

# 一种快速嘴部检测方法在视听语音识别的应用

刘家涛, 陈一民

(上海大学 计算机科学与工程学院, 上海 200072)

**摘要:**在改进噪音环境下的语音识别率中,来自于说话人嘴部的可视化语音信息有着显著的作用。介绍了在视听语音识别(AVSR)中的重要组成部分之一:可视化信息的前端设计;描述了一种用于快速处理图像并能达到较高识别率的人脸嘴部检测的机器学习方法,此方法引入了旋转Harr-like特征在积分图像中的应用,在基于AdaBoost学习算法上通过使用单值分类作为基础特征分类器,以级联的方式合并强分类器,最后划分检测区域用于嘴部定位。将上述方法应用于AVSR系统中,基本上达到了对人脸嘴部实时准确的检测效果。

**关键词:**模态;视听语音识别;Harr-like特征;重要区域;积分图像;区域划分

**中图分类号:**TP391.41

**文献标识码:**A

**文章编号:**1673-629X(2008)10-0016-04

## Fast Mouth Detection Approach Applied in Audio - Visual Speech Recognition

LIU Jia-tao, CHEN Yi-min

(Sch. of Computer Sci. & Eng., Shanghai Univ. of Science and Tech., Shanghai 200072, China)

**Abstract:** The visual information comes from speaker's mouth had proved very useful in improving speech recognition, especially in noise environment. In this paper, first introduced one of the main components in audio - visual speech recognition system: visual front end design then proved a machine learning method for mouth region detection which could rapidly process image with high detection rates. This approach includes the introduction of rotated Harr-like feature in integral image, a learning algorithm based on Adaboost with sign value trees as base classifiers, combination of complex classifiers in cascade and regionalization of the face area. At the end, applied this scheme in AVSR system yield high detection rates which may reaches basically real time requirement.

**Key words:** modality; audio - visual speech recognition; Harr-like feature; region of interest; integral image; regionalization

## 0 引言

近年来,语音识别技术的迅速发展,使得自动语音识别系统有了广泛的应用。较好的有IBM开发的Vi-voice语音系统,微软的语音识别引擎。这些系统在相对安静的环境下能够对连续的单词及词组达到较高的识别率。然而,将其应用到有背景噪声或交互的说话者中的真实环境中,其较差的抗干扰能力,使其根本无法满足广泛应用的要求<sup>[1]</sup>。事实上,在真实的嘈杂环境中,人们对语言的感知是双模态的,它很自然地包含了语音及视觉信息。例如:人们在观看视频时,当听到说话人的声音为/ga/,而看到其嘴部的发声动作为/ba/时,大多数人会感觉对方的声音是/da/。对于一些

单音节,如浊辅音/b/, /d/在普通交流中正常人根本无法分辨,而一些聋哑人却可以通过眼部交流正确无误的理解说话人。这些行为都说明了人脸的视觉信息无论是在有听力障碍还是正常人之间的交流中都起着重要的感知作用。

## 1 视听语音识别系统介绍

通过挖掘说话人的嘴部视觉信息来改善自动语音识别系统的识别率,称之为视听语音识别系统(AVSR)。在嘈杂环境下,添加了视觉特征的识别系统在性能上比传统的单语音识别系统性能要出色得多。同时,视听语音识别技术带来了相对于传统单语音识别技术下更多的研究方向与挑战。在AVSR系统中,除了通常的语音特征提取阶段,来自说话人脸部带有语音信息的视觉特征同样要求在可视化前端设计中被获取。在此阶段中,不但要求准确的人脸检测,而且要求说话人嘴部或唇部的位置估算与跟踪,以此得

收稿日期:2008-01-16

基金项目:上海市科技基金资助项目(7A07094)

作者简介:刘家涛(1980-),男,山东烟台人,硕士研究生,研究方向为多媒体应用技术;陈一民,博士,教授,研究方向为多媒体应用技术、计算机增强现实。

到可适用的视觉特征。相对于单语音或单视觉识别,系统的后端则是对两者的融合,这能明显提升传统单模态识别率。传统语音特征提取及识别技术已相当成熟,新的 AVSR 系统所带来的两大问题是可视化前端设计及视-音融合。其中前者既是国内外各大院校研究的热门方向,也是文中所要详细论述及实现的技术<sup>[2,3]</sup>。

## 2 基于 Harr-like 特征分类器的检测方法

### 2.1 旋转 Harr-like 特征在积分图像中的应用

图像中任何矩形特征值都可以通过间接的表示方式得到快速的计算,称此图像为积分图像。如图 1 所示,积分图像在位置  $(X, Y)$  点的值为原始图像在此点矩形中所有像素的累加值,  $ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y')$ , 其中  $ii(x, y)$  是积分图像,  $i(x, y)$  是原始图像,通过下述两个方程:

$$s(x, y) = s(x, y-1) + i(x, y) \quad (1)$$

$$ii(x, y) = ii(x-1, y) + s(x, y) \quad (2)$$

积分图像完全可以忽略原始图像值,其中  $s(x, y)$  是累计的行值,  $s(x, -1) = 0$ ,  $ii(-1, y) = 0$ 。这样,在积分图像中任意矩形的值都可以参考 4 个数组计算得到,如图 2 所示。位置 1 的值为矩形 A 的累计值,位置 2 为  $A + B$ , 位置 3 为  $A + C$ , 位置 4 为  $A + B + C + D$ 。其中在矩形 D 中的值可通过  $4 + 1 - (2 + 3)$  得到。

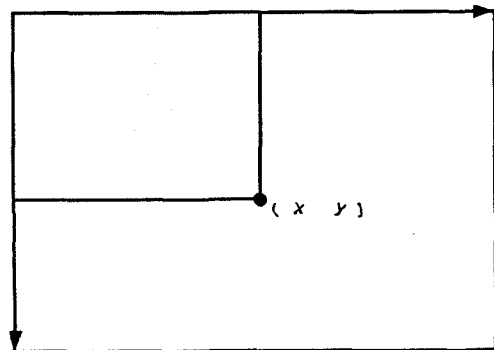


图 1 积分图像在位置  $(x, y)$  的值

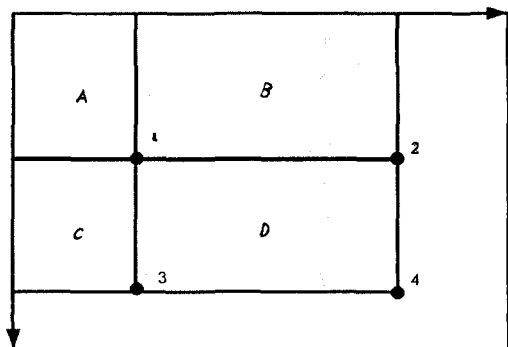


图 2 矩形特征 D 中值的计算

这里之所以使用 Harr-like 特征代替图像的像素

值作为机器学习算法的输入值最主要的目的是因为前者既可以减少类内的依赖性,同时又提高了类间的耦合性,这使得分类器的建立相对简单。并且特征值可以对某些领域上的知识进行编码,当大量简单的特征池通过特征选择时可以提高机器学习算法的效率。同时,特征状态估值的复杂性也很重要,所有的物体检测算法都是通过固定大小的窗口在输入图像上滑动进行的,而滑动窗口可以有不同的比例。

Viola<sup>[4]</sup>使用了 Papageorgiou 在文献[5]中的 Harr-like 特征作人脸检测,特征形状单一、结构简单,分类器通过它们很难达到很好的识别率。文中通过将上述三类特征旋转一定角度,扩展出新的特征形状。根据公式(3),旋转后的单一特征值计算时间在相同大小下与垂直角度的特征计算时间基本不变。只是在相同  $24 \times 24$  大小检测器下遍历所有的垂直及非垂直矩形特征数达 117941。设元组  $r = (x, y, w, h, a)$  表示图 3 各类 Harr-like 矩形原型,  $0 \leq x, x+w \leq W, 0 \leq y, y+h \leq H, x, y \geq 0, w, h \geq 0, a \in \{0^\circ, 45^\circ\}$ 。RecSum( $r$ ) 表示积分图像特征值。则任意特征  $I$  可以由以下公式表示:

$$\text{feature}_I = \sum_{i \in \{1, \dots, N\}} \omega_i \cdot \text{RecSum}(r_i) \quad (3)$$

其中权值  $\omega_i \in R, r_i$  和  $N$  为任意值。例如图 3 的线性特征 2(a) 高度为 2, 宽度为 6, 左上角坐标  $(5, 3)$ , 则其特征表示为:  $\text{feature}_I = -1 \cdot \text{RecSum}(5, 3, 6, 2, 0^\circ) + 3 \cdot \text{RecSum}(7, 3, 2, 2, 0^\circ)$ 。将先前提到的计算积分图像的垂直特征值  $ii(x, y)$  称为 SAT(Summed Area Table), 即图 3 的垂直矩形  $r = (x, y, w, h, 0)$  值可以通过 4 张 SAT 查找表快速计算得到。文献[4]给出的计算公式如下:

$$\text{RecSum}(r) = \text{SAT}(x-1, y-1) + \text{SAT}(x+w-1, y+h-1) - \text{SAT}(x-1, y+h-1) - \text{SAT}(x+w-1, y-1) \quad (4)$$

在此基础上可以推出如图 3(d) 旋转特征  $r$  值 RSAT(Rotated Summed Area Table) 的计算公式。

$$\text{RecSum}(r) = \text{RSAT}(x+w, y+w) + \text{RSAT}(x-h, y+h) - \text{RSAT}(x, y) - \text{RSAT}(x+w-h, y+w+h) \quad (5)$$

新扩展的旋转矩形特征值亦可通过 4 张 RSAT 表快速计算得到。至此,在遍历图 3 中所有 14 个 Harr-like 特征值时仅仅需要 8 张查找表即可获得快速的计算时间。

### 2.2 基于 AdaBoost 的分类器学习算法

在给出了 Harr-like 特征原型后,通过使用 AdaBoost 机器学习算法<sup>[6]</sup>,能够很好地推进简单特征分类

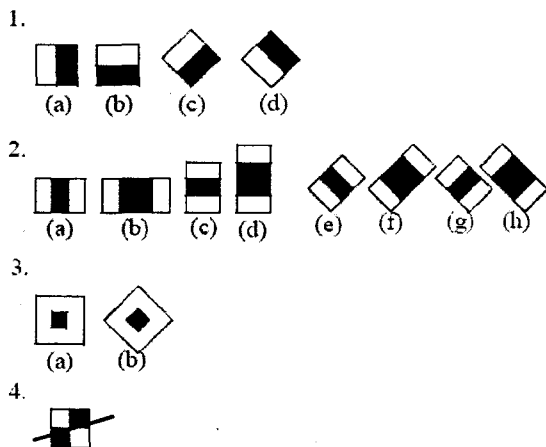


图 3 15 个 Harr-like 矩形特征原型

器的分类效果。将最初弱分类器设计为简单的区分单矩形特征正确与否的分类器,每一个 Harr-like 特征,由其弱学习机决定其最理想的阈值,以使其错误分类最小化。因此,一个弱分类器  $h_j(x)$  包含了一个特征  $f_j$ ,一个阈值  $\theta_j$  和一个同位偶值  $p_j$ 。实际应用中,没有一个单个特征能有很好的完成分类效果。早期推进选择的特征的错误率在 0.1 与 0.3 之间,后期则愈加困难,仅有 0.4, 0.5。

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < \theta_j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

此推进算法每一轮仅选择一个最有效的特征值。人脸检测中,初始的矩形特征选择基本上都是简单有效的。例如,由于人眼深于脸颊,最先选择的特征区域集中在人眼与脸颊、鼻子之间,此块区域相对子窗口的检测比较大,且对人脸的位置、尺寸并不敏感。当人脸正面部位的分类器构成的特征数为 200 个时,其检测率为 95%。随着特征数量的增加,检测效率与计算量也同时递增。

### 2.3 强分类器的级联化

为了在提高检测性能的同时能降低计算时间,将上节中推进选择的分类器级联化,以便于更有效地区分正、负图像。即在更加复杂的分类器使用前,先通过简单的分类器丢弃了无用的子窗口,称此过程为分类器的级联化。整个检测分类的过程类似于一个退化的判定树<sup>[7]</sup>。判别树起始的分类器分类产生的正确结果触发后续的分类器的判断,依此类推,任一级分类器的判别失败将直接丢弃此子窗口。假如每一级能消除 50% 的错误子窗口,则一个 20 级的级联分类器检测率将达到  $0.5^{20} \approx 9.6 \times 10^{-7}$ 。基本上能达到传统的单一多特征分类器效果。由于通过早期的分类判别使其在分类速度上有着显著提高。

其中,每一级级联的分类器训练是通过使用上一节所提到的 AdaBoost 算法得到的,这里首先使用的是

一个仅含有二个特征的强分类器作为判定树的起始点。有效的人脸过滤亦可以通过调节强分类器的阈值来达到最小的错误率。

如将初始的 AdaBoost 阈值设定为  $\frac{1}{2} \sum_{i=1}^T \alpha_i$ 。由两个人脸最基本特征组成的强分类器可以 100% 地检测到含有人脸的子窗口,其在负片图像上检测率为 40%。虽然其检测效果仍未达到系统的要求,但这一过程仅仅需要极少的计算量。上述含有两个特征的强分类器的计算量仅需大约 60 个微处理指令即可完成,效果远远高于一些简单的图像模板检测方法。

### 2.4 划分检测区域

由于在降低分类器错误率的同时会降低正片图像的识别率,根据文献[8],通过划分含有所要分析特征的图像区域来提高检测的准确性。当要检测的图像区域减少时,无论是积分图像的面积计算量,特征数都会大幅度减少,这些计算量的降低都可以提高目标检测的有效性。例如在人脸嘴部检测时,并不需要直接从整张图片中开始目标检测。为了划分有效的检测区域,可以首先对图像进行人脸检测,在所检测到的范围中再进一步地作嘴部检测,这样可以明显地提高嘴部检测的准确性。根据正常人脸的结构,嘴部仅位于人脸的下半部分,通过所检测到的人脸,仅选择其下部五分之二范围作嘴部检测。

## 3 人脸嘴部检测在 AVSR 系统中的应用

通过上节所述方法,分别建立了一个 21 层人脸特征和一个 18 层嘴部特征级联分类器,每一层强分类器阈值存储在 xml 文件中。在 Xeon 3.00GHz 处理器上测试分辨率为 352x288 的视频流时,图 4 给出了各类目标检测下达到的正确率。可以看出 2.4 小节提出的划分区域下的方法完全能检测到人脸嘴部区域。如图 5 所示,前一张为普通函数库下的嘴部检测,后一张为采用文中方法下相同目标检测的效果图。由于视频片段为 25f/s 的 AVI 格式,要达到实时目标的检测效果

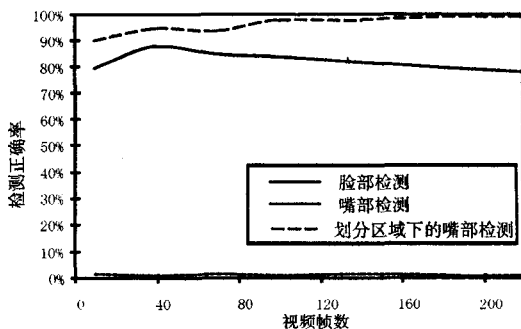


图 4 不同区域下的检测正确率

至少要 40ms,理想下每帧检测时间不得高于 10ms。普通 OpenCV 函数库下  $scale = 1.0$  的检测时间基本都在 10ms 以上,无法满足系统要求。采用文中所述方法,当检测窗口  $scale = 1.3$  时基本满足系统实时检测要求,在检测窗口  $scale > 2.0$  时其检测时间基本保持在 0.06(ms)至 1.2(ms)之间。



图 5 使用文中方法前后的嘴部检测效果图

#### 4 结束语

文中针对 AVSR 系统下可视化前端设计中人脸嘴部快速检测的要求,采用了基于 Harr-like 特征推进级联分类的方法,有效降低了其目标检测的计算时间并提高了准确性。此方法首先引入了积分图像的概念,通过查找表的方法对大量垂直、旋转 Harr-like 特征值进行了快速的计算。在基于 AdaBoost 学习算法上通过推进选择单值分类的基础分类器,以级联的方式合并强分类器从而丢弃了大量检测中无用的子窗口,最后通过人脸嘴部的定位进一步降低了其检测的目标区域。实验表明,此方法在 AVSR 系统上的检测

效果基本上可以达到实时应用的目地,为进一步的视-音融合及识别提供了有效的视觉特征。

#### 参考文献:

- [1] Gong. Speech recognition in noisy environments: a survey[J]. Speech Communication, 1995, 16: 261 - 291.
- [2] Potamianos G, Luetttin J. Audio - visual speech recognition [R]. Final Workshop 2000 Report, Center for Language and Speech Processing. Baltimore, MD: The Johns Hopkins University, 2000.
- [3] Liang H, Liu X X, Zhao Y B, et al. Speaker independent audio - visual continuous speech recognition[C]//In Proc. of IEEE ICME. Lausanne, Switzerland: [s. n.], 2002.
- [4] Viola P, Jones M J. Rapid Object Detection using a Boosted Cascade of Simple Features[J]. IEEE CVPR, 2001(1): 511 - 518.
- [5] Papageorgiou C, Oren M, Poggio T. A general framework for Object Detection[C]//In International Conference on Computer Vision. [s. l.]: [s. n.], 1998.
- [6] Freund Y, Schapire R E. A decision - theoretic generalization of on - line learning and an application to boosting[C]//In Computational Learning Theory: Eurocolt ' 95. [s. l.]: Springer - Verlag, 1995: 23 - 37.
- [7] Amit Y, Geman D, Wilder K. Joint induction of shape features and tree classifiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(11): 1300 - 1305.
- [8] Cristinacce D, Cootes T. Facial feature detection using AdaBoost with shape constraints[C]//British Machine Vision Conference. [s. l.]: [s. n.], 2003.

(上接第 15 页)

表 2 支持度、置信度与规则数对应表

minSupp	minConf	Rules
10%	15%	169
10%	20%	126
10%	25%	89
10%	30%	37
10%	35%	16

包含项 wine 的事务有 287 条,若隐藏项 wine,由于最小置信度和最小支持度的不同组合,更改事务的数量占总事务数的 20% ~ 45%,当修改更多的事务时,相应的时间开销也将增加。文中,研究了由数据挖掘技术引起的事务数据库隐私保护问题,提出了隐藏那些包含敏感项目的关联规则的算法。

#### 6 结束语

目前处理含有隐私项目的关联规则的方法大都是基于支持度或置信度的减少的。文中提出的算法能够

隐藏包含敏感项目的关联规则,而且不需要预知关联规则的种类,并在随后给予了证明。对于给定的事务数据库、敏感项目集,在将来,还需要改善算法的效率,如:减少扫描事务数据库的次数。此外,如果数据库频繁更新,如何在这种情况下,充分保持已作隐藏处理的那些关联规则的不可见性,是有待研究的问题。

#### 参考文献:

- [1] 陈子阳,马朝虹,李宇佳,等. 量化关联规则的隐私保持挖掘方法[J]. 计算机工程, 2005, 31(11): 74 - 76.
- [2] 罗永龙,黄刘生. 一个保护私有信息的布尔关联规则挖掘算法[J]. 电子学报, 2005, 33(5): 900 - 903.
- [3] Evfimievski A, Srikant R, Agrawal R. Privacy preserving mining of association rules[J]. Information Systems, 2004, 29: 343 - 364.
- [4] Seifert J W. Data mining and the search for security[J]. Government Information Quarterly, 2004, 21: 461 - 480.