第28卷 第2期
2002年3月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 2
Mar., 2002

短文

复杂背景下的手势分割与识别[1)]

任海兵 祝远新 徐光祐 张晓平 林学訚
(清华大学计算机科学与技术系媒体所 北京 100084)
(E-mail: Renhb@media.cs.tsinghua.edu.cn)

摘 要 目前在基于单目视觉的手势识别中,手势分割技术几乎都是基于简单的背景或者要求手势者带有特殊颜色的手套,给人机交互增加了一定的限制.本文融合人手颜色信息和手势运动信息,两次利用种子算法对复杂背景下的手势进行分割.根据分割出的手区域大大加速了运动特征参数的提取,并结合手区域的形状特征,建立手势的时空表观模型.识别时,采用独立分布的多状态高斯概率模型,进行时间规整.手势训练集和测试集的识别率分别为97.8%和95.6%.

关键词 复杂背景,手势分割,手势识别,独立分布,多状态高斯概率模型
中图分类号 TP291

HAND GESTURE SEGMENTATION AND RECOGNITION WITH COMPLEX BACKGROUNDS

REN Hai-Bing ZHU Yuan-Xin XU Guang-You ZHANG Xiao-Ping LIN Xue-Yin
(*Media Institute, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
(E-mail: Renhb@media.cs.tsinghua.edu.cn)

Abstract Currently, in the vision-based hand gesture recognition, almost all the technologies on hand gesture segmentation are based on simple backgrounds or on gloves in special colors, which give the human-computer interaction some limitation. This paper presents a new method, which segments hand gestures with complex backgrounds by fusing skin chrominance and coarse image motion, and by using the seed algorithm twice. With the segmented hand areas, the algorithm for motion appearance parameters is accelerated greatly. By integrating temporal information, motion and shape appearances, a spatio-temporal appearance model is proposed for representing dynamic hand gestures. This paper also presents an independent distributed multi-state Gaussian probability model(IDMGPM) for recognition. In this system the average recognition rate is 97.8% on the training set and 95.6% on the testing set.

Key words Complex background, hand gesture segmentation, hand gesture recognition, independent distributed, multi-state Gaussian probability model

1) 国家自然科学基金(69873022)和国家"八六三"(863-306-ZT03-01-1)资助
收稿日期 1999-10-20 收修改稿日期 2000-09-15

1　引言

在多模态的人机交互方面,除了大家所熟知的语音识别外,还有人脸识别,面部表情解释,唇读,头部运动跟踪,注视跟踪,三维手指定位,手势识别和体势识别等方面[1]. 手势识别是其中一个崭新的领域.

基于计算机视觉的手势识别方法使用摄像机直接拍摄手势运动过程,从手势图像序列中分割出人手,根据图像序列前后帧之间的关系计算出运动参数. 如美国 MIT 媒体实验室的 Starner[2],通过提取左右手质心的运动轨迹、手的形状等特征参数,结合语法规则识别 40 个美国手语,正确率达到 97%(没有语法规则的正确率为 90.7%). 另外,Microsoft Korea 的 Hyeon-Kyu Lee[3],利用基于 HMM 的阈值模型(HMM-Based Threshold Model)从单手的运动轨迹识别出 9 种手势命令,平均识别率达到 98.19%.

在基于单目视觉的手势识别方法中,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. 这主要是由于背景各种各样、环境因素也不可预见,所以实现起来困难重重,非常复杂. 不少研究人员采用对手势图像加上种种限制的方法,如使用黑色或白色的墙壁、深色的服装等简化背景[2],或要求人手戴特殊颜色的手套等强调前景[3],来简化手区域与背景区域的划分,降低手势分割的难度和计算复杂度. 但是,人为增加了诸多的限制,不利于自然的人机交互.

基于计算机视觉的手势识别方法的优点在于人手能处于自然状态,使人能够以自然的方法进行人机交互,因此是手势识别技术发展的趋势和目标. 其缺点是理论不够成熟,实现上也非常复杂,识别集都比较小.

本文论述了单目视觉技术中一种复杂背景下的手势分割和识别方法. 首先把运动着的手从复杂的背景中分割出来,然后提取运动和形状特征参数,建立手势的时空表观模型[4,5],采用独立分布的多状态高斯概率模型进行时间规整,最后得到识别结果. 本文的内容安排如下:第二节论述复杂背景下的手势分割,第三节论述手势形状和运动特征的获取,第四节介绍了独立分布的多状态高斯概率模型,第五节给出实验结果和结论.

2　复杂背景下的手势分割

为了使基于计算机视觉的手势识别方法能够付诸实用,本文着重研究在复杂背景下将人手区域分割出来的方法. 单独使用手的运动信息与颜色信息进行人手分割都存在着明显的不足. 所以根据手势图像的特点,将两种信息综合地加以运用,本文提出一种新的手势分割的方法.

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,将其转换为二值皮肤色讯图像序列 I_{skin},其中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_{skin}对应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},本文认为序列 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运动(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区域，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区域，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，所以根据区域连通性，在I_{mov_skin}中应用种子算法，找到最大的连通域B，把这个连通域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. 把连通域B映射到I_{skin}中的相同位置，应用种子算法以此位置为种子，在I_{skin}中扩展得到完整手区域的图像序列I_{hand}.

在拍摄的360个手势图像序列样本中(每个序列平均有12幅图像，总共约有360×12幅)，分割出的手区域只有9幅图像手区域不完整或者手区域多出来一部分. 其它99.79%的图像分割结果都是正确的. 最后，根据I_{mov_skin}中手区域的运动判断手势的开始和结束.

3 特征参数的提取

由第二部分得到一个分割好的手势序列，包括手势的灰度图像序列I_{gray}、手区域图像序列I_{hand}等. 针对手区域图像序列I_{hand}，提取手区域的形状特征；结合I_{gray}和I_{hand}，在相邻两帧的手区域内，计算运动参数，作为帧间运动特征.

3.1 手区域形状特征的分析

I_{hand}中每一帧图像，只存在人手区域. 对每一帧图像，我们用椭圆去拟合图像中的人手区域，通过二阶矩参数(moment)计算出该椭圆的长半轴($a/2$)、长短轴之比(a/b)、以及长轴跟图像平面的x轴之间的夹角(θ)，作为手区域的形状特征.

坐标轴为x,y轴的二维图像坐标系中，区域R的二维p,q阶中心矩$\tilde{m}_{p,q}$以及二阶中心矩矩阵M分别如下：

$$\tilde{m}_{p,q}=\frac{1}{m_{0,0}}\int_R(x-\bar{x})^p(y-\bar{y})^q\mathrm{d}x\mathrm{d}y \tag{1}$$

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

其中，$m_{0,0}$是区域R的二维0,0阶矩，也是区域R的有效面积(即有效像素的数目)，$(\bar{x},\bar{y})$为区域R的质心坐标. 二阶中心矩矩阵M的两个特征值λ_1,λ_2为该拟合椭圆的长半轴$a/2$和短半轴$b/2$的平方. a/b为椭圆长短轴之比，也是椭圆的离心率. 矩阵主特征向量对应于椭圆长轴与图像y轴夹角$\frac{\pi}{2}-\theta$.

3.2 帧间运动参数的估计

以仿射模型近似手势图像帧间的运动. 通过运动图像的光流方程得到图像灰度的残差，进行迭代计算相邻帧间的运动参数. 并且应用鲁棒统计学提高系统的鲁棒性和准确性，结合

手势运动灰度图像序列 I_{gray}和手区域图像序列 I_{hand},加速图像运动场的计算.

设在 t 时刻图像点(x,y)处的辐照度为 $E(x,y,t)$,根据光流方程可得

$$\frac{\partial E}{\partial x}u(x,y)+\frac{\partial E}{\partial y}v(x,y)+\frac{\partial E}{\partial t}=0 \tag{3}$$

$u(x,y)$和 $v(x,y)$是 t 时刻(x,y)处的速度函数.使用图像运动的二维仿射模型近似 $u(x,y)$,$v(x,y)$与(x,y)的关系,表示如下

$$U(x,y)=\begin{bmatrix}u(x,y)\\ v(x,y)\end{bmatrix}=\begin{bmatrix}a_0\\ a_3\end{bmatrix}+\begin{bmatrix}a_1 & a_2\\ a_4 & a_5\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} \tag{4}$$

其中,仿射模型的参数 $a_i(i=0,1,\cdots,5)$对于整个运动区域来说是常量,其组合可以表示图像平面内的多种运动(包括形变).具体地说,有水平平移$(u=a_0)$,垂直平移$(v=a_3)$,各向同性的膨胀$(e=a_1+a_5)$,错切形变$(d=a_1-a_5)$,以及旋转$(r=-a_2+a_4)$.定义一个五维运动参数向量 $\boldsymbol{m}[t](\boldsymbol{m}[t]=[u,v,e,d,r])$就可以描述第 t 帧和第 $t+1$ 帧之间的图像运动.

以 $U(x,y)=[u(x,y),v(x,y)]^{\mathrm{T}}$ 近似像素点(x,y)处的运动,则此时等式(3)的残差 $\tilde{r}(x,y)$为

$$\tilde{r}(x,y)=\begin{bmatrix}\frac{\partial E}{\partial x} & \frac{\partial E}{\partial y}\end{bmatrix}\begin{bmatrix}u(x,y)\\ v(x,y)\end{bmatrix}+\frac{\partial E}{\partial t} \tag{5}$$

结合第二节中分割得到的手区域图像序列 I_{hand},仅在 I_{hand}的手区域 R 内,采用 Geman-McClure 鲁棒函数

$$\rho(\tilde{r},\sigma)=\frac{\tilde{r}^2}{\sigma^2+\tilde{r}^2} \tag{6}$$

其中,$\tilde{r}$ 是残差,σ 是尺度参数,$\sigma=\sqrt{3}\tau$,τ 是残差的最大期望值,迭代计算运动参数 $a_i(i=0,1,\cdots,5)$,使区域 R 内所有残差之和最小.这样,不仅大大降低了计算量,而且采用鲁棒统计学,提高了计算的鲁棒性和准确性.同时,在迭代中逐步减小尺度参数 σ,这样的好处是:随着参数 $a_i(i=0,1,\cdots,5)$迭代变得精确,内外点的划分逐渐趋于合理,反过来使得参数 $a_i(i=0,1,\cdots,5)$更加精确.

4　独立分布的多状态高斯概率模型

4.1　手势的时空表观模型

令 L 表示手势的时间长度(即图像序列的帧数).令第 $t(t=0,1,\cdots,L-1)$帧的形状特征是 $\boldsymbol{s}[t]$(3 维),第 t 帧和第 $t+1$ 帧之间的运动特征是 $\boldsymbol{m}[t]$(5 维).定义一个 8 维特征向量 $\boldsymbol{f}[t](\boldsymbol{f}[t]=[\boldsymbol{m}[t],\boldsymbol{s}[t]]^{\mathrm{T}})$,用来统一描述手势的表观特征.该手势的时空表观特征 A,就定义为特征向量 $\boldsymbol{f}[t]$随时间的变化

$$A=[\boldsymbol{f}[0],\boldsymbol{f}[1],\cdots,\boldsymbol{f}[L-2]]^{\mathrm{T}} \tag{7}$$

4.2　独立分布的多状态高斯概率模型

各个手势运动的时间长度 L 各不相同,为了消除时间轴上长度 L 不同的影响,我们对 L 进行规整,统一为 $n(n=4)$个状态.即把手势的时空表观特征 A,按时间等分为 n 个状态,每个状态有 $m(m=8)$个特征参数,这 m 个参数的意义同 $\boldsymbol{f}[t]$.这样,就组成了 $n*m$ 的矩阵 $X=(x_{i,j})_{n*m}$,表示时间规一化手势序列的运动特征和形状特征.

假设各状态独立分布，每个状态的各个特征参数也独立且服从正态分布. 设任意一个手势类模型为 $\lambda(\mu,\sigma)$，$\mu=(\mu_{i,j})_{n*m}$为各个特征参数的数学期望矩阵，$\sigma=(\sigma_{i,j})_{n*m}$为各个特征参数的标准差矩阵. 其中，$\mu_{i,j}$为第 i 个状态第 j 个特征参数的数学期望，$\sigma_{i,j}$为标准差.

由于每个状态的各个特征参数独立分布，且服从正态分布，所以在手势类 λ 中，第 i 个状态第 j 个特征参数出现观测 $x_{i,j}$的概率为

$$p_{i,j}(x_{i,j}|\lambda)=\frac{1}{\sqrt{2\pi}\sigma_{i,j}}\exp\left(-\frac{(x_{i,j}-\mu_{i,j})^2}{2\sigma_{i,j}^2}\right),\quad 1\leqslant i\leqslant n\ 且\ 1\leqslant j\leqslant m \tag{8}$$

则对于手势模型 $\lambda(\mu,\sigma)$，出现完整手势序列观测 X 的概率为

$$p(X|\lambda)=\left(\frac{1}{\sqrt{2\pi}}\right)^{n*m}\exp\left(-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(x_{i,j}-\mu_{i,j})^2}{2\sigma_{i,j}^2}\right)\prod_{i=1}^{n}\prod_{j=1}^{m}\frac{1}{\sigma_{i,j}} \tag{9}$$

$$-2\ln p(X|\lambda)=mn\ln 2\pi+\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(x_{i,j}-\mu_{i,j})^2}{\sigma_{i,j}^2}+2\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\sigma_{i,j} \tag{10}$$

要使 $p(X|\lambda)$最大，也就是使$-2\ln p(X|\lambda)$最小. 所以识别时，对各手势模型 $\lambda(\mu,\sigma)$，只需计算$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(x_{i,j}-\mu_{i,j})^2}{\sigma_{i,j}^2}+2\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\sigma_{i,j}$，得到值最小的即为归属的类别.

5 实验结果与结论

本实验系统中，总共有 12 种手势来控制全景图浏览器中摄像机的运动，包括 6 种平移手势："向上平移 move up (MU)"，"向下平移 move down (MD)"，"向左平移 move left (ML)"，"向右平移 move right (MR)"，"向前平移 move forward (MF)"，"向后平移 move backward (MB)"；6 种旋转手势："向右偏转 yaw right (YR)"，"向左偏转 yaw left (YL)"，"顺时针旋转 roll clockwise (RC)"，"逆时针旋转 roll counterclockwise (RCC)"，"向下翻转 pitch down (PD)"，"向上翻转 pitch up (PU)".

系统采样频率为 10Hz，图像分辨率为 160×120，24 位真彩色，平均每个手势 12 帧. 在 Pentium Ⅱ 266MHz 计算机上，每个手势序列进行手势分割需要 5s，计算帧间运动参数和帧内形状参数需要 7.3s. 采用独立分布的多状态高斯概率模型识别每个手势需要0.012s，这相对于手势的分割和特征参数的计算可以忽略不计. 在实际系统中，手势的分割可以与采样同步进行，这样计算特征参数和识别只需要 7.312s.

12 种手势，每个手势有 30 个样本，15 个用做训练集，另外 15 个用做测试集. 训练集和测试集的识别率如表 1 和表 2.

表 1 训练集识别正确率

手势种类	MU	MD	ML	MR	MF	MB	YR	YL	RC	RCC	PD	PU	average
正确率(%)	100	100	100	100	100	93.3	93.3	100	100	100	93.3	93.3	97.8

表 2 测试集识别正确率

手势种类	MU	MD	ML	MR	MF	MB	YR	YL	RC	RCC	PD	PU	average
正确率(%)	100	100	100	100	93.3	93.3	93.3	100	93.3	93.3	93.3	86.7	95.6

与目前其它基于计算机视觉的手势识别系统相比,本系统对背景几乎没有什么限制,对手势的起始位置也没有要求,并且估计运动参数时采用鲁棒统计学.这些都大大提高了系统的鲁棒性,增强了系统的实用性.独立分布的多状态高斯概率模型,以贝叶斯决策为依据,充分利用形状和运动参数在时间轴上的变化信息,明显提高了识别效果.

当然,在手势分割与识别中也存在一些问题.例如,在手势分割中,如果背景中有皮肤颜色的区域和做手势的手区域相交,那么种子算法就会将该区域也作为手的一部分;还有皮肤颜色的检测容易受到光照环境的影响等等.特征参数提取中,算法耗费时间太多.在手势识别过程中,假定模型 $\lambda(\mu,\sigma)$ 中各状态和状态内各参数概率分布独立,没有考虑它们之间的相关性,在本系统中识别效果已经很不错了,但是如果考虑各个分量之间的相关性,我们相信效果会更好.以后,随着识别手势集的扩大,这一点是需要完善的.

参 考 文 献

1　任海兵,祝远新,徐光祐,林学闾,张晓平.基于视觉手势识别的研究——综述.电子学报,2000,**28**(2):118～121

2　Starner T, Weaver J, Pentland A. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 1998, **20**(12):1371～1375

3　Hyeon-Kyu Lee, Jin H Kim. An HMM-based threshold model approach for gesture recognition. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 1999, **21**(10):

4　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

5　Haibing Ren, Guangyou Xu, Yuanxin Zhu, Xueyin Lin, Linmi Tao. Motion-and-color based hand segmentation and hand gesture recognition. Proceedings of the First International Conference on Image and Graphics, Journal of Image and Graphics(JIG), 5(Suppl):384～388

任海兵　1998 年于清华大学计算机科学与技术系获工学学士学位,同年被保送到清华大学计算机应用专业直读博士学位.

祝远新　1995 年于清华大学计算机科学与技术系获工学学士学位,同年被保送到清华大学计算机应用专业直读博士学位,1999 年获得清华大学计算机应用专业博士学位.同年,进入 University of Missouri-Columbia 读博士后.

徐光祐　教授,博士生导师,1963 年毕业于清华大学自动控制系.曾在美国 Purdue 大学和 Illinois 大学从事访问研究.曾获多项国家和部委的科技进步奖.现是清华大学计算机系人机交互与媒体集成研究所责任教授.目前主要研究领域为计算机视觉、人机交互技术和多媒体技术.

张晓平　1998 年毕业于清华大学计算机科学与技术系,获工学学士学位,同年被保送到清华大学计算机系读硕士学位.

林学闾　教授,博士生导师,目前任清华大学计算机系书记.主要研究领域为计算机视觉和人机交互技术.