基于组合特征提取技术的手势识别

摘 要

手势在视觉交流上是一个热门的研究领域，主要用于手语识别和人机交互的目的。在本论文中，我们提出了一个通过使用隐马尔可夫模型（HMM 模型）能够实时从彩色图像的序列中识别字母字符（A-Z）和数字（0-9）的系统。我们的系统有三个主要阶段：自动分割和手势区域的预处理、特征提取和分类。在自动分割和手势区域的预处理阶段，通过使用均值漂移算法和卡尔曼滤波，颜色和 3D 深度图是用来探测手将出现的轨迹。在特征提取阶段，笛卡尔系统的使用让我们得到三维组合特征的位置、方向和速度。然后，K-均值聚类采用隐马尔可夫模型。最后阶段所谓的分类，Baum - Welch 算法是用来做一个完整的隐马尔可夫模型参数训练。通过使用左-右手型与 Viterbi 算法结合的方法字母和数字的手势被识别。

实验表明我们的系统能成功识别手势的概率是 98.33%。

关键字：手势识别，计算机视觉，图像处理，模式识别

摘　要

1. 前言

从手势运动和手势位置得到的手语识别是一个用于人机交互的热门研究领域。一个手势是时空格局，这可能是静态或动态或两者。静态的手被称为姿势，而动态的手称作手势。手势解释的目的是推动人机交互从而使人机交互性能接近人际交往。这是由于手势跟踪存在的复杂性，如手的外观，光照变化，跨手闭塞。这些问题损害了跟踪算法的性能和效率。在过去的十年中，几种方法在先进的手势交互的应用前景[1][2][3][4][5]已经被提出来了，但这些差异在他们的模型中又都不相同。这些模型是神经网络[1]，隐马尔可夫模型[2]与模糊系统[5]。

Liu and Lovell 介绍了一种基于 Camshift 实时算法和复合恒定加速度卡尔曼滤波算法的实时手势跟踪系统。而 Nobuhiko 等人用 HSV 颜色空间来追踪非复杂背景下的手和脸，那里的重叠的手和脸通过先前的手和脸模板匹配能较好地分离开。Comaniciu 等人提出了一种使用均值漂移算法和卡尔曼滤波跟踪从摄像头获得的移动物体的技术，这一技术主要获得了实时跟踪性能。先前的技术没有考虑到许多点，如双手的准确分割组合，包括手和脸重叠的鲁棒性跟踪和系统实时高分辨率的运行能力。

Vassilia 等人开发了一种系统，可以识别孤立和连续的希腊手势语言，其中方向向量是从图像中提取，然后作为参数输入到隐马尔可夫模型在句子中被使用。Ho-Sub 等人介绍了手势识别方法，该方法使用位置，角度和速度的组合特征确定作为输入到隐马尔可夫模型的离散向量。这种方法在字母（A-Z），数字（0-9），六编辑命令和六个绘图元素上可以实现。Nianjun 等人通过使用不同的隐马尔可夫模型的拓扑结构数提出一个方法来识别不同国家的从 A 到 Z 的 26 个字母。但是，这些方法运行在一个非复杂离线背景。

Nguyen 等人提出一个手势识别系统，在这个系统里通过卡尔曼滤波和手斑点分析，以获得手部区域动作的描述来跟踪手势。这个系统对背景聚簇和使用皮肤颜色跟踪和识别手势相当强大。此外该系统用包括美国手语拼写字母和数字的 36 个词汇来测试。但是这种方法在我们的系统中研究手的姿势而不是手的运动轨迹。其中有这样一个问题，它提高了手势识别的实时性，是由事实所引起的同样的手势如形状、轨迹和持续时间，甚至是同一个人变化引起的。所以，隐马尔可夫模型是在我们的系统用在它有能力建模时空的时间序列。

本文的主要贡献是研究用于手势识别的位置、方向和速度的组合特征的作用，这个特征是从时空手势路径获得的。此外，它提出了一个能够从三维颜色图像序列中通过使用隐马尔可夫模型模型跟踪单个手势运动轨迹来识别字母字符（A-Z）和数字（0-9）的实时系统。颜色和 3D 深度图是用来检测手。此外，手的轨迹采用均值漂移算法[13]和卡尔曼滤波[14]与 3D 深度图结合的办法来确定。手和脸从来自立体相机、高斯混合模型（GMM）的和颜色信息的三维深度图从复杂背景分割出来，这相对于不利的照明和部分遮挡是更强大的。深度信息解决了手和脸重叠问题。该系统是用来自笛卡尔系统的变化的特征在不同的实验上测试以决定哪个特征能得到最好的结果。每个字母和数字用 30 帧测试（20 帧用来训练和 10 帧用来测试）。测试的手势在识别率上有 98.33%。本文的其余部分如下：

第二部分介绍基本隐马尔可夫模型技术。

第三部分在三个小节中证明这个系统。

第四部分说明实验结果。

最后，第五部分提出总结和结论。

2．隐马尔可夫模型

马尔可夫模型是一个随机过程的数学模型，它在处理过程中产生一个具有相应概率密度分布的状态序列。一个隐马尔可夫模型是由三元组参数$\lambda=(A,B,\prod)$表示如下：

• 一个状态集$S=\{s_1,s_2,...,s_N\}$，N 为常数。

• 一个初始可能每一状态$\prod_i, i=1,2,...,N$，这样第一步$\prod_i=P(s_i)$。

•一个 N 到 N 转移矩阵$A=\{a_{ij}\}$，其中a_{ij}是从状态s_i到s_j的转移的可能性；$1\le i,j\le N$和矩阵 A 的每行之和必须是 1，因为这是让一个给定状态到每一其他状态转移的可能性总和。

•可能的观察序列集$O=\{o_1,o_2,...,o_T\}$，其中 T 是手势路径的长度。

•离散的信号集$V=\{v_1,v_2,...,v_N\}$，其中 M 是离散的信号。

•一个 N 到 M 的观察矩阵$B=\{b_{im}\}$，其中b_{im}给出来至状态s_i的信号v_m的可能值，而矩阵 B 的每行值总和必须是 1，原因和前面的一样。

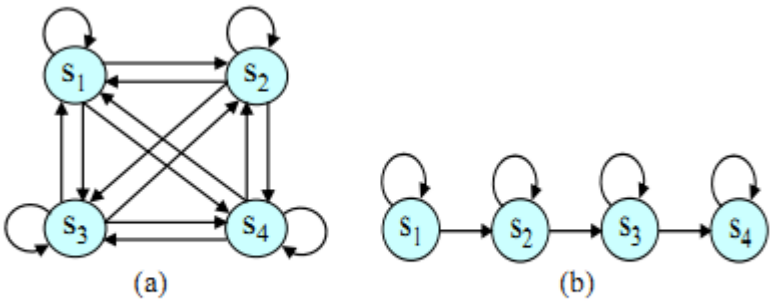

Fig. 1. HMMs topologies with 4 states (a) Ergodic topology (b) LRB topology.

对隐马尔可夫模型来说有三个主要问题：计算问题、解读问题、训练问题。这三个问题可以分别通过前向和后向算法、Viterbi 算法和 Baum- Welch 算法解决。此外，隐马尔可夫模型的拓扑结构有三种：完全连接（遍历模型），在这个结构里可以从任一状态到达其他状态；左-右模型，在这个模型里每个状态只能到达自己状态和下一状态；左-右带状模型，在这个模型里每个状态只能到达自己状态和下一状态（图 1）。

3．推荐系统

我们开发了一个自动识别系统，这个系统可以从三维色彩图片序列中通过单个手的运动轨迹使用隐马尔可夫模型实时识别代表字母（A-Z）和数字（0-9）的手势。特别是，这个系统包括三个主要阶段：手势自动分割和追踪阶段、特征提取阶段和分类阶段（图 2）。

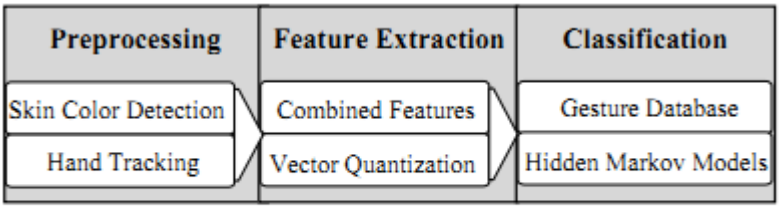

Fig. 2. Simplified structure showing the main computational modules for gesture recognition system.

•预处理：定位和跟踪的手以产生它的运动轨迹（手势路径）。

•特征提取：聚类提取的特征生成作为输入参数到隐马尔可夫模型识别使用的离散向量。

•分类：通过使用离散向量和左-右带状拓扑结构识别手势路径。

3.1 手势自动分割和追踪

本文描述了探测和分割复杂背景下的三维色彩图片里的手势的方法，在这个方法里使用 3D 深度图和色彩信息来分割手势。皮肤颜色区域的分割只有在将色度应用在分析中才能变强大。所以，在我们的系统中使用 YCbCr 颜色空间，其中 Y 分量代表亮度，而（Cb，Cr）分量是指色度。我们忽略 Y 分量，以减少亮度变化的影响，仅使用色度分量，这样充分得到颜色信息。一个大型肤色和非肤色像素数据库被用来训练高斯模型。高斯混合模型使用皮肤数据库开始建立皮肤模型，其中大量 k-均值聚类算法用来模型的训练，以确定 GMM 参数的初始配置。对于三位色彩图像序列中的手和脸的肤色分割，我们计算肤色深度值以增加肤色信息。深度信息（图 3（c））解决了由基于相互关联交叉和已知坐标的照相机位置数据而测量获得的手和脸的重叠问题。几组值组成了最终的三维坐标点。聚类算法可以看成是种在三维空间中的使用了两种准则的区域生长，这两种准则是：皮肤颜色和欧氏距离。此外，这种方法对于实时环境中发生的不利照明和部分遮挡有更好

的鲁棒性。还有，点分析被用来推测边界面积、周长和形心。欲了解更多详情，读者可以阅读[2]，[21]。

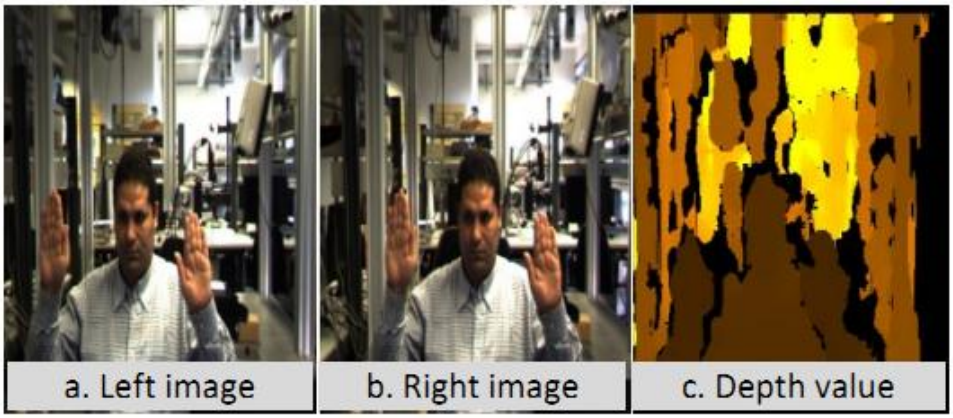

Fig. 3. (a) Left image frame of video stream. (b) Right image. (c) The depth value of left and right image via the Bumblebee stereo camera.

经过来至分割步骤的手势目标定位后，我们发现手的颜色直方图和 Epanechnikov 核。这个核分配来至中心的父像素较小的权重以增加密度估计的鲁棒性。为了找到在连续帧中手势目标的最佳匹配，我们使用 Bhattacharyya 系数通过从手势目标和模板的比较得到的贝叶斯误差来测量相似程度。我们对先前的帧计算手势区域的平均深度值以解决手和脸的重叠问题。平均偏移过程被定义为递归地和执行优化计算平均偏移向量。经过每个均值偏移优化，系统给出手势目标的测量位置。能够计算不确定的估计值，然后通过卡尔曼迭代迭代得到手势目标。因此，我们可以通过检测连续的图像帧之间的手的关系得到手势路径（图 5（d））。有关详细信息，读者可以参考[2]，[8]，[21]。

3.2 特征提取

毫无疑问，选择好特征来识别手势路径在系统性能起着重要的作用。手势路径有三个基本特征：位置、方向和速度。我们分析这些从手势轨迹提取的特征的有效性，同时将它们结合起来测试它们的识别速率。手势路径是一个由手质心（xhand，yhand）组成的时空坐标。

在笛卡尔空间的该坐标可以直接从手势帧中提取。我们考虑两种位置特征。第一种位置特征是从质心到手势路径的各个点的距离 Lc ，因为同一手势根据不同的起点形成不同位置特征（公式 1）。第二种特征是计算从起点到手势路径上当前点的

距离 Lsc（公式 3）。

$$Lc_t = \sqrt{(x_{t+1} - C_x)^2 + (y_{t+1} - C_y)^2} \quad (1)$$

$$(C_x, C_y) = \frac{1}{n}(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n} y_t) \quad (2)$$

$$Lsc_t = \sqrt{(x_{t+1} - x_1)^2 + (y_{t+1} - y_1)^2} \quad (3)$$

其中 t=1,2,…,T-1，T 是手势路径的长度。(C_x, C_y)指在 n 点的重心。为了验证实时性，我们计算每帧后的手势路径的重心点。

第二个基本特征是方向，它给出了在手势处理过程中贯穿空间的手势走向。如上所述，方向特征是基于每个点上的手势位移向量的计算，它代表着手势路径质心的方向(θ_{1t})、连续两个点的方向(θ_{2t})和起点与当前手势点的方向(θ_{3t})。

$$\theta_{1t} = \arctan\left(\frac{y_{t+1} - C_y}{x_{t+1} - C_x}\right) \quad (4)$$

$$\theta_{2t} = \arctan\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right) \quad (5)$$

$$\theta_{3t} = \arctan\left(\frac{y_{t+1} - y_1}{x_{t+1} - x_1}\right) \quad (6)$$

第三个基本特征是速度，它在手势识别阶段特别是在一些关键情形中起很重要的作用。速度是基于一个现实，在这现实里手势路径的拐点处手的速度降低。速度是通过欧氏距离除以视频帧里面的两个连续点的时间如下：

$$V_t = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} \quad (7)$$

在直角坐标系中，我们使用不同的特征组合，以获得各种特征向量。例如，在帧 t+1 的特征向量可以通过位置特征(Lc_t, Lsc_t)、位置特征和方向特征(Lc_t, Lsc_t, V_t)、方向特征$(\theta_{1t}, \theta_{2t}, \theta_{3t})$、方向特征和速度特征$(\theta_{1t}, \theta_{2t}, \theta_{3t}, V_t)$、位置特征和速度特征与方向特征$(Lc_t, Lsc_t, \theta_{1t}, \theta_{2t}, \theta_{3t}, V_t)$来获得。

在时刻 t 每帧都包含一个特征向量集，其中的空间维数正比于特征向量的大小。在这种方式下，手势被描绘成一个有序的向量特征序列，它在三维中被处理和聚簇成一个离散值作为隐马尔可夫模型的输入。这可以通过使用 K-均值聚类算法[19]，[20]完成，它将手势模型在特征空间中分为 k 集群。

矢量量化：量化提取的特征得到了离散值。当诸如位置和速度等基本特征单独使用时，这些特征被规范化和乘上从 10 到 30 的不同标量值。另一方面，方向特征的规范值除以 10，20，30 和 40 已获得它的码字。除了直角坐标系中的组合特征，我们使用 K -均值聚类算法将手势特征分类到特征空间中的 k 集群。该算法是基于每个集群的中心到特征点的最小距离。我们将特征向量集分成集群集。这使我们能过通过一个集群在特征空间模拟手势轨迹。这计算得来的集群指数用来当做隐马尔可夫模型的输入。此外，在数据集中我们通常不知道集群的个数最好是多少个。为了得到在每个 K-均值算法中的每个执行中集群 K 的个数，我们假定 K=28，29，…，37，这样的假定是基于在所有的字母（A-Z）和数字（0-9）的分段部分的数量，其中每个直线段被编入同一集群。

假设我们有 n 个来自同一类得训练特征向量 x1,x2, ..., xn 的样本，同时我们将它们分为 k 集群，k<n。在集群 i 中我们让m_i为向量的平均值。如果这些集群被很好的分离，那么一个最小距离分离器被用来分离它们。也就是，如果$\|x-m_i\|$是所有 k 距离中的最小一个是，我们可以认为 x 是在集群 i 中。

•对均值 m1,m2, ...,mk 建立一个随机初始向量量化编码本

•在任何情况下都没有改变

我们使用估计方法来给每个训练特征样本分类到一个集群m_i中

For i=1 to k

对集群 i 用所有训练好的特征样本的均值代替m_i

end(for)

•end(until)

一个总的看法是，不同的手势有不同集群中的空间运动轨迹，而同样的姿态有非常相似的轨迹。

3.3 分类

在我们系统的最后阶段就是分类。在本阶段，Baum - Welch 算法[15]是用来对初始化隐马尔可夫模型参数做一个完整的训练来构建手势库。手势库中的字母（A-Z）和数字(0-9)的每个参考模型是通过左-右 Banded 模型根据它的复杂性用从 3 到 6 个不同状态来模拟。正如，如果训练样本数量跟模型参数相比不足的话，过

多的状态会形成过拟合问题。通过选择最大观察手势模型可能性分类手势路径。最可能的手势模型是所有 36 手势中观察可能性最大的手势。所观察到得手势 O 一帧一帧通过 Viterbi 算法识别（即累计，直到它收到手势信号结束）。下列步骤展示了 Viterbi 算法如何在手势模型 $\lambda_g(a^g, b^g, \Pi^g)$ 中工作的。（图 4）

1. 初始：

$$for\ 1 \le i \le N,$$

$$\delta_1^g(i) = \Pi_i^g . b_i^g(o_1)$$

2. 递回（累计观察概率计算）：

$$for\ 2 \le t \le T,\ 1 \le j \le N.$$

$$\delta_t^g(j) = \max_i[\delta_{t-1}^g(i).a_{ij}^g].b_j^g(o_t)$$

3.终止：

$$P(O|\lambda_g) = \max_i[\delta_T^g(i)]$$

其中，N 是状态个数，Π_i^g 是状态 i 的初始值，a_{ij}^g 是状态 i 变成状态 j 的权值，$b_j^g(o_t)$ 指在在 t 时刻在状态 j 下 o 发散的权值，而 $\delta_t^g(j)$ 指在 t 时刻状态 j 下最大可能值。

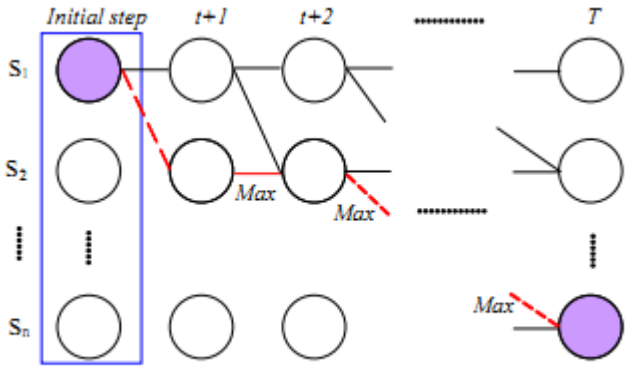

Fig. 4. The best path for LRB model with N states where it starts from S_1 to S_N, N=3, 4, ..., 6 and t=1.

4. 实验结果

我们系统已经用不同的视频序列包括重叠和部分遮挡视频来测试过。我们提出的系统通过隐马尔可夫模型使用单只手的运动轨迹能够实施执行和展示良好识别三维彩色图片里的字母和数字结果的能力。输入图像通过拥有有 6 毫米 15fps 的 240×320 像素的图像分辨率的大黄蜂立体摄像机系统捕获，Matlab 实现。在我们的实验结果，每一个孤立的手势是基于 30 的视频序列，其中 20 个样本进行训练和 10 个样本进行测试（总共，我们的数据库包含 720 个训练样本和 360 个测试样本）。手势识别模块将手势路径跟参考手势库进行匹配，以识别出手势属于哪类。通过 Viterbi 算法采用 LRB 拓扑从 3 到 6 不同数量状态一帧一帧地实时识别字母和数字以计算更好的优先级。

我们测试了直角坐标系中三个基本特征（位置、方向、速度）的重要性。此外，对隐马尔可夫模型观测序列或在单独特征下正常量化或在组合特征下采用 k-均值聚类算法量化。从表 1，孤立的手势识别率在采用特征$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$下获得了最佳效果 98.33%。识别率是正确识别手势数目与测试手势的数量（公式 8）。

$$Reco.\ ratio = \frac{\#\ recognized\ gestures}{\#\ test\ gestures} \times 100\% \qquad (8)$$

TABLE I
RESULTS OF HAND GESTURES ACCORDING TO DIFFERENT FEATURES EXTRACTION IN CARTESIAN SYSTEM WITH THE BEST FEATURE CODE NUMBER.

Feature type	Feature space	# Feature code	Training data	Hand gestures results		
				Testing data	Correct data	Recognition (%)
Separated in Cartesian coordinates	Lc	20	720	360	187	51.94
	Lsc	25	720	360	118	32.78
	v	25	720	360	206	57.22
	θ_1	18 ; 36	720	360	349	96.94
	θ_2	18	720	360	321	89.17
	θ_3	18	720	360	236	82.22
Union in Cartesian coordinates	(Lc, Lsc)	35	720	360	293	81.39
	(Lc, Lsc, V)	33	720	360	309	85.83
	$(\theta_1, \theta_2, \theta_3)$	33	720	360	338	93.89
	$(\theta_1, \theta_2, \theta_3, V)$	34	720	360	339	94.16
	$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$	33	720	360	354	98.33

表 1

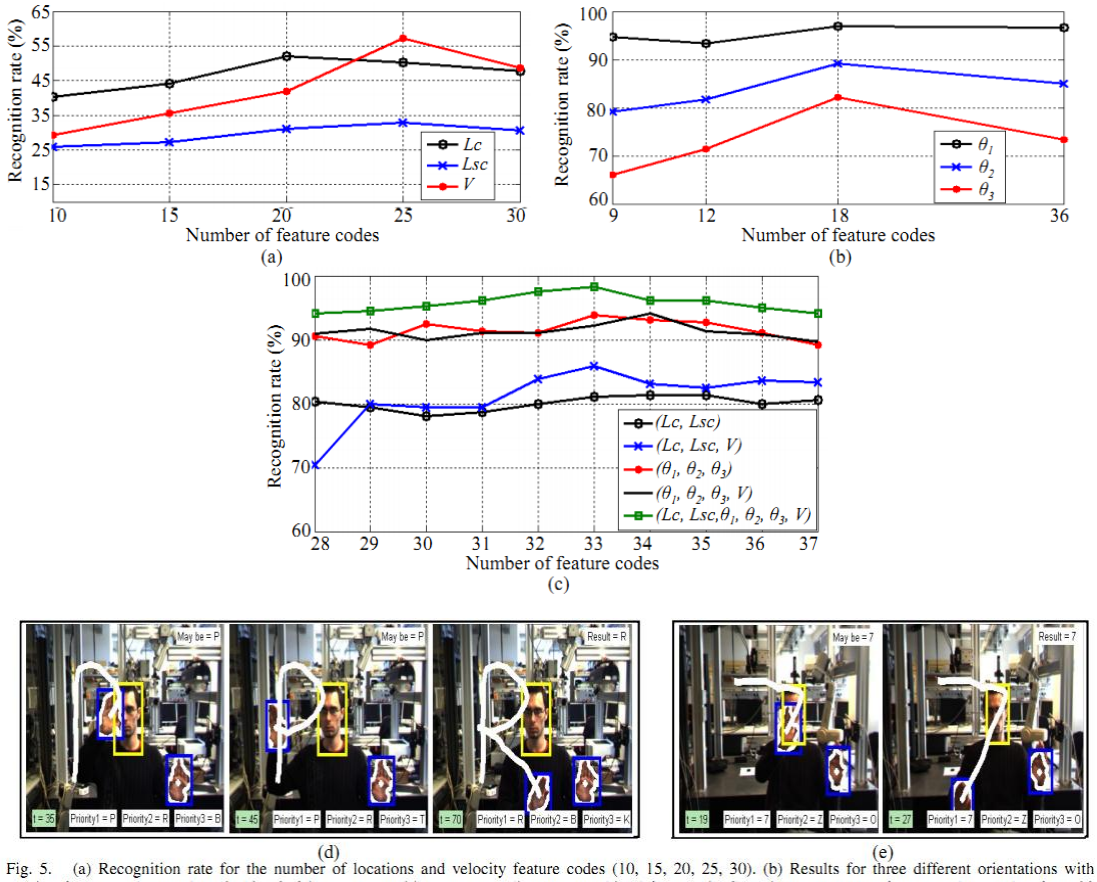

Fig. 5. (a) Recognition rate for the number of locations and velocity feature codes (10, 15, 20, 25, 30). (b) Results for three different orientations with varying feature codes number (9, 12, 18, 36). (c) Recognition rate according to a combined features in Cartesian system over feature codes number from 28 to 37. (d) The high priority is alphabet 'P' at t=45 and at t=70 the result is 'R'. (e) Solving overlap problem between hand and face at t=19 and the high priority is '7' at t=27.

图 5

根据图 5(a)和 5(b)中单独特征，方向特征$(\theta_1, \theta_2, \theta_3)$比位置特征$(Lc, Lsc)$或速度特征(V)的识别率要好。这导致方向特征$(\theta_1 = 96.94\%)$是三个基本特征中最有效的一个。此外，速度特征展示比方向特征更低的识别力（57.22%）。同时，位置特征Lsc有最低的识别率 32.78%。此外，包括速度信息的特征(Lc, Lsc, V), $(\theta_1, \theta_2, \theta_3, V)$和$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$比使用速度特征有更高的识别率（图 5(c)）。简而言之，图 5 说明了实验的结果表明最好的特征码号（对特征$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$来说，最好的特征码号是 33）。相应的，图 5(d)和图 5(e)展示了手势字母'R'和数字'7'的系统输出，同时用 3D 深度图解决了手和脸的重叠问题。

5. 结论

在本论文中，我们提出了一个通过使用隐马尔可夫模型（HMM 模型）能够实时从彩色图像的序列中识别字母字符（A-Z）和数字（0-9）的系统。这个系统使用笛卡尔坐标系统中的位置、方向和速度的组合特征。我们已经表明，这些特性的有效性可以产生合理的识别率。数据库包含孤立手势的 720 帧训练样本和 360 帧测试样本。当应用于几个包括诸如部分遮挡和重叠的复杂情形视频样本时，这个系统能表现出良好的性能。结果表明，本系统是适合实时应用，并且具有 98.33％的手势识别率。未来研究将采用指尖与多摄像机系统结合来识别手势点和句子而不是手势重心点。

鸣谢

这项工作由 Bernstein-Group BMBF(FKZ: 01GQ0702), 文理学院助学金(C4-NIMITEK2,FKZ:UC4-3704M),Forschungspraemie(BMBF-Frderung,FKZ:03FPB00213)和 DFG 建立的合作研究中心”Companion-Technology SFB / TRR 62”支持。