{}^{l}\hat{a}_{ij})\times(w_j\,{}^{l}\hat{b}_{ij})}{\sqrt{\sum_{i=0}^{K-1}\sum_{j=0}^{9}(w_j\,{}^{l}\hat{a}_{ij})^2}\sqrt{\sum_{i=0}^{K-1}\sum_{j=0}^{9}(w_j\,{}^{l}\hat{b}_{ij})^2}} \tag{4}$$

式中, $w_j(j=0,1,\cdots,9)$是权重. 我们根据等方差的原则来确定各特征分量的权重.

5　实验结果

本文设计的手势命令集由"向上平移(1)","向下平移(2)","向左平移(3)","向右平移(4)","向前平移(5)","向后平移(6)","向右偏转(7)","向左偏转(8)","顺时针转(9)","逆时针转(10)","向下翻转(11)",和"向上翻转(12)"12 种手势组成. 我们做了大量实验来验证和评价本文提出时空表观模型、多模式信息的分层次融合策略以及动态时空规整算法的性能. 我们邀请 2 位实验者坐在摄像机前分别做上面的 12 种手势, 每个手势重复 5 次, 于是得到 120 个手势样本(即每种手势有 10 个样本). 每个手势持续时间约 1.2s(采样率为 10Hz). 图像是 24 位真彩色, 大小是 160×120. 分别从每种手势的 10 个样本中取前 6 个构成训练集, 剩下的 48 个样本构成测试集. 实验表明在 PIII(600MHz)PC 机上, 一个 12 帧图像的手势序列, 分割平均需要 1.1s; 计算特征参数平均 0.9s, 识别平均时间约 0.07s.

图 2 所示是分割出的一个"向上翻转"手势样本. 图中(a), (b), (c), (d), (e), (f)分别是原始彩色图像序列、灰度图像序列、皮肤颜色区域的二值图像序列、运动区域的二值图像序列、表示人手区域的二值图像序列以及对人手图像区域进行椭圆拟合的结果示意.

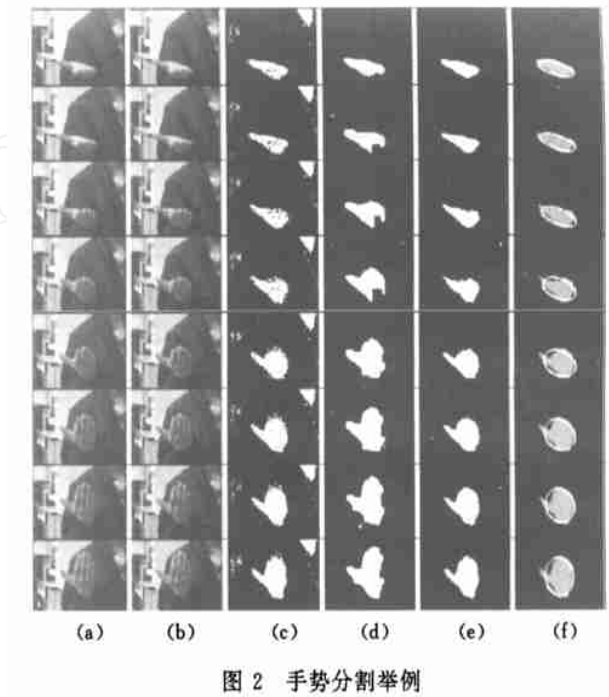

图 2　手势分割举例

构造手势的时空表观模型时, 到底是选取图像运动的平面模型还是仿射模型, 还是最简单的平移运动模型, 这取决于手势命令集的大小、手势的做法、对识别率的要求以及对处理速度的要求等诸多因素. 我们分别选取平移模型、仿射模型以及平面模型进行了实验. 实验表明, 对于本文的手势命令集选用图像运动的仿射模型既能满足高识别率的要求, 又能满足处理速度的要求. 另外, 还必须选用合适的规整长度. 为了选取合适的规整长度, 我们需要考察训练集里样本长度分布的统计特性, 并以识别率和计算量为判别准则选取合理的规整长度. 实验结果表明, 当规整长度为 4 时, 不仅识别正确率高, 而且计算量小.

采用图像运动的仿射模型参数构造时空表观中的运动表观部分, 并取规整长度为 4 时, 识别训练集手势样本的实验结果如表 1 所示. 表 2 所示是在同样条件下系统识别测试集样本的实验结果.

表 1　识别训练集手势样本得到的识别率

手势编号	1	2	3	4	5	6	7	8	9	10	11	12	平均
识别率(%)	100	100	100	83.3	100	100	100	100	100	100	100	83.3	97.2

表 2 识别测试集手势样本得到的识别率

手势编号	1	2	3	4	5	6	7	8	9	10	11	12	平均
识别率(%)	100	100	100	75	75	100	100	75	100	100	100	75	91.7

从表 1 和表 2 可知,训练集上的平均识别率为 97.2%,测试集上的平均识别率为 91.7%.与之相比,文献[5]在无语法规则的情况下,测试集和训练集的识别率分别是 93.5%和 90.7%;文献[10]的识别率为 93.14%.

为了考察时空表观模型中的运动表观部分或者形状表观部分的独立描述能力,本文分别仅其中的运动表观或形状表观作为手势特征进行实验.实验表明,仅用运动表观时平均识别率为 88.9%,而仅用形状表观时的平均识别率为 73.6%.这证明了用于描述动态手势的时空表观模型可以实现运动表观与形状表观之间的有效融合.

6 结论及展望

本文围绕连续动态手势的分割、建模、特征抽取以及识别深入研究了基于视觉连续动态手势识别的各个重要环节.并通过大量实验来验证和评价本文提出的时空表观模型、多模式信息的分层次融合策略、动态时空规整算法以及训练和识别算法的性能.实验表明这些算法是有效的,也是可靠的.

当然,在我们提出的方法里,也存在一些困难.例如,我们目前假设景物中运动且具有人体皮肤色度特征的物体就是做手势的人手,当景物中出现大面积运动的人脸时这个假设就不成立了.显然,引入简单的人体几何模型约束能够一定程度上克服这个问题,此时可能有必要引入多个摄像机.手势命令集的大小取决于具体应用,对于某些简单的应用很小的手势命令集就足够了,例如,用手势遥控家电.而对于其它的一些应用,则要求尽可能大的手势命令集,例如聋哑手语识别.对于本文设计的应用,要精确地控制全景图的运动,可能需要更丰富的手势命令.例如,"开始"、"结束"以及表示数量的手势命令等等.当然也可以用语音命令来输入数量信息,实现多模式的人机接口.成功的手势识别策略应该考虑手势的时间-空间上下文,这要求把语法规则引入到识别过程.研究人际交流中手势的语法规则并把它应用到识别过程也是我们的进一步方向.

参 考 文 献

1 Ren Hai-Bing, Zhu Yuan-Xin, Xu Guang-You *et al*. Vision-based recognition of and gestures: A survey. Acta Electronnica Sinica, 2000, 28(2): 118-121(in Chinese)
(任海兵,祝远新,徐光祐等.基于视觉手势识别的研究.电子学报,2000,28(2):118-121)

2 Zhu Yuan-Xin, Ren Hai-Bing, Xu Guang-You *et al*. Extracting spatio-temporal appearance by hierarchically integrating multi-modal information for hand gesture recognition. In: Proceedings of the Asian Conference of Computer Vision, Taiwan, 2000. 282-287

3 Marcus B A, Churchill P J. Sensing human hand motions for controlling dexterous robots. In: Proceedings of the Second Annual Space Operations Automation and Robotics Workshop. Wright State University, 1988

4 Tao Lin-Mi, Peng Zheng-Yun, Xu Guang-You *et al*. The color feature of human skin and face detection in complex background. In: Proceedings of the National Conference on Intelligent Interface and Intelligent Application, Taiyuan, 1999. 205-211(in Chinese)
(陶霖密,彭振云,徐光祐等.人类的肤色特征及复杂环境下的人脸检测.见:第四届中国计算机智能接口与智能应用学术会议论文集,太原,1999.205-211)

5 Thad Starner, Alex Pentland. Real-time American sign language recognition using desktop and wearable computer based video. MIT Media Laboratory: Technical Report 466, 1998

6 Quek F K H. Eyes in the interface. Image and Vision Computing, 1995, 13(6): 511-525

7 Vaillant R, Darmon D. Vision-based hand pose estimation. In: Proceedings of the IEEE 1st International Workshop on Automatic Face and Gesture Recognition, 1995. 356-361

8 Gluckman J, Nayar S K. Planar catadioptric stereo: Geometry and calibration. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins Colorado, 1999. 23-25

9 Cui Yun-Tao, Weng J J. View-based hand segmentation and hand-sequence recognition with complex backgrounds. In: Proceedings of the IEEE International Conference of Pattern Recognition, 1997. 617-621

10 Lee Hyeon-Kyu, Kim J H. An HMM-based threshold model approach for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(10): 961-972

11 Gavrila D M, Davis L S. Towards 3D model-based tracking and recognition of human movement: A multi-view approach. In: Proceedings of the International Workshop on Automatic Face and Gesture Recognition, Switzerland, 1995. 272-277