ISSN 1000-9825 Journal of Software 软件学报 2000, 11(1): 54~61

基于表观的动态孤立手势识别*

祝远新 徐光祐 黄浴

(清华大学计算机科学与技术系 北京 100084)

E-mail: yuanxinzhu@263.net

摘要 给出一种基于表观的动态孤立手势识别技术. 借助于图像运动的变阶参数模型和鲁棒回归分析, 提出一种基于运动分割的图像运动估计方法. 基于图像运动参数, 构造了两种表观变化模型分别作为手势的表观特征, 利用最大最小优化算法来创建手势参考模板, 并利用基于模板的分类技术进行识别. 对120个手势样本所做的大量实验表明, 这种动态孤立手势识别技术具有识别率高、计算量小、算法稳定性好等优点.

关键词 计算机视觉, 人机接口, 手势识别, 图像运动模型, 鲁棒回归, 动态规划匹配

中图法分类号 TP391

尽管语音识别作为一个研究热点已经有几十年的历史, 但直到最近几年, 基于计算机视觉的手势识别才逐渐引起研究人员的关注和兴趣. 手势识别研究的主要目的就是把手势这种既自然又直观的交流方式引入人机接口(human computer interface)中, 实现更符合人类行为习惯的人机接口[1]. 此外, 手势识别还可用于虚拟现实、三维设计、临场感、可视化、医学研究、手语理解等领域. 手势识别问题的解决方法对于表情识别、唇读、步态识别、时空纹理分类、视觉导航、图像拼接和基于内容的视频检索等研究都有直接推广的意义.

现有的手势识别方法可以分为两大类: 基于三维手/手臂建模的方法和基于表观建模的方法[2]. 在手势建模方面, 最常用的方法是给手或手指建立三维模型[3], 但是, 繁重的计算使得这类方法非常困难, 而且离实际应用也很遥远. 基于表观的方法主要研究如何直接利用图像序列里的表观变化来识别手势, 它的着重点不是手或手臂的静止三维结构, 而是运动所引起的图像序列里的表观变化. 从文献来看, 已经有人开始研究基于表观的手势识别[4~6]. 文献[4]提出用一组相似值(通过与一组二维空间视图或三维时空视图作相关运算得到)构成的向量作为作手势的表观特征. 然而, 当有多个用户或者单个用户所作手势不能用过去的时空模式准确描述时, 基于这种表观特征的识别方法就会失效, 因此他们的方法缺乏通用性. 文献[5]通过抽取图像序列里的运动边界、跟踪边界的运动并作聚类分析进行手势识别, 但是没有给出识别结果, 也没有利用手势的形状、纹理等其他信息.

本文提出另一种基于表观的手势识别技术. 首先, 基于图像运动的变阶参数模型和鲁棒回归, 本文提出一种基于运动分割的帧间图像运动估计方法. 其次, 基于帧间图像运动参数, 本文创建了两种不同的表观变化模型——时空表示和总体图像运动表示, 分别用于手势识别, 并依据实验结果对二者的性能进行比较分析. 然后, 本文指出了一条如何把运动、形状、颜色、纹理等信息统一起来进行手势识别的途径. 最后, 我们设计并实现了对12种手势进行在线识别的实验系统, 识别率超过90%.

1 基于运动分割的帧间图像运动估计

1.1 图像运动模型

对于基于视觉的孤立手势识别系统, 用户作手势时一般是面向系统的, 因而我们不必考虑运动估计过程中

* 本文研究得到国家863高科技项目基金(No. 863-306-03-01)资助. 作者祝远新, 1972年生, 博士, 主要研究领域为计算机视觉与模式识别, 多模式人机交互, 多媒体技术与系统. 徐光祐, 1940年生, 教授, 博士生导师, 主要研究领域为计算机视觉, 分布式多媒体, 人机交互技术. 黄浴, 1967年生, 博士后, 主要研究领域为计算机视觉, 人机交互技术.

本文通讯联系人: 祝远新, 北京 100084, 清华大学计算机科学与技术系信息处理与应用教研组

本文 1998-11-10 收到原稿, 1999-02-01 收到修改稿

存在的三维遮挡或重现问题. 参数化的图像运动模型对一个区域里的图像运动作明确假设并通常用一个低阶多项式来表示该运动[7]. 常用的参数化图像运动模型有平移模型、仿射模型和平面模型. 例如, 式(1)给出的就是图像运动的平面模型

$$\mathrm{u}(\mathrm{x})=\begin{bmatrix}u(x,y)\\ v(x,y)\end{bmatrix}=\begin{bmatrix}a_0+a_1x+a_2y+a_6x^2+a_7xy\\ a_3+a_4x+a_5y+a_6xy+a_7y^2\end{bmatrix}, \tag{1}$$

其中 a_i 是常数(对一个运动区域而言), $\mathrm{u}(\mathrm{x})$ 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. 为了便于书写和描述, 我们定义

$$\begin{aligned}&\mathrm{X}(\mathrm{x})=\begin{bmatrix}1&x&y&0&0&0&x^2&xy\\ 0&0&0&1&x&y&xy&y^2\end{bmatrix},\quad \mathrm{T}=(a_0,0,0,a_3,0,0,0,0)^T,\\ &\mathrm{A}=(a_0,a_1,a_2,a_3,a_4,a_5,0,0)^T,\quad \mathrm{P}=(a_0,a_1,a_2,a_3,a_4,a_5,a_6,a_7)^T,\end{aligned} \tag{2}$$

其中 T, A 和 P 分别表示图像运动的平移、仿射以及平面模型的参数. 就图像运动而言, 各参数 $a_i(i=0,1,\ldots,7)$ 有各自的物理意义, 即可以用这些参数或它们的组合分别表示平移运动(u)、垂直运动(v)、各项同性的膨胀(e)、形变(d), 以及绕观察方向的旋转(r)、偏转(y)、俯仰(p)等. 具体意义如式(3)所示

$$u=a_0,\quad v=a_3,\quad e=a_1+a_5,\quad d=a_1-a_5,\quad r=-a_2+a_4,\quad y=a_6,\quad p=a_7. \tag{3}$$

1.2 鲁棒回归

给定图像区域的运动模型之后, 我们采用鲁棒回归策略去回归模型参数[8]. 鲁棒统计学的主要目的是回归那种能最好地拟合绝大多数数据的模型, 并同时检测出参数回归分析的"内点"(即符合模型的数据点)和"外点"(即与模型出入较大的数据点).

令 $\mathbf{R}$ 是分析区域内像素点的集合, Θ 是模型参数向量, 那么光流估计中的亮度恒定假设就可表述为

$$I(x,t)=I(x-\mathrm{X}(x)\Theta,t+1),\quad \forall x\in\mathbf{R}. \tag{4}$$

这里, I 是亮度函数, t 表示时间. 把式(4)右边泰勒展开、化简, 并去掉高于一次的项, 得到式(5).

$$\nabla\mathrm{I}^T(\mathrm{X}(x)\Theta)+I_t=0,\quad \forall x\in\mathbf{R}. \tag{5}$$

这里, $\nabla\mathrm{I}=[I_x,I_y]^T$ 是图像亮度梯度; I_x, I_y 和 I_t 分别是图像亮度关于空间维和时间的偏导数. 为估计参数 Θ, 我们需要针对某个误差范数 ρ 最小化式(6)中的目标函数

$$E(\Theta)=\sum_{x\in\mathbf{R}}\rho(\nabla\mathrm{I}^T(\mathrm{X}(x)\Theta)+I_t,\sigma), \tag{6}$$

其中 σ 是一个尺度参数. 因为人手既不是平面也不是刚体, 如果在这种情况下运用最小二乘方法, 那么估计出来的运动参数必然是不准确的, 因而 ρ 最好是某种能容许一定总体误差或外点的误差范数. 这里, 我们采用 Geman-McClure 函数, 如式(7)所示, 式中残差 $r=\nabla\mathrm{I}^T(\mathrm{X}(x)\Theta)+I_t$.

$$\rho(r,\sigma)=\frac{r^2}{\sigma^2+r^2}. \tag{7}$$

与截断二次函数相比, Geman-McClure 函数在内点和外点之间提供了更平滑的过渡.

我们用带有连续策略的超松弛算法使式(6)里的目标函数达到最小. 概括地说, 如果考虑目标函数 $E(\Theta)$ 关于 a_i 的导数, 那么为最小化 $E(\Theta)$, 第 $n+1$ 步的迭代更新方程就是

$$a_i^{n+1}=a_i^n-\omega\frac{1}{T(a_i)}\cdot\frac{\partial E(\Theta)}{\partial a_i}. \tag{8}$$

其中 $0<\omega<2$ 是超松弛参数, 它用于在第 $n+1$ 步对 a_i^{n+1} 的估计进行"过分"修正. 当 $0<\omega<2$ 时, 可以证明该方法是收敛的, 但是收敛速率对 ω 的具体取值很敏感. $T(a_i)$ 是 $E(\Theta)$ 的二阶偏导数的上界, 即

$$T(a_i)\geq\frac{\partial^2E(\Theta)}{\partial a_i^2}. \tag{9}$$

我们在迭代过程中使用了连续策略, 即在每一次迭代中, 根据 $\sigma_{n+1}=0.95\sigma_n$ 来减小尺度参数 σ 的值. 这样做的结果是, 起始时分析区域内的所有数据点都被看作内点, 而随着迭代的进行开始出现外点, 而且外点的影响逐渐被减小.

1.3 多分辨率策略

为了准确地估计帧间大运动量, 我们引入多分辨率策略[7]. 首先构造高斯金字塔, 然后从最粗一级空间分辨率开始, 初始运动参数 Θ_0 设为零, 估计出运动参数增量 $\Delta\Theta_0$, 把得到的运动参数 $\Theta_0+\Delta\Theta_0$ 投射到下一个分辨率级就得出下一个分辨率级的初始运动参数 Θ_1, 根据 Θ_1 把该级 t 时刻的图像向 $t+1$ 时刻的图像配准, 然后再利用配准后的图像和 $t+1$ 时刻的图像估计该级的运动参数增量 $\Delta\Theta_1$. 重复这个过程, 直至估计出最高分辨率级的运动参数

1.4 基于运动分割的帧间运动估计

我们知道, 尺度参数 σ 在鲁棒回归中发挥着重要作用. 文献[8]指出, 选择 $\sigma=C\tau$ 作为 σ 的初始值将导致一个凸优化问题. 对特定鲁棒误差范数而言, C 是一个常数, 例如, 当 ρ 是 Geman-McClure 函数时, $C=\sqrt{3}$. τ 是最大期望残差. 然而, 在具体实验时, 他们并没有这么做, 也没有给出解释, 只是指出对于 σ 在一定幅度内的变化, 求得的解是相当稳定的. 在我们看来, 有两点理由促使他们放弃把 $C\tau$ 作为 σ 的初始值: 首先, 初始残差分布无法预料, 因此得出的初始解可能很糟糕; 其次, 选择 $C\tau$ 作为 σ 的初始值将需要更多的迭代次数, 因而消耗更多的时间(当分析区域较大, 如整幅图像时尤其如此). 本文给出了一个更合理的方案来确定 σ 的初始取值.

本文提出的基于运动分割的图像运动估计方法由两步组成, 每一步包含一次鲁棒回归.

第 1 步的目标是把图像里的运动区域(物体)从复杂背景里初步分割出来. 我们假设背景保持基本静止(对于手势识别系统, 该假设在绝大多数情况下是成立的), 然后选择整幅图像作为分析区域, 并选用平移运动模型进行第 1 次鲁棒回归. 为了自动确定 σ 的初始取值并进行高效率的迭代, 我们选择 CT 作为 σ 的初始值, 其中, C 的意义同上, 而 T 则是根据文献[9]提出的阈值选择算法从两帧图像的差图像的灰度直方图计算出来的一个阈值, 它是把差图像里的灰度分为两类的最优阈值. 这样做, 可以从一开始就把绝大多数运动点(物体点)作为外点处理, 因而迭代效率更高, 收敛更快. 因为绝大多数数据点都是背景像素点, 所以估计出的运动(尽管运动量非常小)是背景的运动. 根据回归出的运动参数以及 σ 的最终值对残差进行分析(即检测外点), 就可以得到运动物体的粗糙分割(其中还包括一些噪声点).

第 2 步的目标是实现准确的帧间运动估计, 并且得到运动物体的精细分割. 我们使用二维仿射模型或平面模型来近似物体在图像平面的运动, 当物体运动引起的深度变化与成像距离相比很小时, 这种近似是合理的. 我们把第 1 步得到的粗糙分割作为新的分析区域, 并选用仿射或平面模型进行第 2 次鲁棒回归. 由于分析区域较小, 而且绝大多数数据点都属于运动物体, 因此我们把 $\sqrt{3}\tau$ 作为 σ 的初始值(τ 的意义同上, 即最大期望残差). 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. 然后, 通过分析残差(即检测内点)就可以得到运动物体的精细分割.

2 手势的表观特征

手势识别系统的性能很大程度上取决于手势特征的抽取与构造. 基于帧间运动参数, 本文构造出两种不同的表观变化模型(时空表示和总体图像运动表示), 分别作为手势的表观特征用于识别.

2.1 手势的时空表示

给定一个包含孤立手势的图像序列, $g^T=\{I_1, I_2, \ldots, I_T\}$, 其中 T 表示该序列的时间长度, $I(t)$ ($t=1, 2, \ldots, T$) 是序列里的第 t 帧亮度图像. 如果把第 $t-1$ 帧和第 t 帧之间的图像运动(仿射或平面)参数作为分量构成的参数向量记为 $\mathrm{P}[t]$, 那么对应手势的时空表示 P^T 就可以被建模为

$$\mathrm{P}^T=\{\mathrm{P}[1], \mathrm{P}[2], \ldots, \mathrm{P}[T-1]\}. \tag{10}$$

2.2 手势的总体图像运动表示

如前所述, 图像运动模型的参数 a_i ($i=0, 1, 2, \ldots, 7$) 都有明确的物理意义, 因此我们可以定义一个有确定物理意义的运动向量来描述帧间图像运动. 记 $I(t-1)$ 和 $I(t)$ 之间的运动向量为 $\mathrm{m}[t]$, 则 $\mathrm{m}[t]$ 的定义为

$$\mathrm{m}[t]=[u, v, e, d, r, y, p]^T. \tag{11}$$

为消除不同用户作手势时存在的速率差异, 我们为上面的图像序列(g^T)创建另一个表观变化模型: 总体图

像运动表示. 式(12)给出了总体图像运动表示的定义

$$\Sigma = [U, V, E, D, R, Y, P]^T = \hat{m} \cdot \sum_{t=1}^{T-1} \mathrm{m}[t], \quad \mathrm{M} = \frac{1}{\|\Sigma\|}\Sigma \tag{12}$$

其中M 是序列的总体图像运动表示, 它是把序列里所有(每两)帧间运动向量的分量加权后的累加和向量规一化以后得到的向量; $\hat{m}$ 是加权因子"·"表示向量内积运算; U, V, E, D, R 和 P 分别表示总体图像运动所表示的水平位移、垂直位移、各向同性膨胀、变形、绕观察方向的旋转、偏转以及俯仰等分量.

3 手势识别

3.1 动态规划匹配和ODPM 距离

不同用户作手势时存在的速率差异会在时空表示的时间轴上引起非线性波动, 如何消除这些非线性波动是手势识别中的一个重要问题. 语音识别的实践表明, 任何线性变换从本质上说都不能很好地处理高度复杂的非线性波动. 我们引入在语音识别里所用的最优动态规划匹配(optimal dynamic programming matching, 简称ODPM)[10]来度量两个时空表示之间的最小距离. 动态规划匹配是具有非线性时间规一化效果的模式匹配算法, 使用某种指定属性的非线性规整函数对时间轴上的波动近似建模, 通过弯曲其中一个模式的时间轴使之与另一个模式达到最大程度的重叠(此时的残差距离最小), 从而消除两个时空表示模式之间的时间差别. 两个模式之间的最小化的残差距离就是它们之间的时间规一化后的距离(本文称为ODPM 距离). 利用动态规划技术可以高效率地实现上面的最小化过程.

3.2 创建各手势的参考模板

假设给定手势训练集中有 L 种孤立手势, 每种手势有 J 个样本. 对于训练中的每个样本, 我们分别计算它的时空表示 $\mathrm{P}^T[l, j]$ 和总体图像运动表示 $\mathrm{M}(l, j)$, 其中 $l(l = 1, 2, \ldots, L)$ 是手势类编号; $j(j = 1, 2, \ldots, J)$ 是样本编号; T 表示第 l 种手势的第 j 个样本的时空表示所包含的参数向量的个数. 创建手势的参考模板就是从每种手势的 J 个时空表示和 J 总体图像运动表示出发, 找出最能反应该手势表观特征分布的一个时空表示和一个总体图像运动表示分别作为该手势的时空表示模板和总体图像运动表示模板.

如果把第 l 种手势的时空表示模板记为 $\hat{\mathrm{P}}^{\hat{T}}[l]$, 那么创建时空表示模板问题就转化为选取使 $\hat{\mathrm{P}}^{\hat{T}}[l]$ 在给定的训练集上满足某种最优标准. 基于两个手势的时空表示之间的ODPM 距离度量, 我们用最小最大优化算法[11]为每种手势创建一个时空表示模板.

由于一个手势样本的总体图像运动表示反映了贯穿(包含该手势的)整个图像序列的各种图像运动, 因此, 我们把第 l 种手势的总体图像运动表示模板定义为该手势的 J 个样本的总体图像运动表示的平均值 $\hat{\mathrm{M}}[l]$, 即

$$\hat{\mathrm{M}}[l] = \frac{1}{J}\sum_{j=1}^{J} \mathrm{M}(l, j). \tag{13}$$

3.3 基于模板分类的识别算法

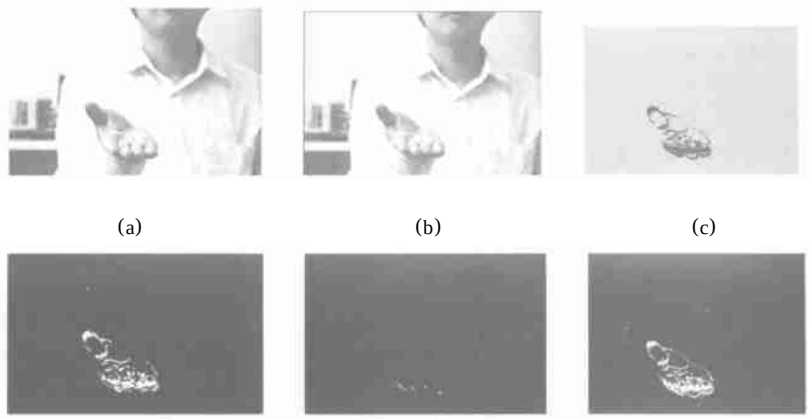
有了 $\hat{\mathrm{P}}^{\hat{T}}[l]$ 和 $\hat{\mathrm{M}}[l](l = 1, 2, \ldots, L)$, 我们利用模板分类技术进行手势识别. 假设取手势的时空表示模板作为识别时的参考模板, 那么, 当输入一个未知手势时, 系统首先计算它的时空表示, 然后计算该时空表示与库中各时空表示模板之间的ODPM 距离, 并找出最小距离. 当最小距离不超过某个预置的门限时, 该手势就与距离最近的那个时空表示模板所对应的手势属于同一类, 否则拒识.

如果我们选取手势的总体图像运动表示模板作为识别时的参考模板, 那么识别过程基本不变, 只是需要计算未知手势的总体图像运动表示, 并且使用欧氏空间距离而不是ODPM 距离. 在实验时, 对于训练集里的每一种手势, 我们既计算它的时空表示模板, 也计算它的总体图像运动表示模板, 而且对分别用两种模板作为参考模板而得到的识别结果进行比较分析.

4 运动估计举例

图 1 是帧间运动估计中的一个实例, 图 1(a)和(b)分别是一个包含"向上"手势的图像序列里的第 2 帧和第

3 帧图像. 选用平移模型对整幅图像进行第 1 次鲁棒回归得到的运动参数是{- 0.06184, 0, 0, - 0.1446, 0, 0, 0, 0}. 图 1(c)中的灰色像素点就是根据第 1 次回归结果检测出的内点, 它包括静止的和有微小扰动的背景点以及某些物体点. 其中的黑色像素点是根据第 1 次回归结果检测出的外点, 显然都是一些运动量较大的物体点, 它们的集合就构成了运动物体的粗糙分割. 把物体的粗糙分割作为新的分析区域, 并选用平面模型进行第 2 次鲁棒回归得出的运动参数是{- 0.3644, 0.004, 0.0066, - 2.2835, - 0.0012, - 0.0086, 0, 0}. 根据第 2 次回归结果检测出的内点就构成了运动物体的精细分割, 如图 1(d)所示, 外点如图 1(e)所示, 它们是一些噪声点. 图 1(f)是把图 1(a)和图 1(b)之间的差图像二值化以后得到的结果图像. 比较图 1(d)和(f), 不难看出, 图 1(f)中有很多噪声轮廓点, 如头部轮廓点、手臂轮廓点等, 而在图 1(d)中则没有. 图 1 表明, 我们提出的基于两次鲁棒回归的运动估计方法既能准确地估计帧间图像运动, 又能同时得到运动物体的精细分割.



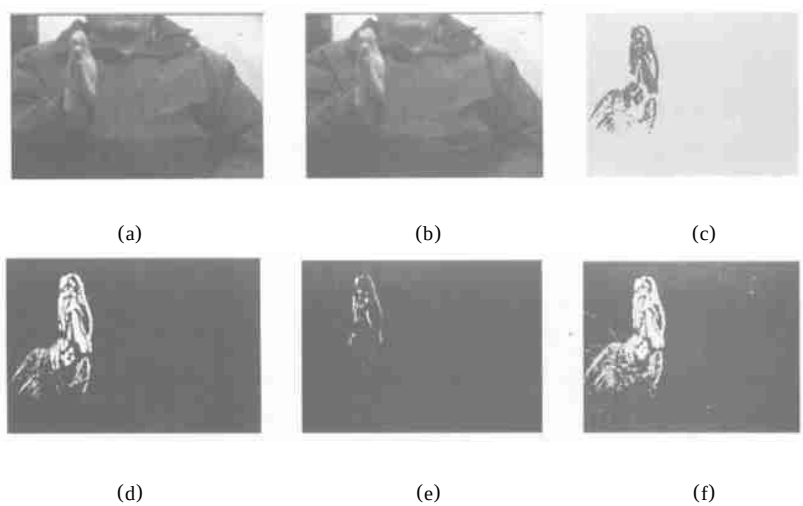
(a)　(b)　(c)

(d)　(e)　(f)

Fig. 1　An instance of the motion estimation between frames

图 1　帧间运动估计的一个实例

图 2 是帧间运动估计的另一个实例. 图 2(a)和(b)分别是一个包含"向右"手势的图像序列里的第 2 帧和第 3 帧图像. 第 1 次回归出的参数值为{- 0.1217, 0, 0, - 0.0042, 0, 0, 0, 0}, 第 2 次回归出的参数值为{- 2.7344, 0.0008, - 0.01120, - 0.1343, - 0.0037, - 0.0160, 0, - 0.0003}. 图 2(c)～(f)表示的含义分别与图 1 中对应图像的含义相同.



(a)　(b)　(c)

(d)　(e)　(f)

Fig. 2　Another instance of the motion estimation between frames

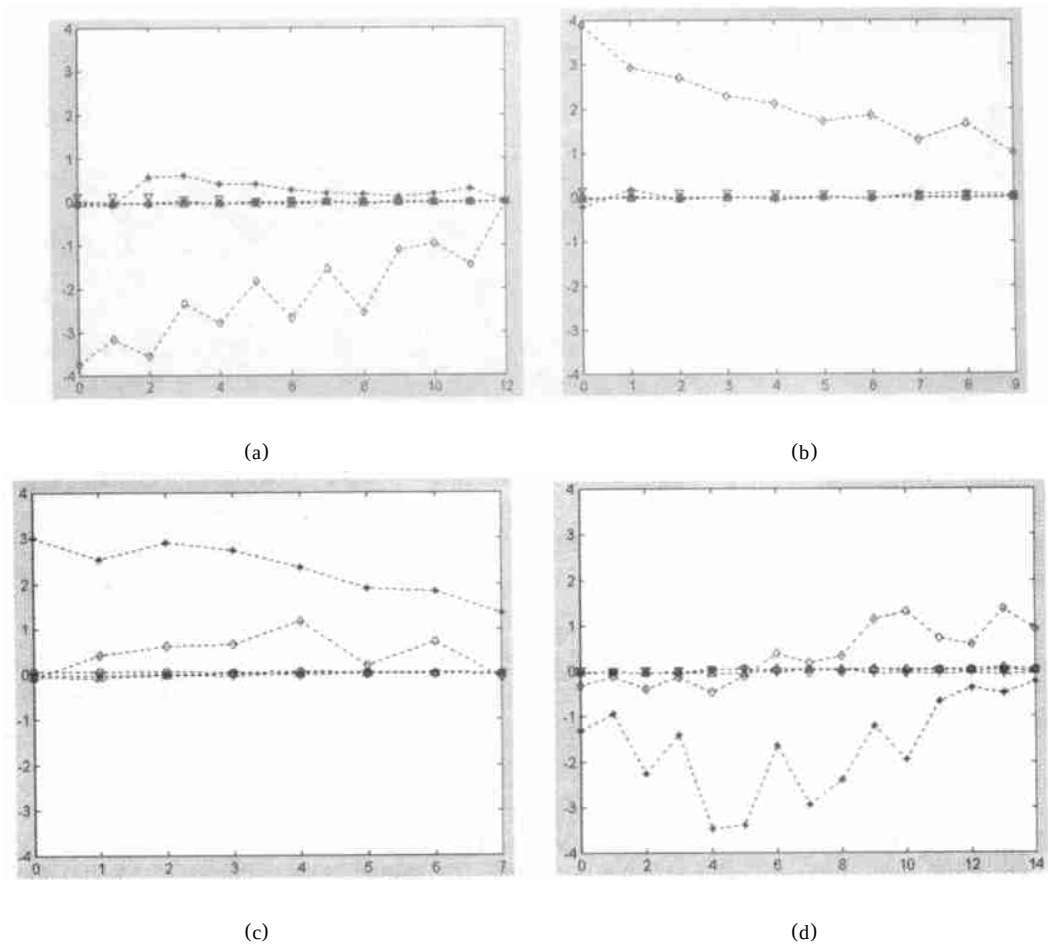
图 2　帧间运动估计的另一个实例

5 识别结果

我们设想用一组手势实现三维鼠标的功能, 从而实现更自然、更符合人类行为习惯的人机接口. 因此, 在我们的实验系统里选用的 12 种手势分别是“向上”(MU)、“向下”(MD)、“向左”(ML)、“向右”(MR)、“向前”(MF)、“向后”、“向左偏转”(YL)、“向右偏转”(YR)、“顺时针旋转”(CC)、“反时针旋转”(CCC)、“向下俯”(PD)和“向上仰”(PU).

我们做了大量的实验来验证和评价本文提出的时空表示模型、总体图像运动表示模型以及训练和识别算法的有效性. 我们邀请 10 位实验者坐在摄像机前作上面的 12 种手势, 于是得到包含 120 个图像序列(即每种手势有 10 个样本)的训练集. 每个序列包含一个手势, 持续时间约 1s(采样率为 10Hz). 图像是 256 级的灰度图像, 大小是 160×120. 在 PII(266MHz)的 PC 机上, 帧间运动估计的平均时间约 450μs, 识别一个手势的时间约 4s.

图 3 以帧间运动参数的轨迹形式描述了从 4 个图像序列计算出的时空表示. 每个图像序列包含 1 个手势, 4 个序列包含的手势分别是MU,MD,ML 和MR. 参数 a_0, a_1, a_2, a_3, a_4 和 a_5 的轨迹(a_6, a_7 的值很小, 故忽略)都是点划线, 其不同之处在于, 不同参数轨迹的点划线上每个数据点所用的符号不同, 分别是“*”、“°”、“+”、“◇”、“△”和“▽”. 从图 4 可以看出, 本文所创建的手势时空表示是高度可区分的, 或者说时空表示的类间离散性大. 图中(a), (b), (c)和(d)分别对应于MU,MD,ML 和MR 手势的某个样本的时空表示.



(a) (b) (c) (d)

Fig. 3 Spatial-Temperal representation of hand gesture described by trajectory of motion parameters between frames

图 3 以帧间运动参数的轨迹形式描述的手势时空表示

选取时空表示作为手势表观特征进行识别, 在训练集上实现的识别率见表 1. 表 2 是用总体图像运动表示作为手势表观特征时在训练集上得到的识别率. 比较表 1 和表 2, 不难看出, 把总体图像运动表示作为手势表观特征时, 识别率更高. 除此之外, 后者作为手势表观特征时识别速度更快. 表 3 是用总体图像运动表示作为手势表观特征时在测试集(每种手势有 10 个测试样本)上得到的识别率.

Table 1 Recognition rates achieved on training set with spatio-temporal representations as features of hand gestures

表 1 把时空表示作为手势表观特征时在训练集上得到的识别率

Hand gestures①	MU	MD	ML	MR	MF	MN	YL	YR	CC	CCC	PD	PU	Average③
Recognition rates②(%)	90	90	90	100	80	80	100	80	90	90	80	90	88.3

①手势, ②识别率, ③平均.

Table 2 Recognition rates achieved on training set with overall image motion representations as features of hand gestures

表 2 把总体图像运动表示作为手势表观特征在训练集上得到的识别率

Hand gestures①	MU	MD	ML	MR	MF	MN	YL	YR	CC	CCC	PD	PU	Average③
Recognition rates②(%)	90	100	100	100	90	80	90	100	90	100	100	100	95

①手势, ②识别率, ③平均.

Table 3 Recognition rates achieved on testing set with oveall image motion representations as features of hand gestures

表 3 把总体图像运动表示作为手势表观特征在测试集上得到的识别率

Hand gestures①	MU	MD	ML	MR	MF	MN	YL	YR	CC	CCC	PD	PU	Average③
Recognition rates②(%)	100	100	100	100	90	80	80	90	90	100	70	80	90

①手势, ②识别率, ③平均.

6 结论和展望

利用图像运动的变阶参数模型和鲁棒回归, 本文论述了一种基于运动分割策略的帧间运动估计方法. 由于分析区域的像素点很多(因而约束多), 而要估计的参数又很少(最多 8 个), 所以参数化的图像运动模型对分析区域的图像运动提供了充分约束, 数值稳定性高. 选择变阶参数模型(平移、仿射和平面)使得计算量小, 计算效率高. 此外, 利用鲁棒回归求解运动参数时, 位于运动边界的错误度量被看作"外点", 因而它们对解的影响被减弱了.

利用帧间运动参数, 本文创建了时空表示和总体图像运动表示两种表观变化模型作为手势的表观特征, 分别用于手势识别. 大量实验表明, 把两种模型作为手势的表观特征所得到的识别率都超过了 88%, 但是在使用后者时所得识别率更高, 而且识别速度更快.

既然作为运动分析的"副产品", 我们得到了运动物体的精细分割, 我们将对分割出的运动物体的形状、颜色以及纹理进行分析, 以便创建更富有描述能力的表观变化模型, 从而进一步提高手势的识别率. 要想把手势应用于人机接口, 还必须研究手势流的自动检测和自动分割.

参考文献

1 Zhu Yuan-xin, Xu Guang-you, Yu Zhi-he. Vision-Based hand gesture interpretation: a survey. In: Wu Quan-yuan, Zhang Yue-liang eds. Proceedings of the 3rd National Conference on Artificial Interface and Artificial Application. Beijing: Publishing House of Electronics Industry, 1997. 279~284.
(祝远新, 徐光祐, 俞志和. 基于计算机视觉的手势解释方法研究. 见: 吴泉源, 张跃良编. 智能接口与智能应用新进展——第 2 届全国智能接口与智能应用学术会议论文集. 北京: 电子工业出版社, 1997. 279~284)

2 Pavlovic V, Sharma R, Huang T S. Visual interpretation of hand gestures for human-computer interaction: a review.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 677～695

3 Heap T, Hogg D. Towards 3D hand tracking using a deformable model. In: Irfan Essa ed. Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition. Los Alamitors, CA: IEEE Computer Society Press, 1996. 140～1452

4 Darrell T J, Essa I A, Pentland A P. Task-Specific gesture analysis in real-time using interpolated views. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 16(12): 1236～1242

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

6 Zhu Yuan-xin, Huang Yu, Xu Guang-you *et al*. Motion-Based segmentation scheme to feature extraction of hand gestures. In: Zhou J, Jain A K, Zhang Tian-xu *et al* eds. Proceedings of SPIE, Vol 3545. Washington, DC: SPIE, 1998. 228～231

7 Bergen J R, Keith P, Hanna J *et al*. Hierarchical model-based motion estimation. In: Sandini G ed. Proceedings of the 2nd European Conference on Computer Vision. Berlin: Springer-Verlag, 1992. 237～252

8 Black M J, Anandan P. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Computer Vision and Image Understanding, 1996, 63(1): 75～104

9 Otsu N. A threshold selection method from gray-level histogram. IEEE Transactions on Systems, Man and Cybernetics, 1979, 9(1): 62～66

10 Sakoe H, Chiba S. Dynamic programming optimization for spoken word interpretation. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): 43～49

11 Rabiner L R. On creating reference templates for speaker independent interpretation of isolated words. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): 34～41

Appearance-Based Dynamic Hand Gesture Recognition from Image Sequences with Complex Background

ZHU Yuan-xin XU Guang-you HUANG Yu

(*Department of Computer Science and Technology Tsinghua University Beijing* 100084)

Abstract In this paper, the authors present an appearance-based approach to dynamic hand gesture recognition. A motion-based segmentation scheme for image motion estimation is proposed using variable-order parameterized models of image motion and robust regression. Based on image motion parameters, two different appearance change models of hand gestures are created. Template-Based classification technique is then employed to perform hand gesture recognition in which reference templates are created with a min-max type of optimization. A series of experiments on 120 image sequences show that high recognition rate, low computation load, and high stability can be achieved with the proposed methods.

Key words Computer vision, human computer interface, hand gesture recognition, image motion model, robust regression, dynamic programming matching.