基于色度分析的唇动特征提取与识别

姚鸿勋[1],吕雅娟[1],高 文[1,2]
(1. 哈尔滨工业大学计算机科学与工程系,黑龙江哈尔滨 150001;2. 中国科学院计算技术研究所,北京 100080)

摘 要: 本文提出了一种基于色度滤波的唇动特征提取与识别方法,它通过唇的色度滤波,得到增强的唇动图像,再利用可变模板,描述口型轮廓并提取特征参数,并用 HMM 模型进行唇运动序列图像识别. 该方法鲁棒性强,对光照没有苛刻的要求,且针对非特定人,适用于自然条件下的实用环境,解决了可变模板对目标边缘有较高分辨率的要求,使方法更实用化. 本文的实验是基于单纯的视觉信息(没有声音信道的信息)的唇动识别,不加语音信息,实验集合只限于单韵母,识别率可达 95.8 %.

关键词: 唇读; 唇动; 色度; 可变模板; HMM 模型
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2002) 02-0168-05

Lip-Movement Features Extraction and Recognition Based on Chroma Analysis

YAO Hong-xun[1], LÜ Ya-juan[1], GAO Wen[1,2]
(1. *Dept. Computer Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*;
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: This paper presents an approach of lip-movement features extraction and recognition based on chroma analysis. By the lip chromatic filter, enhanced lip-movement images are obtained. The contours are described and feature parameters are extracted using deformable templates. Sequence images of lip-movement with the features have been recognized by employing the HMM model. This method is quite robust to illumination and speaker independent, and it is fit for realistic surroundings under natural condition. It works out the problem of high resolution on object contours which deformable template needs, and makes itself more practicable. All lip-movement recognition experiments are based on visual information alone (without using acoustic information), the experimental set is limited within single Chinese vowels, and the accuracy rate of recognition gets up to 95.8 %.

Key words: lip-reading; lip-movement; chroma; deformable template; HMM model

1 引言

人类的语言认知过程本身就是一个多通道的感知过程. 生活经验告诉我们,在人与人交流的过程中,人们对于他人讲话的内容不仅仅需要通过声音来感知,往往还需要眼睛观察其口型,表情等的变化,才能准确地理解对方所讲的内容. 以往的语音识别系统忽略了语言感知的视觉特性,仅仅利用了听觉特性,使得现有的语音识别系统在噪声环境或多话者条件下,其识别率都大大下降,限制了它的应用领域. 近年来,唇读作为语音识别的辅助手段引起了越来越多的研究人员的关注,初步的研究结果表明,将唇读与语音进行融合能有效地改善识别率,特别在噪声环境下,效果更为明显[1~3].

在唇读研究中,最重要的环节就是唇的定位跟踪及特征提取,这一问题能否良好的解决直接影响唇读识别的结果. 但是唇的准确定位是非常困难的,这是因为不同人唇的形状差异较大,而且唇型受说话形变、头部运动、光照等因素的影响也较大. 以往的系统为了能够精确提取口型轮廓采用手动的办法定位唇区域,或将唇涂上深色口红,而且必须保证特定的光照条件,或者在唇的周围插上发光二极管来跟踪唇动,但这些方法都使得唇读不能满足实际应用. 因此,研究自然条件下的唇定位与识别方法是非常必要的.

本文提出了一个唇色滤波器的算子,在滤波后的图像上再用可变模板方法实现口型轮廓的提取与跟踪,并将跟踪后的结果(曲线参数)送入识别器,用 HMM 模型识别图像序列的发音类. 实验结果表明,该方法能够适应自然条件下的口型描述,不受口型缩放、变形、旋转的影响,对不同唇型有很好的鲁棒性,并且由于增加了唇色自适应学习模块,使得系统能够对不同的唇色、光照和摄像机色调进行自适应调整,为唇动正确识别提供了保证.

收稿日期:2001-02-08;修回日期:2001-09-16
基金项目:国家 863 计划青年基金(No. 863-306-QN99-4);国家 863 计划项目(No. 863-306-ZT03-01-2);国家自然科学基金重点项目(No. 69789301);中科院百人计划的资助

2　唇色滤波器的设制与特征提取

先前的很多唇读系统采用灰度图像进行唇动检测[4,5],利用直方图的峰、谷分割目标,提取口型轮廓,但是这些信息受胡须、阴影、光照等影响严重,很难得到理想的结果. 分析其主要原因是一般的灰度图像仅仅使用了色彩的亮度信息,忽略了色度信息,而亮度信息随光照,阴影的变化有较大的变动;相反,色调信息却相对稳定,它恰恰是唇色和肤色的一个重要特征.

2.1　唇色分析与唇色滤波器

我们的系统中采用肤色模型与特征脸相结合的方法进行人脸检测和定位[6],在人脸定位以后,取脸的下半部区域作为感兴趣的区域. 由于面部像素主要表现的肤色和唇色值比较接近,为了有效地把唇色从肤色中区分出来,寻找一个合适的彩色滤波器,通过滤波器只保留感兴趣的唇部,这是一个对特征提取和识别很有意义的想法.

通过对不同性别、不同年龄及不同肤色的人脸图像的肤色和唇色分布情况的统计,可以得出亮度信息归一化以后肤色和唇色各自具有相对稳定的色度聚类特性的结论. 因此,我们想要的滤波器应该是一个色度空间的量,它与 YUV 彩色表示法中的 U、V 分量有关,是它们的一个线性表达式,与色调有关,因此有一个 α 旋转角度,另与坐标原点有一个偏移量 X_0,因此,唇色滤波器 Z 可用下式表示:

$$Z = U\cos\alpha + V\sin\alpha + X_0 \tag{1}$$

通过对性别、年龄、肤色的 1800 幅人脸图像的肤色和唇色实验的统计分析,分别求得旋转角 α 为 66°,平移量 X_0 为 -27,即

$$Z = 0.407U + 0.914V - 27 \tag{2}$$

通常,U、V 取

$$\begin{cases} U = -0.147R - 0.289G + 0.436B \\ V = +0.615R - 0.515G - 0.100B \end{cases} \tag{3}$$

将式(3)代入式(2),得

$$Z = 0.493R - 0.589G + 0.026B - 27 \tag{4}$$

对关心区域的每个像素点用式(4)进行滤波,并将滤波值分布区间为[0,255],因此,

$$\begin{cases} \mu = K \cdot \dfrac{(Z - Z_0)^2}{r_z^2}, & \text{当} |Z - Z_0| \le r_z \\ \mu = 1, & \text{其它} \end{cases} \tag{5}$$

其中,μ 为系数,$\mu \in [0,1]$,Z_0 为唇色样本的 Z 值分布中心,即滤波中心. r_z 为滤波半径. 滤波后的灰度值 F 计算表达式为:

$$F = 255\mu \tag{6}$$

图 1 为滤波前后的图像对照. 可见,通过彩色滤波后唇色明显突出出来,这使后面的唇动定位与跟踪操作相对容易并且提高了准确度.

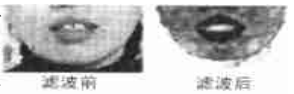

图 1　滤波前后图像对照

2.2　唇的粗定位

对滤波后的图像进行二值化处理,得二值化的参考图像,用一窗口沿参考图像滑动,计算窗口内唇色区域的占有率,当占有率满足一定条件,且窗口中心坐标 (X_0, Y_0) 与唇色区域重心 (X_1, Y_1) 坐标距离较小时,则认为窗口内的区域为唇部区域. 由于没有唇部区域大小的先验知识,必须选取不同大小的滑动窗口,窗口按从大到小的顺序选取,以避免选定的唇部区域不够完整. 因为对于人脸的下半部区域,唇部大致居中的概率较大,故窗口的运动轨迹采用从中心向四周扩散的运动方式. 粗定位结果如图 2 所示.

图 2　利用唇色模型得到的唇部粗定位结果

2.3　有导师的色滤波器

为了适应特殊的目标和使用环境,色度滤波器可以进行自动调整,以学习和适应新的环境.

也可以人工施教,告诉系统学习的对象为哪部分,用鼠标在其上拖动,系统实时获取相应对象的色度信息,以导出适当的色度滤波器.

2.4　唇特征提取

在口型轮廓提取中,可变模板方法是值得推荐的[5,9],它的优点是不受口型的变形、缩放、旋转的影响,能够有效的描述口型轮廓,与口型匹配的曲线参数直接给出了口型的形状特征,给进一步的识别带来了方便. 图 3 表示了口型模板,主要由两种曲线构成:抛物线和四次曲线. 内唇由两个抛物线描述,外唇由四次曲线描述,因四次曲线能更精确地反映外唇的形状. 唇型模板可以用以下的参数确定:(X_c, Y_c), θ, w_0, w_1, h_1, h_2, h_3, h_4, a_{off}, q_0, q_1. 其中前 10 个参数如图 3 标示,后 2 个参数中 q_0, q_1 分别是外唇上下轮廓线四次曲线的四次项系数,它表示四次曲线偏离抛物线的距离. (事实上,上唇的外轮廓线是两条四次曲线来描述的,由于模板的对称性,两个四次项系数是相同的,均为 q_0.) 但是,可变模板方法的缺点是对自然条件下摄取的图像应用起来有困难,原因是它是依靠图像的轮廓信息来进行口型和曲线的匹配,而自然(非特殊光照)条件下的口型轮廓的精确提取是非常困难的,因为肤色和唇色相当的接近,这使得很多情况下可变模板方法不能达到令人满意的结果.

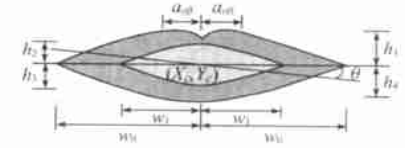

图 3　口型模板

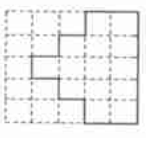

图 4　左嘴角模式

由于我们的方法首先对图像进行了唇色滤波处理,处理后的图像唇色被很好地保留了下来,或者说明显地增强了,这

使得唇部区域的二值化处理及边缘提取变得容易了.

® 特征点定位 为了确定匹配口型的曲线,首先确定六个特征点,它们是:两个嘴角点,上唇外边缘中点,下唇外边缘中点,上唇内边缘中点,下唇内边缘中点.这些特征点决定了曲线的初始位置.我们知道,合理的初始位置可以大大加快匹配的过程.特征点的定位是在滤波处理后的图像上进行的,采用局部模式匹配和局部投影特征相结合的方法.

以左嘴角点的定位为例,使用的模板如图 4 所示.垂直扫描线从左到右依次检测,在垂直扫描线上若模板内像素个数与模板外像素个数的差大于一定值时,匹配成功,此时模板中心为左嘴角的位置.同理,可找到右嘴角.再根据左右嘴角的坐标,可以计算出可变模板的中心坐标 (X_c, Y_c),外唇宽 w_0 及旋转角 θ 的初始值.类似地,利用窗口内像素点在垂直方向上投影图灰度变化阶跃点位置,可以分别求出另四个特征点:上唇外边缘中点,下唇外边缘中点,上唇内边缘中点,下唇内边缘中点.这样就可以确定可变模板中参数 h_1, h_2, h_3, h_4 的初值.

® 口型的曲线描述 口型的曲线描述就是调整曲线参数使之与口型达到最佳匹配,这是一个使代价函数最小化的过程[5].在我们的系统中,代价函数包含了口型的轮廓信息和时空约束信息及其它一些惩罚项,对口型的形变进行了约束限制.匹配算法采用最速下降法.

由于唇色滤波突出了口型信息,因此增强了口型模板定位的鲁棒性.图 5 为部分具有代表性的图像唇定位结果.其中 (a) 为有胡须的干扰的图像,(b) 为有一定倾角的口型图像,(c) 为从 VCD 中截取的口型图像并带有一定的旋转角,(d) 为有特殊唇型者的口型图像(上唇内边缘向下弯曲).

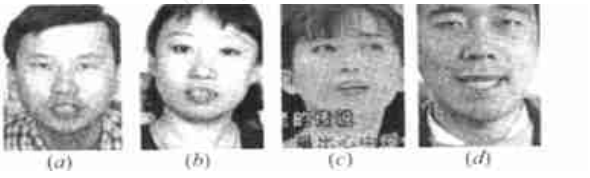

图 5 唇定位结果

® 口型特征参数提取 匹配的曲线,其坐标中心和旋转角对唇动识别不起作用,因此可不必输出.另外实验中发现 a_{off} 对识别也不起作用,它的存在还影响 q_0 的变化,因此,在实际口型轮廓提取中去掉了这一参数.故共得到用于描述口型形状的八个特征参数,w_0, w_1, h_1, h_2, h_3, h_4, q_0, q_1.另外,牙齿和舌的信息对唇读有很大的帮助,因此,根据模板的内唇轮廓,对内唇内部从上到下进行了三等分,得到三部分的灰度均值,用这些值可初略地估计唇内牙齿和舌的状况,如露出上牙、露出下牙、不露牙等.加上整个口腔内的灰度均值,共得到 4 个灰度均值,把这 4 个值与嘴唇的灰度均值相比作为灰度特征,分别用 y_{up}, y_{md}, y_{low}, y_{av} 表示.故在完成特征抽取后,共得到了 12 个特征参数(下面分别用 C_1, C_2, …, C_i, …, C_{12} 表示),实验表明,对识别贡献最大的三个特征为 w_1, y_{av} 和 $h_2 + h_3$,我们分别对不同贡献的参数,给于不同的权值.

3 基于 HMM 模型的唇动识别

由于唇动过程是一个连续过程,在我们的唇动识别系统中,采用了隐马尔可夫模型(Hidden Markov Model),因为它很好地反映了人类发音过程的特点.有关 HMM 的基本原理及其应用在文献[10]中有详细的论述.

® 状态类型的选择 HMM 具有多种结构类型,常用的有各态历经型(ergodic model),无跨越从左向右型(left-right model without skip),有跨越从左向右型(left-right model with skip).对于唇动而言,它具有很强的时间延续性,因此较合理的模型应为自左向右的模型.对于无跨越的自左向右模型,每个状态只能向右侧编号高一位的状态或本状态转移,如图 6所示,因此转移矩阵 A 中只有主对角线或右副对角线上的元素允许非零,因而 A 比较稀疏,大大减少了模型参数估计的计算量,因此在唇动识别中采用了无跨越的自左向右模型.

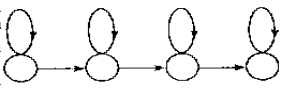

图 6 无跨越的自左向右模型

该模型的状态转移有如下特性:

$$a_{ij} = 0, j < i \quad 或 \quad j > i + 1$$

即每一个状态不能向比它低的状态转移,并且状态间转移步幅不能大于 1.状态"1"与状态"N"分别为源状态和吸收状态,这意味着发音时唇动必须从状态"1"开始到状态"N"结束,这正与人的发音过程对应.这样初始概率为

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases}$$

® B 参数类型的选择 B 参数是 HMM 模型最重要的参数,它直接与观察有关,对模型影响最大.根据 B 参数选择的不同,HMM 可分为离散的(DHMM)、连续的(CHMM)和半连续(SCHMM)三种类型.前面讨论的都是离散型的 HMM,它的主要特点是计算量少,运算速度快.在训练数据充足的情况下,可得到较好的识别效果.但是由于连续信号采用 VQ 量化,又不可避免引入了 VQ 误差,同时由于 VQ 码本和 DHMM 模型是分开产生的,因此它不是一种经过优化的组合.采用 CHMM 提高了识别率,但计算量大,自由参数多,需要大量的训练数据才能得到较好的训练结果.由于训练样本相对较少,因此在唇动识别中采用了 SCHMM 模型.

® 特征归一化 由于采集图像中唇的大小各有不同,因此必须对提取的特征进行归一化处理.以第一帧的特征 w_0 为基准,把它归一到某一特定值 K,其它特征按这一比例进行规正.规正公式如下:

$$\overline{C}_i^{(t)} = \frac{K}{w_0^{(1)}} \cdot C_i^{(t)}, \quad i = 1, 2, 3; \; t = 1, 2, \cdots, T$$

其中,K 为第一帧特征 w_0 的规正值,T 为特征序列的长度.

为了充分体现唇动的动态特性,对规正后的特征进行了处理,对每一帧特征分别与第一帧特征做差值,如下式所示:

$$\begin{cases} \Delta \overline{C}_i^{(1)} = 0 \\ \Delta \overline{C}_i^{(t)} = \overline{C}_i^{(t)} - \overline{C}_i^{(1)}, \; t = 2, 3, \cdots, T \end{cases}$$

实验表明,用差值特征 $\Delta \overline{C}_i$ 比绝对特征 $\overline{C}_i$ 具有更好的识别效果,因为 $\Delta \overline{C}_i$ 比 $\overline{C}_i$ 能更好的体现唇动的动态特性.在后

面的实验中,用 $\Delta\bar{C}_i$ 作为训练和识别的特征.

® 训练过程　对于 HMM 模型,首先需要通过训练来得到每一个音口形变化的模型参数,然后进行识别过程匹配. 通常可以用五元组模型 $\lambda=(A, c_{jm}, \mu_m, \sum_m, \pi)$ 来表示 SCHMM 模型. 它们的含义分别是:状态转移概率矩阵 $A=\{a_{ij}\}$,其中 $a_{ij}=P\{q_{t+1}=S_j \mid q_t=S_i\}$,初始状态概率集合 $\pi=\{\pi_i\}$,而 SCHMM 模型中观测向量的概率密度函数为 m-阶混合高斯分布函数,其具有如下形式:

$$b_j(O_t)=\sum_{m=1}^{M} c_{jm} N(O_t, \mu_m, \sum_m) \tag{7}$$

其中,c_{jm}为混合比系数,μ_m 为均矢量,$\sum_m$ 为协方差矩阵. 训练过程的任务是:根据给定的序列训练数据来估计参数 $a_{ij}, c_{jm}, \mu_m, \sum_m$. 可用 Baum-Welch 迭代法进行参数训练. 具体步骤如下:

步骤 1:确定初值:

$$\begin{cases} a_{ij}=0.5, & i=j \text{ 或 } j=i+1,\ a_{NN}=1; \\ c_{jm}=1/M, & 1\le j\le N, 1\le m\le M; \\ \sum_{mp}=0.25, & 1\le m\le M, 1\le p\le P, \\ & P \text{ 为特征的维数}; \end{cases}$$

μ_m 的初值利用聚类算法 LBG 产生;

步骤 2:根据引入的前向概率和后向概率公式[10]分别计算前向概率 $\alpha_t(i)$ 和后向概率 $\beta_t(i)$;

步骤 3:用式(8)～(11)分别求 a_{ij}, c_{jm}, μ_m 及 $\sum_m$;

$$a_{ij}=\frac{\sum_{k=1}^{K}\sum_{t=1}^{T-1}\alpha_t(i)\, a_{ij} b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{k=1}^{K}\sum_{t=1}^{T-1}\alpha_t(i)\beta_t(i)}, \text{其中} \begin{matrix}1\le i\le N\\ 1\le j\le N\end{matrix} \tag{8}$$

$$c_{jm}=\frac{\sum_{k=1}^{K}\sum_{t=1}^{T}\gamma_t(j,m)}{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t(j,m)}, \text{其中} \begin{matrix}1\le j\le N\\ 1\le k\le M\end{matrix} \tag{9}$$

$\gamma_t(j,m)$ 为 t 时刻处于状态 j 的第 m 个混合项的概率,其具有如下形式:

$$\gamma_t(j,m)=\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}\cdot\frac{c_{jm}N(O_t,\mu_m,\sum_m)}{b_j(O_t)}$$

而 $O=O_1O_2\cdots O_T$ 表示所观察到的一段序列,O_t 为 t 时刻的观察值,T 为观察序列长度.

$$\mu_m=\frac{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j,m)\cdot O_t}{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j,m)}, \text{其中}\quad 1\le m\le M \tag{10}$$

$$\sum_m=\frac{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j,m)\cdot(O_t-\mu_m)\cdot(O_t-\mu_m)'}{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma_t(j,m)} \tag{11}$$

其中 $1\le m\le M$

步骤 4:收敛性判断:求 $P\{O\mid\lambda\}$,并对其作

$$\left|\frac{\Delta\ln P}{\ln P}<\varepsilon\right|$$

判断,若上式不成立并且循环次数在一定范围内,则退到步骤 2,否则,参数训练结束.

® 识别过程　假定训练过程中已经对训练集中每类发音的唇动变化建立了各自相应的 HMM 模型,设为 $\lambda_1, \lambda_2, \cdots, \lambda_K$, K 为类数. 我们用 Viterbi 算法进行识别. 识别过程如下:

步骤 1:求测试唇动序列图像的特征向量;

步骤 2:对每一个 HMM 模型 $\lambda_1, \lambda_2, \cdots, \lambda_K$,用 Viterbi 算法计算;

$$P\{O\mid\lambda_i\}=\max_{1\le i\le N}[\delta_T(i)]$$

步骤 3:取 $I=\arg\max_{1\le i\le K}[P\{O\mid\lambda_i\}]$,则测试序列属于 I 类;

步骤 4:对测试序列在模型 λ_i 下用 Viterbi 算法进行状态解码.

如果只需给出识别结果,则可略去步骤 4.

4　实验结果和结论

在我们的实验中,采用的是 CPE-1000 图像卡,JVC TK-1070 彩色摄像机和 MIMTRON MTV-3301CB 彩色摄像机,采集图像为 24 位真彩色图像,采集速度每秒 25 帧,CPU 为 Pentium-II 300.

实验中共采集了 10 个人分别发汉语拼音 a, o, e, i, u 的序列口型图像,其中每人每个音发 5 组,共得到 250 组序列图像. 序列图像的平均长度为 22 帧,即平均每个音的发音长度为 22 帧. 对这些图像序列用前面的方法进行了唇动检测、定位与特征提取. 用规格化后的差值特征 $\Delta\bar{C}_i$ 进行 SCHMM 模型训练和识别实验.

在实验中,对所有人每人选四组数据进行训练,另一组用于识别.

状态数 N 和混合项 M 的选取直接影响识别的结果和识别速度. 我们对状态数 N 和混合项 M 选取不同的值进行了实验,实验结果如图 7 所示.

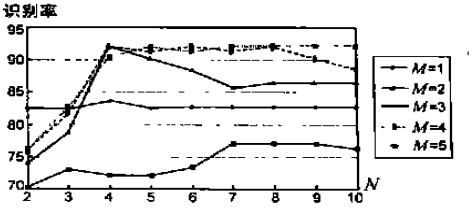

图 7　状态数 M 和混合项数 N 与识别率的关系

从图中可以看出适当地增加 N 和 M 可以提高识别率,但这也会增加训练和识别的参数,增加系统开销,并且需要较多的训练数据. 如果 N 和 M 选择较小则不能充分体现出唇动的特征分布,影响识别率. 从图中可以看出当 N 取 4, M 取 3 时系统对所选用的识别集具有较高的识别率和较少的参数. 表 1 为 $N=4, M=3$ 时的识别矩阵,其中每个发音有 50 组测试数据,此时的识别率为 92.4 %. 行中为实际的发音,列中为识别结果.

表 1　$N=4, M=3$ 时的识别矩阵

	a	*o*	*e*	*i*	*u*
a	50	0	0	0	0
o	0	44	0	0	6
e	1	0	47	2	0
i	0	0	8	42	0
u	0	2	0	0	48

上面是对非特定人进行动态识别的情况,同样我们对特定人进行了实验,实验中对每个人的四组数据进行训练,另一

组进行识别,重复该过程 5 次. 对所有人进行上述操作,则得到特定人的识别结果. 在我们先前的系统中曾对相同的识别集进行特定人与非特定人静态口型图像的识别实验,实验中选取每个发音的最具代表性的一帧作为识别数据. 表 2 给出这四个实验的识别结果.

表 2 不同方法的识别结果比较

	特定人	非特定人
静态识别	95.2 %	76.8 %
动态识别	95.8 %	92.4 %

从表中可以看出对于特定人,静态识别和动态识别的识别结果差异不大,而对于非特定人,动态识别的识别率明显高于静态识别. 这是因为对于特定人,对特定音的发音口型变化不大,因此静态的单帧图像具有很好的稳定性. 而不同的人由于发音习惯不同,对同一个音发音时的口型可能有很大的不同,但是唇的动态运动趋势是大致相同的,因此,动态识别才能取得较好的识别结果. 实验结果表明,本文提出基于色度分析的唇动特征提取和识别方法在自然条件下取得了令人满意的结果.

5 结论

本文对研究对象的色度进行了分析,找到了一个唇色滤波器,增强了唇目标,解决了可变模板方法的关键问题,使之成功地应用到唇动问题的特征提取和识别上. 下一步的工作是进一步扩大发音的集合,以期望逐渐满足辅助语音识别的要求. 唇读问题的解决,对增加与聋哑人的沟通有很大帮助;对聋哑人的教学,有更积极的意义.

参考文献:

[1] Stork D G, Wolff GJ, Levine W P. Neural network lipreading system for improved speech recognition [A]. Proc. of Inter. Joint Conf. on Neural Networks [C], 1992, 2: 289 - 295.

[2] Duchnowski P, Meier U, Waibel A. See me, hear me: integrating automatic speech recognition and lip-reading [A]. Proc. of Inter. Conf. on Spoken Language, ICSLP94 [C], 1994: 547 - 550.

[3] Hennecke M E, Prasad K V, Stork D G. Visionary speech: looking ahead to practical speechreading systems [A]. Speechreading by Humans and Machines [C], Berlin: Springer Verlag Press, 1996, Volume150 of NATO ASI Series F: Computer and Systems Sciences, 1996: 331 - 350.

[4] Luettin J. Towards speaker independent continuous speechreading [A]. Proc. of European Conf. on Speech Communication and Technology [C], Rhodes (Greece), 1997: 1 - 4.

[5] Hennecke M E, Prasad K V, Stork D G, Using deformable templates to infer visual speech dynamics [A]. 28th Asilomar Conf. on Signals, Systems, and Computers [C], Pacific Grove: IEEE Computer Society Press, 1994: 578 - 582.

[6] Gao W, Liu M B. A hierarchical approach to human face detection in complex background [A]. The First Inter. Conf. on Multimedia Interface [C], Beijing, 1996: 289 - 292.

[7] 姚鸿勋,刘明宝,高文,等. 基于彩色图像的色系坐标变换的面部定位与跟踪法 [J]. 计算机学报, 2000, 23(2): 158 - 165.

[8] 姚鸿勋,高文,李静梅,吕雅娟,等. 用于口型识别的实时唇定位方法 [J]. 软件学报, 2000, 11(8): 1126 - 1132.

[9] Coianiz T, Torresani L, Caprile B. 2D deformable model for visual speech analysis [A]. Speechreading by Humans and Machines [C]. Berlin: Springer Verlag Press, 1996, volume 150 of NATO ASI Series, Series F: Computer of Systems Sciences, 1996: 391 - 398.

[10] Rabiner L R. A Tutorial on hidden Markov models and selected applications in speech recognition [A]. Proc. of the IEEE [C]. 1989: 77 (2).

作者简介:

姚鸿勋 女. 1965 年生于浙江省杭州市. 1987 年 7 月、1990 年 3 月先后获得哈尔滨船舶工程学院计算机与信息科学系计算机应用专业学士和硕士学位. 现为哈尔滨工业大学计算机系副教授,从事图像处理、模式识别、多媒体技术及自然人机交互方面的研究工作. 已发表论文 20 余篇.

吕雅娟 女. 1972 年出生于黑龙江省齐齐哈尔市. 1994 年、1999 年分别在哈尔滨工业大学获得学士和硕士学位,现在哈尔滨工业大学计算机应用技术专业攻读博士学位,研究方向为图像处理、模式识别、自然语言理解、机器翻译等.