# 在DCT域进行LDA的唇读特征提取方法

何 俊,张 华,刘继忠
HE Jun,ZHANG Hua,LIU Ji-zhong

南昌大学 江西省机器人焊接重点实验室,南昌 330031
Key Laboratory of Robot & Welding,Nanchang University,Nanchang 330031,China
E-mail:welding@ncu.edu.cn

**HE Jun,ZHANG Hua,LIU Ji-zhong.LDA based feature extraction method in DCT domain in lipreading.Computer Engineering and Applications,2009,45(32):150-152.**

**Abstract**: To solve the key problem of extracting visual speech feature in lipreading,a method based on DCT and LDA is proposed.To extract most discriminative visual feature among different mouth classes,first,DCT is performed on the visual speech region;and then based on LDA the most discriminative feature vector is extracted from DCT coefficients.The experiments on speaker-dependent,speaker-independent database and in real-time lipreading environment show that this method is more effective than traditional manual DCT coefficients extraction method and PCA feature extraction method.

**Key words**: lipreading;feature extraction;Discrete Cosine Transformation(DCT);Linear Discriminative Analysis(LDA)

**摘 要**:为解决视觉语言特征提取这个唇读技术中最关键的难题,提出一种新的基于DCT和LDA的特征提取方法。为提取对不同口型最具分类能力的特征矢量,首先基于DCT对视觉语言部位变换降维,然后基于LDA算法从DCT系数提取对口型分类性能最优的特征矢量。在特定人与非特定人的唇读数据库上以及实时唇读识别的实验都表明,该方法唇读识别率比传统的人工直接选择DCT系数法以及PCA提取法有明显提高。

**关键词**:唇读;特征提取;离散余弦变换(DCT);线性判别分析(LDA)
**DOI**:10.3778/j.issn.1002-8331.2009.32.047 **文章编号**:1002-8331(2009)32-0150-03 **文献标识码**:A **中图分类号**:TP391.1

## 1 引言

从本质上说,人类语言的产生和感知都是多模态的,早在1954年就证明了视觉通道对噪声环境下语音识别的有效性。这给噪声环境下的语音识别以及有听觉障碍的残疾人、老年人提供了更好的人机交互方式。如图1是双模态唇读语音识别的流程图。唇读技术主要包括以下几个环节:人脸检测、视觉语言特征定位、特征提取、识别(或和语音融合后再识别),其中视觉语言特征提取是最关键的环节。特征提取技术可以分为以下三种方法[1]:基于形状的方法、基于像素的方法和混合方法。基于形状的方法认为大部分语言信息包含在嘴唇的轮廓中,或者更广泛一点,包含在脸部轮廓中,如嘴唇、下巴、面颊等,有的采用嘴唇的几何参数如宽度、高度、面积、圆度等描述嘴唇[2],有的采用傅里叶描述子、图像矩、主动形状模型(ASM)描述轮廓特征[3];基于像素的方法把所有视觉语言特征区域的每个像素都作为语言特征,但缺点是维数太高,一般采用图像变换的方法降维,例如PCA(Principal Component Analysis)、DWT(Discrete Wavelet Transformation)、DCT(Discrete Cosine Transformation)等;所谓混合法,是把前两种方法得到的特征组合在一起形成共同的特征矢量,主动面模型(AAM)就是基于这组特征训练得到的。有研究表明基于像素的方法比基于形状的方法有更高的识别率[1]。该文主要基于像素法的唇读特征提取进行了研究,在完成人脸定位后,首先对视觉语言特征图像进行离散余弦变换(DCT),然后提出一种基于线性判别分析(Linear Discriminative Analysis,LDA)提取最优特征矢量的方法,在特定人、非特定人数据库上的实验均证明了该方法的有效性。

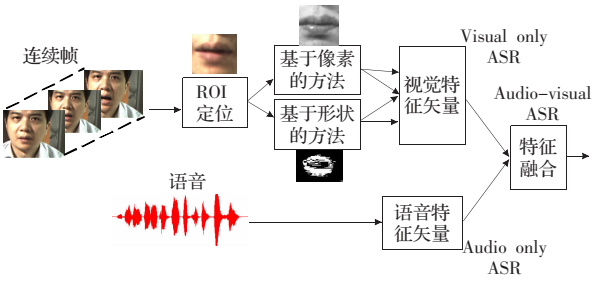


图1 语音唇读识别流程示意图

## 2 视觉语言特征提取

大多数文献一般认为视觉语言特征区域(Region of Interest,ROI)是一块以嘴唇为中心的嘴部区域[1]。也有的文献将下

**基金项目**:江西省教育厅基金(the Educational Foundation of Jiangxi Province of China under Grant No.GJJ08012);南昌大学创新团队基金。
**作者简介**:何俊(1969-),男,博士生,副教授,主要研究领域为模式识别,图像处理;张华(1964-),男,博士生导师,教授,主要研究领域为机器人焊接自动化,人工智能;刘继忠(1974-),男,博士,副教授,主要研究领域为信号处理,模式识别。
**收稿日期**:2008-06-18 **修回日期**:2008-10-15

巴和面颊包含进来[4]，甚至是整张脸。如图 2，分别为以嘴唇为中心的 ROI 和包含了下巴的 ROI。对于不同口型类别，希望从 ROI 提取出对不同口型最具分类能力的特征矢量。

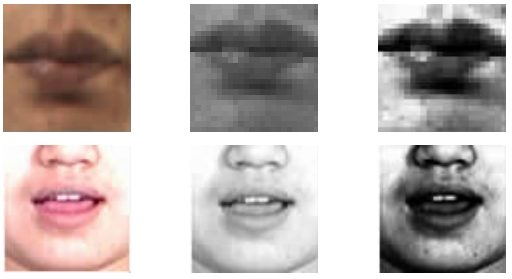


**图 2　ROI 灰度化并直方图均衡**

## 2.1　ROI 预处理

首先根据肤色唇色自适应检测算法[5]检测出嘴唇区域，得到左右嘴角、上唇最高点和下唇最低点 4 个关键点，进而计算出嘴唇的宽度和高度，得到嘴部图像，为消除人脸和镜头距离变化的影响，把 ROI 归一为 24×24 的灰度图像，并进行直方图均衡后得到归一化的 ROI，如图 2。

## 2.2　DCT 降维

由于归一化后 ROI 区域 496 维的特征维数太高，一般采用图像变换的方法降维。该文就常用的两种图像变换方法 DCT 法、PCA 法进行了比较。二者都是基于正交变换的原理将输入矢量在正交坐标系中投影，不仅可以减少各个向量的相关性，而且投影后能量会集中在少数变换系数上，若去除较小的分量保留较大的分量，还原出的信号失真就很小并能保留绝大部分的原信号特征，这就是基于图像变换法降维的原理。和 DCT 法比较，PCA 法（也称 KL 变换）是在最小均方误差意义上的最佳变换，能完全去除原信号的相关性，但是 PCA 法也有三个明显的缺点：一是变换的基向量依赖于原信号的协方差矩阵，每次增加样本训练时，要重新计算该矩阵；第二个缺点是若原信号维数较高，则特征值和特征向量的计算比较困难；第三个缺点是 PCA 法提取的是主分量，那些具有分类信息的较小分量因此可能被忽略掉。而 DCT 有类似于离散傅里叶变换（FFT）的快速算法，数据压缩能力在一定条件下近似 KL 变换，在实时场合下更有应用价值，而且 DCT 系数有明确的物理意义。因此，采用 DCT 法对 ROI 降维。

如图 3，是从哈尔滨工业大学制作的 HIT Bi-CAV Database Ⅱ唇读数据库中提取的以嘴唇为中心的 ROI 区域进行 DCT 降维示意图。首先，归一化为 24×24 的 ROI 被分成 9 个 8×8 的子图像，分别进行 DCT 降维，如图 3(g)。图 3 左下图是用 DCT 的各个系数恢复的图像，区域 0 是直流分量，代表了图像的粗轮廓特征，如图 3(d)；区域 1 是低频分量，代表了 ROI 的轮廓特征，如图 3(e)；区域 2 是中频分量，代表了 ROI 的细节特征，如图 3(f)；区域 3 是高频分量，代表了 ROI 区域更细微的特征。即不同的 DCT 系数代表了 ROI 区域的不同特征，一般认为DCT 法是一种基于像素的特征提取方法，但根据图 3 可以认为它的各个 DCT 系数实际也是轮廓特征的一种描述。

大多数文献都将那些较大的 DCT 系数直接作为用于唇读识别的特征矢量[1]。传统的 DCT 提取系数方法有很多，如典型的 Zig-Zag 法等，但基本原理都是保留左上角能量较大的低频值。如图 3(a)、(b)、(c)是分别用左上角 27、54、81 个 DCT 系数

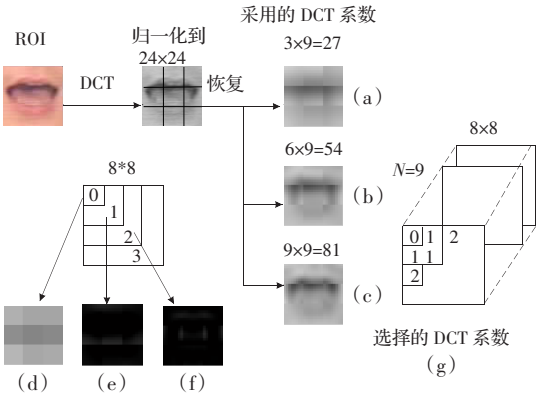


**图 3　ROI 进行 DCT 变换的示意图**

(a)用左上角 27 个 DCT 系数还原的 ROI 图像；(b)用左上角 54 个 DCT 系数还原的 ROI 图像；(c)用左上角 81 个 DCT 系数还原的 ROI 图像；(d)0 区域 DCT 系数对应的图像；(e)1 区域 DCT 系数对应的图像；(f)2 区域 DCT 系数对应的图像；(g)分块 DCT 示意图

还原的 ROI 图像，显然保留 54 个 DCT 系数已经有较好的还原效果。这个方法的思路是提取 ROI 区域的主分量，但对唇读口型识别来说，主分量并不表示具有最好的口型分类效果，有些次要分量也可能包含重要的分类信息，如图 3，虽然图 3(d)与图 3(e)、(f)相比是 ROI 主分量，具有更大的 DCT 系数，但显然图 3(e)、(f)比图 3(d)包含了更重要的口型信息。如何从这些 DCT 系数中提取最具有分类效果的特征并实现进一步降维？提出一种新的基于线性判别分析（LDA）的唇读特征提取法。

## 2.3　基于 LDA 的特征提取

首先，根据嘴唇的长、宽、长宽比、是否露齿将基本口型聚类为 10 类。设每幅图像的特征矢量为 $\boldsymbol{V}_i=[v_{i1},v_{i2},\cdots,v_{i496}]$，为简化计算，首先从 496 个 DCT 系数中按 Zig-Zag 法选择较大的前 135 个 DCT 系数作为特征矢量，然后设法把这 135 维特征投影到新的空间中提取 $d$ 维对 10 种口型分类性能最好的特征矢量。

根据类别可分离判据的特征提取原理，要实现特征降维后仍具有很好的分类能力，就是要找到一个变换矩阵 $\boldsymbol{W}$，使得变换后的特征矢量总的类间距离和总类内距离的比值最大化，以下是变换矩阵 $\boldsymbol{W}$ 的求解过程：

设原始特征总的类内散布矩阵和总体散布矩阵分别为式(1)和式(2)：

$$\boldsymbol{S}_w=\sum_{i=1}^{c}\sum_{x\in Di}(x-m_i)(x-m_i)' \tag{1}$$

$$\boldsymbol{S}_T=\sum_{x}(x-m)(x-m)' \tag{2}$$

其中，$m=\frac{1}{n}\sum_{x}x=\frac{1}{n}\sum_{i=1}^{c}n_im_i$，特征提取是从 135 维空间向 $d$ 维空间投影，即通过 $d$ 个方程来进行：

$y_i=w_i'x\quad i=1,2,\cdots,d$

写成矩阵形式：

$$\boldsymbol{y}=\boldsymbol{w}'\boldsymbol{x} \tag{3}$$

则变换后特征矢量的类内散布矩阵为式(4)，类间散布矩阵为式(5)。

$$\overline{\boldsymbol{S}_w}=\sum_{i=1}^{c}\sum_{y\in Yi}(y-\overline{m_i})(y-\overline{m_i})' \tag{4}$$

$$\overline{\boldsymbol{S}_b}=\sum_{i=1}^{c}n_i(\overline{m_i}-\overline{m})(\overline{m_i}-\overline{m})' \tag{5}$$

容易证明：

$\overline{S_w}=W'S_wW \quad \overline{S_b}=W'S_bW$

通过变换矩阵 $W$，应使得变换后类间离散度和类内离散度的比值最大。离散度的一种简单标量度量是散布矩阵的行列式的值。由于行列式的值等于矩阵的本征值的乘积，也就是在各个主要分布方向上的方差的积，因此，得到准则函数如下：

$$J(W)=\frac{|\overline{S_b}|}{|\overline{S_w}|}=\frac{|W'S_bW|}{|W'S_wW|} \tag{6}$$

使准则函数 $J(W)$ 最大解出变换矩阵 $W$，可参阅文献[6]，通过求解方程(7)得到变换矩阵 $W$。

$$(S_b-\lambda_i S_w)W_i=0 \tag{7}$$

最后，根据式(3)提取到最具分类能力的 $d$ 维特征矢量。

## 3 实验

分别采用了哈尔滨工业大学制作的 HIT Bi-CAV Database Ⅱ特定人唇读数据库和华南理工大学制作的非特定人唇读数据库检测算法的有效性。

### 3.1 DCT 系数归一化

为尽量消除照度的影响，需要对 DCT 系数大小归一化。在把 ROI 归一化到 24×24 大小的图像后，进行分块 DCT 变换，为简化计算，按 Zig-Zag 法保留每个子图像左上角较大的 15 个系数，一共得到 15×9=135 个系数留作下一步特征提取。设第 $i$ 幅图像的任一维特征矢量 $v_{ij}(j=1,2,\cdots,135)$ 服从正态分布，对每维特征矢量按式(8)进行归一化：

$$v_{ij}'=\frac{v_{ij}-\mu_i}{3\sigma_i} \tag{8}$$

其中，$v_{ij}$ 表示第 $i$ 幅图像的第 $j$ 个 DCT 系数，$\mu_i$ 为该系数的均值，$\sigma_i$ 为方差，把 DCT 系数都归一化到(0,1)区间。最后，根据 LDA 算法从 135 维中提取 $d$ 维最具分类能力的特征矢量。

### 3.2 特定人唇读识别

特定人唇读识别实验数据库选用的是哈尔滨工业大学制作的 HIT Bi-CAV Database Ⅱ唇读数据库，该库包含 1 个人，约 1 000 个词，每个词发音 3 遍，在结构光下拍摄。随机选择了数据库中的 10 个词，采取了“留一识别法”，即用 2 遍发音用于训练，剩下的 1 遍发音用于识别。所有的训练和识别采用连续结构的隐马尔可夫模型(CHMM)，每个词用 3 个状态，考虑到每个词包含的帧数有限，用于训练混合高斯模型的数据量较小，在实验中每个状态只采用 2 个混合高斯模型。

实验中分别采用了人工 Zig-Zag 法、PCA 法、LDA 法 3 种不同方法对 DCT 系数进行特征提取。图 4 显示了当提取的特征维数 $d$ 分别为 9 维、27 维、54 维、90 维时的唇读识别结果。实验显示，人工法和 PCA 法的唇读识别率相差不大，而 LDA法的唇读识别率较其他两种方法有较大提高，在特征维数较低时表现得尤为明显。显然这是因为前两种方法提取的都是 DCT 系数的主分量，而 LDA 提取的是 DCT 系数中最具分类能力的分量。实验也显示提取的特征维数 $d$ 过小或者过大都会影响到唇读识别率，当选择 $d$=54 维特征时，基于 LDA 取得了 68.4% 的最好唇读识别结果。

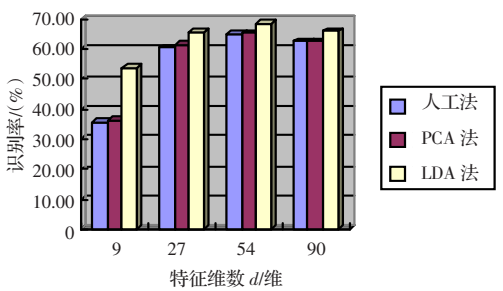


图 4 三种特征提取方法识别结果比较

### 3.3 非特定人唇读识别和实时唇读识别

非特定人数据库选用的是华南理工大学制作的唇读数据库，共包含了 40 人，约 100 个词，每个词说一遍，在结构光下摄制，从中选择了拍摄效果较好的 10 个样本；特定人实时唇读识别实验照明采用的是普通日光灯，实验采用的软件平台是 VC6.0 和 MATLAB7.1，硬件是北京大横公司的数字摄像头 DH-HV1301UC。在 VC 上实现实时采集 ROI 图像，在 MATLAB 上完成训练和识别。此外，为了检测不同 ROI 区域对唇读识别的影响，分别选择了嘴唇和下半脸区域作为 ROI，基于 LDA 提取 54 维特征，实验结果如表 1。

表 1 基于 LDA 法在不同条件下的实验结果

| 实验对象 | | 识别率/(%) | |
|---|---|---|---|
| | | 嘴唇 | 下半脸 |
| 特定人 | 数据库 | 68.4 | 70.4 |
| | 实时 | 38.4 | 39.3 |
| 非特定人 | 数据库 | 23.8 | 22.3 |

表 1 显示基于 DCT+LDA 算法，特定人的唇读识别率明显高于非特定人，据分析主要有两个原因：一是因为基于像素法的特征提取方法利用了全部 ROI 的像素信息，即提取的是口型的低级特征，而实际上即使相同发音每个人口型结构仍然会有差异，像素特征提取法缺乏形状特征提取法的高级特征描述能力；另一个原因是光线对像素法特征提取有较大影响，嘴唇是立体的，即使同一个人发相同的音在不同的光照下提取的特征矢量也会不相同，这点在实时唇读实验中对识别率的影响尤为明显。

此外，表 1 显示当把 ROI 从嘴唇扩大到下半个脸时，对特定人而言，唇读识别率有一定提高，但对非特定人的唇读识别率反而略有下降。据分析这是因为包含了下巴等更多的个性特征后，对不同人而言同一个词发音的特征差异会更大，导致非特定人的唇读识别率下降。表 1 还显示实际环境下的特定人实时唇读识别率明显低于数据库上的识别率，据分析，这主要是由于实时环境下训练和识别是分时进行的，而不同时刻日光灯下的 ROI 定位会出现偏差，导致同一个词会提取出不同的特征矢量，从而影响唇读识别率。

## 4 结论

针对唇读技术中视觉语言特征提取，提出一种新的基于 DCT 和 LDA 的二级特征提取方法。不同于传统的提取 DCT 系数主分量的特征提取思路，该文采用 LDA 算法在 DCT 系数中提取出具有最好分类性能的特征分量。在特定人唇读数据库上对人工选择 Zig-Zag 法、PCA 提取法以及 LDA 法三种 DCT 系数特征提取方法作了比较实验，基于特定人的 LDA 特征提取法取得了最好的唇读识别结果。此外，基于该特征提取法分别作了非特定人和特定人实时环境下的孤立词唇读识别实验，通过

(下转 155 页)