|  |  |
| --- | --- |
| 项目编号 | KYCX140111B |
| 项目类型 | 创新训练 |



**大学生创新创业训练计划**

**结题报告书**

项 目 名 称 纸币图像字符识别算法研究

项 目 级 别 □国家级 □省级 🗹校级

项 目 负 责 人 占必强

负责人所在学院 电子信息工程学院

联 系 电 话 15071440565

创新创业中心制表

2015年3月

* 1. 基本情况

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | | 纸币图像字符识别算法研究 | | | 项目类型 | 创新训练  创业训练  创业实践 | |
| 成果形式 | | 程序+专利 | | 立项时间 | 2014 年 4 月 | | |
| 完成时间 | | 2015年4月 | | 验收时间 | 2015 年4月 | | |
| 项  目  主  要  研  究  人  员 | 序号 | 姓 名 | 学号 | 联系电话 | E-MAIL | | 项目分工 |
| 1 | 占必强 | 11111027 | 15071440565 | 2583105532@qq.com | | 编码、算法 |
| 2 | 蒋祥祥 | 11112026 | 18171264786 | 1103316911@qq.com | | 编码 |
| 3 | 廖填兴 | 11071302 | 13007125698 | 421130992  @qq.com | | 数学算法 |

* 1. 项目执行情况简介

|  |
| --- |
| **1. 研究目的及意义**  纸币作为一种货币，是国家金融的命脉，现代经济生活中必不可少的东  西。随着中国经济的繁荣，纸币的大量使用增加了它的发行量。随着这一情况的发生，一些不法分子也从中找到机会，导致大量假币在市场上流通。市面上的点钞机不能识别出全部的假币（比如对于检测近几年出现的“HD90”这种超级假币），也不提供自动识别纸币冠号的功能，传统方法大多采用手工抄号的方法辅助识别点钞机没有识别出来的假币。如果在点钞机上增加号码识别功能，就能够在点钞的过程中实时地识别纸币的冠号，并根据纸币冠号的唯一性，辅助假币的检测。当出现重复的冠号或者未使用的冠号时，就可以认为存在假币。在点钞过程中，不但可以动态识别纸币冠号，而且还能够将识别的冠号存储起来，用来取代手工抄号，不但减少人力物力的浪费、降低成本，而且也提高了效率，适应社会经济发展的需求。  正是由于纸币冠号的识别具有方便纸币管理和辅助假币识别这两个特点，而且新国标对点钞机有了号码识别功能的要求，带有号码识别功能的点钞机在不久的将来会有着巨大的实用价值和广泛的应用前景。  本课题根据市场需求，研究开发出一套针对第5版人民币的纸币序列号识别程序，其采用数字图像处理技术和模式识别技术，实现了纸币冠号的动态识别。通过仿真结果来看，该算法具有较高的识别率。  **2. 研究内容及成果**  **2.1 算法总体设计**  整个算法识别包括预处理和识别两大部分。图像预处理是完成后续识别工作的重要前提，包括边缘检测、倾斜校正、灰度图像二值化、平滑处理、单字符切割等。识别算法有模板匹配、支持向量机算法。本文将模板匹配和支持向量机这两种算法相结合，使用复合识别算法，利用这两种算法各自的优点，提高算法整体性能。  整个号码识别过程包括：纸币的边缘检测，图像倾斜校正，号码区的判定，灰度图像二值化，平滑处理，单字符切割，单字符归一化，特征提取，模板匹配和支持向量机模块。号码识别算法系统结构图如下：    图1 算法系统结构图  **2.2 分模块设计**  如2.1算法总体设计中所讲，整个识别算法分为预处理和字符识别两大部分，每一部分中又分为多个模块。预处理部分包含纸币图像的倾斜校正、纸币号码区的判定、灰度图像二值化、平滑处理和单字符切割、单字符归一化这几个独立的功能模块。字符识别部分包含特征提取、模板匹配和后验的SVM这三个独立的功能模块。各个模块相对独立，且相邻模块之间存在服务关系，所有模块组合在一起实现了纸币号码识别功能。  **2.2.1 预处理模块设计**  样本图像是一张带有背景的完整纸币灰度图像，纸币的冠字号码只占其中很小一部分。要想识别纸币的冠号，必须首先将冠号（前景图像）从带有背景的整张纸币图像（背景图像）当中提取出来。由于票面情况复杂，多有褶皱、污染，加之图像传感器自身在扫描图像时也会带入一定量的噪声，不能直接对采集的图像进行冠号识别。因此，在识别之前要对采集到的图像进行预处理，得到易于识别的字符图像。预处理效果的好坏直接影响后续字符识别算法的性能，想要得到一个识别率较高的号码识别算法就必须提高预处理阶段的性能。  由算法整体设计中所述，预处理阶段包括纸币的边缘检测、图像倾斜校正、号码区的判定、灰度图像二值化、单字符切割和归一化，具体结构图如下：    图2 预处理模块结构图1    图3 预处理模块结构图2  **2.2.2 识别模块设计**  在字符识别中，通过对模板匹配算法和SVM做理论方针分析，结果发现单独使用任何一种识别算法识别率都不能满足系统的要求。模板匹配算法稳定，但时间复杂度较高，且对形状相近的字符（如“2”和“Z”）识别时错误率较高。支持向量机SVM对两类问题识别率高，但对多分类问题则效果很差。  本算法结合模板匹配算法和支持向量机各自的优点，提出一种基于模板匹配和支持向量机相结合的复合号码识别算法。首先用主成分分析PCA对原字符降维，提取主要的特征，接着进行模板匹配识别，分别计算相关和欧式距离，将这两种指标的识别结果作比较，如果相同就作为最终的识别结果，若不同，则将这两种不同的识别结果作为支持向量机SVM的训练样本，进行训练，计算分隔超平面，再输入原字符的特征，进行类判决，从而做出最终的识别判定。识别结构图如下：    图4 识别模块结构图1    图5 识别模块结构图2    图6 识别模块结构图3  **2.3 识别结果示例**  这里列举一张纸币样本进行仿真，以下是仿真各个阶段的截图：    从识别结果我们观察到第一个字符’O’ 识别错为’Q’，这里有两个原因，一是单字符归一化图像右下方有污迹噪声，干扰了识别；二是在制作模板过程中漏掉了这种形态的’O’，导致模板缺失这种形态的特征，比对出错。针对原因一，只有进一步优化去噪算法，减少噪声干扰；对原因二，我们应该训练更多的样本，从而获取更多特征，保证匹配准确。 |

* 1. 研究总结报告

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1. **研究中遇到的困难**   本课题主要需要图像处理和模式识别方面的知识，然而我们并未系统的学习过相关课程，因此整个研究过程几乎是从头学起。首先，我们系统的自学了图像处理方面的知识，如图像去噪、均衡化、图像分割、平滑、连接断点等等，有了这些基本功后，我们开始研究课题该怎么做，期间我们查阅大量论文，借鉴前人的研究方案，启发自己，一步一步摸索出我们自己的方法。遇到的问题清单如下：   1. 倾斜矫正：角度的检测、形变矫正。 2. 单字符切分：准确切分的同时不受噪点干扰。 3. 字符归一化：因光照、污迹、印刷大小等原因，导致同一字符具有不同灰度特征，归一化后可以使相同字符的特征尽可能相似，从而提高识别准确度。归一化包括大小归一化、形态归一化等。 4. 支持向量机SVM的数学理论：凸优化、核函数。 5. 如何选择特征以区分相似的字符：如(Q、O、D、0)，(E、F)，(C、G)、(S、5、6)。 6. 算法的速度和识别准确率存在矛盾：例如模板匹配的准确率高但计算量也大，特征值法速度快但难以找到稳定的特征(如连通域个数、交点个数、局部纹理特征)，SVM对于两类问题的识别准确率高但对多类问题则力不从心。 7. **研究成果及创新点**   **1）预处理**  **a.**字符分割算法能判断较小的噪点和字符，只将字符分割出来，而舍弃噪点。  **b.**倾斜矫正算法能将线性形变图像很好地纠正成标准图像，为后面提取号码区提供了保证。该算法通过阈值设定先分割出纸币和背景的横向边缘，通过Hough变换计算纸币倾斜角度，旋转水平后再检测纸币和背景的纵向边缘，计算形变角度，然后逐像素点移位，最终将倾斜形变图像变换成标准图像。  **2) 准确率**  识别部分采用先模板匹配，将欧氏距离和相关度分别作为比对依据得出两种识别结果，再比较这两种识别结果异同。如果不同就将这两种结果作为SVM的输入，通过SVM的最终判段出该字符到底是这两种结果中的哪一种。SVM的识别结果作为最终结果。这种模板匹配和SVM相结合的复合识别方法能提高识别准确度。  **3) 识别速度**  **a.**支持向量机识别模块，预先制作了一张35x36的包元矩阵，每个包元是一个结构体，存放着两种字符组合PCA的特征向量、平均值、组合名、类别等信息，为后面SVM训练提供数据。因为36种字符两两组合数为36\*35/2 ，因此该包元矩阵只有一半有值，另一半为空，从而节省了空间。这张表是事先生成的，作为离线数据包存放，在识别字符时只需从表中检索到对应的组合，就可直接使用计算好的参数，而不必再次计算，这样大大节省了识别速度。  **b.**所有程序分为两部分，一部分是生成离线数据的程序，存放于un\_used/目录下，另一部分是识别过程用到的程序，存放于当前目录下。这样更节省CPU资源，提高了识别的速度。下面是所有程序名以及存放目录的截图：  **3.研究工作的不足**  由于项目时间短，自身知识积累不够，导致以下问题还没来得及解决：  **1) 预处理**  a.纸币面向问题未解决。  b.纸币污染、残缺严重，无法提取号码区。  c.预处理部分的倾斜矫正只能矫正线性形变图像，不能处理非线性形变，如褶皱、卷曲等情况。  **2) 特征提取**  本算法直接通过PCA提取主特征识别，未结合其他简易有效的特征，如闭环数、交汇点数、局部纹理(LBP)等特征。  **3） 识别**  a.由于试验样本不足，训练特征还不全面，导致遇到特殊形态的字符识别错误。  b.没统计识别准确率和耗时  **4）程序设计漏洞**  a.算法未设计防错检测处理，遇到特殊情况可能出错  b.研究过程中，没有详细的规划，导致项目进度不能保证；同时，没有注意撰写文档、流程图，导致事后调试程序、修改代码效率低下。  **4.研究的收获**  研究过程中，我有了很多数学上的收获，例如差值法归一化；Hough变化的思想；PCA中对矩阵的特征值有了形象的理解；SVM中对凸优化、集合间隔、函数间隔、核函数将低维空间中无法分类的特征映射到高维空间但仍在低维空间计算内积的思想；模式识别中贝叶斯决策、线性判别函数、优化算法（梯度下降法、牛顿法、lagrange乘子法）等等。带着课题背景和解决问题的目的，原本枯燥的数学变得生动有趣、实用。让我们体会到数学的强大威力。  除了数学和专业理论上的收获，我们还收获了许多实践经验。最初看论文，推导公式只是让我们从原理上了解了是什么，具体怎么做呢？我们还真有点棘手，我们通过在MATLAB上仿真逐渐学会了图像矩阵的操作(转制、重排、切割等等)，后期又在matlab file exchange 网站下载源代码，学习matlab 程序怎么写。从最初对矩阵操作一无所知、程序不分文件和函数到后来有了自己的函数命名、文件分类等编程习惯。最后，我还学会了怎么画算法结构图和程序流程图，让自己的项目管理能力有了提升。 | | | | | | | |
| 四、经费开支情况 | | | | | |
| 序号 | 开支项目 | | | 开支金额（元） | |
| 1 | 电脑配件 | | | 518 | |
| 2 | 图书 | | | 234.7 | |
| 3 |  | | |  | |
| 4 |  | | |  | |
| 5 |  | | |  | |
| 6 |  | | |  | |
| 7 |  | | |  | |
| 合计 | 742.7 | | | | |
| 五、成果名称及形式 | | | | | |
| 序号 | 成果项目 | | 成果形式 | | |
| 1 | 《人民币纸币图像序列号的组合识别方法》 | | 专利 | | |
| 2 | BankNoteOCR | | 源码 | | |
| 3 |  | |  | | |
| 指导老师意见 | **序号** | **评分标准** | | | **得分** |
| 1 | 项目执行情况（10分） | | |  |
| 2 | 项目总结（30分） | | |  |
| 3 | 项目成果（30分） | | |  |
| 4 | 经费开支使用（20分） | | |  |
| 5 | 项目记载记录情况（10分） | | |  |
|  | | | |  |
| 指导老师签名：  年 月 日 | | | | |
| 项目负责人所在学院意见 | 内容提示：学院专家组对结题的意见，包括对项目研究工作和研究成果的评价等。  单位公章  单位负责人签字：  年 月 日 | | | | |
| 学校审核意见 | 公 章  年 月 日 | | | | |

注：填写此表时，不要任意改变栏目和规格，内容简明扼要。如因篇幅原因需对表格进行调整，应当以“整页设计”为原则。用A4纸双面打印装订。