基于属性网格计算的验证码识别研究

闫懋申,冯嘉礼
(上海海事大学 信息工程学院,上海 浦东 201306)

摘 要:文中旨在基于对属性网格计算的研究基础上,提出一种新的验证码识别研究方式,并通过实验验证该方法的可行性。方法如下:在验证码识别之前,需要先进行预处理,其中包括:图像的灰度化、二值化、去噪、分割以及归一化。在识别部分中,会用到属性网格计算器,这是一种基于定性映射的新型计算器。以对象的不同属性特征为维度,建立属性网格,并根据属性网格做出计算。首先对标准字符建立属性网格,然后在每次识别字符时,对识别字符的属性进行网格计算,并得出结果。实验结果成功率较高。该结果证明了该方法的可行性,同时表明了识别成功率与不同属性的权值分配有关。

关键词:验证码识别;网络安全;属性网格计算器;定性映射
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2013)01-0034-03
doi:10.3969/j.issn.1673-629X.2013.01.009

Study on Recognition of CAPTCHA Based on Attribute Grid Computing

YAN Mao-shen, FENG Jia-li
(College of Information and Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: A new method is expected for CAPTCHA recognition in this paper, which is based on research on attribute grid computing, and the experiments show this method is feasible. There are several steps in the preprocessing of CAPTCHA, which include transforming image to grayscale image, image binarization, image denoising, image segmentation and normalization. In recognition part, use attribute grid computer, which is a kind of new computer based on qualitative mapping, build a attribute grid for object with its attributes and compute according to it. In this paper, build a attribute grid for standard characters, then, once need to recognize character, this grid can be used for computing and getting the results. The success rates of experiments are high, which verified the feasibility of this method, and also veified that the success rates of recognition are related with the allocation of the weights among different attributes.

Key words: CAPTCHA recognition; Web security; attribute grid computer; qualitative mapping

0 引 言

验证码是当前的一种为对抗自动访问网站的机器人程序而设立的人机辨识机制。其一般实现方式为:将数字、字母或汉字以扭曲甚至粘连的方式写入图片文件中,并对文件进行背景加噪声操作,之后将图片文件显示在网页中,以供网页浏览者使用。验证码的特点是:由于其扭曲性及背景噪声,一般程序很难进行内容识别,而人脑的思考机制却可以凭借其对以往所了解的字符信息进行特征比对,轻易得到验证码的正确内容。基于此种机制,验证码很好地做到了对人与机器区分。

图 1 为验证码示例。

图 1 验证码示例(从某论坛下载的 5 张验证码)

事实上,验证码是可以通过程序来进行识别的,目前国内外已有很多学者在验证码识别方面进行了研究,且已经取得了不同程度的进展。步骤上,验证码的识别可以分为以下几个部分:预处理、特征提取和分类识别。

但具体的方式是多种多样的,具体效果方面也是各有所长,如:A. A. Chandavale 等人提出的基于字符特征值的方法[1]用于对简单验证码进行识别,贺强、晏立提出的基于形状上下文的方法[2]用于对复杂验证码的轮廓分析,贾磊磊等提出的基于 BP 神经网络和基于支持向量机方法[3]用于在学习中提高识别率。文中

收稿日期:2012-04-15;**修回日期**:2012-07-25
基金项目:国家自然科学基金资助项目(60075016)
作者简介:闫懋申(1983-),男,山东嘉祥人,硕士研究生,主要从事模式识别与智能计算方面的研究;冯嘉礼,博士,教授,博士生导师,主要从事人工智能方面的研究,在我国首次提出了人脑思维的"属性论"的观点。

提出了一种基于属性网格计算的验证码识别方法。

事物在与其他事物发生相互关系时表现出来的质叫作属性。事物属性是携带事物运动变化信息的载体,因此,通过属性及其变化规律的研究,可以从中破译出事物本身的运动变化规律。而属性网格正是建立在属性划分的基础上的,以多维网格的区域判定来进行定性评估的一套多维模型[4,5]。

图 2 为属性网格计算器示意图。

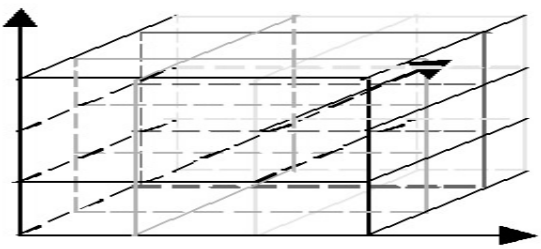

图 2　属性网格计算器示意图

以血三脂为例,可通过建立属性网格示意图来评判患者的血三脂是否在合理范围,并进一步判定患者哪些指标可能不在合理区域内[6],以便对进一步的诊断做出智能预估。

类似地,本中将数字、字母的特征建立属性网格,并对检验到的图片信息进行属性网格判定,以判断图片内容特征与哪个字符的属性网格最近似,以之得出图片内容信息。具体做法是:

(1)对可能用到的标准字符(如:数字、大小写字母)进行特征提取;

(2)以提取到的特征值做加权运算,建立属性网格;

(3)对验证码的字符进行特征提取;

(4)将验证码字符的特征值与属性网格进行加权绝对差值运算;

(5)依照④的结果进行字符识别判断。

1　图像的预处理

图像预处理的过程一般包括以下几步。

1.1　图像灰度化

在验证码识别中,颜色信息是无用处的,只需要黑白二值图像即可。为了得到二值图像,先把彩色图像转化为灰度图像。

设某像素点的红、绿、蓝色的色值分别为 R、G、B,则该点的灰度值计算公式为:

灰度值 $=R\times0.299+G\times0.587+B\times0.114$。

1.2　图像二值化

由上面灰度值图像可知,背景噪声仍然存在,而且256 级别的灰度值对下一步的处理而言仍是比较麻烦的。因而继续处理,对其做二值化运算。

二值化所得图像如图 3 所示。

图 3　二值化运算所得图像

二值化过程中最重要的是阈值选择的问题[7],对于大多数情况下的验证码,适用于局部阈值法。在对噪声进行滤波操作之后,将图片在宽度上划分为 5 部分,在忽略纯白色(背景色)像素点的前提下,选取每个部分中占像素点数最多的灰度值为阈值,以这些局部的阈值进行二值化。

1.3　图像分割

由于验证码是由一组字母和数字组成的,而识别需要一个一个来识别,因此首先要把图像分割为独立的字母或数字。这看上去简单,但实际上因为噪声的原因和数字字母的部分断裂,以及数字字母之间都可能部分相接,分割变得十分麻烦,只好根据特殊情况来特殊处理了。文中采用连通域算法进行字符分割[8]。

算法如下:

(1)初始化一个队列,队列节点有两个数据域,可分别用来存储行列坐标;

(2)以行优先的顺序遍历获得第一个值为 1 的点,将其入队列,并标记该点位置;

(3)取队头元素,检测其 8 邻域,若发现有未标记的值为 1 的点,则将其入队并标记该点位置;

(4)重复执行第③步,直到队列为空;

(5)检测已标记区域大小,若小于 20 个像素,则已标记区域只视为某字符的部分区域,跳至第②步,否则,第一个字符的连通域标记完成;

(6)同理可以对后面的字符进行连通域切割。

图 4 为第一字符切割效果示意图。

图 4　第一字符切割效果示意图

1.4　图像归一化

把每个数字或字母像素都尽可能饱满地放入一个 16×16 的数组中。需要图像缩放,缩放算法如下:

(1)设一个 16×16 的二维数组 Image[16][16];

(2)计算原字符的宽高,设为 x,y;

(3)设抽行间隔 xTemp,yTemp,则:xTemp $= x/(x-16)$,yTemp $= y/(y-16)$;

(4)读取像素点信息,若为 0,则对应数组位置的值置为 1,若为 255,则对应值置为 0,其间,每个 xTemp

行则忽略一行像素,同样,每隔 yTemp 列也忽略一列像素。

2 属性网格建立及字符识别

2.1 属性网格建立

选取全部大小写英文字母以及 0 ~ 9 的数字的标准字符,并根据标准字符的以下几个属性特征建立属性网格:直线数、环数、拐点数以及八线法交点情况。

图 5 为细化效果示意图。

图 5 细化效果示意图

(1)字符细化处理:为便于计算字符属性,需要先将字符进行细化处理[9]。主要思想为:逐步将非骨架部分的 1 变为 0。具体说来,对满足以下条件的点,将其变换成 0 点[10]:

条件 1 值为 1;

条件 2 其 4 邻域($x1$,$x3$,$x5$,$x7$)中至少有一个 0 点;

条件 3 其 8 邻域($x1$ ~ $x8$)中至少有两个以上的 1 点;

条件 4 8 联通连接数为 1;

条件 5 设 $x3$ 为 0,8 联通连接数仍为 1;

条件 6 设 $x5$ 为 0,8 联通连接数仍为 1。

(2)直线存在计算:对于连续 Num(Num ≥5)个点,取第 $i(i \in [0,\text{Num}-3])$ 个点与第 $i+3$ 个点连线,所连成的各线的角度差值小于 30 度。

(3)环存在计算:对于细化后的图像,若沿某点可以经不重复的路线回到原点,则记为存在环。

(4)拐点计算:拐点是指字符线的左右走向发生变化之处。计算连续点之间的斜率,斜率发生正负值变化处即为拐点。

(5)八线法计算交点:八线法是指把均匀引出 3 条横线、3 条垂线以及沿对角的顶点引出 2 条斜线,共计 8 条直线,然后分别计算出每条直线与字符的交点个数[11]。

计算交点的算法:

a. 竖线:在 i =4,7,10 时,使 j 从 0 到 14 递增,依次扫描 Image $[j][i]$,每次遇到 Image $[j][i]$ =0 而 Image $[j+1][i]$ =1,则交点值加 1;

b. 横线:与竖线类似,在 i =4,7,10 时,使 j 从 0 到 14 递增,依次扫描 Image $[i][j]$,每遇到 Image $[i][j]$ =0 而 Image $[i][j+1]$ =1,则交点值加 1;

c. 斜线:使 i 从 0 到 14 递增,依次扫描 Image $[i][i]$ 与 Image $[i][15-i]$,每次遇到 Image $[i][i]$ =0 而 Image $[i+1][i+1]$ =1,或 Image $[i][15-i]$ =0 而 Image $[i+1][15-(i+1)]$ =1,则交点值加 1。

对每一个标准字符进行以上 5 步运算,然后以每个字符的上面各步运算结果为对应的字符属性,建立属性网格。

2.2 待识别字符的属性网格计算

对每一个切割得到的待识别字符,分别按照 2.1 中的方法计算其:直线数、环数、拐点数以及八线法交点情况。

对待识别字符与每一个标准字符库的直线数、环数、拐点数以及八线法交点数做加权绝对差值的和值运算,即对四个属性分别赋以权值,再计算每项的绝对差值,并乘以对应权值,并将每项的结果求和。

选取总和值(即加权绝对差值之和[12])最小的字符作为识别结果。

3 实验结果及分析

分别以三种不同赋权值方式,对 40 张图片,共计 200 个字符进行计算,所得正确率情况如表 1:

表 1 权值分配与正确率对应表

直线数	环数	拐点数	八线法交点数	正确率 单位(%)
20	20	20	40	64.5
26	26	26	22	74
30	30	30	10	57

由以上数据可知:

(1)以属性网格计算的方式进行验证码字符识别是可行的;

(2)对同样属性网格下,权值分配方式会对识别结果有较大影响;

(3)八线法交点数对于总体的影响很大,但由于字符的倾斜性,八线法不可或缺又不能过于依赖;

(4)对于形近字符,如:1 与 1,8 与 B 等,为保证判断正确率,需做进一步的判断,着重比较其区别之处。

(5)与其他常见识别算法相比,在对较复杂验证码的识别方面,正确率已达到较高水平。以文中所提到的方法为例,已超过了文献[1]和文献[3]的正确率,但未达到文献[2]的正确率。

4 结束语

通过以上实验,可以看到以属性网格计算的方式来研究验证码识别是可行的,但当前的程度还不够成熟,仍需进行进一步的改善。

本方法当前存在的缺陷有:

(下转第 42 页)

到了多项式时间算法,也就证明了 NP 等于 P。而从历史上看,某个高难度的问题突然被人发现具有美妙的算法的事例是时有发生的。最典型的就是那两位印度科学家找到了关于质数判定的多项式时间算法,而这样的算法在他们之前一直被认为是不太可能的。

5 结束语

NP 问题是理论计算机领域最重要的问题之一,其相关概念原理相当复杂而难以理解,不少这方面的研究者包括一些基金及论文评审专家、科技期刊的掌控者,对其中一些重要概念的理解都很模糊。文中用通俗的语言,从多个方面论述了与 NP 问题相关的最核心概念,分析了不同角度的研究方法和研究途径,包括列举分析了一些失败或错误的研究方法以及对基本概念最常见的一些错误理解,同时对该问题的研究前景进行了分析预判,可供相关人员参考。

参考文献:

[1] Arora S, Barak B. Complexity Theory: A Modern Approach [M]. Cambridge: Cambridge University Press, 2009.

[2] Aaronson S. Is P versus NP formally independent? [J]. Bulletin of the European Association for Theoretical Computer Science, 2003, 81: 109-136.

[3] 杜丁柱,葛可一,王 洁. 计算复杂性导引[M]. 北京:高等教育出版社,2002:35-57.

[4] Sahni S. Data Structures, Algorithms and Applications in C++ [M]. [s. l.]: McGraw-Hill, 1998.

[5] Cook S A. The complexity of theorem proving procedures [C]//Proceedings of Third Annual ACM Symposium on Theory of Computing. New York: Association for Computing Machinery, 1971: 151-158.

[6] Karp R M. Reducibility among combinatorial problems[M]// Complexity of Computer Computations. New York: Plenum Press, 1972: 85-104.

[7] 杨正瓴. 密码学与非确定型图灵机[J]. 中国电子科学研究院学报,2008,3(6):558-562.

[8] 杨正瓴. 第二类计算机构想[J]. 中国电子科学研究院学报,2011,6(4):368-374.

[9] Yang Zhengling. A non-canonical example to support that P is not equal to NP[J]. Transactions of Tianjin University, 2011, 17(6): 446-449.

[10] Fortnow L. The Status of the P Versus NP Problem[J]. Communications of the ACM, 2010, 52(9): 78-86.

[11] Posa L. Hamiltonian circuits in random graphs[J]. Discrete Math., 1976, 14(4): 359-364.

[12] 俞 慧,吴 巍,黄 潇,等. 基于改进的蚁群算法的组播路由问题的研究[J]. 计算机技术与发展,2012,22(1):107-110.

[13] 胡 俊,洪 龙,沈春来. 分布式系统节点负载动态平衡策略研究[J]. 计算机技术与发展,2012,22(2):93-95.

[14] 张建成,宋丽华,鹿全礼,等. 云计算方案分析研究[J]. 计算机技术与发展,2012,22(1):165-167.

[15] 郭 怡,茅 苏. 云计算下基于 CRP 算法的资源提供策略[J]. 计算机技术与发展,2012,22(2):80-84.

(上接第 36 页)

①处理步骤较复杂,耗时较长,在 1.66GHz 的 CPU,2G 内存的机器上处理一个字符耗时近 1s;

②正确率仍不够高。

下一步计划的改进方向为:

①优化算法效率;

②引入机器学习机制,增强程序对特殊样本空间的适应性;

③针对易混淆字符的易混部分,进行着重比较。

参考文献:

[1] Chandavale A A, Sapkal A M, Jalnekar R M. Algorithm to Break Visual CAPTCHA[C]//International Conference on Emerging Trends in Engineering and Technology. Nagpur, Maharashtra, India: [s. n.], 2009: 258-262.

[2] 贺 强,晏 立. 基于形状上下文的复杂验证码识别算法[J]. 计算机工程,2011(2):200-202.

[3] 贾磊磊,陈锡华,熊 川. 验证码的模糊识别[J]. 西昌学院院报(自然科学版),2010(1):60-62.

[4] Feng Jiali. Attribute network computing based on qualitative mapping and its application in pattern recognition[J]. Journal of Intelligent & Fuzzy Systems, 2008, 19(1): 243-258.

[5] 许广林,冯嘉礼,刘永昌. 基于属性计算网络的模式识别[J]. 计算机科学,2008(4):200-202.

[6] 石伟峰. 智能计算在中医诊断系统中的应用研究[J]. 电脑知识与技术,2009(5):3472-3473.

[7] 吴 忠,朱国龙,黄葛峰,等. 基于图像识别技术的手写数字识别方法[J]. 计算机技术与发展,2011,21(12):48-51.

[8] 许 明. 验证码的识别与反识别[D]. 南京:南京理工大学,2007.

[9] 张 充,史青宣,苗秀芬,等. 基于 BP 神经网络的手写体数字识别[J]. 计算机技术与发展,2008,18(6):128-130.

[10] 安 然,张少军,陈 华,等. 字符识别中毛刺的去除方法[J]. 计算机技术与发展,2007,17(9):136-138.

[11] 张东辉. 数字验证码识别技术初探[J]. 黑客防线,2008(5):119-122.

[12] Lu Gang, Hao Ping. Recognition of CAPTCHA Based on Weighted Template Matching and Supervised Learning[J]. Computer and Modernization, 2010(12): 40-43.

基于属性网格计算的验证码识别研究

作者： 闫懋申， 冯嘉礼
作者单位： 上海海事大学 信息工程学院,上海 浦东 201306
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(1)