DOI:10.11721/cqnuj20150521

基于随机森林分类的快速标签检测*

章　沛，陈小瑜
(广州医科大学 信息技术中心，广州 510182)

摘要:主要解决了传统的人工标签检测中存在的成本高、速度慢以及检测率低的问题。与传统算法相比,提出了一种基于随机森林分类的快速标签检测算法。文章首先设计了自适应特征选择方法,能够在大规模的特征中自动选择鉴别力最高的特征,较好地区分前景和背景。在此引入了基于金字塔随机撒块的随机森林分类器,能够快速有效地完成多种标签的检测。实验结果验证了本算法的高速、高精度特性,适用于高速自动化生产线上的标签检测。

关键词:随机森林;自适应特征选择;标签检测;金字塔加速
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1672-6693(2015)05-0131-06

标签作为产品的主要标志之一,不仅代表了产品自身和生产厂家的形象,还承担着吸引消费者的作用,因此,如何保证标签的完整性与正确性至关重要。在传统的检测算法中,大多通过人工检测来完成,但人工检测无法满足日益增长的生产速度的要求,人工检测对应的低检测率、高成本的问题也限制了生产行业的进一步发展[1]。计算机视觉技术利用图像处理和机器学习[2]的成熟技术,能够显著提高标签检测的速度和精度[3]。

邢堃等人[4]从图像处理的角度出发,利用 Canny 算法作为掩码图像进行缺陷检测,能够克服图像抖动和非同步性带来的检测问题。在此基础上,陈府庭等人[5]将色彩信息拓展到 LAB 空间,从而获得了精度更高的检测算法。张树君等人[6]主要设计了基于边缘的图像处理算法,通过手动指定矩形框并统计其中像素数目的变化来识别有缺陷的标签。针对手动指定困难的问题,覃希等人[7]利用 SVM 分类器采集大量的矩形框特征,利用机器学习技术来完成合格和不合格标签的分类。但传统算法中仍然存在一些问题,比如特征选择困难、特征数目过多以及二值分类器无法较好地满足多类标签的分类要求等[8]。

针对上述问题,本文提出了基于随机森林分类的快速标签检测算法。主要包括两个创新点。第一,为了更好地区分前景和背景,获得更加具有鉴别意义的特征,本文设计了自适应特征选择方法,在大规模的像素点特征中进行自动选择。该方法在保证计算效率的前提下,大大提高了分类器的检测能力。第二,针对大数据量、多类别的标签图像,本算法引入了随机森林的分类技术,能够有效地完成标签的检测。在此基础上,提出了金字塔随机撒块算法,大大提高了标签检测的速度。实验部分证明了本算法精度高、速度快,能够满足高速自动化生产线的需求。

本文的安排如下。第二章给出了本算法的主要流程和框架图。在此基础上,第三章和第四章分别对自适应特征选择模块和快速随机森林分类模块进行了讨论。在第四章和第五章中,本文完成了实验论证和全文总结。

1 算法框架

标签图像中存在的缺陷包括标签缺失、标签偏移以及标签破损等,图 1 给出了本算法的主要框架示意图,主要包括两个部分:训练和检测。在训练部分,首先构建了自适应的色彩空间池,用于最大化前景和背景的差异;其次,创新地以随机撒矩形框的方式作为分类器种子,并用上述多色彩空间值作为代表特征;最后,利用随机森林分类器对待训练图像集进行训练,得到合适的分类器。在检测部分,首先对待检测图像进行处理,构建多色彩空间下的复合图像;再次,利用训练部分得到的分类器完成该图像的分类。

* **收稿日期**:2014-11-26　**修回日期**:2015-05-26　**网络出版时间**:2015-06-08 12:29
作者简介:章沛,男,高级工程师,研究方向为数据挖掘,计算机网络等,E-mail:qq17568145@126.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.n.20150608.1229.012.html

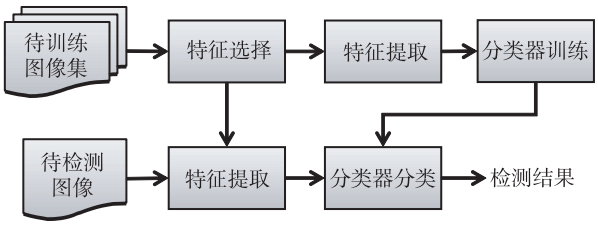


图1 算法框架

2 自适应特征选择

传统的特征选择主要利用了RGB、HSV等色彩信息[9],但存在以下两个问题。第一,不同的色彩信息对于不同瓶体的区分能力有差异,如果使用相同的色彩信息无法获得较好的性能。第二,色彩信息的人工选择难以最大化前景和背景的差异,且存在的冗余色彩信息会提高算法的计算复杂度。因此,本文提出了自适应特征选择算法,能够获得显著性最高的特征序列用于检测。

2.1 特征构建

原则上,色彩、纹理、运动以及形状信息等都可以作为特征供检测使用,且每种特征都存在很多参数可供调整,因此存在的特征数目巨大。本文主要利用了色彩信息,通过设置局部空间的色彩滤波器来完成特征的构建。构建的色彩滤波器需要满足3个条件:首先,滤波器系数简单,保证组合特征可以有效计算得到;其次,需要涵盖现有的常用色彩信息,包括RGB、YCrCb等;最后,由于现有相机采集的都是RGB色彩信息,因此以该空间的一维映射组合为宜。

本文利用的特征池主要通过R、G、B色彩值得线性组合得到的,如下式所示:

$$f_1=\{w_1\cdot r+w_2\cdot g+w_3\cdot b|w_*\in\{-2,-1,0,1,2\}\}。\tag{1}$$

也就是说组合后的特征由$-2\sim2$的加权系数组成,即组合特征为125维。但除去冗余的特征(即$(w_1',w_2',w_3')=k(w_1,w_2,w_3)$)和全0特征,还剩余49维特征。当计算特征时,需要将特征值域归一化到$0\sim255$之间。不难看出,本算法还可以很容易地和其他特征相结合,从而获得更加优异的检测性能。

2.2 特征度量

在实际应用中,前景标签部分和背景的液体部分可能存在多种颜色(由于光照、角度以及材质的影响),因此无法用单一高斯模型进行描述。本文利用直方图统计信息,计算特征对于前景和背景的分离度完成度量过程。主要分成3个步骤:1) 根据特征分别计算前景和背景的分布直方图;2) 计算这些分布对应的最大似然估计;3) 根据最大似然估计的统计值来决定特征的选择。

设H_f和H_b分别表示前景和背景的直方图。将直方图分成n个块并利用直方图归一化技术,那么可以得到前景和背景分布直方图中第i个块的概率分别为$p(i)$和$q(i)$。因此,可以得到最大似然估计的结果为:

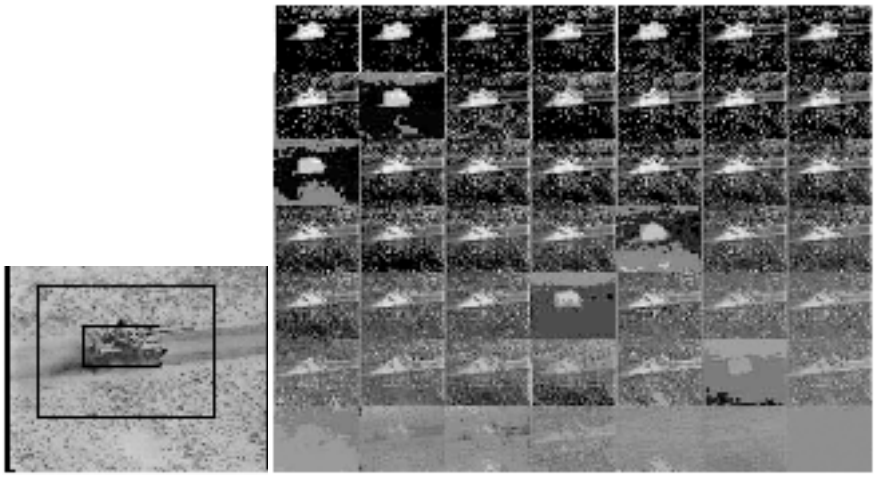
$$L(i)=\log\frac{\max\{p(i),\delta\}}{\max\{q(i),\delta\}},\tag{2}$$

其中δ为无穷小量,能够避免最大似然估计产生奇异值。不难看出,最大似然估计也可以认为是一种直方图,反映的是前景和背景的差异性。最终,为了得到每个特征的鉴别能力,本文利用方差来决定特征度量:

$$\mathrm{var}(L)=E(L^2(i))-E^2(L(i))。\tag{3}$$

2.3 特征筛选

本文从49维特征中进行筛选,图2给出了一个例子。其中,图2b给出了经过公式3的特征度量后得到的排序结果,最有鉴别力的特征位于左上方,而最不具有鉴别力的特征位于右下方。图中反映的是公式2计算出的最大似然估计图,不难看出,鉴别力的排名符合人们的视觉效果,能够很好地帮助后续的标签分类。



a　原始图　　b　特征鉴别力结果

图 2　特征筛选的结果

在本算法中，综合考虑精度和执行效率的影响，筛选出排名前 10 的特征用于标签分类。如果使用全部的特征，不仅使得计算复杂度大大提高，而且冗余特征会造成分类器产生过拟合现象，降低识别率。

3 基于随机森林的标签检测

3.1 随机森林检测算法

决策树算法[10]是一种传统的数据挖掘分类器，通常构成的是一个二叉树结构。其中每一个非叶子节点表示一个分类器，而每一个叶子节点则表示样本的所属类别。不难看出，决策树的构建在于一个个阈值判断问题，包括分类器之间特征值不应当重复、阈值如何选择能够较好地分裂样本集以及如何分配叶子节点的所属类别。从 boosting 得到借鉴，可以通过构建多个树分类器，再根据分类器的联合综合确定森林分类器的结果，从而得到更稳定、更鲁棒的分类性能。在此基础上，Breiman[10]提出了随机森林理论，主要在两处有随机的意义：一个是每棵树用的点分类器是随机的，点分类器之间没有联系；另一个是森林用的树分类器也是随机的，树分类器之间也没有联系。图 3 给出了随机森林运行的一个示意图。

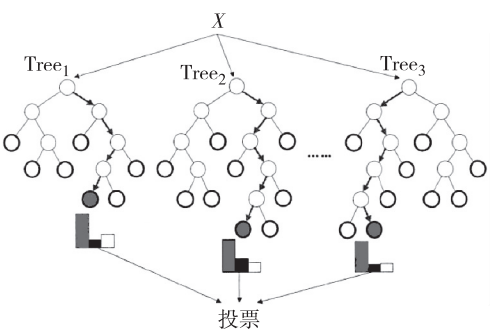


图 3　随机森林运行示意图

综上所述，随机森林分类器的训练过程如算法 1 所述：

算法 1　随机森林分类器训练算法

输入：训练样本集 $\Omega=\{X,Y\}$，其中 $X=\{x_1,x_2,\cdots,x_N\}$，$Y=\{y_1,y_2,\cdots,y_N\}$

输出：随机森林分类器

For $i=1$ to T do

对训练样本集进行随机采样 $\Omega_i=BootstrapSampling(\Omega)$

训练随机树 $TreeRoot_i=GrowRandomizedTree(\Omega_i)$，见算法 2

End for

$Random\ Forest=\{TreeRoot_1, TreeRoot_2, \cdots, TreeRoot_T\}$

其中随机树的训练过程如算法 2 所述：

算法 2　随机树分类器训练算法

输入：训练样本集 $\Omega=\{X,Y\}$，其中 $X=\{x_1,x_2,\cdots,x_N\}$，$Y=\{y_1,y_2,\cdots,y_N\}$

输出：随机树分类器

If Ω_i 中所有训练样本的类别相同 or $\Omega_i=N\leqslant N_0$

Return $LeafNode(p(c|\Omega_i))$

End if

For$t = 1$ to $|\Omega_i|$

If $h(x_t)=1$

将(x_t, y_t)添加到左子节点的数据集中,即 $\Omega_l \leftarrow \Omega_i \cup \{(x_t, y_t)\}$

Else

将(x_t, y_t)添加到右子节点的数据集中,即 $\Omega_r \leftarrow \Omega_i \cup \{(x_t, y_t)\}$

End if

End for

新建左子节点 $LeftNode = GrowRandomizedTree(\Omega_l)$

新建右子节点 $RightNode = GrowRandomizedTree(\Omega_r)$

Return $ParentCode(LeftNode, RightNode)$

本算法选择随机森林分类器的原因主要有3点。首先,待训练的样本集通常很大,随机森林分类器的训练速度、计算效率远远高于传统的SVM、Adaboost等分类器,且不易产生过拟合现象;其次,随机森林能够直接产生多类别的分类结果,而SVM和Adaboost只能通过若干个二类分类器的组合才能完成,因此效率高、结果更加直观;最后,由于森林中的每棵树都是独立生长,因此可以进行并行化加速,完成更加高效的应用。

3.3 检测算法的加速

现有的高速生产线对于分类算法的效率的要求越来越高,为了尽可能地提高算法的拓展性和实用性,本文设计了金字塔随机撒块算法用于分类算法的加速。利用筛选出的5维特征,可以对任意大小的矩形框的像素点进行统计,利用均值作为每个矩形框的特征。在现有算法中,都是利用经验值选择若干指定大小、位置的矩形框用于分类。该类方法主要存在3个问题:首先,该方法对于工程人员的经验和素质要求较高,很难适用于新手使用;其次,人工选择难以达到全局最优;最后,对于矩形框的数目没有一个显性的指导,难以便于后续的维护和拓展。针对上述问题,本文在归一化后的图像上随机撒一些起始点,将其作为矩形框的中心;在此基础上,指定若干大小的尺度,通过构建金字塔确定矩形框用于分类。图4给出了金字塔随机撒块算法的示意。其中,圆点表示了选择了随机撒点的结果;而矩形框簇表示了经过金字塔拓展后的矩形框的位置

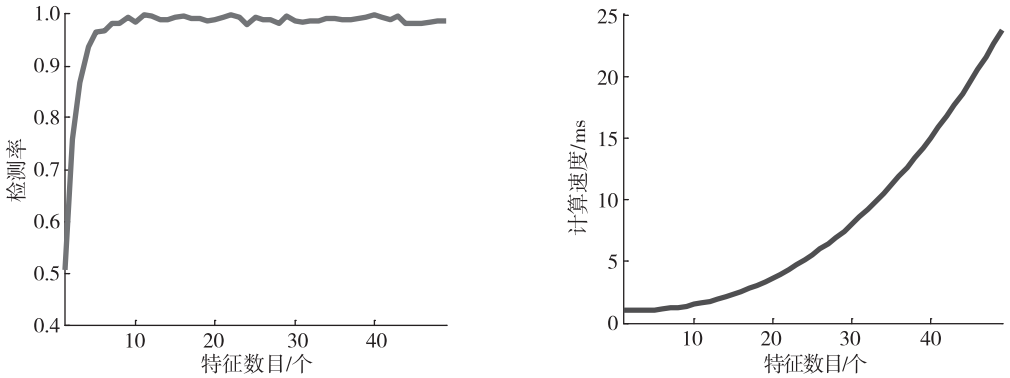
图4 金字塔随机撒块算法的示意

4 实验与结果

下面对本算法的性能进行验证,实验环境为Windows 7主机,Intel Core i3四核处理器,主频2.6 GHz。实验部分主要包括两节:第一节对特征筛选的效率进行验证;第二节对算法的整体性能予以总结与讨论。实验中准备了100张待检测的标签图像,其中包含80张正常图像、5张标签缺失、5张标签偏移、5张标签破损以及5张标签错误。每组实验重复50次,用于验证本算法的鲁棒性。

4.1 特征筛选

图5给出了特征数目和检测率、计算速度的变化关系。不难看出,首先,在本算法中,当特征数目大于7时,已经可以达到较高的检测率;其次,如果继续增加特征数目,将不会有更好的效果,反而会大大提高计算速度,显著增大算法的开销;最后,本算法使用了10个特征,在计算速度上仅为2 ms不到,能够很好地满足高速自动化生产线的要求。



a　检测率随着特征数目的变化曲线　　b　计算速度随着特征数目的变化曲线

图 5　特征筛选的实验结果

4.2 检测结果

性能比较中主要采用了漏检率和误检率两个指标，设正样本被判为正和负的数目为 TP 和 TN，而负样本被判为正和负的数目为 FP 和 FN，那么对应的漏检率和误检率为：

$$p_{漏}=\frac{TN}{TP+TN},\ p_{误}=\frac{FP}{FP+FN}。\tag{4}$$

表 1 中给出了本算法和覃希等人[7]在精度和速度上的比较结果。第一，由于设计的自适应特征选择机制能够选择出适合的特征最大程度上区分前景和背景，因此本算法能够获得更低的错误率。第二，引入的随机森林分类更加适合于标签这种多类别的分类，本算法不仅精度上优于 SVM，还能显著提高计算速度。第三，本文还设计了金字塔随机撒块方法，能够进一步提高随机森林的训练效率、获得更低的计算复杂度。总体来说，本算法能够比覃希等人[7]获得更高的精度和更快的计算速度。

表 1　本算法和覃[7]的结果比较

	$p_{漏}$/%	$p_{误}$/%	时间/ms
覃[7]	3.2	4.7	47
本算法	0.4	0.3	6

5 结论

传统的人工标签检测主要存在成本高、速度慢以及检测率低等问题。针对上述问题，本文提出了一种基于随机森林分类的快速标签算法。一方面，设计了自适应特征选择方法，能够在大规模的特征中自动选择鉴别力最高的特征，较好地区分前景和背景。另一方面，引入了基于金字塔随机撒块的随机森林分类器，能够快速、有效地完成多种标签的检测。实验结果验证了本算法的高速、高精度特性，适用于高速自动化生产线上的标签检测。

参考文献：

[1] 席斌，王振雷，钱锋. 机器视觉工业检测系统的应用与发展[J]. 控制工程，2008 (S1)：220-222.
Xi B, Wang Z L, Qian F. Development and application of machine vision in industrial detection system[J]. Control Engineering of China, 2008 (S1): 220-222.

[2] 李玲俐. 数据挖掘中分类算法综述[J]. 重庆师范大学学报：自然科学版，2011，28(4)：44-47.
Li L L. Review of classification methods in data mining[J]. Journal of Chongqing Normal University: Natural Science, 2011, 28(4): 44-47.

[3] 谭智仁，卢军. 基于图像传感器的标签缺陷检测方法[J]. 组合机床与自动化加工技术，2014 (3)：127-130.
Tan Z R, Lu J. Detectionmethod on flaw of label based on image sensor[J]. Modular Machine Tool & Automatic Mnufacturing Technique, 2014 (3): 127-130.

[4] 邢堃，韩汉光，吴怡之. 基于机器视觉的印刷标签检测系统的改进[J]. 计算机工程与应用，2014(11)：197-201.
Xing F, Han H G, Wu Y Z. Improvement of machine vision based defect detection system for printed labels[J]. Computer Engineering and Application, 2014(11): 197-201.

[5] 陈府庭，汪仁煌. 基于 CIE LAB 的标签检测技术[J]. 计算机与现代化，2011 (10)：74-75.

Chen F T, Wang R H. Label detection technology based on CIR LAB [J]. Computer and Modernization, 2011(10): 74-75.

[6] 张树君,辛莹莹,陈大千. 基于机器视觉的饮料瓶标签检测设备[J]. 食品研究与开发,2014 (3):134-136.
Zhang S J, Xin Y Y, Chen D Q. The bottle of beveage label detection device based on machine vision[J]. Food Research and Development, 2014 (3): 134-136.

[7] 覃希,夏宁霞,苏一丹. 基于支持向量机的垃圾标签检测模型[J]. 计算机应用研究,2010 (10):3893-3895.
Qin X, Xia N X, Su Y D. SVM-based social spam detection model [J]. Application Research of Computers, 2010(10): 3893-3895.

[8] Lin J, Liao Q, He B, et al. Label inspection of approximate cylinder based on adverse cylinder panorama[C]//International conference on optical instruments and technology (OIT2013). International Society for Optics and Photonics, USA: SPIE, 2013: 90451Y-90451Y-7.

[9] 叶利华. 视频标签检测与识别[J]. 制造业自动化,2011,33(6):95-98.
Ye L H. Video label detection and recognition[J]. Manufacturing Automation, 2011, 33(6): 95-98.

[10] 任秀春,贺亚吉. 基于决策树的网络客户分类方法研究[J]. 电子设计工程,2014(5):20-22.
Ren X C, He Y J. Method of network customer classification based on decision tree[J]. Electronic Design Engineering, 2014(5): 20-22.

[11] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

Fast Label Inspection Based on Random Forest Classifer

ZHANG Pei, CHEN Xiaoyu

(Information Technology Center, Guangzhou Medical University, Guangzhou 510182, China)

Abstract: Traditional manual inspection for label confronts problems such as high economic cost, low speed and low accuracy. Aiming at those, this paper proposes a quick inspection methods based on random forest. First, we design the adaptive feature selection strategy, which can select the most distinctive ones from large number of features to separate the foreground and background. Then, we introduce the random forest classifier added with hierarchy random blocks and inspect different type of labels quickly and accurately. Experimental results demonstrate the high accuracy and speed of our algorithm, which is suitable for high-speed production lines.

Key words: label inspection; random forest; hierarchy speedup; feature selection

(责任编辑 游中胜)