计算机科学2002Vol. 29№. 2

唇读识别中的基本口型分类*)

Basic Mouth Shape Classification for Speechreading

柴秀娟 姚鸿勋 高 文 王 瑞

(哈尔滨工业大学计算机科学与工程系 哈尔滨150001)

Abstract This paper has put forward a concept of mouth-shape basic unit, and described an approach of obtaining basic units by mouth-shape images classifying and clustering. The basic units are the gist of breakaway different states during continuous speech recognition or sequence images speechreading, which can definitely measure off statuses for mouth shape changing in sequence images. The method based on mouth-shape basic unit, compared with the approach based on feature vector directly, can rather reduce the number of state space branch, shrink searching space and expedite convergence rate. This paper introduces the preprocessing of mouth shape classification, the approach of real-time lip movement detection and classification, and gives experiment results of how to select the number of original clustering center in order to fit mouth shape classification and how to select features to be propitious to mouth shape classification, then gets a conclusion.

Keywords Speechreading, Clustering analysis, Automatic speech recognition

1 引言

自然人机交互方式使得人同计算机的交流不再局限于键盘、鼠标等外设,而是通过语言及手势、表情、唇动等形体语言来进行,从而使得人机交互变得像人与人之间的交流一样轻松自如。唇读通常被视为说话过程中伴随的辅助信息,它有助于对说话者提供信息的更准确理解,减弱噪音干扰。计算机唇读是指通过建立口型模型和分析运动参数,定量地处理唇动信息辅助进行语音识别[1,2],或者是直接对序列图像进行分类和识别。

发音是一个唇部和喉部肌肉连续运动的过程,人在发相同的音时,肌肉运动是极为相似的。构成运动的各个状态的描述是问题的关键。描述不同的状态就必须明确各种口型。因此在人类说话过程中究竟存在多少种不同口型是研究唇读问题中的一个基本问题,也是一个共同的问题。本文就是试图解决这一问题。口型分类明确了各种状态对应的口型,去掉了状态变化的不确定性,缩小了状态空间,提高了最佳状态匹配的收敛速度。

目前,国内外学者对口型分类进行研究的甚少,只有Fisher曾在68年提出了视素(Viseme)的概念,即语音在视觉意义上的最小可区分单元。大多数学者的做法是直接用特征向量序列来进行HMM模型的状态匹配,而非图像序列的HMM模型状态匹配,这样做一方面可能会因为特征向量各维选择的准确性和充分性与否而影响到匹配结果的正确性,另一方面状态匹配的搜索速度会因为状态的不确定性而大大增加时间和空间的开销,只不过目前研究唇读的集合都相当小,尚未充分体现出这方面的弊端来。如果口型图像有分类,就可以避免特征选取的片面性和状态空间转移的不确定性。我们利用聚类分析算法对口型进行了分类。

针对全部汉语的发音,可区分的口型究竟有多少种,即完成基本的视觉类别的判别。这样如果在对序列图像进行处理时,就可以知道每一幅图像对应的口型是属于哪一类的,也就知道发一个音时它的口型一共经历了哪些变化,即得到发音时的口型变化序列,然后进行识别。口型分类的正确与否直接影响着对视觉类别的判别,从而决定着识别的正确与否,这是唇读与语音识别工作的基础。

2 口型分类的预处理工作

唇的检测是实现唇读的基本前提。首先要进行人脸检测。由于人脸肤色在颜色空间中的分布相对比较集中,因此利用彩色信息检测与跟踪人脸,具有特征提取简单、计算速度快、姿态及尺寸不敏感的特点。但由于图像背景的非限定性,图像中不可避免地存在一些与人脸肤色相近的甚至相同的背景(如人体其它部位的皮肤)。因此,肤色模型只能作为一个粗检测手段。要想在图像中精确地定位人脸,必须充分利用人脸这一特定模式类的各种共性特征。我们采用特征脸(Eigenface)技术进行了人脸精检测[3]。

在人脸检测实现以后,取脸的下半部分区域作为研究对象进行唇动检测。

从彩色空间中,我们不难发现不同色调的各类物体聚类在各自不同的色带中。由此,我们可以利用彩色信息在图像中快速地检测出相应的目标[4,5]。为了节省对序列图像目标检测的时间,可采用运动预测技术,即对时间图像序列中相邻帧目标的运动量进行估计。采用运动检测技术,可以极大地提高搜索效率,加快定位速度,使系统具有更高的实时性。

3 口型分类

3.1 用于口型分类的聚类算法

*)本文的研究工作得到国家863计划项目(863-306-QN99-4)、国家863计划项目(863-306-ZT03-01-2)、国家自然科学基金重点项目(69789301)和中科院百人计划的资助。柴秀娟 硕士生,主要研究领域:多媒体技术等。姚鸿勋 副教授,主要研究领域:人脸面部感知,身份识别,信息处理,多媒体技术等。高 文 教授,博士生导师,主要研究领域:多媒体数据压缩,图像处理,计算机视觉,多模式接口,人工智能,虚拟现实等。王 瑞 博士生,主要研究领域:唇读,多模态人机交互等。

用于口型分类的算法是动态聚类算法。所谓动态聚类算法是先选择若干样品作为聚类中心，再按某种聚类准则，例如最小距离准则使其余样品向各聚类中心聚焦，从而得到初始分类，然后判断初始分类是否合理，若不合理就修改分类。如此反复迭代，直到获得合理的分类。

我们采用 k-均值算法作为口型分类的聚类算法，采用误差平方和准则。

若 N_i 是第 i 聚类 Γ_i 中的样本数目，m_i 是这些样本的均值，即

$$m_i=\frac{1}{N_i}\sum_{y\in\Gamma_i}y \tag{1}$$

把 Γ_i 中的各样本 y 与均值 m_i 间的误差平方和对所有类相加后为

$$J_c=\sum_{i=1}^{c}\sum_{y\in\Gamma_i}\|y-m_i\|^2 \tag{2}$$

J_c 是误差平方和聚类准则，它是样本集合类别集 Ω 的函数。J_c 度量了用 k 个聚类中心 $m_1,m_2,\cdots,m_k$ 代表 k 个样本子集 $\Gamma_1,\Gamma_2,\cdots,\Gamma_k$ 时所产生的总的误差平方。对于不同的聚类，J_c 的值当然是不同的，使 J_c 极小的聚类是误差平方和准则下的最优结果。

为得到最优结果，首先要对样本集进行初始划分(分类)。我们采取了选取前 k 个样本点作为代表点的方法。这种方法最好是前 k 个代表点之间差异较大，能够尽可能地代表几个不同的类，因为初始代表点的选择往往会影响迭代的结果(得到的可能是局部最优解而不是全局最优解)。选定代表点后，计算其余的点离这些代表点的欧式距离，离哪个代表点的距离最小就归入哪一类。

有了初始分类之后，要进一步迭代以谋求聚类中心的准确性。即把聚类 Γ_c 中的一个样本 y 从 Γ_c 移入 Γ_j 对误差平方和所带来的影响。设从 Γ_c 中移出 y 后的集合是 $\tilde{\Gamma}_c$，它相应的均值是 $\tilde{m}_c$，显然

$$\tilde{m}_c=m_c+\frac{1}{N_c-1}[m_c-y] \tag{3}$$

式中的 m_c 和 N_c 是 Γ_j 的样本均值和样本数。

设 Γ_j 接受 y 后的集合是 $\tilde{\Gamma}_j$，它相应的均值是 $\tilde{m}_j$：

$$\tilde{m}_j=m_j+\frac{1}{N_j+1}[y-m_j] \tag{4}$$

式中的 m_j 和 N_j 是 Γ_j 中的样本均值和样本数。由于 y 的移动只影响 Γ_c 和 Γ_j 两类，而对其他的类是无任何影响的，因此我们只需计算这两类的新的误差平方和 $\tilde{J}_c$ 和 $\tilde{J}_j$。

$$\tilde{J}_c=J_c-\frac{N_c}{N_c-1}\|y-m_c\|^2 \tag{5}$$

$$\tilde{J}_j=J_j+\frac{N_j}{N_j+1}\|y-m_j\|^2 \tag{6}$$

假使

$$\frac{N_j}{N_j+1}\|y-m_j\|^2<\frac{N_c}{N_c-1}\|y-m_c\|^2 \tag{7}$$

则把样本从 Γ_c 移入到 Γ_j 就会使误差平方和减少。只有当 y 离 m_j 的距离比离 m_k 的距离更近时才满足上述不等式。

实验证明该算法存在以下两方面的缺陷：

1)算法的结果受到所选聚类中心的个数 k 及初始聚类中心选择的影响，也受到样品的几何性质即排列次序的影响。实际上需试探不同的 k 值和选择不同的初始聚类中心。如果样品的几何特性表明它们不能形成几个相距较远的小块孤立区，则其收敛效果就不很好。

2)算法采用均值作为一个类的代表点，这只有当类的自然分布为球状或接近于球状时，即每类中各分量的方差接近相等时，才能达到最好的效果。

为了能够比较口型分类的结果是否正确，以及口型分类的结果是否能很好地用于唇读系统中，我们又采用了改进的 IRPCL 算法。

改进的 IRPCL 算法，基本上与 RPCL 算法相同，只是加入反例以加快聚类速度，从而改进了原有算法的效率。其算法的步骤如下：

(1)计算各样本的密度。

(2)进行初始化，选择代表点。即求出所有样本中各维的最大值和最小值，以此选择出 k 个初始聚类中心，并记录下来。

(3)统计远离中心的样本集和样本数。

(4)IRPCL 算法收敛检验。各权相对变化是否均小于规定值。若小于，并且尚未达到最大迭代次数，则收敛。若不小于，进行第(5)步。

(5)去掉活化次数小于阈值的空元，确定最优子集数。

我们将算法的理论与实验的结果结合起来分析，可以发现这两种算法之间的异同。

(1)在计算时间上，k-均值算法明显要比改进的 IRPCL 算法快很多。但在结果上，改进的 IRPCL 算法却可能使 k-均值算法中有些明显分错类的情况纠正过来。总体来说，k-均值算法是一种原理简单，时间复杂度低，计算快捷的分类算法，因此也是模式识别中较常用的一种方法。

(2)通过实验及理论分析，我们发现，点集的数据构造、被分析的点集中样本点的数量、所采用的距离度量和相似性度量、所用的聚类准则以及最终的聚类数都会影响到分类的结果。此外，样本各分量之间的尺度比例的确定也是一个十分重要的问题。改进的 IRPCL 算法不仅能够通过调整样本所属类别完成聚类分析，而且还能自动地进行类的合并和分裂，从而得到类数较为合理的各个聚类。

(3)我们在使用 k-均值算法时，是选择前 k 个向量作为初始聚类中心。其前提必须是这 k 个向量基本上能代表一定的类别。而在改进的 IRPCL 算法中，我们首先计算了样本密度，求出了所有样本中各维最大值和最小值，从而根据各维的最大和最小值，选择了 k 个初始聚类中心。相对来说，这种选择初始聚类中心的方法比较选择前 k 个点的方法一般来说都会产生更好的结果。

3.2 分类算法中参数的选择

如何才能得到较为合理的口型分类结果，以便于正确引导唇读，最关键的问题是模式应该分成几类，用哪些特征进行分类，特征向量多少维合适，每维的权值应该是多少。通过实验，可以找到这些答案。

3.2.1 *初始聚类中心个数的确定* 口型分类算法的结果由于所选聚类中心的个数(即 ck)不同而有差异，如果 ck 值取得过小，必然导致原本距离很大的样本被分到同一类中；而 ck 值如果取得过大，也会影响分类的精度，导致过多的合并处理，并可能引起较大的误差，造成分类结果的不合理性。

那么，初始聚类中心的个数要如何选定呢？

在类别数未知的情况下，使用 k-均值算法时，可以假设类别数是逐渐增加的。例如，对 k=1，k=2，k=3…分别使用该算法，然后结合对该问题的知识，分析不同的聚类结果以判断哪一个结果比较好。据此，选择最佳的初始聚类中心个数。

在进行分类算法实现中(ck 表示初始聚类中心的个数)，

ck 越大，表示可以利用更多的样本信息以确定它的类别。并且，ck 取大些有利于减少噪声影响。但是，ck 取得过大，各类之间的相似性会混淆了实际科学的分类结果，因此一定存在一个使误识率最小的 ck 值。

通过大量的实验，并观察分类结果，从中找出被错分类的图像，记录下来，算出当 ck 取不同值时，算法分类的正确率，从而确定出适合的 ck 值。实验表明，$ck=8$ 最为合适。实验结果如图1所示。

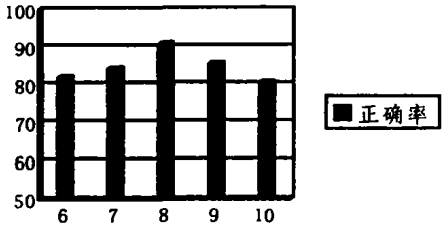

图1　确定初始聚类中心个数的实验结果

其中，横坐标表示 ck 的值，相应的纵坐标表示当 ck 取不同值时对口型的正确分类率。

3.2.2　*用于分类的特征参数的选取*　要想实现对口型的正确分类，就一定要有能够区分不同口型的特征参数。口型模板如图2所示。

图2表示刻画口型一些重要参数：上外唇高度 h_1、上内唇高度 h_2、下内唇高度 h_3、下外唇高度 h_4、内唇宽度 w_1、外唇宽度 w_0 等。

图3给出了选取不同的特征参数组进行分类时的正确率的情况。图3横坐标表示取的五组不同的参数，纵坐标为相应的取不同参数组的分类正确率。

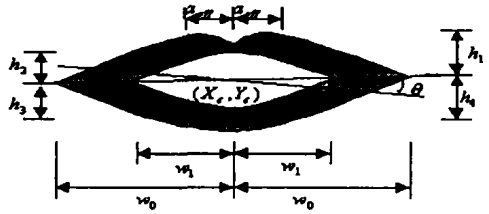

图2　刻画口型的参数

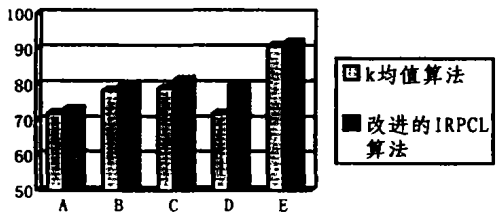

图3　选取不同特征参数分类正确率结果

表1给出了五个参数组包含的具体参数。

实验结果表明，E 组参数的口型分类结果是比较准确的。它能把相似的口型分到同一类中，并形成了8个类。通过大量的实验发现，上唇内高和下唇内高虽然作为特征参数进行分类都有一定的效果，但是如果把这两个参数相加，其分类效果明显好于它们单独作参数时的情形。在选择参数的时候，因为口型的面积是一个重要的参数，它可以决定口型张开的大小，最根本的是明显地区分出张嘴和闭嘴，但实际上要区分张嘴和闭嘴，上唇内高与下唇内高相加也可达到这个目的。当其等于0时，就是闭嘴的情况。实验还发现，选择面积作参数的一个更大的弊病，就是它会混淆两种面积相同但形状却相差很多的口型，这样就达不到对口型正确分类的目的。

虽然一个参数不能决定分类结果，但它会对其他的参数产生影响，同时牵制其他的参数发挥作用，故这个特征不适合作为分类参数。因为面积参数在理论上有一定的积极作用，所以我们尝试着对这个参数做些限定，选择面积与内唇宽的比值作为一新的特征参数，实验结果表明对面积参数的选择并没有达到我们预先设想的目标。

表1　五组参数

A 组	1. 张开的唇包围的面积	4. 内唇宽
	2. 内唇高度	5. 唇的外边缘高度
	3. 内唇宽/原唇宽	6. 两嘴角间距离
B 组	1. 内唇高度	5. 外唇宽
	2. 内唇宽/原唇宽	6. 两嘴角间距离
	3. 内唇宽	7. 图像的7种灰度特征
	4. 唇的外边缘高度	
C 组	1. 模板中心横坐标	5. 下唇内高
	2. 横板中心纵坐标	6. 下唇外高
	3. 上唇外高	7. 外唇宽度
	4. 下唇内高	8. 内唇宽度
D 组	1. 内唇高度	
	2. 内唇宽/原唇宽	
	3. 内唇宽唇的外边缘高度	
	4. 外唇宽	
	5. 张开的唇包围的面积/唇内宽	
E 组	1. 内唇高度	4. 唇的外边缘高度
	2. 内唇宽/原唇宽	5. 外唇宽
	3. 内唇宽	6. 两嘴角间距离

B 组选择了图像的7种灰度特征作为特征参数来分类。但实验证明灰度特征不适合进行口型分类，它们不会提高分类的正确率。而且由于参数过多，还会牵制其他的参数发挥作用，因此决定不采用灰度特征进行分类。

C 组加入了口型模板中的两个参数，模板中心横坐标，以及模板中心纵坐标。实验证明，这两个参数只是用于唇的定位，对唇型的分类没有积极的作用，相反会因为这两个参数的加入，而导致分类结果的不准确，因此也不被采用。

E 组参数加入了两嘴角间的距离这个参数，实际上我们的聚类算法中没有考虑参数加权的问题，假定各参数在分类中所起的作用是不同的。但实际上对于分类各参数所起的作用肯定有所区别。我们认为外唇的宽度对口型分类应该有较大的作用，而两嘴角间的距离和外唇宽度有相似的意义，实验表明，这个参数组的分类效果有较大的提高。故嘴的宽度是口型分类的一个很重要的参数。

4　分类实验与结果

对采集的具有典型口型分布的246帧中文单韵母发音图像进行口型基元提取。图像分辨率为256×256×24bits，基本含有简单背景的正面人脸图像，平面旋转不超过30°，深度旋转不超过15°。人脸大小比例变化范围为30%。采集时没有特殊的光线限定，均为自然光条件。

分别用 k-均值方法和改进的 IRPCL 算法对上述图像进行分类实验，在全部246幅图像中，分类正确的比率分别是90.75%和91.7%，通过实验，发现这两种算法的结果很稳定，

将口型分成八种基本口型。现在选择八种基本口型中的几幅图像作为示例。如图4所示。

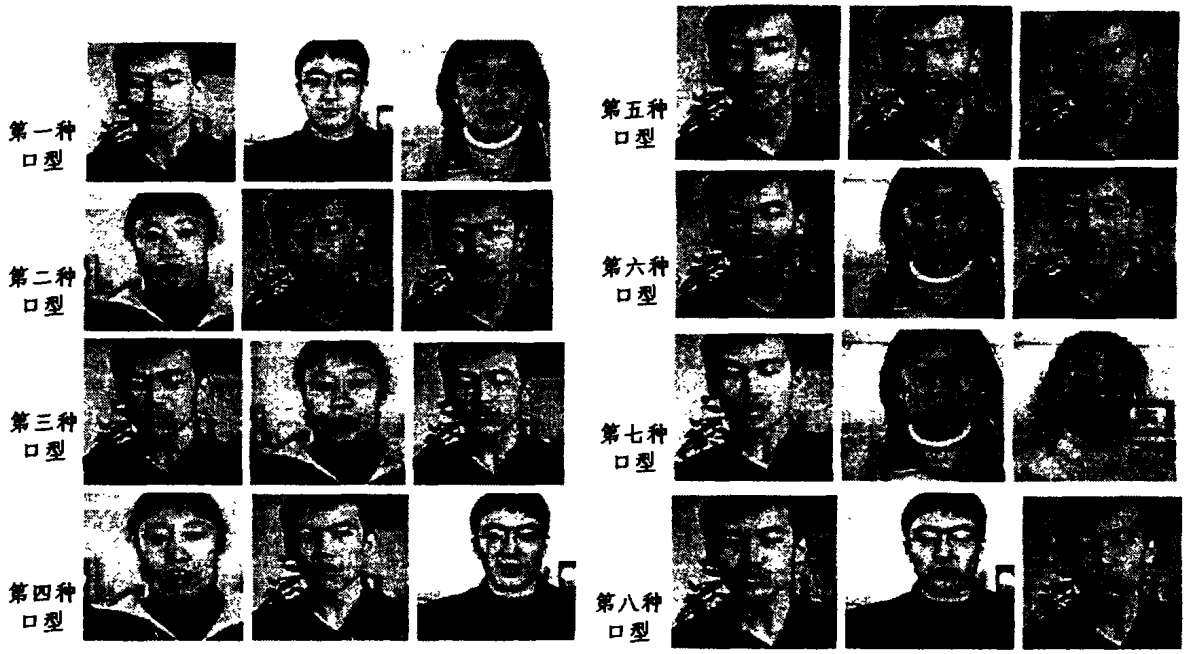

图4 基本口型分类结果之例

通过分析发现，导致产生某些口型误分类的原因，主要有两个方面。第一是因为原始图像嘴角定位不准确，误分类的图像的嘴角常常处于阴影之中，使其颜色与正常唇色有很大差异，而被排除在唇色之外；第二是因为在某些口型中，唇色与露出来的舌头的颜色相近，因此导致口型定位不准确，致使参数并不能描述它所代表的口型，从而导致误分类情况。

本文给出了中文单韵母口型分类的具体结果，产生了衡量中文单韵母发音过程中各幅图像所处状态的明确的准则，减小了发生状态转移时的不确定性，为在唇读识别中序列图像的状态转移提供了一个较为清晰的配准。

5 识别结果和结论

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，采集的图像为24位真彩色图像，采集速度每秒25帧，其中每人每个音发5组，共得到250组序列图像。对图像序列进行时间归正，然后用前面介绍的方法进行了唇检测与分类，并用 SCHMM 模型训练和识别。

表2 识别结果

特定人	非特定人
95.8%	92.4%

实验证明，口型分类对唇读识别确实有着极其重要的意义，它为连续语音识别各个状态的确定提供了明确的依据。把基本口型作为唇读识别的基元用于连续语句的识别是一种新的行之有效的方法。虽然其识别率还不算很高，但那是由于口型序列到语音序列的映射不是唯一的，如果我们的语言模型能够再复杂一些，我们的序列对应关系会找得更好，即识别率会更高。这正是下一步我们要做的工作。

参 考 文 献

1 Stork D G, Wolff G J, Levine E P. Neural Network Lipreading System for Improved Speech Recognition. In: Proc. Intl. Joint Conf. on Neural Networks, 1992, 2: 289～295

2 Hennecke M E, Stork D G, Prasad K V. Visionary Speech: Looking ahead to Practical Speechreading Systems. In: David G. Stork and Marcus E. Hennecke, eds. Speechreading by Humans and Machines, Springer and Systems Sciences. 1996. 331～350

3 Gao W, Liu M B. A Hierarchical Approach to Human Face Detection in Complex Background. the First International Conference on Multimodal Interface, Beijing, 1996

4 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. 软件学报，2000，11(8)：1126～1132

5 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. 计算机学报，2000，23(2)：158～165

(上接第145页)

3 Lakhtakia A, Varadan V V, Messier R, Varadan V K. On the symmetries of the Julia sets for the process $Z \leftarrow Z^p + c$. J Phys A: Math Gen., 1987, 20: 3533～3535

4 Gujar U G, Bhavsar V C. Fractals from $Z \leftarrow Z^a + c$ in the Complex c-plane. Computers & Graphics, 1991, 15(3): 441～449

5 Gujar U G, Bhavsar V C, Vangala N. Fractals images from $Z \leftarrow Z^a + c$ in the Complex z-plane. Computers & Graphics, 1992, 16(1): 45～49

6 Dhurandhar S V, Bhavsar V C, Gujar U G. Analysis of z-plane fractals images from $Z \leftarrow Z^a + c$ for $\alpha < 0$. Computers & Graphics, 1993, 17(1): 89～94

7 王兴元，朱伟勇. 正实数阶广义 J 集的嵌套拓扑分布定理. 东北大学学报(自然科学版)，1999，20(5)：489～492

8 Wang Xingyuan, Liu Xiangdong, Zhu Weiyong, Gu Shusheng. Analysis of c-plane fractal images from $Z \leftarrow Z^a + c$ for . $\alpha < 0$ Fractals, 2000, 8(3)

9 王兴元，刘向东，朱伟勇. 由复映射 $z \leftarrow z^a + c(\alpha < 0)$ 所构造的广义 M-集的研究. 数学物理学报，1999，19(1)：73～79

10 Falconer K J. 分形几何——数学基础及其应用. 曾文曲，刘世耀译. 沈阳：东北大学出版社，1991. 238～245

11 Blancharel P. Complex analytic dynamics on the riemann sphere. Bulletin of the american mathematical society, 1984, 11: 88～144