文章编号:1001 - 9081(2006)09 - 2099 - 02

AdaBoost 算法在喷码图像识别的应用

王 倩,陈 斌,黄文杰
(中国科学院 成都计算机应用研究所,四川 成都 610041)
(wangqian1016@126.com)

摘 要:喷码在印刷中使用比较普遍,其识别通常采用模板匹配的方法,但是由于喷码常出现误差,为模板匹配方法识别带来难度。AdaBoost 是一个建构准确分类器的学习算法,文中将此算法应用于喷码图像的识别,不仅提高了识别的准确率,速度也更为理想。

关键词:喷码; AdaBoost 算法; 号码识别
中图分类号: TP391.41 **文献标识码**:A

Application of AdaBoost algorithm in spray code recognition

WANG Qian, CHEN Bin, HUANG Wen-jie
(*Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan* 610041, *China*)

Abstract: Spray codes are widely used in printing. Template matching is usually adopted to recognize them. But because the spray codes often have errors, it is not easy to recognise the codes exactly with template matching. AdaBoost is a learning algorithm to construct classers. In this paper, the algorithm was used in the recognition of spray codes. It can not only improve the recognition accuracy, but also get more satisfactory speed.

Key words: spray code; AdaBoost algonithm; recognition of the code

0 引言

喷码是一种应用广泛的印刷方式,号码一般由计算机控制喷码机点状喷射出。喷射印刷后需要进行印刷质量的检测,利用图像处理,计算机视觉和模式识别等技术计算机能自动识别喷码图像的号码,判断其印刷是否正确。一般的识别方法是利用模板匹配,其方法是提取待识别喷码图像的特征,与每类标准模板图像的特征进行匹配,根据匹配度大小进行判断,归入匹配度最大的那一类。然而由于种种原因,喷码时常常出现错点,断点,重点和漏点,导致提取的图像特征出现误差,为模板的精确匹配带来难度。AdaBoost 算法是一个建构准确分类器的学习算法,预先把出错的样本加入待学习样本,通过机器学习,能构造出识别能力更强,识别范围更广的分类器。基于算法的上述优势,把它应用到喷码图像的识别,能有效提高计算机号码自动识别能力。

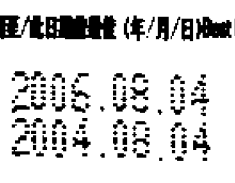

图 1 喷码图像

1 AdaBoost 算法

在机器学习领域,1990 年,Schapire 提出了 boosting 算法,通过构造此方法证明可以将一组弱学习算法提升为一个强学习方法。强学习算法和弱学习算法的概念是:如果一个学习算法通过学习一组样本,识别率很高,则称其为强学习算法;如果识别率仅比随机猜想略高,则称其为弱学习算法。1997 年,Freund 和 Schapire 提出了 AdaBoost 算法。算法基本过程是:

首先给出样本集合,然后对该样本集合进行循环操作,每次循环首先得到一个假设(即 weak learning),然后计算该假设的错误率,根据该错误率改变权重进入下一个循环。具体算法如下:

已知训练样本集:$(x_1,y_1),\cdots,(x_n,y_n)$,$x_i$ 为样本图像,y_i 为分类结果,$y_i \varepsilon \{0,1\}$; 取 m 个特征作为弱分类器:$h_j(x_I)$ $(j=1,\cdots,m)$; 设正例和反例样本数分别为 n,l; 初始化权重分别为:$w_{1,1}=\frac{1}{2n},\frac{1}{2l}$ 循环 for $t=1,2,\cdots,T$:

(1) 规格化权重:$w_{t,j}=\frac{w_{t,j}}{\sum_{j=1}^{n} w_{t,j}}$;

(2) 对于每一个特征 j,训练弱分类器 h_j,只使用这一个特征,计算加了权值误差,

$s_j=\sum w_{t,j}\left|h_j(x_i)-y_i\right|$;

(3) 选择误差 s_t 最小的弱分类器 $h_{t,j}$;

(4) 更新权重 $w_{t+1,j}=w_{t,j}\beta_t^{1-e_i}$,如果 x_i 被分类正确 $e_i=0$,否则 $e_i=1$,其中 $\beta_t=\frac{s_t}{1-s_t}$;

最后形成强分类器为($a_t=\log(1/\beta_t)$):

收稿日期:2006 - 03 - 21; **修订日期**:2006 - 06 - 21

作者简介:王倩(1981 -),女,四川自贡人,硕士研究生, 主要研究方向:图像处理、模式识别; 陈斌(1970 -),男,四川广汉人,研究员,主要研究方向:图像处理,模式识别,工业视觉; 黄文杰(1981 -),男,四川西昌人,硕士研究生,主要研究方向:图像处理、模式识别.

$$H(x)=\begin{cases}1, & \sum_{t=1}^{T}a_t h_t(x)\geqslant\frac{1}{2}\sum_{t=1}^{T}a_t\\ 0, & \text{其他}\end{cases}$$

算法从 m 个弱特征中选取 T 个强特征,组合成一个强分类器。AdaBoost 算法针对二分类问题,可以把它推广到多分类问题,加之本身学习性的特点,考虑把它应用到喷码图像识别。

2 AdaBoost 算法的应用

2.1 弱分类特征的选取

观察喷码图像,单号码图像之间其实有一些典型的区分区域(见图 2)例图中标出 6 和 8 的一个典型区分区域,由此考虑通过一组典型区分区域识别喷码图像。2001 年,Violat 提出基于 Harr 型的 AdaBoost 算法[3],并应用于人脸检测领域取得了成功。借鉴其成功经验,我们选用简单矩形特征,即 Harr-like 特征(见图 3)作为待识别喷码图像特征。Harr-like 特征用五元组表示 $(x,y,w,h,style)$,记录矩形位置,长宽和类型。特征值为矩形中白色区域灰度值和与黑色区域灰度值灰度值和的差值,反映图像局部灰度变化。用快速积分图方法[3]计算特征值。预选取 Harr-like 特征时对矩形面积进行限制:A,B,D 型矩形特征要求白色区域面积和黑色区域面积相等;C 型矩形特征要求白色区域面积是黑色区域面积的两倍。把选取的 Harr-like 特征作为弱分类器(特征)。

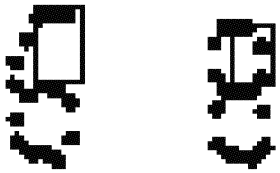

图 2 典型区分区域

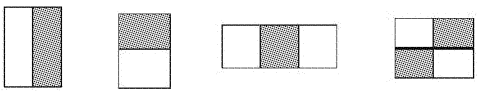

图 3 Harr-like 特征

2.2 分类器的构造

AdaBoost 算法针对的是二分类问题,可以把多个二分类问题组合和一个多分类问题。数字 0 至 9 是一个多分类问题,通过决策树可以把这个多分类问题细化为 9 个二分类问题,这样 AdaBoost 算法就可应用于此。对已拍摄到的喷码图像进行预处理和号码分割,得到大小相等的单号码图像,作为训练样本。利用预取的弱分类器(特征)对每个二分类问题中的正反样本进行 AdaBoost 学习,其中训练轮数 T 根据实际需要选择。在此 AdaBoost 算法应用中,弱分类器

$$h_j(x_i)=\begin{cases}1, & p_j f_j(x)<p_j\theta_j\\ 0, & \text{其他}\end{cases}$$

θ 为弱学习寻找出的域值为用此弱特征计算正例和反例样本的特征值聚类中心之均值;$f_j(x)$ 表示特征值,即 Harr-like 特征值;p_j 为不等号方向,根据实际需要设定;x 为 Harr-like 特征。通过 AdaBoost 算法学习,得到 9 个强的二分类器,其中每个强二分类器由训练选择出的 T 个强特征组合而成。

喷码图像通过这 9 个二分类器的分级分类判定,快速准确地识别出号码。由于喷码常出现错点,断点,重点和漏点,我们故意在训练样本中加入错码样本,以提高分类器对错喷码情况的识别能力。

3 实验结果

选取分辨率为 78×40 的 4 号码喷码图像为实验素材(见图 4)。对现有喷码图像进行预处理和图像分割,得到 10 类各约 100 张分辨率为 19×40 的单号码喷码图像(见图 5)。选取约 6000 个 Harr-like 特征作为弱分类器(特征),AdaBoost 学习轮数选择为 30,利用上述应用方法构造分类器。与模板匹配方法作比较,将单号码喷码图像的等分为 5×5 份,计算每个小区域黑色像素点所占比例,结果作为图像特征值,采用最小距离法进行模板匹配。用两种方法对 30 幅分辨率为 78×40的 4 号码喷码图像进行识别测试,测试环境为:PIII 667MHz CPU, 256MB 内存,VC++6.0 编程,测试结果如表 1。实验结果表明,应用 AdaBoost 算法在喷码图像识别较之传统模板匹配方法识别率更高,识别速度更快。

表 1 测试结果

识别方法	平均识别时间/ms	识别率
AdaBoost	3.8	0.97
模板匹配	5.1	0.93

图 4 4 号码喷码图像

图 5 单号码喷码图像

4 结语

本文介绍了 AdaBoost 算法及其在喷码图像识别的应用,通过实验表明此应用较之传统识别方法的优势。在实际在线识别中,只需根据不同待识别喷码图像获取不同的学习样本构造分类器。沿着此应用思路,还可以把 AdaBoost 学习算法应用到更多的识别、检测等领域。

参考文献:

[1] 何斌, 马天予. Visual C++数字图像处理[M]. 第 2 版. 北京: 人民邮电出版社, 2002.

[2] 杨淑莹. 图像模式识别[M]. 北京: 清华大学出版社; 北京交通大学出版社, 2005.

[3] VIOLA P. Rapid object detection using a Boosted cascade of simple feature[J]. Proc IEEE Conference on Computer Vision and Pattern Recognition, 2001.511 - 518.

[4] YANG MH, KRIEGMAN DJ, AHUJA N. Detecting Faces in Image A Survey[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002, 24(1): 34 - 58.

[5] 黄文杰, 陈斌. 一种快速图像处理的积分图方法[J]. 计算机应用, 2005, 25(Z1), 266 - 268.