一种基于拒识的高可靠性 CAPTCHA 识别算法

张 亮 [1] 张 亮 [1] 黄曙光 [1] 石昭祥 [1]

摘 要 CAPTCHA 是一种阻止机器人滥用自然人资源的网络安全机制. 研究 CAPTCHA 识别技术有助于发现 CAPTCHA 自身的缺陷, 促使其变得更加安全. 针对现有方法难以识别的高粘着 CAPTCHA, 本文提出了一种新的识别算法. 该算法首先使用递归神经网络 (Recurrent neural network, RNN) 对 CAPTCHA 进行识别, 然后为了提高识别结果的可靠性, 提出了一种基于 SVM 的拒识新算法, 并使用数据降维方法对拒识特征进行降维. 实验结果表明: 1) 本文所提识别算法能够识别高粘着型 CAPTCHA, 并且识别结果具有高可靠性; 2) 新的拒识算法相对于其他拒识算法具有明显优势; 3) 数据降维方法能够进一步改善拒识算法的性能, 从而取得更高的可靠性.

关键词 人工智能, 网络安全, CAPTCHA 识别, 拒识算法, 可靠性

DOI 10.3724/SP.J.1004.2011.00891

A Highly Reliable CAPTCHA Recognition Algorithm Based on Rejection

ZHANG Liang[1] ZHANG Liang[1] HUANG Shu-Guang[1] SHI Zhao-Xiang[1]

Abstract CAPTCHA is a kind of network security mechanism that blocks machines from abusing network resource owned by human. Studying the recognition of CAPTCHA can help to find its hidden defects, and thus make it securer. To read closely-connected CAPTCHA that can hardly be recognized by methods of state of art, this paper brought up a new recognition algorithm based on rejection. During the process of this algorithm, recurrent neural network (RNN) was first used to recognize the unknown CAPTCHA. Then, to make the recognition results reliable, a new rejection algorithm was brought up. Data dimension reduction was also performed on rejection features. Experiment results show the following three points: Firstly, our new recognition algorithm can recognize closely-connected CAPTCHA with high reliability. Secondly, the new rejection algorithm is superior to other methods of state of art. Lastly, data dimension reduction algorithm can improve the performance of the rejection algorithm, thus making the recognition results more reliable.

Key words Artificial intelligence, network security, CAPTCHA recognition, rejection algorithm, reliability

随着网络技术的发展, 电子商务、电子政务应用不断深入, 越来越多的传统服务可以网上办理, 在为人们带来巨大便利的同时, 也产生了很多网络安全问题, 其中一个被网络服务设计人员以及网络安全研究人员普遍忽视的问题是: 设计的服务默认都是给自然人使用的, 但是 "On the Internet, no body knows you are a dog"[1] , 没有人能够保证使用者一定是个自然人. 大规模垃圾邮件发送机、论坛灌水机、游戏外挂等就是典型的机器人使用人类资源进行非法行为的例子, 因此需要一种安全机制解决机器人滥用自然人资源的问题.

CAPTCHA (Completely automated public turing test to tell computers and humans apart)[2−5] 就是一种阻止机器人滥用自然人资源的网络安全机制, 目前已经被广泛应用到互联网的各个领域, 包括电子邮件、网络论坛、网上银行等, 已经成为互联网的一个标准防范措施. 鉴于它广阔的应用前景, 目前很多学术机构以及商业公司例如 CMU, PARC, Stanford, NSF, Yahoo, Microsoft, Google, Bell Labs, IBM T.J. Watson, RSA Security Laboratories 等都在对其进行研究. CAPTCHA 存在很多种类型, 最常见的是文字型, 如图 1.

图 1 Yahoo 的 CAPTCHA

Fig. 1 A sample of Yahoo′s CAPTCHA

CAPTCHA 主要基于现有的一些公认的人工智能难题进行实现, 因此研究 CAPTCHA 识别技术不仅有助于发现现有 CAPTCHA 的缺陷, 促进 CAPTCHA 安全性的提高, 而且有助于一些人工智能难题的求解. 文字型 CAPTCHA 主要是针对现有 OCR 技术的弱点而设计的, 因此研究 CAPTCHA 识别还有助于提高现有的 OCR 技术.

本文的识别对象为粘着严重的文字型 CAPTCHA. 针对文字型 CAPTCHA 的识别, 目前的主要方法有模板匹配、神经网络、支持向量机 (Support

收稿日期 2010-06-08 录用日期 2011-03-10
Manuscript received June 8, 2010; accepted March 10, 2011
1. 电子工程学院 合肥 230037
1. Electronic Engineering Institute, Hefei 230037

vector machine, SVM) 等方法[6−10]. 模板匹配法的基本流程是: 模板库建立, 图像预处理, 图像分割, 图像识别. 基于神经网络和 SVM 的识别方法的基本流程是: 图像预处理, 图像分割, 特征提取, 分类器训练 (识别). 分析现有方法可知, 它们存在一个共同阶段: 图像分割. 图像分割对于早期的 CAPTCHA 来说并不是一个难题, 目前研究人员也主要是基于分割的方法来识别 CAPTCHA. 但是随着 CAPTCHA 技术的发展, 高级 CAPTCHA 已经普遍使用粘着严重的字符, 如图 1. 使用基于分割的方法识别高级 CAPTCHA 非常困难, 这已被文献 [11] 所证实. 该网站曾检测到有两台主机同时对 Google 的 CAPTCHA 进行了大量的识别, 其中一台使用了基于分割的识别方法. 监测显示, 该主机基本上没有一次识别成功. 因此, 针对这类 CAPTCHA 的识别, 本文从不分割的角度对其进行研究.

本文的基本结构如下: 首先介绍了高可靠性识别算法的工作流程, 然后对其中的 RNN 识别模块进行了描述; 为了提高识别结果的可靠性, 在分析现有拒识算法的基础上, 提出了一种新的基于 SVM 的拒识算法, 并使用数据降维方法对拒识特征进行降维处理; 最后通过实验证明了新识别算法和新拒识算法的有效性.

1 一种基于拒识的高可靠性 CAPTCHA 识别算法

1.1 基本定义和概念

分类器对未知样本进行识别以后, 如果不使用拒识机制, 则直接返回识别结果. 使用拒识机制时, 则根据一定的策略判断识别结果是否可信. 如果拒识机制认为不可信, 则拒绝识别该样本, 并指出识别失败; 如果可信, 则返回识别结果. 设 N_{test} 为测试集的大小, N_{reco} 为正确识别的样本数, N_{rej} 为拒绝识别的样本数 (也即拒识机制认为不可信的样本数), N_{err} 为识别错误的样本数 (也即拒识机制认为识别正确而实际上是错误的样本数), 则

$$\text{识别率} = \frac{N_{\text{reco}}}{N_{\text{test}}} \tag{1}$$

$$\text{可靠性} = \frac{N_{\text{reco}}}{N_{\text{reco}} + N_{\text{err}}} \tag{2}$$

$$\text{拒识率} = \frac{N_{\text{rej}}}{N_{\text{test}}} \tag{3}$$

$$\text{错误率} = \frac{N_{\text{err}}}{N_{\text{test}}} \tag{4}$$

$$N_{test} = N_{\text{rej}} + N_{\text{reco}} + N_{\text{err}} \tag{5}$$

分析可知 $N_{\text{reco}} + N_{\text{err}}$ 即为拒识机制认为分类器识别正确的样本数. 可靠性反映了识别结果的可信赖程度, 是识别系统的一个重要指标[12−13]. 可靠性的公式可以变形为

$$\begin{aligned}\text{可靠性} = \frac{N_{\text{reco}}}{N_{\text{reco}} + N_{\text{err}}} = \frac{\text{识别率}}{\text{识别率} + \text{错误率}} = \\ \frac{N_{\text{reco}}}{N_{\text{test}} - N_{\text{rej}}} = \frac{\text{识别率}}{1 - \text{拒识率}} = \\ \frac{N_{\text{test}} - N_{\text{rej}} - N_{\text{err}}}{N_{\text{test}} - N_{\text{rej}}} = 1 - \frac{\text{错误率}}{1 - \text{拒识率}}\end{aligned} \tag{6}$$

因此, 对于同一个分类器的识别结果 (N_{test} 相等), 两个不同的拒识算法在相同识别率的情况下, 可靠性越高的拒识算法拒绝的样本数越多 (拒识率越高), 即可靠性也反映了拒识算法拒绝错误识别结果的能力. 在相同识别率下, 可靠性越高的拒识算法性能越好. 从式 (6) 的最后一部分也可以看出: 在相同的拒识率下, 错误率越低越可靠.

在 CAPTCHA 识别中, 可靠性是一个非常重要的因素. 目前很多 CAPTCHA 都有 “看不清, 换一张” 的功能, 如图 1. 使用 CAPTCHA 识别算法的外部程序 (例如自动化 web 应用程序) 可以很容易重新获取 CAPTCHA 图片. 因此即使识别算法的识别率不是很高, 但是如果它能够比较准确的拒绝识别错误的图片, 那么在发生拒识时, 外部程序即可使用 “换一张” 的功能, 指示识别算法对另外一张 CAPTCHA 图片进行识别, 从而间接提高了 CAPTCHA 的识别率. 外部程序对结果的可靠性会有较高的要求, 因为错误的识别结果会影响外部程序的运行效率.

1.2 高可靠性 CAPTCHA 识别算法的基本流程

本文提出了一种基于拒识的高可靠性 CAPTCHA 识别算法, 算法的基本流程为:

1) 对 CAPTCHA 图片进行空白裁剪、灰度化、高度归一化等基本预处理;

2) 使用滑动窗口[14] 提取图像的灰度值作为特征序列;

3) 使用 RNN 识别模块对特征序列进行识别;

4) 使用一种基于 SVM 的拒识新算法判断 RNN 识别结果是否可信. 如果可信, 则输出 RNN 的识别结果, 否则拒绝返回识别结果并指示识别失败.

图 2 是本文算法流程图. 从该图可以看出, 算法中存在两个重要模块: RNN 识别模块和 SVM 拒识模块. 下面分别对这两个模块进行阐述.

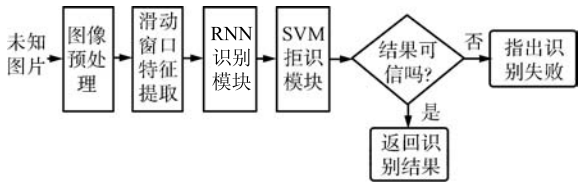


图 2　高可靠性 CAPTCHA 识别算法流程图

Fig. 2　Flow chart of the highly reliable CAPTCHA recognition algorithm

2　基于 RNN 的 CAPTCHA 识别

递归神经网络 (Recurrent neural network, RNN) 是 Graves 等提出的一种递归神经网络[15−16]. 它识别时不依赖于字典, 适合 CAPTCHA 识别的特点.

2.1　RNN 基本结构

用于 CAPTCHA 识别的 RNN 为 3 层结构: 一个输入层、一个隐层以及一个输出层[17], 如图 3 所示. 输入层的输入为 CAPTCHA 图像经过滑动窗口处理后形成的特征值序列 $O_1O_2\cdots O_T$, 输入层的节点数为输入数据的维度, 也即滑动窗口的高度. 隐层包含多个 LSTM (Long short-term memory) 单元[15]. RNN 的特性之一在于 LSTM 单元具有很强的记忆能力, 能够在时间间隔很长的两个特征值之间建立上下文联系, 而没有普通递归神经网络的梯度消亡现象. 在时刻 t $(1 \le t \le T)$, 某个 LSTM 单元的输入不仅包括所有输入节点对它的输入, 而且包括时刻 $t-1$ 所有 LSTM 单元的输出, 如图 3 隐层中最上方的 LSTM 单元. LSTM 单元的具体结构和相应公式可参见文献 [15].

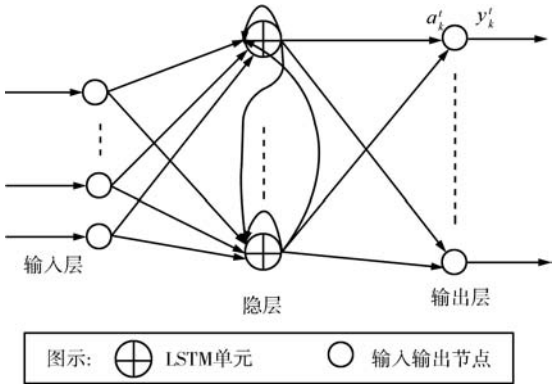


图 3　RNN 结构图

Fig. 3　Architecture of RNN

输出层的目的是求出 $O_1O_2\cdots O_T$ 对应的字符串 $z_1z_2\cdots z_L$ (L 为字符串的长度). 输出层共有 $M+1$ 个输出端, 其中 M 为 CAPTCHA 中包含的字符类别数, 1 对应空白字符. 第 k 个输出端在时刻 t 的输出 y_k^t 为

$$y_k^t = \frac{e^{a_k^t}}{\sum\limits_{k'} e^{a_{k'}^t}},\ 1 \le k \le M+1,\ 1 \le t \le T \quad (7)$$

式 (7) 中 y_k^t 实质上表示时刻 t 输出第 k 个字符的归一化概率, 其中 a_k^t 是第 k 个输出端在时刻 t 接收到的输入之和, 如图 3. 设 b_h^t 表示隐层中第 h 个 LSTM 单元在时刻 t 的输出, w_{hk} 表示第 h 个 LSTM 单元与第 k 个输出端之间的连接权值, 则

$$a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t \quad (8)$$

在每个时刻 $t = 1, \cdots, T$, 选择一个输出端作为当前时刻的输出结果. 将 T 个时刻的选择结果连接起来, 则形成一条路径 π, π 经过函数 $B(\pi)$ 的进一步处理以后得到输出字符串 z. $B(\pi)$ 的作用是先将 π 中相邻的相同字符合并成一个字符, 然后移除字符间的空格[17]. 例如当 π =“AA − AB − CCD−” 时 (“−” 表示空格, 下同), $B(\pi)$ 的输出为 “AABCD”. 分析可知, 同一个字符串对应着多条不同的路径, 其中后验概率最大的路径称为字符串的概率最大路径. 路径的后验概率为 $p(\pi|x)$:

$$p(\pi|x) = \prod_{t=1}^{T} p(\pi^t, t|x) = \prod_{t=1}^{T} y_{\text{sel}}^t \quad (9)$$

式 (9) 中 π^t 是在时刻 t 选择的输出端, y_{sel}^t 表示 π^t 上的输出值.

2.2　RNN 的训练和识别

RNN 训练和识别时均不需要进行字符分割. 设某个 CAPTCHA 样本对应的特征值序列为 $x = O_1O_2\cdots O_T$, 对应的字符串为 $z = z_1z_2\cdots z_L$, 那么训练的目标是使所有训练样本 (设为 Tra) 的字符串在对应的特征值序列之上的概率之积最大, 可以使用 ln 函数将其描述为 (加负号变为求最小)[17]:

$$O = -\ln\Big(\prod_{(x,z)\in Tra} p(z|x)\Big) = -\sum_{(x,z)\in Tra} \ln p(z|x) \quad (10)$$

目标函数确定以后, 使用 BPTT 算法对网络进行训练[15].

RNN 训练好以后, 即可用于识别. 对于某个未知 CAPTCHA, 首先使用滑动窗口提取它的特征值序列, 然后将特征值序列输入到 RNN 中. 通过使用邻域解码算法[17], RNN 即输出一个前 N 个候选词

列表 $TOPN\_LIST$:

$$TOPN\_LIST = \{\langle word_1, prob_1\rangle, \cdots, \langle word_N, prob_N\rangle\} \quad (11)$$

其中, $word_i$ 表示第 i 个候选词, $prob_i$ 表示其对应的后验概率.

3 基于 SVM 的拒识新算法

提高 CAPTCHA 识别结果的可靠性有两种方法: 1) 提高 RNN 自身的识别能力; 2) 使用拒识模块. 本文使用拒识模块. 拒识模块的主要功能是判断 RNN 是否识别正确, 属于一个 2 分类问题, 因此使用 SVM 作为分类器.

对于给出 $TOPN\_LIST$ 下的拒识, 目前有基于归一化识别得分及其差分值的方法[18]、基于概率阈值的方法[19]、LDAM[12] 等方法, 这些方法都存在一个共同点: 只根据 $prob_1, \cdots, prob_N$ 及其变换形式进行拒识, 而没有考虑输入输出以及分类器本身的特性, 但是这些特性对于判断分类器识别是否正确是非常有效的, 因此本文提出了一种综合这些因素的拒识新算法.

3.1 拒识新特征的提取

设 $word_1, \cdots, word_N$ 对应的概率最大路径为 $\pi_1, \cdots, \pi_N$, π_i 的求法可使用文献 [17] 提出的邻域解码算法. 并设 $\pi_i = \pi_i^1 \cdots \pi_i^T$, 其中 π_i^t 为在 π_i 路径中时刻 t 的输出字符. 通过分析, 选取如下拒识新特征:

1) $word_1, \cdots, word_N$ 的归一化后验概率, 其计算公式为

$$prob_i = \frac{prob_i}{\sum_{j=1}^{N} prob_j} \quad (12)$$

后验概率是目前大多数拒识算法主要使用的特征, 它们具有较强的分辨能力.

2) 输入序列的长度 T 和 $word_i$ 的长度 $|word_i|$ $(1 \le i \le N)$.

较长的输入序列对应较长的 CAPTCHA, 因此, 如果一个很长的序列却产生一个很短的单词, 或者很短的序列产生很长的单词输出都是值得怀疑的.

3) 在 $\pi_1, \cdots, \pi_N$ 中, 字符之间的最小间距 $\min Dist_j$ $(1 \le j \le N)$.

字符都需要占用一定的宽度, 两个相邻字符之间间隔的时间步数越短, 则输出结果越不可靠. 对于 π_j, $\min Dist_j$ 的计算公式为

$$\begin{aligned}\min Dist_j = {} & \\ & \arg\min_d(\pi_j^i \pi_j^{i+1} \cdots \pi_j^{i+d} | \pi_j^i \,!=' -' \text{ and} \\ & \pi_j^{i+d} \,!= \,' -' \text{ and} \\ & \pi_j^{i-1} \,!= \pi_j^i \text{ and} \\ & i + d = firstPosSince(i, \pi_j^i, ' -', \pi_j))\end{aligned} \quad (13)$$

其中, $firstPosSince(i, \pi_j^i, ' -', \pi_j)$ 表示在 π_j 中, 从 $i+1$ 位置开始第一个有效字符的位置, 它包括两种情况: a) 第一个不等于 π_j^i 并且不等于 "–" 的位置; b) 第一个等于 π_j^i 但是与 π_j^i 之间存在 "–" 相隔字符的位置.

4) 在 $\pi_1, \cdots, \pi_N$ 中, 字符之间的最大间距 $\max Dist_j$ $(1 \le j \le N)$

字符占用的宽度是有限的, 两个相邻字符之间间隔的时间步数越长, 输出结果越不可靠. 对于 π_j, $\max Dist_j$ 的计算公式为

$$\begin{aligned}\max Dist_j = {} & \\ & \arg\max_d(\pi_j^i \pi_j^{i+1} \cdots \pi_j^{i+d} | \pi_j^i \,!= \,' -' \text{ and} \\ & \pi_j^{i+d} \,!=' -' \text{ and} \\ & \pi_j^{i-1} \,!= \pi_j^i \text{ and} \\ & i + d = firstPosSince(i, \pi_j^i, ' -', \pi_j))\end{aligned} \quad (14)$$

5) 在 $\pi_1, \cdots, \pi_N$ 中, 第一个字符的首次出现位置 $firstCharPos_j$ $(1 \le j \le N)$

使用 $firstCharPos_j$ 主要是为了防止 RNN 在第一个字符处发生识别错误. 图像经过空白裁剪后, 第一个观测值 O_1 就属于第一个字符的一部分, 因此如果某个路径经历很长的时间后都没有将第一个字符输出, 那么可以相信在该路径中第一个字符识别不正确, $firstCharPos_j$ 的计算公式为

$$\begin{aligned}firstCharPos_j = {} & \\ & \arg\min_d(\pi_j^1 \pi_j^2 \cdots \pi_j^d | \pi_j^d \,!= \,' -') = \\ & firstPosSince(1, ' -', ' -', \pi_j)\end{aligned} \quad (15)$$

6) 在 $\pi_1, \cdots, \pi_N$ 中, 最后一个字符的首次出现位置 $lastCharPos_j$ $(1 \le j \le N)$

使用 $lastCharPos_j$ 是为了防止路径在最后一个字符处发生识别错误. 如果最后一个字符出现的越靠前, 则路径错误的可能性越大, $lastCharPos_j$ 的计算公式为

$$lastCharPos_j =$$

$$\begin{cases} T, & \text{若 } \pi_j^T != \text{'}-\text{'} \text{ and } \pi_j^T = \pi_j^{T-1} \\ \arg\min\limits_d \{\pi_j^d \cdots \pi_j^T | \pi_j^d != \text{'}-\text{'} \text{ and} \\ \quad \pi_j^d != \pi_j^{d-1} \text{ and} \\ \quad \exists k(\pi_j^d = \pi_j^{d+1} = \cdots = \pi_j^k \text{ and} \\ \quad \pi_j^{k+1} = \cdots = \pi_j^T = \text{'}-\text{'} \text{ and} \\ \quad d \le k \le T)\}, & \text{其他} \end{cases} \tag{16}$$

7) 在 $\pi_1, \cdots, \pi_N$ 中, 相似字符的对数 $simiCharCount_j$ $(1 \le j \le N)$.

如果在 π_j 中存在两个字符 π_j^i 和 π_j^k 满足如下条件, 则称它们是一对相似字符:

$$\begin{cases} k - i = 2 \\ \pi_j^i = \pi_j^k \text{ and } \pi_j^i != \pi_j^{i+1} \end{cases} \tag{17}$$

$simiCharCount_j$ 指的是 π_j 中相似字符的对数. 使用该特征的原因在于通过实验发现, RNN 输出错误在字符中间产生的概率比较大, 例如对于输出 "−A − A−", 两个 A 之间的空格一般是错误的, 它是 A 的可能性更大.

8) $word_1, \cdots, word_N$ 中宽字符的个数 $wideCharCount_j$ $(1 \le j \le N)$.

这里指的是 $word_j$ 中包含 "W", "w", "M", "m" 4 个字符的个数, 这些字符的宽度要比其他字符大. $wideCharCount_j$ 主要是为了避免过分拟合特征 2). $word_j$ 包含的长字符越多, 则即使它的长度较短, 仍然可能会对应一个较长的输入序列. 设 $word_j^i$ 表示 $word_j$ 中的第 i 个字符, 则 $wideCharCount_j$ 的计算公式为

$$\begin{aligned} &wideCharCount_j = \\ &\quad \text{count}(word_j^i = \text{'}W\text{'} \parallel word_j^i = \text{'}w\text{'} \\ &\quad \parallel word_j^i = \text{'}M\text{'} \parallel word_j^i = \text{'}m\text{'}, \\ &\quad 1 \le i \le |word_j|) \ (1 \le j \le N) \end{aligned} \tag{18}$$

9) 在 $word_1$ 对应的路径中, T 个时刻对应的最小置信度 $\min Conf$.

在每个时刻 t, 每个输出端 k 上都有输出, 其数值大小代表时刻 t 输出字符 k 的概率, $word_1$ 对应的路径是通过在每个时刻 t 选择概率最大的输出端而形成的, 但是如果在某个时刻, 次优输出端和最优输出端之间的概率相差不大, 那么在这个时刻选择次优输出端也是很有可能的, 因此定义 $\min Conf$:

$$\min Conf = \arg\min_t \left(\frac{y_{\max}^t}{y_{\text{sec}}^t} \right), \ 1 \le t \le T \tag{19}$$

其中, $y_{\max}^t$ 为时刻 t 最优输出端的概率, y_{sec}^t 为时刻 t 次优输出端的概率. 分析可知 $\min Conf$ 的最小值为 1, $\min Conf$ 越小, 则 $word_1$ 结果越不可靠.

分析可知, 新的拒识特征为 $8N + 2$ 维.

特征提取以后, 使用式 (20) 进行归一化:

$$f_i = \frac{f_i - u_i}{\sigma_i} \tag{20}$$

其中, f_i 表示第 i 个特征值, u_i, σ_i 分别表示训练集上该特征的均值和方差.

3.2　拒识特征的降维分析

新的拒识特征为 $8N + 2$ 维, N 越大, 维数越高, 由此可能带来维数灾难, 因此考虑使用降维方法是必要的. 降维的过程本身也属于特征生成的过程, 好的降维方法能够生成更有效的特征, 从而提高识别率.

PCA 是一种常见的降维方法, 被广泛地应用到人脸识别等领域[20−21], 但是实际中很多实验指出 PCA 在降低数据维度的同时, 也经常导致识别率的降低[22]. 原因在于 PCA 主要是基于样本集总体散度最大化准则来进行的, 而总体散度由类内散度和类间散度组成, 最大化总体散度不仅使类间散度最大化, 还有可能使得类内散度趋于最大, 而增大类内散度将必然增加样本集类别划分的难度. 鉴于此, 本文不采用 PCA 方法.

LDA 是另外一种常用的降维手段, 与 PCA 不同, LDA 寻找最能区分各类样本的方向, 经过 LDA 变换后, 获得的新特征将更具区分性, 从而能够提高样本的识别率, 得到了广泛的应用[23]. 但是 LDA 对于本文的拒识特征存在一个严重的问题. LDA 处理以后的最高维度为 $C - 1$, 其中 C 为样本类别数. 而拒识问题属于 0-1 分类问题, 因此如果使用 LDA, 那么得到的新特征将为 1 维, 而这 1 维数据用于拒识判断是不够的. 因此我们使用 Duin 等提出的一种改进型 LDA 方法[24], 不妨将其指代为 LoogLDA. 通过该方法得到的样本最高维度为 $n - 1$, 其中 n 为样本的原维度.

LoogLDA 的基本思想是在类间散度矩阵 S_B 中使用 Chernoff 距离代替 LDA 中的欧式距离. 两个概率密度函数 d_1, d_2 之间的 Chernoff 距离 ∂_c 定义为

$$\partial_c = -\log \int d_1^a(x) d_2^{1-a}(x) \mathrm{d}x \tag{21}$$

其中, a 是一个属于 $(0, 1)$ 之间的常量.

在两类的情况下, LoogLDA 的目标为使以下函数值最大:

$$J_c(A) = \mathrm{tr}((AS_WA^{\mathrm{T}})^{-1}(p_1p_2A(m_1-m_2)\times (m_1-m_2)^{\mathrm{T}}A^{\mathrm{T}}- AS_w^{0.5}(p_1\log(S_w^{-0.5}S_1S_w^{-0.5}))+ p_2\log(S_w^{-0.5}S_2S_w^{-0.5}))S_w^{0.5}A^{\mathrm{T}})) \quad (22)$$

其中, A 是一个待求的 $d\times n$ 维矩阵, n 是原数据维度, d 是用户指定的新维度. S_w 是类内散度, m_1 和 m_2 分别是两个类的类均值向量, S_1 和 S_2 是两个类的散度, p_1 和 p_2 是两个类的先验概率. 求出 A 以后, 即可使用它将原样本从 n 维空间变换到 d 维空间.

3.3 SVM 拒识模块的训练和识别

通过 RNN 的训练集训练 SVM 拒识模块是不合适的, 它带来的一个问题就是类别数严重不平衡, 因此使用 RNN 的验证集对 SVM 进行训练. SVM 训练和识别时, 使用基于概率的方法, 这样它可以给出类别对应的概率.

训练 SVM 拒识模块的主要流程为:

1) 在验证集的第 j 个样本 V_j 上提取拒识特征向量 $\boldsymbol{F}_j=(f_1^j,f_2^j,\cdots,f_{8N+2}^j)$, 对应的类别为 C_j, 当 RNN 正确识别 V_j 时 $C_j=1$, 否则 $C_j=0$, 则 $\langle \boldsymbol{F}_j, C_j\rangle$ 构成一个拒识训练样本. 设验证集为 Val, 则可以提取 $l=|Val|$ 个拒识特征向量, 形成一个 $l\times n$ $(n=8N+2)$ 的矩阵 Tra, 其对应的类标签为 $traLabel=(C_1,C_2,\cdots,C_l)^{\mathrm{T}}$;

2) 对 Tra 的各维属性进行归一化;

3) 使用 LoogLDA 在 Tra 上提取变换矩阵 $A(d\times n)$;

4) $Tra=Tra\cdot A^{\mathrm{T}}$;

5) 使用 Tra 和 $traLabel$ 训练 SVM.

使用 SVM 拒识模块进行识别的主要流程为:

1) 提取第 j 个测试样本的拒识特征向量 $\boldsymbol{F}_j$;

2) 使用验证集上各维属性的均值和方差对 $\boldsymbol{F}_j$ 进行归一化;

3) $\boldsymbol{F}_j=\boldsymbol{F}_j\cdot A^{\mathrm{T}}$;

4) 使用训练好的 SVM 对 $\boldsymbol{F}_j$ 进行识别. 设 SVM 输出 1 的概率为 y_1, 则当 y_1 满足式 (23) 时, 认为 RNN 识别正确, 否则认为识别错误, 应该拒识.

$$y_1 > Threshold \quad (23)$$

分析可知, 通过设置不同的拒识阈值 $Threshold$, 可以实现不同的拒识强度, 从而得到不同的可靠性.

4 实验和结果分析

实验的主要目的有: 1) 证明本文的识别算法能够识别高粘着型 CAPTCHA, 并且识别结果具有较高可靠性; 2) 证明本文提出的拒识方法相对于现有拒识方法在提高可靠性方面具有优势; 3) 证明经过 LoogLDA 降维能够进一步提高识别结果的可靠性.

4.1 实验数据

测试一个算法的性能最好是在一个公共数据集上进行. 目前对于手写识别有 IAMDB, CEDAR 等数据集, 但是对于 CAPTCHA 识别, 并不存在相应的数据集. 公布一个网站的 CAPTCHA 识别算法会给该网站带来安全风险, 也会给公布者带来法律问题, 因此, 与文献 [25] 相似, 本文使用 CAPTCHA 生成程序生成大量图片进行实验. 图 4 是本文识别的 CAPTCHA 示例, 从图中可以看出这类 CAPTCHA 有如下特点: 开放字典, 字符粘着紧密, 存在字符变形, 字体大小不一致, 每个 CAPTCHA 的字符数也不一致.

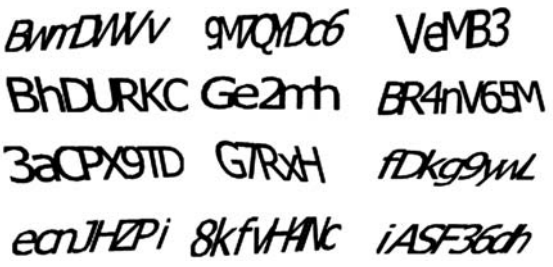


图 4 CAPTCHA 样本示例

Fig. 4 Samples of CAPTCHA in experiments

4.2 实验设置

实验中共采集到 3 000 张 CAPTCHA 图片, 将其分成 3 个不相交的数据集: 训练集为 1 000 张, 验证集为 500 张, 测试集为 1 500 张. 由于图像非常干净, 因此不需要进行去噪等预处理操作, 但需要使用式 (24) 对图像高度进行归一化:

$$\begin{cases} h\_new = h\_mean \\ w\_new = w\_old\cdot\dfrac{h\_new}{h\_old} \end{cases} \quad (24)$$

式中 w_new, h_new 分别表示新图像的宽度和高度, h_mean 是训练集图像的平均高度, h_old 是原图像的高度. 在高度上进行归一化主要是为了使所有的图像对 RNN 有相同的输入维度.

使用滑动窗口机制进行特征提取, 窗口的宽度为 1, 窗口的高度为图像的高度, 直接使用图像的原始灰度值作为特征数据.

RNN 输入层的单元数为图像的高度, 隐层含 100 个 LSTM 单元, 每个 LSTM 中含一个细胞 (Cell), 输出层含 57 个输出单元. 训练时为了提高 RNN 的抗噪能力, 对训练数据加入 $\mu=0$, $\sigma=1$

的高斯噪声. 使用带有冲量项的随机梯度下降法进行训练, 学习速率为 1E−4, 冲量为 0.9. 每 10 epoch 检测 1 次验证集上的识别率, 如果在检测 10 次之后验证集上没有出现更好的结果, 则停止训练.

拒识模块中使用 libsvm[1] 进行训练和识别, 其参数使用 libsvm 的默认设置; 提取前 3 个候选词的拒识特征用于拒识判断.

4.3　不降维时的结果分析

表 1 是使用 4 种拒识算法后得出的 CAPTCHA 识别算法的识别率和可靠性. 其中方法 0 为本文提出的拒识方法, 方法 1 为基于归一化概率及差值的方法[18], 方法 2 为 avg_top[19], 方法 3 为 LDAM[12]. 表的横向为使用不同拒识阈值以后得到的识别率, 纵向表示在给定识别率下 CAPTCHA 识别算法的可靠性. 表中第一列对应拒识阈值为 0 的情况, 此时拒识模块不发生作用, 识别率和可靠性都为 0.55. 从表中可以看出如下几点:

1) 本文算法能够对开放字典下的粘着型 CAPTCHA 取得一定的识别率, 说明本文提出的识别算法是有效的.

2) 不使用拒识方法前, CAPTCHA 识别算法的可靠性为 0.55; 使用拒识方法 0 后, 可靠性得到了大幅提高. 当然识别率也有所降低, 但是降低幅度远小于可靠性的增加幅度, 在识别率 ≤ 0.41 时, 可靠性 ≥ 0.89, 说明本文提出的识别算法能够取得高可靠性.

3) 在相同的识别率下, 方法 0 能够取得最高的可靠性, 特别是在识别率较高时 (0.53 ∼ 0.43), 其优势尤其明显. 这说明本文提出的拒识方法相对于其他拒识方法具有明显的优势, 同时也说明在设计拒识算法时, 考虑输入输出以及分类器自身的特性对于提高算法的性能是非常有帮助的.

表 2 是各种拒识方法得到的拒识率和错误率。从该表同样可以看出: 在相同的拒识率下, 方法 0 具有最低的错误率, 也即具有最高的可靠性, 相对于其他方法具有优势. 经过测试, 在测试集上, 不使用拒识模块时识别单个 CAPTCHA 平均需要 0.362 s, 使用拒识模块以后平均需要 0.408 s, 识别时间增加幅度不大, 并且在可以接受的范围内.

4.4　降维后的结果分析

图 5 是在不同降维方法下, 使用本文拒识方法得出的识别率和可靠性之间的关系图. 其中方法 0 表示不进行降维, d-LoogLDA 表示使用 LoogLDA 并且将原数据降低到 d 维, 1LDA 表示使用 LDA 降维, 并且新维度为 1. 从图中可以看出, LDA 在 0.35 ∼ 0.52 之间的曲线要优于原曲线 (方法 0 的曲线), 但是在 0.52 ∼ 0.55 以及 0.10 ∼ 0.35 之间则显著低于原曲线. 对于 LoogLDA, 当维度为 2, 3, 4 时, 其曲线在绝大部分区间上均显著优于原曲线和 LDA 的曲线, 说明本文使用 LoogLDA 进行降维是正确有效的. 从图中还可以看出, 随着 d 的增加, 经过 LoogLDA 处理以后, 拒识算法的性能并不是也随之

表 1　各种拒识方法得到的识别率和可靠性

Table 1　Recognition rates and corresponding reliability under different rejection methods

识别率/可靠性	0.55	0.53	0.51	0.49	0.47	0.45	0.43	0.41	0.39	0.37	0.35
方法 0	0.55	0.71	0.77	0.80	0.84	0.85	0.88	0.89	0.91	0.91	0.92
方法 1	0.55	0.56	0.66	0.73	0.78	0.80	0.84	0.86	0.87	0.89	0.90
方法 2	0.55	0.65	0.69	0.73	0.76	0.80	0.82	0.85	0.88	0.89	0.90
方法 3	0.55	0.57	0.63	0.68	0.69	0.72	0.73	0.77	0.80	0.82	0.85

表 2　各种拒识方法得到的拒识率和错误率

Table 2　Word rejection rates and error rates under different rejection methods

拒识率/错误率	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70
方法 0	0.45	0.38	0.32	0.26	0.19	0.13	0.09	0.06
方法 1	0.45	0.41	0.35	0.28	0.24	0.15	0.11	0.08
方法 2	0.45	0.40	0.34	0.28	0.25	0.16	0.11	0.09
方法 3	0.45	0.41	0.36	0.32	0.27	0.21	0.14	0.12

[1]Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines [Online], available: http://www.csie.ntu.edu.tw/ cjlin/libsvm. 2011-03-02

增加. 3LoogLDA 的曲线略优于 2LoogLDA 的曲线, 但是 4LoogLDA 的曲线相对于 3LoogLDA 的优越性并不明显, 说明对于 LoogLDA 存在一个最佳的 d, 可以根据验证集确定 d.

总之, 通过对原特征进行 LoogLDA 处理, 拒识算法的性能得到了改善, 识别结果的可靠性得到了进一步的提高.

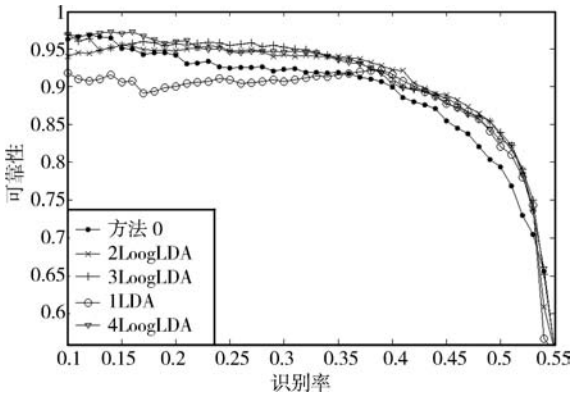


图 5 不同降维方法下识别率和可靠性之间的关系

Fig. 5 Relationships between recognition rate and reliability under different data dimension reduction methods

4.5 与现有识别方法的对比

我们选择了两种现有的 CAPTCHA 识别方法与本文的方法进行了对比实验. 方法 1 来自于文献 [6], 其识别流程为: 首先利用字符位置等信息将单字符完整分割出来, 然后使用模板匹配法进行识别. 方法 2 来自于文献 [7], 它首先使用 CFS (Color filling segmentation) 算法对 CAPTCHA 进行分割, 然后使用 SVM 进行识别. 实验发现, 方法 1 在本文数据集上的识别率为 0, 方法 2 的识别率则仅为 0.05, 而本文算法却取得了 0.55 的识别率.

图 6 和图 7 分别是文献 [6−7] 识别的 CAPTCHA. 从这些图可以看出, 这些 CAPTCHA 虽然存在很严重的噪声干扰或字符变形, 但字符之间基本不粘着, 因此可以找到一种较好的分割算法将这些图像中的字符串分割为单字符. 而单字符的识别已经不是一个难题 (识别率 > 0.95[7]), 因此这些文献在非粘着型 CAPTCHA 上取得了较高的识别率. 但对比本文的识别对象, 图 4 中字符不仅存在变形, 而且紧密粘着并发生一定的重叠, 很难正确地将这种 CAPTCHA 中的各个字符分割开来, 分割的失败导致这些识别方法难以取得成功. 本文算法避免了分割环节, 因此取得了一定的识别率, 并且由于使用了一种新的拒识算法, 识别结果能够取得很高的可靠性.

HMM 模型是目前英文手写识别领域进行隐式分割识别时使用的主流模型, 目前识别精度较高并应用较广的是字典驱动型 HMM (Lexicon-driven HMM, LDHMM)[14], 表 3 是该模型的实验结果[13]. 从表中可以看出, 当字典较小时, LDHMM 识别率高, 识别时间短, 但是随着字典的增大, 其识别率逐渐降低, 识别时间急剧变长, 字典为 80 K 时, 识别单个单词的时间为 14.46 s. 而本文识别的 CAPTCHA 的字典大小为 M^L, M 为字符的类别数, L 为 CAPTCHA 的长度. 实验中 M 为 56 (通过对 CAPTCHA 图片中字符进行统计而获得), $L = 5 \sim 8$, 因此字典大小 $> 56^8 = 96\,717\,311\,574\,016 \approx 96\,717$ G. 在如此庞大的字典下, 使用 LDHMM 不仅识别率不高, 而且识别时间过长, 但本文算法识别单个 CAPTCHA 所需的平均时间仅为 0.408 s, 并且取得了一定的识别率, 相对于 LDHMM 在时间上具有明显优势.

图 6 文献 [6] 识别的 CAPTCHA

Fig. 6 Samples of CAPTCHA recognized by [6]

图 7 文献 [7] 识别的 CAPTCHA

Fig. 7 Samples of CAPTCHA recognized by [7]

表 3 不同字典大小下 LDHMM 的识别率和识别时间

Table 3 Recognition rates and corresponding recognition time of LDHMM under different sizes of lexicon

字典大小	识别率	识别时间 (s/word)
10	0.98	0.010
1 000	0.91	0.273
10 000	0.81	1.992
40 000	0.73	7.516
80 000	0.68	14.46

4.6 对于加强 CAPTCHA 安全性的启示

CAPTCHA 的 “看不清, 换一张” 的功能本来是为了方便自然人而开发的, 但是本文的研究结果表明这种功能实际上也是一个安全漏洞. 本文提出的识别算法可以在一定的识别率时取得很高的可靠性, 可以在发生拒识时, 模拟人的行为通过 “换一张” 的功能选择另外一张 CAPTCHA 进行识别, 使得针对 CAPTCHA 的识别过程更加隐秘 (不容易

被网站管理人员发现), 并且对依赖于 CAPTCHA 识别的外部 Web 程序而言, 发生拒识时的影响小于识别错误时的影响. 加强 CAPTCHA 生成算法的安全性, 使之产生的 CAPTCHA 被机器自动识别的难度非常大, 并且不会产生忽难忽易的现象 (即让 CAPTCHA 识别程序换不到容易的 CAPTCHA) 是对抗本文识别算法的措施之一.

当然, 简单去掉 "换一张" 功能也是一种解决方法, 但是这就要求 CAPTCHA 生成算法每次都能生成容易被人识别但很难被机器识别的 CAPTCHA, 这通常是一个很难达到的要求.

除文字型外, CAPTCHA 还有图像理解型、声音识别型、手机验证型等多种形式. 其中图像理解型已经在某方面显示出明显的优势[26]. 因此充分发掘自然人与机器人之间的真正差距, 设计出更好更安全的非文字型 CAPTCHA 也是解决问题的重要途径.

5 结束语

CAPTCHA 是人工智能在互联网安全领域的一次大规模应用. 研究 CAPTCHA 识别技术有助于发现 CAPTCHA 自身的缺陷, 从而间接推动 CAPTCHA 技术的发展. 本文研究的主要贡献为:

1) 提出了一种识别新算法, 可以识别高粘着 CAPTCHA, 并且结果具有高可靠性;

2) 提出了一种根据分类器具体特性进行拒识的新算法, 相对于现有的单纯基于后验概率的拒识算法, 这种算法能够在相同的拒识率下取得更低的错误率和更高的可靠性;

3) 对提取的拒识特征使用 LoogLDA 进行处理, 从而进一步提高了 CAPTCHA 识别结果的可靠性.

此外, 本文还结合实验指出了 "换一张" 的功能实际上是 CAPTCHA 的一个安全漏洞, 并指出了几种弥补这个漏洞的方法. 本文的识别算法和拒识算法除了用于 CAPTCHA 识别, 还可以用于语音识别, 在线手写识别, 脱机手写识别等许多领域.

但是, 正如本文中指出的那样, 提高可靠性的方法除拒识外, 还可以从提高 RNN 识别率这方面进行着手, 虽然一般来说这样做是很困难的, 但这也是一个重要的并且非常有意义的研究方向, 目前我们正在进行相应的研究工作.

References

1 Steiner P. On the Internet, nobody knows you′re a dog. *The New Yorker*, 1993, **69**(20): 61−61

2 Ahn L V, Maurer B, McMillen C, Abraham D, Blum M. reCAPTCHA: human-based character recognition via web security measures. *Science*, 2008, **321**(5895): 1465−1468

3 Rusu A, Thomas A, Govindaraju V. Generation and use of handwritten CAPTCHAs. *International Journal on Document Analysis and Recognition*, 2010, **13**(1): 49−64

4 Egele M, Bilge L, Kirda E, Antipolis S, Kruegel C. CAPTCHA smuggling: hijacking web browsing sessions to create CAPTCHA farms. In: Proceedings of the ACM Symposium on Applied Computing. New York, USA: ACM, 2010. 1865−1870

5 Soupionis Y, Gritzalis D. Audio CAPTCHA: existing solutions assessment and a new implementation for VoIP telephony. *Computers and Security*, 2010, **29**(5): 603−618

6 Hocevar S. PWNTCHA-CAPTCHA decoder [Online], available: http://caca.zoy.org/wiki/PWNtcha, March 2, 2011

7 Yan J, Ahmad A S E. A low-cost attack on a Microsoft CAPTCHA. In: Proceedings of the 15th ACM Conference on Computer and Communications Security. New York, USA: ACM, 2008. 543−554

8 Zhang J S, Wang XF. Breaking internet banking CAPTCHA based on instance learning. In: Proceedings of the International Symposium on Computational Intelligence and Design. Hangzhou, China: IEEE, 2010. 39−43

9 Li Yong-Shun. Research on the Cooperation of Soft Computing and Its Application in Captcha Attack [Master dissertation], Anhui University, China, 2010
(李永顺. 软计算融合算法及其在 Captcha 识别方面的应用研究 [硕士学位论文], 安徽大学, 中国, 2010)

10 Li Ying. Investigation on Generation and Recognition of Verification Code [Master dissertation], Nanjing University of Science and Technology, China, 2008
(李颖. Web 验证码的生成与识别 [硕士学位论文], 南京理工大学, 中国, 2008)

11 Google′s CAPTCHA busted in recent spammer tactics [Online], available: http://securitylabs.websense.com/content/Blogs/2919.aspx, March 2, 2011

12 He C L, Lam L, Suen C Y. A novel rejection measurement in handwritten numeral recognition based on linear discriminant analysis. In: Proceedings of the 10th International Conference on Document Analysis and Recognition. Washington D. C., USA: IEEE, 2009. 451−455

13 Koerich A L, Sabourin R, Suen C Y. Recognition and verification of unconstrained handwritten words. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(10): 1509−1522

14 Plotz T, Fink G A. Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition*, 2009, **12**(4): 269−298

15 Graves A. Supervised Sequence Labelling with Recurrent Neural Networks [Ph. D. dissertation], The Technical University of Munich, Germany, 2008

16 Graves A, Liwicki M, Fernandez S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(5): 855−868

17 Zhang Liang, Huang Shu-Guang, Shi Zhao-Xiang, Hu Rong-Gui. CAPTCHA recognition method based on RNN of LSTM. *Pattern Recognition and Artificial Intelligence*, 2011, **24**(1): 40−47
(张亮, 黄曙光, 石昭祥, 胡荣贵. 基于 LSTM 型 RNN 的 CAPTCHA 识别方法. 模式识别与人工智能, 2011, **24**(1): 40−47)

18 Cai Tie, Zhu Jie. Fast out-of-vocabulary rejection algorithm in ASR system. *Computer Engineering*, 2005, **31**(10): 22−24
(蔡铁, 朱杰. 自动语音识别系统中的 OOV 快速拒识算法. 计算机工程, 2005, **31**(10): 22−24)

19 Koerich A L. Rejection strategies for handwritten word recognition. In: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition. Washington D. C., USA: IEEE, 2004. 479−484

20 Li Xiao-Li, Da Fei-Peng. A rapid method for 3D face recognition based on rejection algorithm. *Acta Automatica Sinica*, 2010, **36**(1): 153−158
(李晓莉, 达飞鹏. 基于排除算法的快速三维人脸识别方法. 自动化学报, 2010, **36**(1): 153−158)

21 Zheng Yu-Hui, Sun Quan-Sen, Xia De-Shen. An efficient 2DPCA-based non-local means filter. *Acta Automatica Sinica*, 2010, **36**(10): 1379−1389
(郑钰辉, 孙权森, 夏德深. 基于 2DPCA 的有效非局部滤波方法. 自动化学报, 2010, **36**(10): 1379−1389)

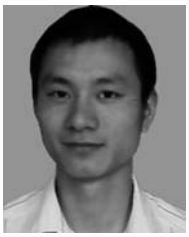
22 Gunter S. Multiple Classifier Systems in Offline Cursive Handwriting Recognition [Ph. D. dissertation], University of Bern, Switzerland, 2004

23 Cheng Zheng-Dong, Zhang Yu-Jin, Fan Xiang, Zhu Bin. Study on discriminant matrices of commonly-used Fisher discriminant functions. *Acta Automatica Sinica*, 2010, **36**(10): 1361−1370
(程正东, 章毓晋, 樊祥, 朱斌. 常用 Fisher 判别函数的判别矩阵研究. 自动化学报, 2010, **36**(10): 1361−1370)

24 Duin R P W, Loog M. Linear dimensionality reduction via a heteroscedastic extension of LDA: the chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, **26**(6): 732−739

25 Wachenfeld S, Klein H U, Jiang X Y. Recognition of screen-rendered text. In: Proceedings of the 18th International Conference on Pattern Recognition. Washington D. C., USA: IEEE, 2006. 1086−1089

26 Gossweiler R, Kamvar M, Baluja S. What′s up CAPTCHA? a CAPTCHA based on image orientation. In: Proceedings of the 18th International Conference on World Wide Web. New York, USA: ACM, 2009. 841−850



张 亮 电子工程学院博士. 主要研究方向为模式识别, 人工智能和信息安全技术. 本文通信作者.
E-mail: mathfun@163.com
(**ZHANG Liang** Ph. D. at Electronic Engineering Institute. His research interest covers pattern recognition, artificial intelligence, and information security. Corresponding author of this paper.)

张 亮 电子工程学院博士. 主要研究方向为 Web 信息处理和信息安全技术.
E-mail: liviocheung@sina.com
(**ZHANG Liang** Ph. D. at Electronic Engineering Institute. His research interest covers Web information processing and information security.)

黄曙光 电子工程学院教授. 主要研究方向为计算机应用技术和信息安全技术.
E-mail: sghuang20020505@sina.com
(**HUANG Shu-Guang** Professor at Electronic Engineering Institute. His research interest covers computer application and information security.)

石昭祥 电子工程学院教授. 主要研究方向为人工智能以及信息安全技术.
E-mail: shizx@gmail.com
(**SHI Zhao-Xiang** Professor at Electronic Engineering Institute. His research interest covers artificial intelligence and information security.)