# 适用于机器人视觉的手势识别系统

汤志彦　冯　哲
（复旦大学计算机科学与工程系，上海 200433）

**摘　要**　手势是一种高效的人机交互和设备控制的方式，基于视觉的手势识别是人机交互、模式识别等领域的一个富有挑战性的研究课题。文章提出并实现了一个可用于与机器人交互的静态手势检测和识别系统。该系统用摇动检测的方法定位人手；用基于现场采样得到的肤色模型进行手的分割；用简化并改进的 CAMSHIFT 算法对手势进行跟踪；最后用模式识别的方法提取简单特征进行识别。实验证明，该系统快速、稳定而有效。

**关键词**　机器人交互　跟踪　手势识别

**文章编号** 1002-8331-(2005)16-0051-04　**文献标识码** A　**中图分类号** TP18

# A Gesture Recognition System Used on Robot Vision

**Tang Zhiyan　Feng Zhe**
(Computer Science and Engineering Dept.,Fudan Univ.,Shanghai 200433)

**Abstract**: Gesture is a kind of effective human-computer interaction and device control techniques.Vision-based recognition of hand gestures is a challenging project in computer science.This paper puts forward and realized a practical gesture system which can be used in controlling robot and HCI.It locates hand by waving detection,segment gesture using present skin color model,track gesture using simplified CAMSHIFT tracker and recognize gesture by simple features finally.Experiments show that the system proposed is fast,stable and effective.

**Keywords**: human-robot interaction,tracking,gesture recognition

## 1　引言

在人机交互和设备控制中，可以选择的方式有控制器、声音、手势等。其中手势自然舒适且符合用户的交互习惯，和声音控制一起逐渐取代了各种控制器用来作为和智能机器人的交互方式。因此手势的检测和识别已成为人机交互及模式识别领域里的一项重要研究内容。

手势可以分为动态手势和静态手势。动态手势定义为手或手指运动的轨迹，是时域上的问题[7]。而静态手势则通过某一时刻的特定手型传递意义，是空域上的问题[8]。文中介绍的系统主要针对静态手势的检测和识别问题。

一个完整的手势系统，必须首先定位做手势的主体即人手。视觉上的定位技术分为基于颜色定位和基于运动定位两种。绝大多数的颜色定位技术依赖于直方图匹配或者利用皮肤的训练数据建立查找表的方法[1]，其主要缺点是在不同的条件下（光照，摄像设备不同）皮肤颜色变化较大，且这种定位开销也很大。基于运动的定位[2]需要有一个假设，即图像里的主要运动分量是手臂运动。

静态手势的识别特别依赖于手势特征的选取。手势特征建立在图像属性的基础上。图像属性参数包括：轮廓、边界、图像矩、图像特征向量以及区域直方图特征等等。文献[3]使用了基于边界的轮廓特征。由于图像矩计算简单，因此也常被用作图像属性参数。其它常被使用的属性参数还包括 Zernike 矩、方向直方图[4]等等。

但是一般来说，在基于机器人的应用中，可以利用的计算资源是非常有限的。目前一些比较有效的手势识别算法所耗费的资源一般都会超出机器人应用中所能容忍的上限。因此，该文主要考虑静态手势识别算法的时间复杂度和空间复杂度，提出了一个适用于机器人控制和交互的静态手势识别系统。

文章其它部分组织如下：第 2 节为系统综述，论述本系统实现的大致框架；第 3 节论述了如何利用人手的摆动检测和定位人手；第 4 节介绍了本系统使用的肤色模型，该模型用于跟踪运动手的位置以及肤色分割；第 5 节使用简化并改进的 CAMSHIFT 算法对人手进行跟踪；第 6 节介绍如何利用手部肤色模型对手势区域进行分割；第 7 节分析了手势识别的方法。最后结合试验对该文提到的方法和系统进行分析和总结。

## 2　系统分析与技术方案

图 1 描述了该文提出的基于视觉的实时静态数字手势识别系统的框架，可用于智能机器人和人之间的交互。

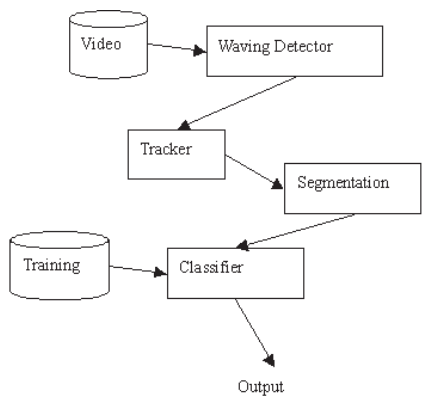


图 1　静态手势系统

一个静态手势系统主要包括两个部分：手势的检测和手势的识别。

和人脸检测类似，在复杂的场景中检测出人手区域同样是

相当复杂和费时的工作。该文中采用了一种结合运动定位和颜色定位的方法对人手区域进行定位。如图 1 所示，系统首先通过摇动检测器在输入的视频中定位手在复杂背景中的大致位置，产生初始的手的候选区域。然后使用简化并改进了 CAMSHIFT 跟踪器对手势进行跟踪，来不断更新准确的手部区域。利用基于现场采样得到的皮肤颜色模型对手部区域进行分割，得到二值化的手势图。

二值化的手势图给出了手部区域的具体位置。在手势识别的时候，系统对手势二值图提取特征，然后利用已训练好的分类器和提取出的特征向量进行手势分类。

## 3 摇动检测与人手定位

考虑到机器人应用中对资源的限制，我们要求用户先通过摇动手给机器人一个信号[2]。如图 2 所示，当人手摇动的时候，人手经过的区域内像素的平均亮度值必将有连续的剧烈的波动变化。这种变化是图像中其他区域所不具备的。这样系统只需要在输入的一段连续的帧的序列上，找出那些变化比较大的区域，就可以获得大致的手部摇动的位置。

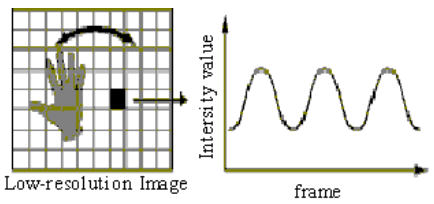


图 2 手部摇动时像素的亮度变化

在系统中首先将每一帧的图像分隔为若干个大小为 $m*n$ 的子块，对第 $t$ 帧的每个子块 $(i,j)$，用下面的式子来计算它在连续 10 帧图像内的变化程度。

$$S(i,j,t)=\sum_{n=0}^{9}\{[w(n)\times|I(i,j,t-n)-I(i,j,t-n-1)|]\}$$

其中，$I(i,j,t)$ 表示了第 $t$ 帧上子块 $(i,j)$ 的平均亮度（如图 3 所示）：

$$I(i,j,t)=\frac{1}{m\times n}\sum_{p=0}^{m-1}\sum_{q=0}^{n-1}luminance(p,q,t),(p,q)\in block_{i,j}$$

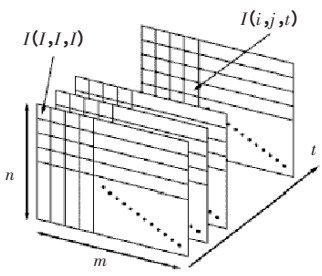


图 3 $I(i,j,t)$ 的直观表示

为了反映时间的影响，在计算每个块在连续 10 帧上的变化程度的累积时，给每一个帧赋予一个随时间递增的权重 $w(n)$，$n=0,\cdots,9$。

经过上述计算，$S(t)$ 最大的块就是最近 10 帧之内变化最为剧烈的区域。

上述过程中使用的完全是基于运动的定位。然后，用一些简单的肤色判定规则，即如果此区域周围有相当的像素近似于经验的肤色，则判定这个块周围为人手的区域。如图 4 所示，Y 方框区域为用摇动检测找到的手部区域。

图 4 摇动检测得到的手部区域

## 4 手部肤色模型

如何把人手和背景分割开呢？最简单的一种办法就是对用户做手势时的场景加上一些限制。例如规定用户必须使用纯色的手套，或者背景一定要使用单色的墙壁等。这种方法由于背景比较简单，并且用户手部区域和背景区域之间的反差比较大，因此实现比较容易，并且可以达到很好的效果。但是，这种方法需要对用户做出太多的限制，适用性不强。

在图 5 所示的 HSV（Hue-Saturation-Value）色彩模型中，颜色用色调、饱和度和亮度这三种特性来刻画。色调和饱和度合称为色度。一般说来，色度和亮度是相互独立的。在不同的光照条件下，虽然物体颜色的亮度会产生很大的差异，但是它的色度具有恒常性，基本保持不变。过去的一些研究和统计表明，所有的人类（尤其是黄种人和白种人）肤色的 Hue 值差别极小，任意两个人的肤色 Hue 值的平均相关系数高达 0.92。Hue 值成为区分肤色最为重要的特征。

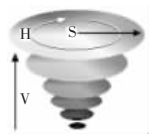


图 5 HSV 色彩模型

考虑到上述 HSV 空间色度值的特点，为了可以减少肤色受到的光照等影响，更好地进行跟踪和分割。将手势图像从 RGB 颜色空间变化到 HSV 颜色空间，完全抛弃对光照极为敏感的 $V$ 分量，重点关注 $H$ 分量。

使用经验的 Hue 肤色颜色分布模型，由于环境因素（摄像设备，人种差异）的影响多种多样，几乎无法预见和避免这些因素的影响。因此系统中，在大致已经通过手势检测得到手的位置之后，并不使用经验的肤色模型，而是由系统从当前检测到的手势区域取得一些肤色样本建立“当前手”的肤色概率模型。如图 6 所示，一般情况下，取得的手掌区域的中心部分大部分是当前人手的肤色。

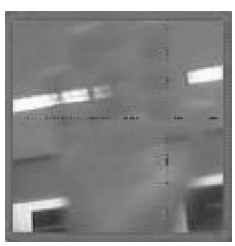

图 6 当前手的肤色

区分皮肤像素与非皮肤像素可以使用不同的色彩统计模型。高斯模型、混合高斯模型和直方图模型都可以用作皮肤的色彩模型。在系统里，采用以直方图统计模型为基础的肤色模型。这里对检测到的手势的中心区域统计关于色度($H$,$S$)的直方图。直方图的 Bin 个数为 64(16*4)。在统计直方图时，对直方图取值进行归一化。

因为 Hue 值在 $S$ 或者 $V$ 值特别大或者特别小的时候会几乎失去其表征色度的意义，将忽略饱和度或亮度在 15%以下和 85%以上的像素，这种情况下将此处肤色概率设置为 0。

根据归一化后的手部肤色直方图，可以为一幅图像建立一张肤色概率图(Skin-color Probability Diagram)。肤色概率图上的每个像素对应了原图像上该像素属于手部肤色的可能性。这个可能性用上述关于色度的直方图来计算。假设像素($x$,$y$)对应的色度值落在直方图的第 $i$ 个 bin，那么像素($x$,$y$)在肤色概率图上的取值 $SPD(x,y)$ 为训练数据中落在直方图第 $i$ 个 bin 内的像素比例。

## 5 基于 CAMSHIFT 的手势跟踪算法

在对摇动检测得到的手势区域进行跟踪时，采用了 CAMSHIFT(Continuous Adaptive Mean SHIFT)跟踪器[5,6]。CAMSHIFT 是一个基于随机颜色概率模型的跟踪器，它的最大好处是基于颜色分布，与跟踪对象的具体模型无关。在手势识别的应用中，要跟踪的人手的肤色在 HSV 空间的色度值上具有很鲜明的分布特点，因此可以用 CAMSHIFT 跟踪器对手进行跟踪。

在用 CAMSHIFT 跟踪手部区域时，使用了第 4 节中提到的基于直方图统计的手部肤色模型。

CAMSHIFT 算法的过程如下所示：

输入：一个初始搜索窗口，即摇动检测器定位到的区域

(1)计算搜索窗口的肤色概率图。

(2)计算肤色概率的 0 阶矩和 1 阶矩：

$$M_{00}=\sum_x\sum_y SPD(x,y)$$

$$M_{01}=\sum_x\sum_y ySPD(x,y)$$

$$M_{10}=\sum_x\sum_y xSPD(x,y)$$

(3)计算搜索窗口中的高概率肤色质心的位置：

$$(x_c,y_c)=\left(\frac{M_{10}}{M_{00}},\frac{M_{01}}{M_{00}}\right)$$

(4)计算搜索窗口中高概率肤色区域的大小。

(5)根据高概率肤色区域的大小调整搜索窗口的中心和尺寸。

(6)重复上述步骤(1)~(5)，直到某次调整中搜索窗口的中心和尺寸的变化小于某个阈值为止。此时，高概率肤色质心的位置就是要跟踪的对象(人手)的位置。

在随后的帧里，对象(人手)移动之后，仍然按照上述过程计算高概率肤色区域，调整人手出现的区域，从而实现了对人手的跟踪。

上述过程中的第(5)步需要在每次的迭代过程中，根据高概率肤色质心的位置，确定搜索窗口的尺寸。这是 CAMSHIFT 算法的核心问题。Bradski 建议在跟踪人脸的情况下，新的搜索窗口的宽和高分别设置为($2\sqrt{M_{00}}$,$2.4\sqrt{M_{00}}$)。考虑到手部肤色和人脸肤色的不同，对上述原则进行了调整。根据对实验结果的比较，在跟踪人手的时候，设置搜索窗口的宽和高为：($1.2\sqrt{M_{00}}$,$1.2\sqrt{M_{00}}$)。

CAMSHIFT 跟踪器跟踪到手的运动之后，就可以相当准确地定位任何时刻手的位置和大小，从而为下面的识别过程提供了良好的基础。此时系统停止摇动检测进程，除非跟踪器丢失了手的位置。

任何跟踪器都避免不了“丢失对象”这个问题。当运动对象移出屏幕或者有其他对象遮挡的情况下，CAMSHIFT 跟踪器的搜索窗口将会越来越小($M_{00}$ 变得极小)，这样即使跟踪对象重新出现在屏幕上，跟踪也无法继续。因此增加了一个启发式的规则解决这个问题。当搜索窗口足够小(比如 5 像素 *5 像素)的时候，把搜索窗口设置为整个图像。这样一来，当对象再次出现的时候，逐渐变小的窗口会很快框住对象。当然这里不可避免地会带来一个新的问题，某些时候跟踪器会把整幅图像中的另一个相似对象误认为是原对象。

## 6 手势区域的分割

在对手部区域进行摇动检测和跟踪后，可以进一步地利用手部肤色模型对得到的手部区域进行分割。手部区域中，肤色概率高于一定阈值的像素标记为 1，其他像素标记为 0。这样，就可以得到一张关于手部区域的二值图，其中手部区域的灰度值设为 255，其他背景区域的灰度值为 0。

在二值图中，手部连通区域里经常会包含一些由于图像噪音引起的空洞。需要对二值图逐步求精。在此系统中，采用了形态学分析中的小结构闭运算算子对分割图进行处理。最后，通过区域合并与标号的算法，可以计算出手部区域二值图中的连通区域，选取面积最大的区域作为人手区域，就得到了平滑后的手部区域二值图。

(a) (b) (c)

图 7 手势区域二值图

图 7 描述了上述手部区域分割的过程。其中(a)为根据摇动检测器或跟踪器得到的手势区域原图像；(b)为利用手部肤色模型对手势区域二值化后得到的初始二值图像；(c)为经过形态处理去掉图像噪声后得到的单连通二值图。

## 7 手势识别

静态手势识别就是在某特定时刻的输入图像中，识别用户做出的是何种手形，这是一个传统的模式识别的问题。目前有很多手势识别的分类算法，如句法模式识别方法、模板匹配和查表的方法、贝叶斯分类器、组合神经网络等。其关键的问题就是使用何种手势模型以及提取怎样的特征。常用的手势图像特征包括图像灰度、二值影像、区域、边界、轮廓指尖等。

考虑到机器人的处理能力有限，而这里要识别的只是简单的数字手势而不是复杂语言手势，所以要求提取的特征应简单快速而有效。如图 8 所示，人手骨骼化(使用一个像素删除查找

表来骨骼化原二值图)之后,手势图像的特征相当明显,可以简单的计算手势骨骼化之后的分支数目,作为一个总体特征 $N$。再对正规化后的 30*40 的骨骼图的上 30 行做行投影计数,得到 $m_1,m_2,\cdots,m_{30}$ 30 个参数。两个特征组成一个 40 维的特征向量 $V=\{N,N,N\cdots,m_1,m_2,\cdots,m_{30}\}$,其中为了突出分支数目的重要性,人为的让 $N$ 占了 10 维。此特征向量作为该系统进行手势识别时的特征。

图 8 人手骨骼化后的手势图像

还试着用 PCA 方法和 Zernike 矩等其他方法提取特征,并与骨骼化后提取的特征进行了比较。综合考虑实时系统上的速度效率,最终我们还是使用了骨骼特征。

在对一定量的训练数据进行标注后,尝试了 k-NN 和 SVM 两种分类器,发现两者的效果极为相近。考虑到资源的限制,在最终的系统中使用了快速简单的 k-NN 分类器。

## 8 实验与总结

这里使用了 2000 张标注好的手势图片作为训练语料。这些图片分别对应了 6 种数字手势:数字 0,1,2,3,4,5。在训练过程,因为输入都是已经处理好的二值手势图,因此制作训练语料时候的环境和个人肤色对系统的识别率没有影响。

在测试时,首先使用 1000 张用摄像机采集的手势图像作为测试集,该文提出的系统的识别模块能对其中的 978 张图片做出了正确的分类,正确率高达 97.8%。若使用 SVM 分类器,正确率约为 98%,但是时间上的开销约为此系统的 5 倍,不符合机器人视觉系统的要求。

然后开始测试在真实环境下系统的性能。实验人员在摄像设备前方给出摇手的信号之后,系统立刻捕捉到这个信号且准确定位了手的大致位置;实验人员移动手,跟踪器也可以有效的跟踪且框住手的位置。对任何时候实验人员用手做出的 0~5 之间的手势信号,该系统的正确分类率为 95%左右。

之后将这个系统移植到某公司的智能机器人上进行了实验,制作了一个与机器人进行猜拳(石头剪子布)的互动游戏。机器人随机出一个手势之后,开始识别人做出的手势,之后判断胜负。效果较好,人出的拳中 10 次约有 9 次机器人可以正确识别。

该文描述了一个基于视觉的态手势识别系统,经过实验证明,该系统快速稳定,可以用在人和机器人的交互上。该系统的定位和跟踪模块已经比较完善,在以后的工作中,笔者将主要对识别模块进行改进,对不同应用有针对性地提取不同的快速有效特征。(收稿日期:2004 年 8 月)

## 参考文献


1.T Starner,J Weaver et al.Real-time American sign language recognition using desk and wearable computer based video[J].IEEE Trans PMAI,1998;20(12):1371~1375

2.Kota Irie,Kazunori Umeda.Detection of Waving Hands from Images Using Time Series of Intensity Values[C].In:The 3rd China-Japan Symposium on Mechatronics(CJSM),2002-09

3.R Cipolla,N J Hollinghurst.Human-robot interface by pointing with uncalibrated stereo vision[J].Image and vision computing,1996-10;14:171~178

4.W T Freeman,K Tanaka,J Ohta et al.Computer vision for computer games[C].In:Proc Int'l Conf Automatic Face and Gesture Recognition,Killington,1996-10:100~105

5.Bradski G R.Computer Video Face Tracking for use in a Perceptual User Interface[J].Intel Technology Journal Q2'98,1998

6.Hunke M,Waibel A.Face Locating and Tracking for Human-Computer Interaction

7.Starner T,Pentland A.Real-time American sign language recognition from video using hidden markov models[R].MIT Media Laboratory:Technical Report 375,1995

8.Cui YT,Weng JJ.View-based hand segmentation and hand-sequence recognition with complex backgrounds[C].In:Proceedings of the IEEE International Conference on Pattern Recognition,Osaka,Japan,1997

9.Cho KM,Jang JH,Hong KS.Adaptive skin-color filter[J].Pattern Recognition,2001;34(5):1067~1073

10.任海兵,祝远新,徐光祐.基于视觉手势识别的研究一综述[J].电子学,2000;28(2):118~121


(上接 31 页)

其它性质和应用。(收稿日期:2004 年 12 月)

## 参考文献


1.M Dalal.Updates in propositional databases[R].Technical report,Rutgers University,1988

2.栾尚敏,戴国忠.有限信念集上修正的一种方法[J].软件学报,2003;14(5):911~917

3.Alchourron C,Gardenfors P,Makinson,D.On the logic of theory change:Partial meet contraction functions and their associated revision functions[J].Journal of Symbolic Logic,50:510~530

4.H Katsuno,A Mendelzon.On the difference between updating a knowledge base and revising it[C].In:Principles of Knowledge Representation and Reasoning:Proc.Second International Conference(KR'91),1991:387~394

5.Liberatore,Schaerf.Arbitration(or How to Merge Knowledge Bases)[J].IEEE Transactions on Knowledge and Data Engineering,1998;10(1):76~90

6.Knoieczny,perez.Merging with Integrity Constraints[J].Lecture Notes in Computer Scienc,1999:233~257

7.Revesz.On the semantics of theory change:arbitration between old and new information[C].In:Proc PODS' 93,12th ACM SIGACT SIGMOD SIGART Symp.Principles of Database Systems,1991:263~294

8.Peter Z Revesz.Model-Theoretic Minimal Change Operators for Constraint Databases[C].In:ICDT 1997,1997:447~460

9.Jinxin Lin.Information Sharing and Knowledge Merging in Cooperative Information Systems[C].In:Proceedings of the 4th Workshop on Information Technologies and Systems,Vancouver,CA,1994