DOI: 10.13382/j.jemi.2017.07.010

基于 P-ReliefF 特征选择方法的带钢表面缺陷识别*

屈尔庆[1] 刘 坤[1] 陈海永[1] 孙鹤旭[1,2]

(1. 河北工业大学控制科学与工程学院 天津 300130;2. 河北科技大学电气工程学院 石家庄 050018)

摘 要:带钢表面缺陷纹理的复杂性和多样性、背景纹理中存在的伪缺陷等给现有的带钢表面缺陷特征提取和识别带来了极大的困难。为此,提出了一种新的带钢表面缺陷选择与识别方法。首先,通过各向异性扩散算法对带钢表面的伪缺陷干扰进行抑制;其次,利用提出的 P-ReliefF 方法对表面缺陷特征进行选择,相比传统的 ReliefF 方法,该方法考虑了不同维度特征之间的关联性;最后,利用筛选的特征集和支持向量机(SVM)核分类器对带钢表面缺陷进行分类与识别。实验结果表明,提出的方法能够提取出具有高区分性和鲁棒性的带钢表面缺陷特征,并且对于划痕、褶皱、凸起和污渍等不同类型的带钢表面缺陷,本方法相比传统的方法可以获得更高的识别率。

关键词: 特征选择;带钢表面缺陷;ReliefF;相关性;缺陷识别

中图分类号: TP391.4 文献标识码: A **国家标准学科分类代码**: 510.40

Steel strip surface defect recognition based on P-ReliefF feature selection method

Qu Erqing[1] Liu kun[1] Chen Haiyong[1] Sun Hexu[1,2]

(1. College of Control Science and Engineering, Hebei University of Technology, Tianjin, 300130, China;
2. College of electrical engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China)

Abstract: The complexity and diversity of the texture of the surface defects on the surface of strip steel, as well as the false defects in the background texture, have brought great difficulties to the feature extraction and recognition of the strip steel surface defects. Therefore, this paper presents a new method for the selection and identification of steel strip surface defects. First of all, through the inhibition of anisotropic diffusion algorithm for suppression of false defect on strip surface; secondly, selection of surface defect features using P-Relief method in this paper, compared with the traditional Relief method, this method considers the different dimensions of feature correlation. Finally, using the feature set and SVM kernel classifier to classify and recognize the surface defects of steel strip. The experimental results show that the proposed method can extract the features of strip surface defect pairwise independence and robustness, and for scratches, bumps and folds, stains and other different types of defects, this method compared with the traditional method can get higher recognition rate.

Keywords: feature selection; strip surface defect; ReliefF; correlation; defect identification

1 引 言

钢铁是国家制造业的支柱产业,我国当前钢铁年产量为11亿吨,其中作为主要产品冷热轧带钢产量占46%,被广泛应用于设备制造业、化工业、建筑业、航空业、船舶业等多个领域。在带钢的生产过程中,由于原材料、轧制设备等原因,会导致带钢表面出现不同类型的缺

收稿日期:2017-04 Received Date: 2017-04

*基金项目:国家自然科学基金(61403119)、河北省自然科学基金(F2014202166)、天津市特派员科技计划(15JCTPJC55500)资助项目

陷,这些缺陷不仅影响产品外观,而且会降低产品的抗腐蚀性、耐磨性和疲劳强度等性能。利用基于视觉信息的智能检测系统自动获取、传递、分析带钢表面信息对于保证带钢表面缺陷的质量,提高企业的竞争力具有重要的意义。

钢铁生产线的恶劣环境容易导致带钢表面缺陷的多样性和复杂性,这将给缺陷特征提取和选择带来极大的困难。当前,带钢表面缺陷的视觉特征包括从不同视角提取的多种类型的特征,例如包括像素、边缘、骨架或图像区域在内的几何结构特征、直方图特征[1]、协方差矩阵[2]、LBP (local binary pattern) 特征[3]和自相关特征[4]在内的统计特征以及通过空域滤波与频域滤波得到的滤波特征等。为了提高对带钢表面缺陷的识别力,需混合运用各种类型特征;而大量特征的运用将增加冗余,降低计算效率,很难满足生产的实际效率需求。除此以外,缺陷的复杂外观导致部分提取的图像特征的弱判别性和弱鲁棒性。因此,从原始特征集中选择具有高判别力的特征是很重要的。

特征选择是模式识别和机器视觉学习中的一个重要问题,它帮助我们将注意力集中在对分类问题最有效的那些特征上,这些方法选择特征集以增加所提取特征的区分性、鲁棒性和简洁性。依据特定标准,特征选择方法从原始特征集中选择特征子集,或者根据分类性能重新排序特征。在评估和选择后删除无关或冗余的特征可以有效地减少特征维度且提高特征的判别力。

研究人员已经提出了一些用于特征选择的技术,以确定最丰富的特征子空间[5]。2002 年,提出了一种基于特征相似度的无监督特征选择方法[6]。最近,又提出了基于 boosting 的一种嵌入式特征选择方法,该方法找到了一个最优权重值,消除了一些可以忽略的小权重值[7]。Chen 等人[8]提出了一种特征选择方法,该方法既揭示了堆栈泛化的泛化能力,又揭示了量级和几何特征的互补判别信息,提升了基于支持向量机(SVM)的分类准确性。另外,随机变量之间的非线性独立测量的监督特征选择方法,Hilbert-Schmidt 独立标准(HSIC)[9]及相关的扩展算法也被相继提出。非参数加权特征提取(NWFE)[10]也是一种广泛使用的图像数据监督降维方法。

研究结果[11]表明,在目标分类中使用基于 SVM 的特征选择方法是非常有用的。递归特性消除(RFE)和 F 值(FS)是两个简单而著名的 SVM 的基础特征选择方法[12-13]。在 RFE 中,基于 SVM 的决策函数的变化是特征选择的标准。在 FS 方法中,通过计算类内和类间信息的比率得到 F 值对每个特征的判别能力进行度量。特征选择的规则是根据特征判别能力的高低来设置的。然而,由于 F 评分没有揭示特征之间的相互信息,所以它只适合于处理两类特征的问题。ReliefF 特征选择方法[14]通过计算类内和类间样本的差异来测量特征的判别能力进而更新不同特征变量的权重,可以针对多类目标分类问题进行特征选择与提取。但 ReliefF 方法不考虑特征变量间的关联性,并且对不同特征变量间的联合效应缺乏评估。为此,本文提出一个新的 P-ReliefF 特征选择法,它可以考虑到不同维度特征间的关联性与冗余性,获得更具区分力和更紧致的特征子集。

为抑制带钢表面图像中背景纹理的干扰,本文采用各向异性扩散对图像进行去噪处理,可以在平滑背景图像的同时有效地保留目标的边缘,保护图像中的细节特征,提高图像质量。然后,通过 P-ReliefF 特征选择法对由特征变量形成的特征对进行联合评估来选择特征子集。最后利用 SVM 对带钢表面缺陷进行分类,采用径向基函数(RBF)作为分类器的内核函数,有效地实现带钢表面缺陷的分类与识别。

2 各向异性扩散去噪处理

各向异性扩散的去噪过程:首先建立缺陷图像的线性偏微分方程的初始条件,然后求解该微分方程,得到不同时刻的解,简称为 PM 扩散模型。PM 模型利用梯度算子来辨别由噪声引起的图像梯度变化和由边缘引起的图像梯度变化,然后用邻域加权平均去除由噪声引起的小梯度变化,同时保留由边缘引起的大梯度变化,这个过程迭代进行,直至图像中的噪声被去除。依据图像特征的强弱,减少或增强扩散系数[15]。

$$\begin{cases} \dfrac{\partial u}{\partial t} = \operatorname{div}[g(|\nabla u|)\nabla u]; t \in (o,T) \\ u(x,y,0) = u(x,y) \end{cases} \tag{1}$$

式中:$u(x,y,t)$是随时间变化的图像,控制着扩散速率,在梯度大的区域扩散速率大,在梯度小的区域扩散速率小,通常选取的图像梯度函数,这样在扩散时保护到图像边缘信息,$|\nabla u|$是梯度的模,$g|\nabla u|$扩散系数函数用于控制扩散速度。理想的扩散系数应使各向异性扩散在灰度变化平缓的区域快速进行,面对灰度变化比较大的位置(即图像特征处)低速扩散乃至不扩散函数,所以,$g|\nabla u|$具有如下性质:

$$g(0) = 1 \tag{2}$$

$$\lim_{|\nabla u \to \infty|} g(|\nabla u|) = 1 \tag{3}$$

式(2)、(3)分别为在非边缘处加强扩散和在边缘处停止扩散。基于以上两个方面的性质,PM 建立扩散系数函数:

$$c(\|\nabla I\|) = \mathrm{e}^{-(\|\nabla I\|/K)^2} \tag{4}$$

$$c(\|\nabla I\|) = \frac{1}{1 + (\dfrac{\|\nabla I\|}{K})^2} \tag{5}$$

常数项 K 为局部梯度阈值,用来控制边缘的灵敏度,通常经验选取或者用图像噪声相关的函数来表示。

在理想状态下,图像的边缘部分通常具有较大的梯度值,通过设置扩散方程 $c(\|\nabla I\|)$ 使模型在图像边缘实行较弱的平滑,以保持边缘信息,平坦区域通常具有较小的梯度值,设置较大的扩散系数使在图像平坦处实行较强的平滑。

针对擦刮痕、凹坑、水滴等部分带钢表面缺陷图像进行各向异性扩散处理后的结果如图 1 所示。

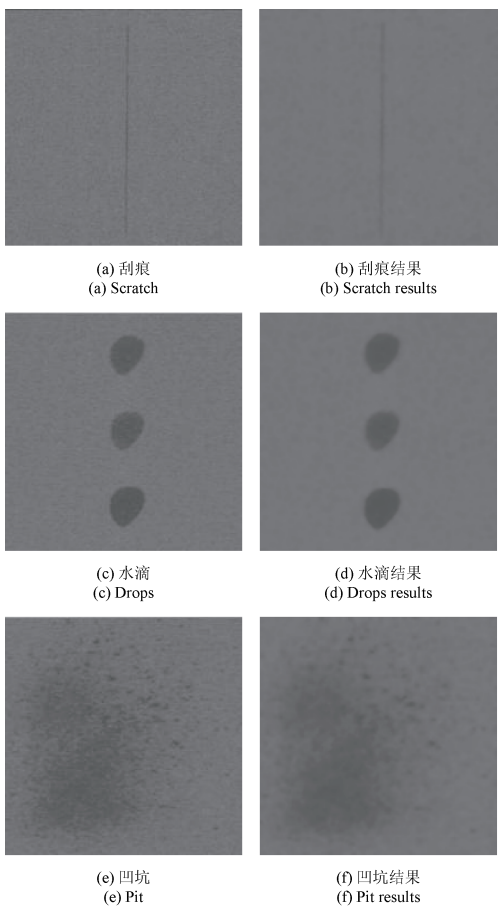


图 1　各向异性图像处理结果

Fig. 1　Anisotropic diffusion image processing results

3　P-ReliefF 特征选择算法

3.1　ReliefF 特征选择算法

Relief 算法是过滤类型特征选择方法之一,最早由 Kira 提出,用于解决两类数据分类的特征维度缩减问题。它使每个维度在特征空间中具有权重,并且通过计算不同维度上样本的类内和类间之间的差异来测量它们的识别力。它是一种基于机器学习的特征评估方法。由于其简单有效,Relief 算法广泛应用于模式识别,并推广至 ReliefF 算法,用于解决多类别分类问题[16]。

在 Relief 算法中不同维度特征的权重通过训练样本的迭代过程来刷新。不同维度特征的识别力是通过计算同类近邻样本和反类近邻样本间的差值来得出的。假设训练样本为 $S=\{s_1,s_2,\cdots,s_m\}$,样本特征为 n 维特征,由 $F=\{f_1,f_2,\cdots,f_n\}$ 表示。然后用全维度特征描述样本,并且第 i 样本由 $s_i=\{s_{i1},s_{i2},\cdots s_{in}\}$,$i\in[1,m]$ 表示。第 1 维度特征的样本差异 s_i 和 s_j 由式(6)确定,称为差异系数。

$$\mathrm{dif}(s_i,s_j,k)=\left|\frac{s_{ik}-s_{jk}}{\max_k-\min_k}\right| \tag{6}$$

式中:$\max_k$ 和 $\min_k$ 是所有样本中第 k 维度特征属性值的最大和最小值。特征差异系数是由不同样本计算得来。如果同类近邻样本同一维度的差异系数很小,而反类近邻样本的差异系数很大,这就意味着此维度特征可识别且特征权重应适当增加;相反,特征权重应适当减少。

简而言之,对于每个采样实例 D 对应的类 C,找到来自相同类的 k 个最近邻域和每个采样实例的相对类,它们分别表示为 H 和 M。假设不同维度的特征权重为 W。特征权重由不断迭代测量训练样本的差异系数计算得出。通过设置阈值来选择较高的权重。然后,获取紧凑特征。Relief 特征选择法如表 1 所示。

表 1　Relief 算法程序

Table 1　Relief algorithm program

算法 1: Relief
W←0
for i = 1 to m do
随机选择一个样本 D
从相同类中寻找 k 最邻近样本 H 并从相对类中选择 k 最邻近样本 M
for j = 1 to n do
W(j) = W(j)-dif(D,H,j) + dif(D,M,j)
end
end

Relief 算法仅适用于两类数据分类问题。为了使 Relief 能够适用于多类别特征选择问题,提出 ReliefF 方法来解决这个问题。二者间的主要区别是权重值更新的方式;ReliefF 法权重由式(7)计算得出。

$$W(j)=W(j)-\mathrm{dif}(D,H,j)+\sum_{class(M)\neq class(D)}\mathrm{dif}(D,M,j) \tag{7}$$

3.2　ReliefF 特征选择算法的不足

传统 ReliefF 算法独立测量每个维度特征的识别力,并将特征权重作为特征评估的标准。然而,标准的应用相当严格且不能保证不同特征的联合效应。而且,由低权重特征组成的联合效应要比独立特征好。如图 2(a)所示,处于正交坐标系中的圆圈和星星分别表示不同等级。坐标系由两个维度特征 f_1 和 f_2 组成。图 2(b)和

(c)表示其在 f_1 和 f_2 上的投影。如果从单独维度特征来看,因为类间差异大于类内差异,无论是 f_1 还是 f_2,样本都不能很好的分类。Relief 算法中 f_1 和 f_2 的权重为负且最终将被忽略。同时,图 2(d)中特征 f_1 和 f_2 联合可有效地区分样本。这意味着两个特性的组合可以获得更好的性能。因此,需要综合考虑不同特征间的联合效应。

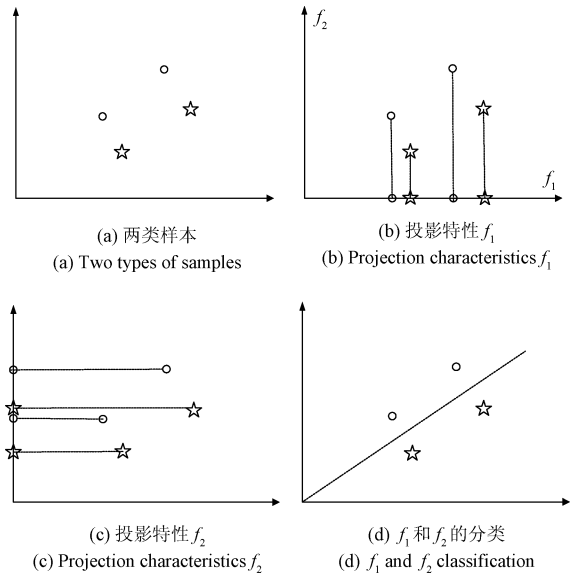


图 2 特征选择示例

Fig. 2 Example of feature selection

3.3 P-ReliefF 特征选择算法

本文提出了一种新的特征评估与选择方法,该方法可以考虑到特征之间的相关性,这里称为 P-ReliefF 算法。该方法是评估组合特征对分类的性能。假设总特征维度为 n,如果在所有维度上组合特征,则计算的复杂性呈指数级。因此,为了减少算法的运算效率,二维特征对被用于评估由最优特征和所有其它特征组成的特征的识别力。即首先选择最优特征,并且选择与其良好协作的其他特征为特征集的伙伴。最优特征可以根据样本的识别力来选择,其可以通过原始 ReliefF 算法获得,还可以通过任何其他方法获得。具体步骤概括如下。首先,利用 ReliefF 特征选择方法对原始的特征集合的贡献度和权重进行计算,找到一个具有最大权重的特征;然后,在随后抽样中,通过测量特征对对分类性能的影响来更新特征权重。特征对由具有最高权重和需要更新的特征组成。目前,将单一特征评估被转换为联合特征对评估。

为了测量特征对的相关性和联合效应,提出一个新的差异系数,其通过计算两个向量的余弦距离来度量特征对之间的相关性,如下:

$$d(A,B,i,j) = \frac{A(i,j)B(i,j)}{|A(i,j)||B(i,j)|} \tag{8}$$

式中:i、j 是特征维度的序号,分别表示第 i、j 个特征;A 和 B 是对样本的标签,$A(i,j)$ 和 $B(i,j)$ 是样本 A 和 B 在特征维度 i 和 j 的特征对构成的二维矢量。

同样假设用符号表示。每个试样实例 D 对应等级 C,从每个样本实例中分别找出同类和不同类的各 k 个最近邻样本,其分别表示为 H 和 M。假设不同维度的特征权重为 W。则 P-ReliefF 算法的过程如表 2 所示。

表 2 P-Relief 算法程序

Table 2 P-Relief algorithm program

算法 2: P-Relief
W←0
i←1
执行 reliefF 特征选择一次,取样权重是 W
for i = 2 to m do
l←为在 W 中的最大特征权重
随机选择一个样本 D
在同类中从最接近样本 H 查找 k,在不同类中从最接近样本 M 查找 k
for j = 1 to n do
W(j) = W(j)-dif(D,H,j,l) + dif(D,M,j, l)
end
end

4 SVM 分类器

SVM 分类器可以利用有限的样本所提供的信息对模型的复杂性和学习能力两者进行了最佳寻优,可以获得较强的泛化能力并已广泛的应用于模式分类应用问题中[17]。SVM 分类器在线性可分情况下,在原空间寻找两类样本的最优分类超平面;在线性不可分的情况下,加入了松弛变量进行分析,通过使用非线性映射将低维输入空间的样本映射到高维属性空间使其变为线性,在高维属性空间采用线性算法对样本的非线性进行分析,在该特征空间中寻找最优分类超平面。通过在属性空间构建最优分类超平面,得到全局最优的分类器,并在整个样本空间的期望风险以某个概率满足一定上界。样本集如下:

$$T = (x_i, m_i)_{i=1}^{N} \tag{9}$$

其中表示大小为 N 的输入数据中的样本,m_i 表示样本的标签,$m_i = 1$ 或 $m_i = -1$。可以表达为:

$$\boldsymbol{w}^{\mathrm{T}} \cdot x + b = 0 \tag{10}$$

式中:$\boldsymbol{w}$ 是超平面的法向量,也称为权重向量,b 是常数偏差。

样本点和超平面之间的距离越大,样本就越可能被正确分类,找到最佳平面或分类器,使来自样本集合中的点的最小几何边界最大化。

$$\min \frac{1}{2}\|\boldsymbol{\omega}\|^2 \tag{11}$$

$$\text{s.t.}\quad y_i(\boldsymbol{\omega}^{\mathrm{T}}\cdot x_i+b)\geqslant 1\quad i=1,2,\cdots,N \tag{12}$$

它是一个标准二次规划(QP)优化,其双重问题可以通过拉格朗日二元性实现:

$$\min_{\omega,b}\max_{a_i\geqslant 0}\left[\frac{1}{2}\|\boldsymbol{\omega}\|-\sum_{i=1}^{N}a_i(m_i(\boldsymbol{\omega}^{\mathrm{T}}\cdot x_i+b)-1)\right] \tag{13}$$

式中:a_i 是拉格朗日乘数,满足 KKT 条件,可以交换最小和最大的位置。

$$m=\operatorname{sgn}\left(\sum_{i=1}^{N}a_i m_i(\boldsymbol{\psi}^{\mathrm{T}}(x_i),\boldsymbol{\psi}^{\mathrm{T}}(x))+b\right) \tag{14}$$

VC 维是对函数的一种度量,和样本的维数无关,VC 维越高,问题越复杂。所以核函数可以在低维度实现内积并在高维度实现分类:

$$m=\operatorname{sgn}\left(\sum_{i=1}^{N}a_i m_i K(x_i,x)+b\right) \tag{15}$$

如果情况是线性不可分离的,优化问题变为:

$$\min\frac{1}{2}\|\boldsymbol{\omega}\|^2+E\sum_{i=1}^{N}\xi_i \tag{16}$$

$$\text{s.t.}\quad m_i(\boldsymbol{\omega}^{\mathrm{T}}\cdot x_i+b)\geqslant 1-\xi_i \tag{17}$$

式中:ξ_i 是松弛变量,$\xi_i\geqslant 0, i=1,2,\cdots,N$,$E$ 是惩罚因子。

由于核函数在设置 SVM 模型中起着重要作用,因此重要的是选择适当的核函数。在本文中,SVM 模型基于径向基函数构建如下:

$$\max\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_i a_j m_i m_j(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j) \tag{18}$$

其中 $0\leqslant a_i\leqslant E$,

$$\sum_{i=1}^{N}a_i m_i=0\quad i=1,2,\cdots,N \tag{19}$$

$$K(x,y)=\exp(-\gamma\times\|\boldsymbol{u}-\boldsymbol{v}\|^2) \tag{20}$$

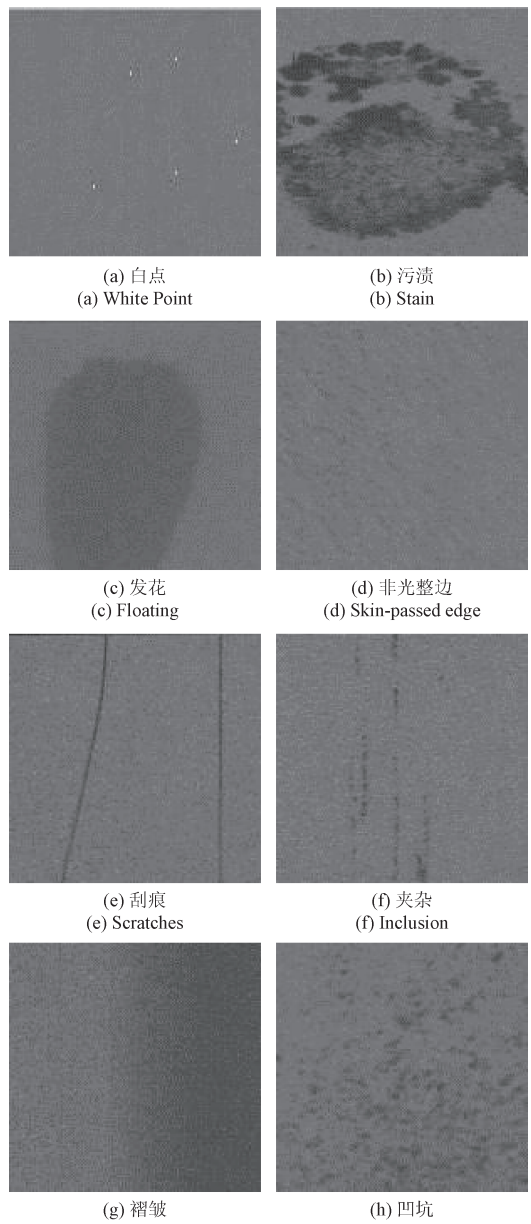
选择径向基核函数需设定两个参数,核函数自身的参数 $\gamma=0.1$,以及错误代价系数 $C=0.1$。SVM 的最终决策函数只由少数的支持向量所确定,计算的复杂性取决于支持向量的数目,而不是样本空间的维数。

5　硬件系统

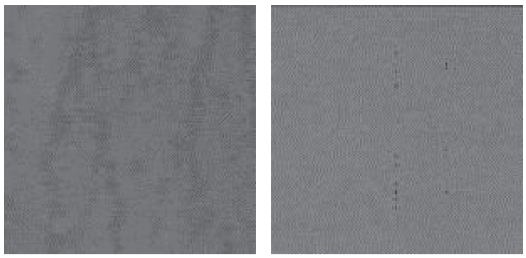
带钢表面检测系统由图像采集装置、光源和输送单元组成。图像采集系统使用的相机是 DALSA 的黑白线阵相机,具有 36 kHz 的水平扫描频率。LED 条形灯在频繁闪光模式下对带钢表面进行照明,保证了清晰、高对比度图像的采集。输送装置采用多辊组合,可增加螺纹以避免设备在某些紧急情况下出现脱节现象,也可增加张力避免长距离传动引起的波动。钢板的厚度从 1.5 ~ 10 min不等。根据所需厚度,钢板的轧制速度范围从 3 ~ 20 m/s。图像处理单元由高性能服务器执行。它包含 48 GB存储器的 GHz 处理器,可以满足高速图像处理的要求。

6　实验分析

为了检验类本文的特征提取与选择方法对缺陷的识别性能,本文建立了带钢表面缺陷样本库,该缺陷库共包含了 10 类缺陷,样本库中的缺陷图像来自于唐钢生产线,通过已经应用于唐山钢铁集团现场的表面缺陷检测系统在线采集并整理后得到的。图 3 所示为 10 类缺陷有代表性的部分样本。样本库共有 2 390 幅图像,包括白点(WP)、污渍(ST)、发花(Fl)、非光整边(SE)、刮痕(SC)、夹杂(IN)、褶皱(FO)、凹坑(PI)、水滴(WD)、网眼(NM)等 10 种缺陷类型。表 3 列出了样本库的缺陷类别和样本数量。



(a) 白点　(a) White Point
(b) 污渍　(b) Stain
(c) 发花　(c) Floating
(d) 非光整边　(d) Skin-passed edge
(e) 刮痕　(e) Scratches
(f) 夹杂　(f) Inclusion
(g) 褶皱　(g) Fold
(h) 凹坑　(h) Pit



(i) 水滴
(i) Water drop

(j) 网眼
(j) Net-mesh

图 3　带钢表面缺陷的样本图像

Fig. 3　Sample images of strip surface defects

表 3　样本库的缺陷类别和样本数目

Table 3　Defect type and sample number of sample library

类型	WP	ST	Fl	SE	SC
数量	160	400	180	200	350
类型	IN	FO	PI	WD	NM
数量	100	180	200	280	340

在特征选择之前，提取灰度特征、纹理特征和频域特征作为原始的带钢表面缺陷特征。灰度特征主要由基于灰度值的统计信息组成，包括均值、方差、峰度和能量等。纹理特征主要描述像素的灰度级的空间布置，这里提取基于灰度共生矩阵（GLCM）的二阶矩、对比度，熵和逆差矩等特征。提取的频域特征主要包括傅里叶变换和小波特征等。利用上述特征组合得到 49 个维的原始特征表示带钢表面缺陷。

为了联合评估不同类型的特征对分类的贡献度，本文首先使用标准化方法对上述特征数据进行归一化以消除由特征向量之间的不同标准导致的附加偏差的影响。数据标准化方法如下：

$$D = \frac{D - \mu}{\sigma} \tag{21}$$

式中：μ 是平均值，σ 是特征的标准偏差。为了验证所提出的 P-ReliefF 特征选择法的有效性，选取其中 6 类缺陷的原始特征向量的箱线图与 P-ReliefF 特征选择向量的箱线图相对比。图 4 所示为原始特征集中随机的两个特征向量的箱线图，图 5 所示为由 R-ReliefF 法选择特征向量的两个箱线图。箱线图中横坐标代表不同的缺陷类型，纵坐标表示在不同特征维度上的度量值。

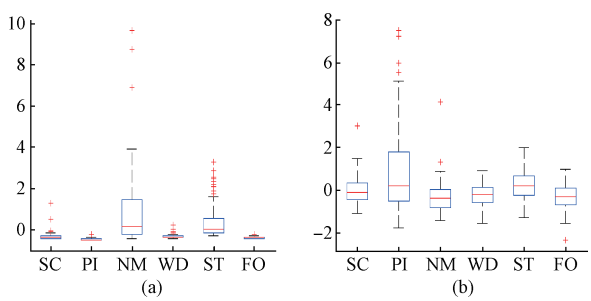


图 4　原始特征向量的箱线图

Fig. 4　The boxplot of original feature vector

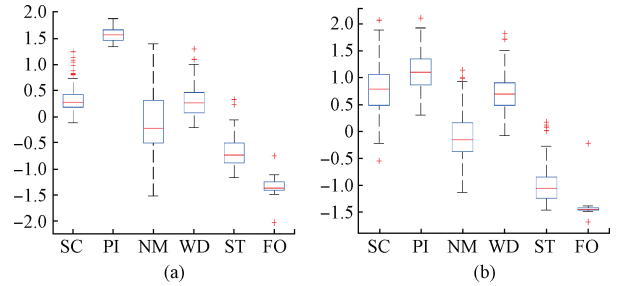


图 5　P-ReliefF 特征向量选择箱线图

Fig. 5　The boxplot of feature vector selected by P-ReliefF

从图 4 可以看出原始特征向量的数据范围在不同缺陷间存在较大的重叠分布区域，而且很难识别不同种类的缺陷，而在图 5 中利用 P-ReliefF 选择的特征具有明显的高识别力。因此，使用 P-ReliefF 算法选取特征向量能更好的识别不同种类的缺陷，能够消除较低的识别力。

为进一步验证本文提出的特征选择方法的有效性，本文利用 P-ReliefF 方法筛选出的带钢表面缺陷特征，结合 SVM 在实际生产线上采集的 10 种不同类型的缺陷上进行测试。

图 6～8 所示为针对 10 种不同类型的带钢表面缺陷，使用原始的特征向量与利用 ReliefF 特征选择算法及本文提出的 P-ReliefF 特征选择算法进行特征选择后进行分类的结果。其中图 6 给出了利用原始特征向量进行缺陷分类时统计得到的混淆矩阵，图 7 给出了利用 ReliefF 特征选择进行缺陷分类时统计得到的混淆矩阵，图 8 给出了利用本文的 P-ReliefF 算法提取的特征进行缺陷分类时统计得到的混淆矩阵。

	WP	DT	Fl	SE	Sc	In	Fo	PS	Mo	SH
WP	90.2			9.8						
DT		97.1						29		
Fl		7.8	81.1						11.1	
SE				68.7				31.3		
Sc					74.3	25.7				
In	2				8.6	89.4				
Fo				2.9			97.1			
PS		50.5						35.7	13.8	
Mo		18.0						10.4	71.6	
SH				20.0						80.0

图 6　原始特征向量分类结果

Fig. 6　Classification results by the original feature

	WP	DT	Fl	SE	Sc	In	Fo	PS	Mo	SH
WP	97.8			2.20						
DT		98.7						1.3		
Fl			80.3						19.7	
SE				92.6				7.4		
Sc					92.7	7.3				
In					8.5	91.5				
Fo		5.9					94.1			
PS					9.0			50.6	40.4	
Mo									71.2	28.8
SH				21.5						78.5

图 7　ReliefF 特征选择方法的分类结果

Fig. 7　Classification results by the ReliefF feature selection method

	WP	DT	Fl	SE	Sc	In	Fo	PS	Mo	SH
WP	98.7			1.3						
DT		89.1						10.9		
Fl			84.7					15.3		
SE				94.2				5.8		
Sc					95.1	4.9				
In					15.5	84.5				
Fo				4.4			95.6			
PS		8.2						61.8	30.0	
Mo								12.7	74.6	12.7
SH				4.5						95.5

图 8　P-ReliefF 特征选择方法的分类结果

Fig. 8　Classification results by the P-ReliefF algorithm

从分类混淆矩阵的对比结果中可以看出，相比原始的特征集合而言，经过 ReliefF 特征选择后的缺陷识别率有所提升。另外，采用本文提出的 P-ReliefF 方法的识别率高于原始的 ReliefF 特征选择后的识别率，验证了本文提出的特征对权重更新方法在特征选择方面的优越性。

为了进一步验证本文提出方法的效果，本文在表 4 给出了针对不同类型的缺陷，利用原始的特征集合、经过 ReliefF 特征选择筛选的特征与利用本文提出的 P-ReliefF 特征选择方法筛选的特征进行缺陷识别时的统计性能。通过计算缺陷识别结果的真正率(TP)、真负率(TN)、假正率(FP)和假负率(FN)得到两个性能指标精度(Precision)和召回率(Recall)，根据精度和召回率计算 F-measure 可以评估算法的整体性能，各指标定义如下：

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

表 4　精度，召回率和 F-measure 性能指标

Table 4　Precision, Recall and F-measure performance index

缺陷	原始特征			ReliefF 特征			P-ReliefF 特征		
	Pr	Re	Fm	Pr	Re	Fm	Pr	Re	Fm
WP	74.8	90.2	82.1	96.5	97.8	97.1	97.0	98.7	97.9
DT	95.2	97.1	96.2	98.2	98.7	98.4	98.9	89.1	93.8
FI	63.8	81.1	71.4	68.1	80.3	73.7	79.0	84.7	81.7
SE	93.9	68.7	79.3	94.4	92.6	93.5	95.6	94.2	94.9
Sc	89.7	74.3	81.3	93.3	92.7	93.0	96.5	95.1	95.8
In	94.6	89.4	91.9	97.2	91.5	94.3	99.1	84.5	91.2
Fo	97.9	97.1	97.5	97.3	94.1	95.7	98.1	95.6	96.8
PS	70.9	35.7	47.5	80.2	50.6	62.1	96.6	61.8	75.4
Mo	83.7	71.6	77.2	70.4	71.2	70.8	97.1	74.6	84.4
SH	89.6	80.0	84.6	91.5	78.5	84.5	97.5	95.5	96.5
平均	85.4	78.5	80.9	88.7	84.8	86.3	95.5	87.4	90.8

从表 4 统计结果可以看出，与原始的特征向量和传统 ReliefF 方法筛选的特征相比，由 P-ReliefF 算法选择的特征具有更好的分类性能，同时也表明本文提出的基于特征对的权重更新方法可以更好地对带钢表面缺陷进行特征选择与识别。

7　结　论

本文针对现有的 ReliefF 特征选择方法中缺乏对特征之间相关性考虑的问题，提出了一种新的 P-ReliefF 特征选择算法，该算法以特征对为基础评估了不同维度特征对之间的相关性和冗余性，进而获得了对特征的区分力的排序，得到更为高效、紧致的特征子集。通过在带钢表面缺陷的分类与识别问题进行验证后表明，本文提出的特征选择方法可以从复杂多变的带钢表面纹理特征中提取更有效的特征子集，提高了带钢表面缺陷的识别率。

目前，考虑到算法的计算效率，本文的特征选择方法主要通过对二维的特征对之间的相关性评估获得对分类的贡献度的计算，主特征选择的次序在一定程度会影响最终的特征子集结果。未来的研究将进一步探索在多维空间下的特征群之间的关联性与冗余性，在获得更高效、紧致的特征子集的同时，去除特征选择次序对结果的影响。

参考文献

[1] ZAHRAN O, KASBAN H, EL-KORDY M. Automatic weld defect identification from radiographic images[J]. NDT & E International, 2013, 57(6):26-35.

[2] 赵永红,张林让,刘楠,等. 采用协方差矩阵稀疏表示的DOA估计方法[J]. 西安电子科技大学学报:自然科学版, 2016, 43(2):58-63.
ZHAO Y H, ZHANG L R, LIU N, et al. A DOA estimation method using sparse representation of covariance matrices[J]. Journal of Xi'an Electronic and Science University: Natural Science Edition, 2016, 43(2):58-63.

[3] XU K, SONG CH. Feature extraction method based on global binary pattern and its application[J]. Pattern Recognition & Artificial Intelligence 2013, 26(9): 872-877.

[4] 周美丽,白宗文. 基于形状特征的图像检索系统的设计[J]. 国外电子测量技术, 2015,34(6): 82-84.
ZHOU M L, BAI Z W. Design of image retrieval system based on shape feature[J]. Computer Engineering, 2015,34(6): 82-84.

[5] JIA X P, KUO B C, CRAWFORD M M. Feature mining for hyperspectral image classification[C]. Proceedings of the IEEE, 2013, 101(3):676-697.

[6] MITRA P, MURTHY C A, PAL S K. Unsupervised feature selection using feature similarity[J]. IEEE Pattern Analysis and Machine Intelligence, 2012, 24(3):301-312.

[7] ARCHIBALD R, FANN G. Feature selection and classification of hyperspectral images with support vector machines[J]. IEEE Transactions on Geoscience and Remote Sensing,2007, 4(4):674-677.

[8] CHEN J, WANG C, WANG R. Using stacked generalization to combine svms in magnitude and shape feature spaces for classification of hyperspectral data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(7):2193-2205.

[9] GRETTON A, BOUSQUET O, SMOLA A J, et al. Measuring statistical dependence with Hilbert-Schmidt norms[C]. International Conference on Algorithmic Learning Theory, 2005:63-77.

[10] KUO B C, LANDGREBE D A. Nonparametric weighted feature extraction for classification[J]. IEEE Transactions on Geoscience & RemoteSensing, 2005, 3809(5):567-576.

[11] PAL M. Supportvector machine-based feature selection for land cover classification: A case study with DAIS hyperspectral data[J]. International Journal of Remote Sensing, 2006, 27(14):2877-2894.

[12] GUYON I, GUNN S, NIKRAVESH M, et al. Feature extraction, foundations and applications[J]. Studies in Fuzziness and Soft Computing, 2006, 205(12):68-84.

[13] 潘启明,符承军. 基于ACO-FS-SVM特征选择加权的网络入侵分类方法[J]. 计算机与数字工程,2014, 42(8):1454-1561.
PAN Q M, FU CH J. Network intrusion classification method based on ACO-FS-SVM feature selection weighting[J]. Computer and Digital Engineering,2014, 42(8):1454-1561.

[14] 何涛,胡洁,夏鹏,等. 基于ReliefF算法与遗传算法的肌电信号特征选择[J]. 上海交通大学学报,2016, 50(2): 204-208.
HE T, HU J, XIA P,et al. EMG signal feature selection based on ReliefF algorithm and genetic algorithm[J]. Journal of Shanghai Jiaotong University,2016, 50(2): 204-208.

[15] 宋晓琳,郑亚奇,曹昊天. 基于HMM-SVM的驾驶员换道意图辨识研究[J]. 电子测量与仪器学报, 2016, 30(1):58-65.
SONG X L, ZHENG Y Q, CAO H T. Research on driver lane change intention identification based on HMM-SVM[J]. Journal of Electronic Measurement and Instrumentation, 2016, 30(1):58-65.

[16] KUO B C, HO H H, LI C H, et al. A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(1): 317-326.

[17] SAH R, SHREEJA B V, MANIKANTAN K, et al. Entropic-GWT based feature extraction and LBPSO based feature selection for enhanced face recognition[C]. International Conference on Communications and Signal Processing (ICCSP), IEEE, 2015:180-184.

作者简介

屈尔庆,1971年出生,2010年于燕山大学获得硕士学位,现为河北工业大学博士研究生,正高级工程师,主要研究方向为冶金信息自动化。
E-mail:querqing@126.com

Qu erqing was born in 1971, received his M.Sc. degree in 2010 from Yanshan University, now he is Ph.D. student in Hebei University of Technology, professor of engineering. His main research interest include metallurgical information and automation.