

# 2017 年全国大学生信息安全大赛

## 参赛作品简介

作品名称: 基于不可见文本水印的纸质文档泄密源头  
管理系统的设计与实现

组 长: 林晶

组 员: 李彦征, 郭楠楠, 赵重源

提交日期: 2017.5.27

电子邮箱: 939566372@qq.com

# 目录

摘要.....	4
1 作品概述.....	5
1.1 研究背景.....	5
1.2 产品功能.....	5
1.3 相关工作.....	6
1.4 产品特色.....	6
1.4.1 系统特色.....	6
1.4.2 核心算法特色.....	7
2 作品设计及实现.....	8
2.1 系统概述.....	8
2.2 系统流程图.....	9
2.3 功能模块.....	10
2.3.1 虚拟打印模块.....	10
2.3.2 水印嵌入模块.....	11
2.3.3 水印提取模块.....	11
2.4 实现原理.....	12
2.4.1 字符切分算法.....	13
2.4.2 图片预处理算法.....	13
2.4.3 水印嵌入算法.....	14
2.4.4 去噪算法.....	14
2.4.5 水印提取算法.....	15
3 作品测试与分析.....	16
3.1 测试目的.....	16
3.2 测试方案.....	16
3.3 测试设备.....	17
3.4 算法测试.....	17
3.4.1 打印扫描前后不变量测试.....	17
3.4.2 水印算法不可见性测试.....	18
3.4.3 测试不同字体对不可见性的影响.....	19
3.4.4 电子版水印提取率分析.....	19
3.4.5 不同字体对水印提取率的影响.....	20
3.4.6 扫描分辨率 DPI 对提取率的影响.....	20
3.4.7 步长 K 对提取率的影响.....	20
3.4.8 算法改进前后水印提取率的改变.....	21
3.6 功能测试.....	21
4 创新性说明.....	24
4.1 核心算法的改进.....	24
4.2 算法改进说明.....	25
4.2.1 字符切割算法.....	25
4.2.2 嵌入部分.....	26
4.2.3 水印提取.....	27
5 总结.....	28

---

参考文献.....	29
-----------	----

## 摘要

针对纸质涉密文件非法传播，涉密责任无从追溯。小组对元文法提出的基于字符翻转的抗打印扫描文本水印算法，进行改进与优化，实现基于文本水印的纸质文档泄密追踪系统。小组设计针对该水印算法的字符切分算法、添加预处理环节、增加去噪模块，以提高算法的鲁棒性。并进行大量实验测试，采集数据，分析各个物理量（打印机扫描仪型号、字体字号、打印扫描使用的分辨率等）对水印提取率的影响，找到最大误差区间，验证系统是否满足用户的需要。

**【关键词】** 数字水印 二值图像 字符切分 纸质文档 泄密追踪

# 1 作品概述

## 1.1 研究背景

随着信息化的发展和计算机的普及,电子文档成为人们日常办公的主要形式。但很多文档主要还是以纸质文档的形式存在。目前对于重要文件的打印主要采用限制打印,复印次数,或指定专人负责,打印文档时需要经过批准,审批,登记等步骤,并定期对纸质文档进行销毁或归档处理。当前通过加密,访问控制,身份认证等手段对电子文档进行保护的文档安全管理技术比较成熟,但电子文档打印之后就很难对其进行控制,信息泄露的风险加大。特别是关键文档的泄露,往往会对一个企业乃至国家带来极大的威胁和风险。

针对文本文档的抗打印扫描的数字文本水印技术能够将打印负责人的身份信息作为水印嵌入到将打印的纸质文件中,在文件泄露被捕获后通过提取水印信息即可对纸质文件泄密源头进行管控追踪。因此,研究此项技术有着迫切的需求与极大的价值。

虽然现在市面上已经出现了大量的水印嵌入的算法,但是嵌入水印的不可见性和鲁棒性并不是很理想。这一定程度上影响了嵌入信息的有效隐藏和正确提取。本系统基于元文法等提出的基于字符翻转的抗打印扫描文本水印算法,对字符分割、水印嵌入算法部分进行了优化。同时,增加了去噪算法,优化了水印提取的算法,大幅度提高了水印提取的正确率。

## 1.2 产品功能

该系统具有较高的普适性。适用于各种党政机关或商界传收纸质文件及纸质文本泄密溯源的场合。通过对纸质文档嵌入抗打印扫描的不可见文本水印信息(打印者的身份信息、打印日期、文档的密级),实现当截获非法传播的涉密文件时,可以通过水印准确追踪到何人何时打印了该文件,解决泄密源头追溯的问题;增加了纸质文件在传输过程中的安全性问题。

该系统具有很好的可操作性。用户登录系统后,选择需要打印的文件,按正常进行打印,无需进行额外的操作,系统自动截获需要打印的文件,对文件嵌入该用户的相关信息,然后发送给打印机,打印进行打印。

该系统具有较高的实用性。通过优化核心算法，在大量的实验测试保证水印提取率达到 95%，并加入校验码，进行译码自动纠错，使打印扫描后，通过该系统可以准确追溯泄密源头。

## 1.3 相关工作

虽然现在市面上已经出现了大量的水印嵌入的算法，但是嵌入水印的不可见性和鲁棒性并不是很理想。这一定程度上影响了嵌入信息的有效隐藏和正确提取。有时候，水印的嵌入还会引起原文的失真。现有的文本水印算法主要分为 5 大类：基于文档结构、基于自然语言处理技术、基于二值图像水印技术、基于字符几何特征以及基于底纹的方法。

基于文档结构的文本水印方法无法抵抗文本内容和格式变化攻击，隐藏的信息量少；基于自然语言处理技术的文本水印方法，但不适用于文本内容不宜更改的情况，从而不具有普适性；基于字符集合特征的文本水印方法，由于汉子等字符的字型结构特点以及字体风格的影响，字符修改后的世界效果不理想，同时该方法的实现需要特殊字符库的支持，这对该技术的应用带来了很大的不便；基于底纹的文本水印方法优点是水印信息容量大，缺点是浪费油墨，从而使打印成本升高，并且在一定程度上还会影响文档的正常阅读；基于二值图像水印技术的文本水印方法对于原件的复印件不能正常提取。

综上所述，现有的文本水印方法不能保持理想视觉效果的前提下，实现从文本打印原件和复印原件中全盲地提取水印信息。元文法提出了一种用于二值文本图像的数字水印算法，具有较好的不可见性，可以抵抗打印和复印的攻击，不依赖于语种和字体字号。但是元文法提出的基于字符翻转的抗打印扫描文本水印算法只提供了理论依据，在多次重复测试下，发现该算法水印提取率仅达到 85% 左右。而且没有进行实现，未给出最佳字符切分方法、编码校验以及最优步长的选取。

## 1.4 产品特点

### 1.4.1 系统特色

实现抗打印扫描的针对于纸质文档泄密追踪系统，泄密者在打印文档时，该

系统自动统嵌入打印者的用户名、文档密级以及打印日期，在追回泄密文档时，可对其进行扫描并提取水印，准确追踪到何人何时打印了该文档。

### 1.4.2 核心算法特色

但是元文法提出的基于字符翻转的抗打印扫描文本水印算法只提供了理论依据，在多次重复测试下，发现该算法水印提取率仅达到 85%左右。而且没有进行实现，未给出最佳字符切分方法、编码校验以及最优步长的选取。小组对元文法进行改进与优化。使水印提取率提高 10%左右，达到 95%。

1. 提取水印前进行去噪跟纠偏。打印扫描时不可避免的会有噪声、图片倾斜、字符粘连等问题。我们设计并实现去噪、纠偏等图像预处理技术,以满足实际应用的需要,提高系统的鲁棒性。
2. 设计一种针对该算法的字符切分算法,来确保打印扫描前后字符序列的一致性,保证水印提取的准确率。同时为了不可见性的需求,尽可能地在切分中去除类似“一”、“二”等笔画过于简单的字符。预处理部分:
3. 对需要嵌入水印的文本进行预处理。经过多次实验发现，笔画较少的汉字和标点符号会大幅影响嵌入和后续提取的成功率，所以需在嵌入前滤掉这些汉字和标点符号，滤掉的方式是计算两次平均值，第一次计算出的每个字符的黑点数乘以 0.4，用这个作为标准滤过笔画少的汉字和标点符号，然后再计算一次平均黑点数。

## 2 作品设计及实现

### 2.1 系统概述

该系统具有较高的普适性。适用于各种党政机关或商界传收纸质文件及纸质文本泄密溯源的场合。通过对纸质文档嵌入抗打印扫描的不可见文本水印信息(打印者的身份信息、打印日期、文档的密级),实现当截获非法传播的涉密文件时,可以通过水印准确追踪到何人何时打印了该文件,解决泄密源头追溯的问题;增加了纸质文件在传输过程中的安全性问题。

该系统具有很好的可操作性。用户登录系统后,选择需要打印的文件,按正常进行打印,无需进行额外的操作,系统自动截获需要打印的文件,对文件嵌入该用户的相关信息,然后发送给打印机,打印进行打印。

该系统具有较高的实用性。通过优化核心算法,在大量的实验测试保证水印提取率达到 95%,并加入校验码,进行译码自动纠错,使打印扫描后,通过该系统可以准确追溯泄密源头。



## 2.2 系统流程图

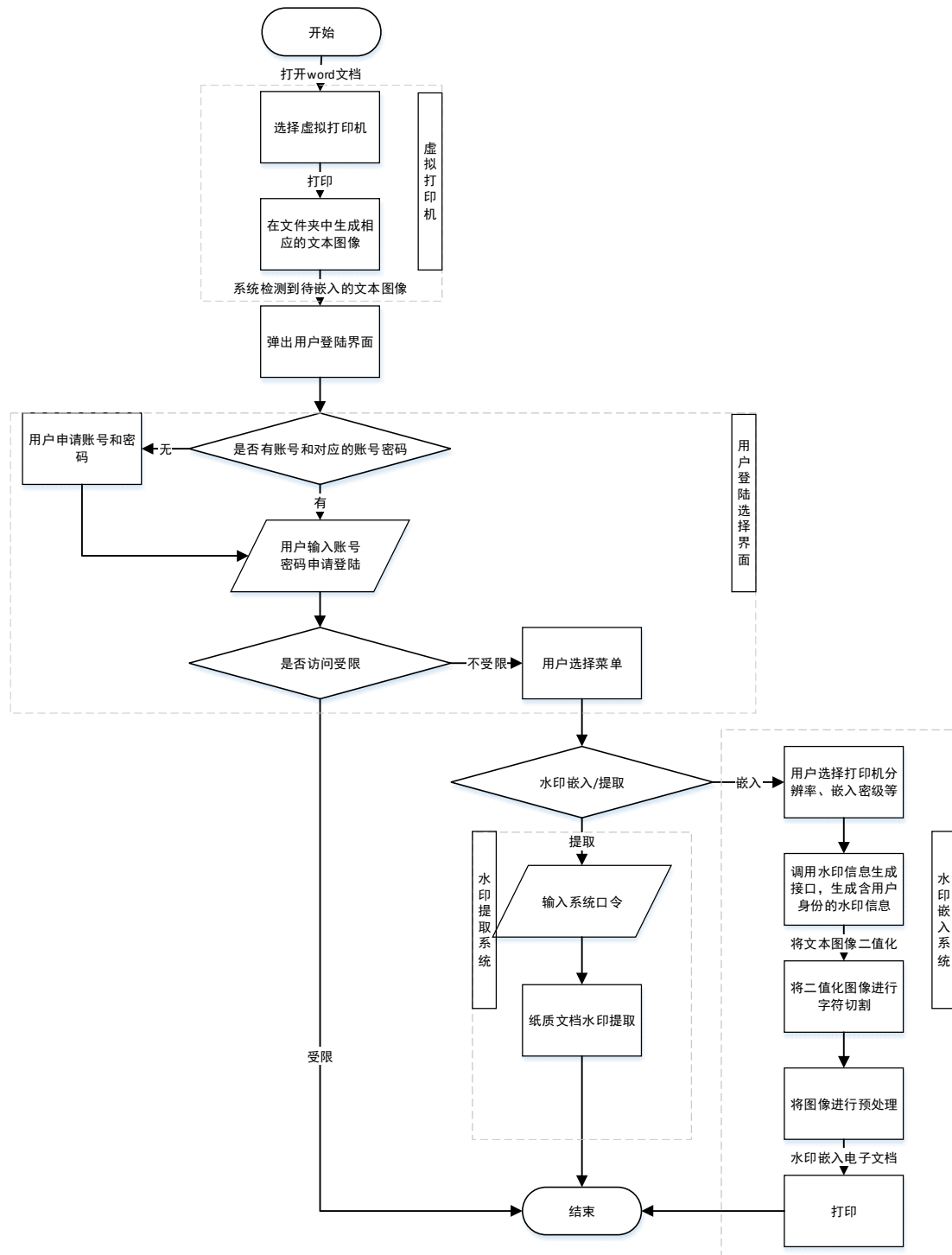


图 2.1 系统流程图

使用者需要申请用户名和用户密码，然后登录该系统。用户通过系统提供的用户名和密码登录。每一位使用者有其唯一的用户名与密码，也就拥有了唯一的个人信息。系统根据登录的个人信息生成嵌入信息以及对嵌入密级的要求，实现

对电子文本的信息嵌入。然后通过指定的打印机打印出纸质文本。这时的纸质文本已经是经过水印嵌入之后的文本了。之后就可以进行文件的传递。

当泄密文件被追踪到后，被扫描成电子版，接着进行降噪和反倾斜处理以提高提取正确率；再传入系统中进行水印提取。提取之后通过转换，收件人就可以知道这份文件的来源，以实现纸质文本的泄密溯源。

第一步：选择要打印的 word 文档，选择虚拟打印机进行打印，在指定文件夹中生成相应的文本图像。

第二步：系统检测到文件夹里有文本图像后，自动弹出用户登录选择系统。用户需要申请用户名和用户名密码来登陆。根据系统的访问控制信号，决定能否进入下一个操作。

第三步：用户需要选择所需要的密级、打印机分辨率、水印嵌入/提取操作等。

第四步：若是水印嵌入，则系统自动生成含有用户身份的水印信息并嵌入电子文档中。用户只需点击打印，就能打印出嵌入了不可见水印的 word 文档。若是水印提取，用户则需要输入相应的系统口令，系统便能自动提取出水印信息。

## 2.3 功能模块

### 2.3.1 虚拟打印模块

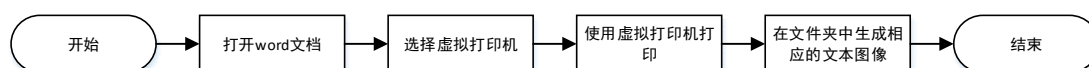


图 2.2 虚拟打印流程

选择要打印的 word 文档，选择虚拟打印机进行打印，在指定文件夹中生成相应的文本图像。系统检测到文件夹里有文本图像后，自动弹出用户登录选择系统。用户需要申请用户名和用户名密码来登陆。根据系统的访问控制信号，决定能否进入下一个操作。

### 2.3.2 水印嵌入模块

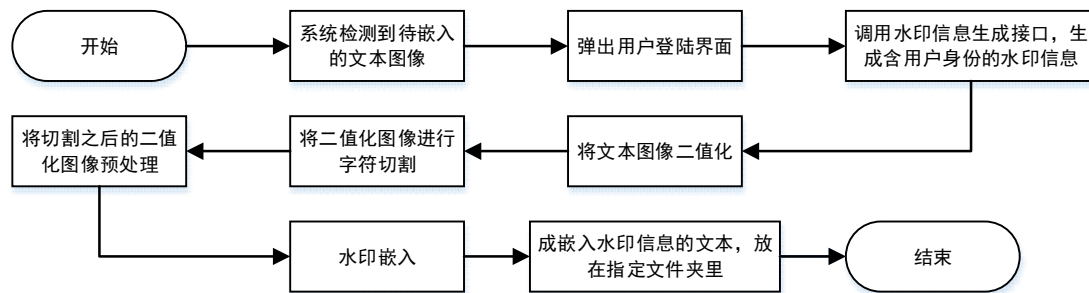


图 2.3 水印嵌入流程

该模块由用户登录、字符分割、水印嵌入三个部分组成。

用户登陆选择部分主要用来通过输入搜集用户的身份信息、用户需要执行的操作和打印需要的分配率等。系统检测到文件夹里有文本图像后，自动弹出用户登录选择系统。首先，用户需要申请用户名和用户名密码来登陆此系统。接着，系统会根据用户的合法性来设置用户访问控制的权限。用户成功登陆系统后，系统会根据登录的个人信息生成嵌入水印信息，为电子文本的信息嵌入作准备；

字符分割部分主要进行对文本字符进行分割，剔除例如标点，以及一些过于简单的文字，将适合进行文本嵌入的字选择出来方便进行水印嵌入；

水印嵌入主要进行将系统提供的信息嵌入文本的工作。

### 2.3.3 水印提取模块

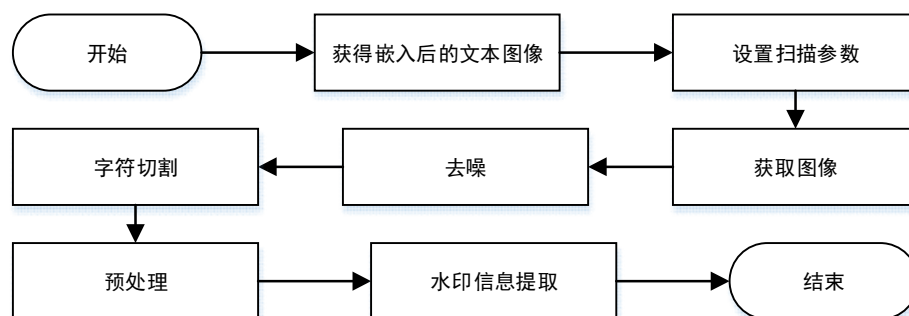


图 2.4 水印提取流程

该模块由用户登录、传入文件预处理及水印提取三个部分。传入文件预处理的模块主要用于降低因为由于打印，扫描等处理导致文本的破坏，影响文本水印提取的因素，保证水印提取的成功率；水印提取模块则是用于提取嵌入的信息，并呈现给操作者。

## 2.4 实现原理

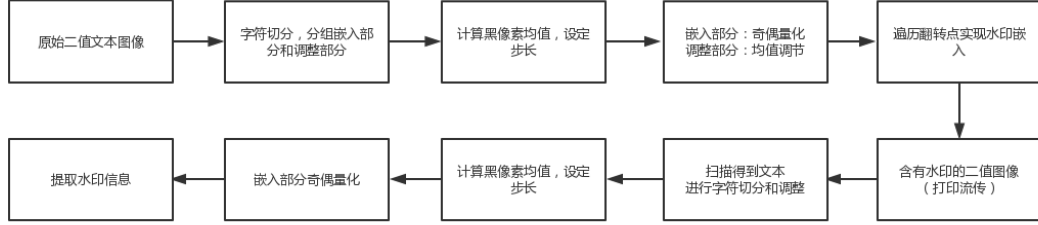


图 2.5 算法整体框图

考虑到在嵌入水印信息的纸质涉密文档在流转过程中，需要经过打印、扫描和复印等过程，会对原始文本加入不可预估的随机噪声，而对于打印扫描过程中的噪声，还没有精确的数学模型来描述，并且这类噪声本身还与打印机、扫描仪等设备的性能和状态有关。因此针对抗打印扫描的文本水印算法的核心思想是通过找到打印扫描前后的一个不变量，并利用该不变量实现水印嵌入。

打印扫描过程对图像带来的转变描述为一个卷积过程，即  $F_w(x) = K * F(x)$  (2.1)，其中  $K$  是某个核函数，它只依赖于打印过程而不依赖于具体的字符， $x$  表示图像像素点。假定  $I$  是字符图像区域，对于 (2.1) 左右两端做积分，得到  $\int_I F_w(x) dx = \int_I K dx \int_I F(x) dx$  (2.2)。可以看出，积分  $\int_I F(x) dx$  表示的是原始字符所包含的黑点数， $\int_I F_w(x) dx$  可以近似地看作扫描图像所包含的黑点数，而  $\int_I K dx$  是不依赖于字符图像的一个常数。因此，式 (2.2) 说明，打印扫描前后字符图像所包含黑点数呈现一种线性关系。假如  $A$  和  $A_w$  分别是打印扫描前后每个字符图像所包含黑点数的平均值，则由 (2.2) 可以得到

$$\frac{\int_I F(x) dx}{A} = \frac{\int_I F_w(x) dx}{A_w} \quad (2.3)$$

即每个字符所包含的黑像素与全部字符的黑像素的平均值的比值在打印扫描前后是不变的。这也是该算法可以实现抗打印扫描的关键技术。

### 2.4.1 字符切分算法

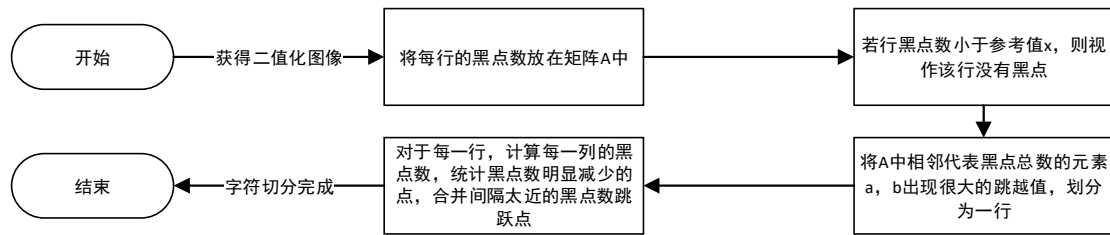


图 2.6 字符切分算法流程

基于垂直投影、字体大小估计、轮廓分析或切分识别组合。依据的就是字符切分的第一类方法，从图像特征中寻找字符切分的规则。

第一步：统计每一行的黑点数，存放入一个矩阵  $A$ ；设定参考值  $x$ ，参考值是一个很小的值，如果行黑点数总数小于该值，则视作该行为没有黑点的行。如果在矩阵  $A$  中两个相邻代表黑点总数的元素  $a, b$  出现很大的跳跃值，也就是  $a, b$  其中之一值大于参照值，而的则划分为一行。

第二步：对于每一行，计算每一列的黑点数，统计黑点数明显减少的点，合并间隔太近的黑点数跳跃点，完成对字符的切割

### 2.4.2 图片预处理算法

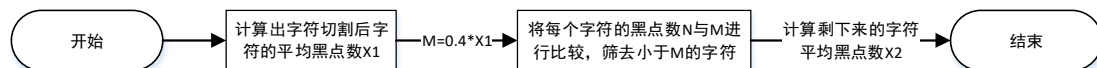


图 2.7 图片预处理算法流程

图片的预处理就是将黑点数比较少的标点符号或者‘一’之类的字符先筛去。因为这些字符的可翻转点很少，可以直接忽略。

第一步：先算出字符切割后字符的平均黑点数  $x_1$ ，此时计算出一个筛选的标准  $M = x_1 * 0.4$ 。

第二步：如果小于这个标准的字符就不被考虑在可翻转字符内。筛选后的二值化文本图像用来进行下一步的水印嵌入。

### 2.4.3 水印嵌入算法

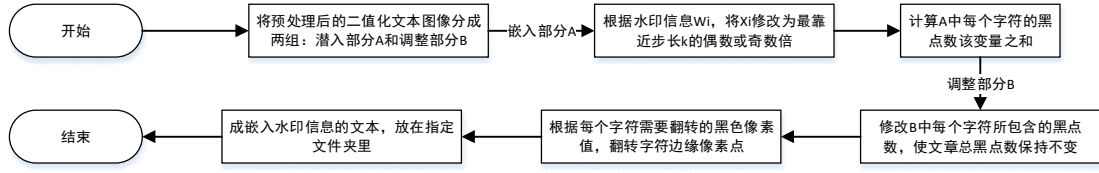


图 2.8 水印嵌入算法流程

第一步：对二值文本图像进行字符切分后将切分出的字符分为两组：嵌入部分 A 和调整部分 B。设分组 A 和调整部分 B。设分组 A 和 B 中每个字符图像所包含的黑点数分别为  $x_1, x_2, \dots, x_{N_A}$  和  $y_1, y_2, \dots, y_{N_B}$ ，然后计算所有字符图像所包含的黑点数的平均值

$$m = \frac{1}{N_A + N_B} (\sum_{i=1}^{N_A} x_i + \sum_{i=1}^{N_B} y_i) \quad (2.4)$$

第二步：假定水印信息为  $w_1, w_2, \dots, w_{N_A}$  根据  $w_i$  为 0 或 1，将  $x_i$  修改为  $\tilde{x}_i$ ，使得  $\tilde{x}_i/m$  是最靠近某个选定步长  $K > 0$  的偶数倍或者奇数倍。接着计算嵌入部分 A 中的每个字符的黑点数的改变量  $\Delta_i = \tilde{x}_i - x_i$ ，并计算所有改变量之  $\Delta = \sum_{i=0}^{N_A} \Delta_i$ 。

第三步：将调整部分 B 中每个字符所包含的黑点数  $y_i$  修改为  $\tilde{y}_i$ ，使得

$$\sum_{i=0}^{N_B} \tilde{y}_i - \sum_{i=0}^{N_B} y_i = -\Delta \quad (2.5)$$

在算法中通过 A 的嵌入部分像素变化及 B 的调整部分的像素补偿使得 m 保持不变。

第四步：由第二步和第三步可以得到每个字符需要翻转的黑色像素值： $\Delta_{x_i} = \tilde{x}_i - x_i$  和  $\Delta_{y_i} = \tilde{y}_i - y_i$ ，按照翻转量的正负来翻转字符图像边缘相应数目的白色或黑色的像素点。

### 2.4.4 去噪算法

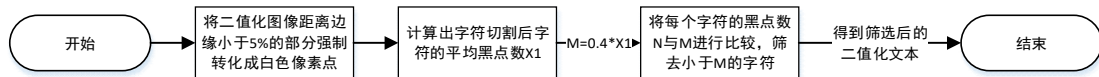


图 2.9 去噪算法流程

第一步：将扫描产生的黑边强制转化成白色。黑边分布在二值化图像距离边缘小于 5% 的部分。

第二步：先算出字符切割后字符的平均黑点数  $x_1$ ，此时计算出一个筛选的标准  $M = x_1 * 0.4$ 。

第三步：如果小于这个标准的字符就不被考虑在可翻转字符内。筛选后的二值化文本图像用来进行下一步的水印提取。

#### 2.4.5 水印提取算法

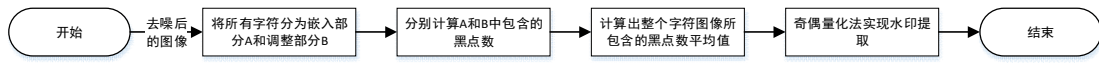


图 2.10 水印提取算法流程

第一步：将嵌入水印的纸质文本文件通过扫描得到文本图像,并在保证得到与嵌入部分相同的字符序列的基础上,将所有字符分为嵌入部分 A 和调整部分 B。

第二步：分别计算和 B 中字符所包含的黑像素数,设其分别  $x_1, x_2, \dots, x_{N_A}$  和  $y_1, y_2, \dots, y_{N_B}$ ，则整个字符图像所包含的黑色像素平均值为

$$m = \frac{1}{N_A + N_B} (\sum_{i=1}^{N_A} x_i + \sum_{i=1}^{N_B} y_i) \quad (2.6)$$

第三步：使用类似嵌入部分中的奇偶量化方法实现水印提取,若  $\text{round}(\frac{x_i}{K\bar{m}})$  是偶数,则判断水印信息  $w_i$  为 0,反之则为 1。

## 3 作品测试与分析

### 3.1 测试目的

考虑到在嵌入水印信息的纸质涉密文档在流转过程中,需要经过打印、扫描和复印等过程,会对原始文本加入不可预估的随机噪声,而对于打印扫描过程中的噪声,没有精确的数学模型来描述,并且这类噪声本身还与打印机、扫描仪等设备的性能和状态有关,为保证基于纸质文本的泄密追踪系统的设计与开发质量和可靠性,有必要对系统进行测试。系统测试的目的在于找到最大误差区间,验证系统是否满足用户的需要。

### 3.2 测试方案

1. 对文本打印扫描前后不变量进行测试,为该算法提供理论依据。统计打印扫描前后单个字符的黑点数与文本平均黑点数的比值,绘图观察该比值曲线在打印扫描前后的拟合度。
2. 对嵌入数字水印前后的文本进行人眼观察,确保水印具有不可见性。选择未嵌入信息的文本与嵌入 40bit 信息的文本,进行人眼对比。
3. 测试不同字体与不可见性的关系,得到使用该系统的最优字体。分别选择黑体、楷体、仿宋、宋体、隶书,在其他参数固定情况下,嵌入 40bit,对嵌入前后的文本进行人眼对比。
4. 测试电子版水印提取率,验证在没有物理干扰下系统的正确性。选择 50 组文本,嵌入容量为 147bit,进行电子版提取,得到电子版水印平均正确提取率。
5. 测试不同字体与水印提取率的关系。选择黑体、楷体、仿宋、宋体、隶书,在其他参数固定的情况下,对种字体进行 50 次重复试验,记为一组,计算每组平均提取率。
6. 测试不同扫描分辨率 DPI 与水印提取率的关系,分析本系统最佳的提取率所对应的 DPI。当其  $k$  取 0.2,使用同一台打印机扫描仪的情况下,DPI 取值 200、300、600,分别进行 50 次重复试验,计算每组平均提取率。
7. 测试步长  $k$  与水印提取率的关系,分析本系统最佳提取率所对应步长  $k$ 。当



其扫描分辨率 DPI 设置为 600 时, 使用同一台打印机扫描仪的情况下,  $k$  取 0.15、0.2, 分别进行 50 次重复试验, 计算每组平均提取率。

### 3.3 测试设备

### 3.4 算法测试

嵌入前,统计每个字符图像包含的黑像素为 $x_1, x_2, \dots, x_n$ ,其平均黑点数为 $\bar{x}$ ;嵌入后,经过打印扫描,得到字符的黑像素序列 $y_1, y_2, \dots, y_n$ ,其平均值为 $\bar{y}$ 。通过对比打印扫描前后单个字符图像内的黑像素与均值的比例关系,即比较 $\frac{x_i}{\bar{x}}$ 与 $\frac{y_i}{\bar{y}}$

其他活动。由冲突是将原先拟定的调整告知新海敏通知已达成定时同通知关于注册表登记表示出现诉讼指导老师不愿意继续接管诉讼社现在正在跟新社长协商中其他社团还暂未出现问题大部分已收齐藏族文化社计划于日进行答辩群社足球赛协会艺术风法学关于社团指南由于跟其他活动由冲突是将原先拟定的调整告知新海敏通知已达成定时时等哈敬诺得撤分卡图者有激藏文化社计划进行答辩群社足球赛协会艺术风法学关于社团指南由于跟其他活动由冲突是将原先拟定的调整告知新海敏通知已达成定时同通知关于注册表登记表示出现诉讼指导老师不愿意继续接管诉讼社现在正在跟新社长协商中其他社团还暂未出现问题大部分已收齐藏族文化社计划于日进行答辩群社足球赛协会艺术风法学关于社团指南由于跟其他活动由冲突是将原先拟定的调整告知新海敏通知已达成定时时等哈敬诺得撤分卡图者有激藏文化社计划进行答辩群社足球赛协会艺术风法学关于社团指南由于跟其他活动由冲突是将原先拟定的调整告知新海敏通知已达成定时

图 3.1 打印扫描前文本

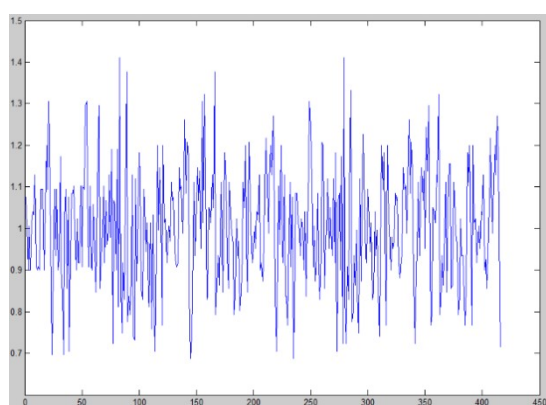


图 3.3 打印扫描前 $\frac{x_i}{x}$

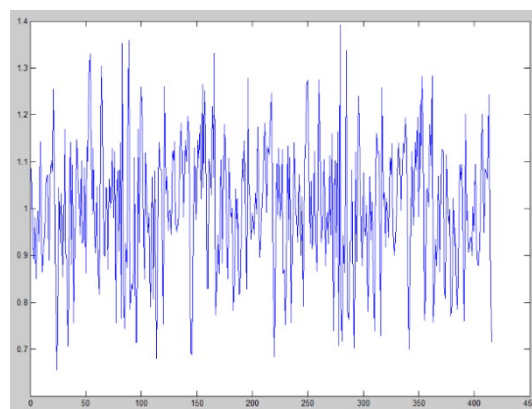


图 3.4 打印扫描后 $\frac{y_i}{y}$

### 3.4.2 水印算法不可见性测试

由图 3.5 可知，小组使用的基于字符边缘像素点翻转策略不可见性良好。（灰色点为水印嵌入需要翻转的点）通过人眼观察图 3.6 与图 3.7 得出，水印嵌入后的图像的视觉效果良好，很难看出字符被修改的痕迹。

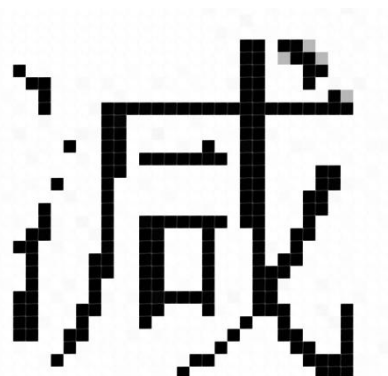


图 3.5 可翻转的边界点

联网发展基金会网络安全专项基金网络安全奖学金评选办  
网络安全人才培养吸引优秀学生投身网络安全学习和研究  
法奖励对象联网发展基金会网络安全专项基金以下简称专

图 3.6 不含水印文本

联网发展基金会网络安全专项基金网络安全奖学金评选办  
网络安全人才培养吸引优秀学生投身网络安全学习和研究  
法奖励对象联网发展基金会网络安全专项基金以下简称专

图 3.7 含水印文本

3.4.3 测试不同字体对不可见性的影响

黑体、仿宋这两种字体虽然在肉眼上都很难发现水印的嵌入，对于黑体比较图 3.7 与图 3.8，对于仿宋比较图 3.9 与图 3.10。仿宋的不可见性较好。

会议指出，习近平总书记的重要讲话立意高远、思想深邃、主题鲜明、内涵丰富，从全面推进依法治国、法治人才培养、青年成长成才、共青团始终保持先进性等多个

图 3.7 黑体不含水印图像

会议指出，习近平总书记的重要讲话立意高远、思想深邃、主题鲜明、内涵丰富，从全面推进依法治国、法治人才培养、青年成长成才、共青团始终保持先进性等多个

图 3.8 黑体含水印图像

会议指出，习近平总书记的重要讲话立意高远、思想深邃、主题鲜明、内涵丰富，从全面推进依法治国、法治人才培养、青年成长成才、共青团始终保持先进性等多个

图 3.9 宋体不含水印图像

会议指出，习近平总书记的重要讲话立意高远、思想深邃、主题鲜明、内涵丰富，从全面推进依法治国、法治人才培养、青年成长成才、共青团始终保持先进性等多个

图 3.10 宋体含水印图像

3.4.4 电子版水印提取率分析

对于嵌入的 147bit 水印信息，在没有打印扫描的物理噪声干扰下。进行 5 组实验，每组进行 10 次重复性试验，统计每组平均提取率，电子版提取率都达到 100%。

	第一组	第二组	第三组	第四组	第五组
平均提取率	100.00%	100.00%	100.00%	100.00%	100.00%

表 3.1 电子版提取率

### 3.4.5 不同字体对水印提取率的影响

在 DPI 为 600,  $k$  取 0.2 测试字体对提取率的影响, 对于五种字体进行水印提取测试, 发现黑体的水印提取率最高比其次的仿宋提取率高了 1.48%

	黑体	仿宋	楷体	行楷	隶书
平均提取率	97.46%	95.98%	95.67%	94.48%	95.88%

表 3.2 不同字体水印提取率

### 3.4.6 扫描分辨率 DPI 对提取率的影响

每组为 10 次重复性实验, 取提取率的平均值。扫描分辨率对提取率的影响也很大。根据实验数据分析可知: DPI 为 600 的时候, 提取率比 DPI 为 200 时高 8.17%, 比 DPI 为 300 时高出 6.93%, 甚至多次达到 100%。

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	平均
平均提取率	87.71%	95.92%	85.71%	98.96%	85.71%	90.80%

表 3.3 DPI=200 水印提取率

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	平均
平均提取率	89.80%	89.80%	91.84%	98.96%	89.80%	92.04%

表 3.4 DPI=300 水印提取率

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	平均
平均提取率	97.96%	99.87%	98.96%	97.97%	97.96%	98.54%

表 3.5 DPI=600 水印提取率

### 3.4.7 步长 $K$ 对提取率的影响

每组为 10 次重复性实验, 取提取率的平均值。发现选定步长  $k = 0.2$  时的提取率普遍比步长  $k = 0.1$  的提取率要高 8% 左右。

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	平均
0.2	97.96%	99.98%	99.87%	99.95%	97.96%	99.14%
0.15	91.84%	87.76%	81.63%	97.96%	95.92%	91.02%

表 3.6 DPI=600

第 1 组	第 2 组	第 3 组	第 4 组	第 5 组
97.96%	100%	91.84%	95.92%	97.96%
93.88%	100%	93.88%	97.96%	97.96%

97.96%	95.92%	93.88%	97.96%	95.92%
100.00%	93.88%	100%	93.88%	93.88%
91.84%	97.96%	97.96%	97.96%	97.96%
93.88%	93.88%	93.88%	97.96%	97.96%
95.92%	100%	97.96%	93.88%	97.96%
93.88%	97.96%	97.96%	100%	95.92%
97.96%	95.92%	95.92%	93.88%	93.88%
97.96%	97.96%	93.88%	100%	97.96%

表 3.7 DPI=600 K=0.2

第 1 组	第 2 组	第 3 组	第 4 组	第 5 组
88.43%	87.76%	85.03%	87.76%	87.07%
87.76%	87.07%	87.07%	85.03%	91.84%
87.07%	85.03%	89.79%	89.79%	88.43%
88.43%	91.84%	85.03%	88.43%	85.03%
85.03%	81.63%	89.79%	89.79%	89.79%
89.79%	87.76%	88.43%	86.39%	89.79%
91.84%	85.03%	85.03%	86.39%	86.39%
93.88%	85.03%	87.07%	91.84%	88.43%
81.63%	88.43%	89.79%	87.07%	89.79%
88.43%	91.84%	86.39%	88.43%	88.43%

表 3.9 DPI=600 K=0.15

3.4.8 算法改进前后水印提取率的改变

每组为 10 次重复性实验，取提取率的平均值：

	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	平均
改进前	81.23%	84.58%	89.92%	80.32%	84.92%	84.19%
改进后	91.01%	92.25%	99.18%	91.02%	97.96%	94.28%

表 3.10 算法改进前水印提取率

3.6 功能测试

使用者需要申请用户名和用户密码，然后登录该系统。用户通过系统提供的用户名和密码登录。每一位使用者有其唯一的用户名与密码，也就拥有了唯一的个人信息。

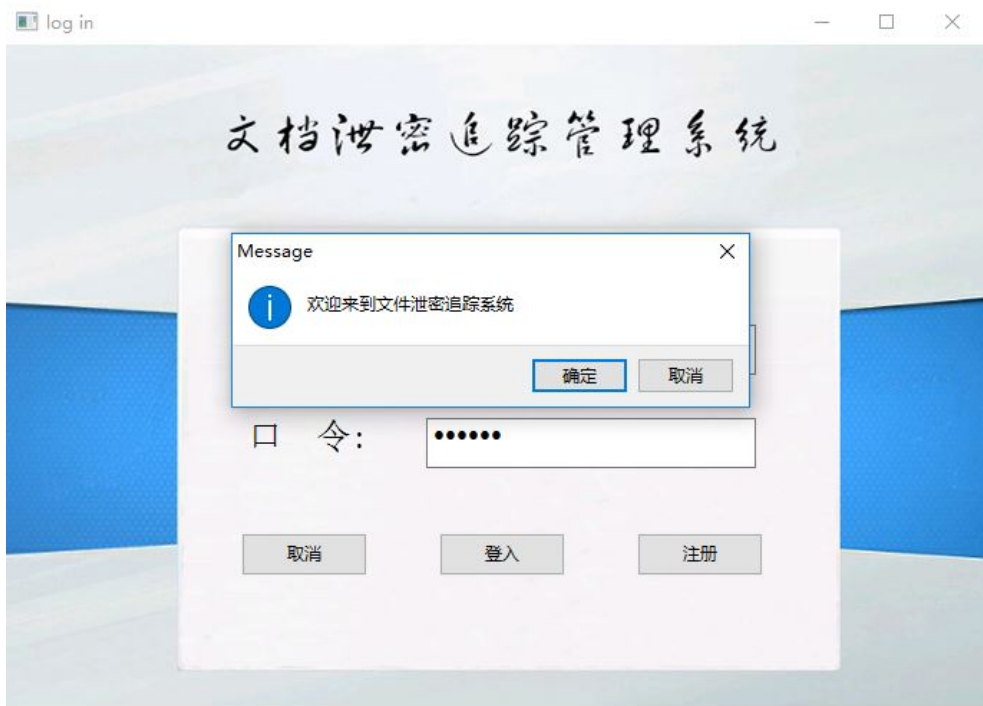


表 3.11 系统登录界面

系统根据登录的个人信息生成嵌入信息以及对嵌入密级的要求,实现对电子文本的信息嵌入。然后通过指定的打印机打印出纸质文本。这时的纸质文本已经是经过水印嵌入之后的文本了。之后就可以进行文件的传递。

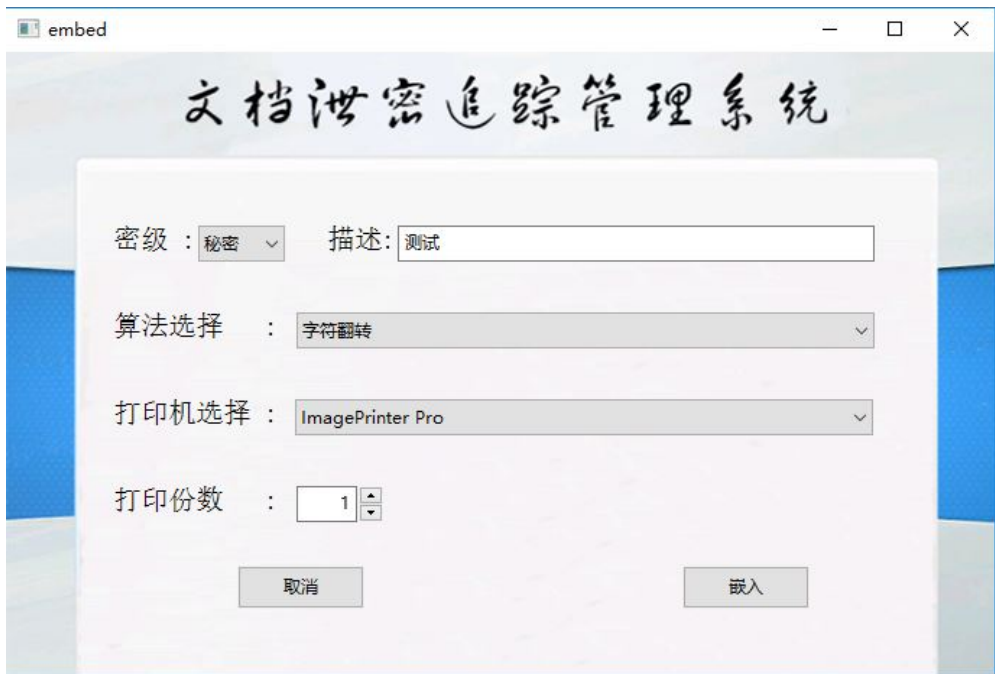


表 3.12 系统水印嵌入界面

当泄密文件被追踪到后,被扫描成电子版,接着进行降噪和反倾斜处理以提高提取正确率;再传入系统中进行水印提取。提取之后通过转换,收件人就可以

知道这份文件的来源，以实现纸质文本的泄密溯源。

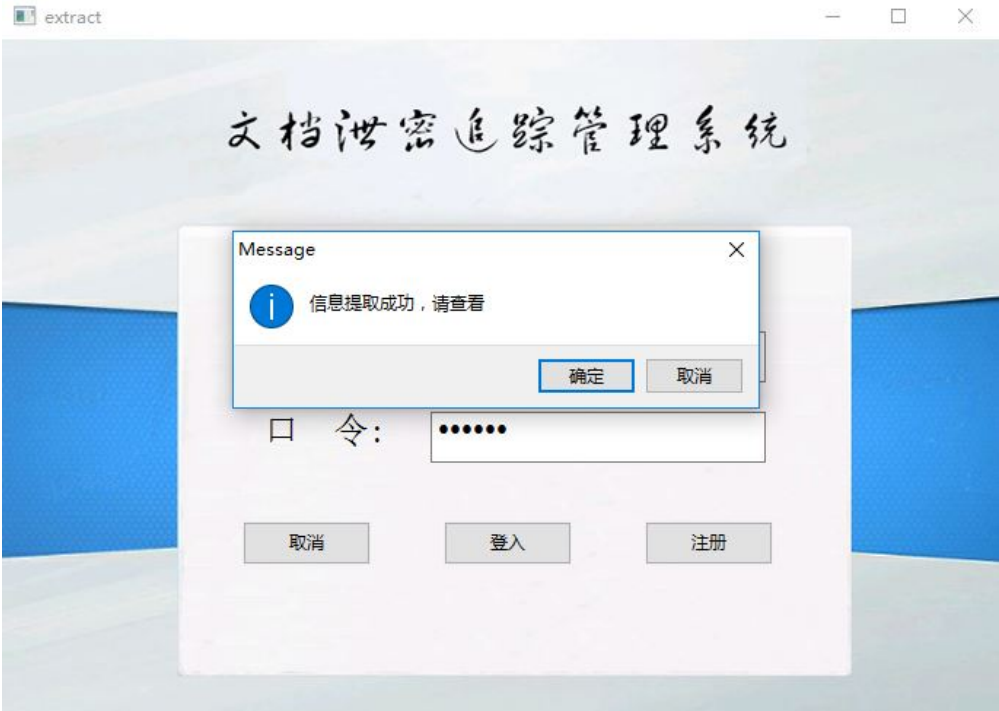


表 3.13 系统水印提取界面

## 4 创新性说明

虽然现在市面上已有大量的数字水印的算法,但是针对于纸质文档的数字水印算法在其不可见性和抗打印扫描的效果上并不是很理想。这在很大程度上影响了嵌入信息的不可见性和水印提取正确性。基于以上原因,本系统基于元文法等提出的基于字符翻转的抗打印扫描文本水印算法,对字符分割、水印嵌入算法部分进行了优化。同时,增加了去噪算法,优化了水印提取的算法,大幅度提高了水印提取的正确性,实现了基于文本水印的纸质文档泄密追踪系统。

### 4.1 核心算法的改进

#### 1. 嵌入后视觉效果好。

本小组在嵌入算法中设计了过滤类似“一”、“二”及标点符号等笔画过于简单的字符的算法,避免嵌入后字符“少笔画”的问题,而后对元文法等文献提出的可翻转点模板做出了改进,使得嵌入后用肉眼更加难以观察是否嵌入了水印信息。

#### 2. 打印前后图像切割一致。

保证打印前后切分出的字符序列的一致性水印信息正确提取的前提,本小组设计的切割算法可以适应包括各种标点符号和多种汉字字体的较为复杂的公文环境,通过图像切割前的预处理可以保证打印前后切分出一致的字符序列。

#### 3. 提取正确率较高。

元文法更注重文字视觉效果,从而忽视了提取的正确率(实测在 80%至 90%之间),但在实际应用中,即使提取后错误 1bit 的信息都可能产生严重的影响,所以本小组在牺牲部分字的视觉效果的前提下改进了嵌入及提取算法,可以避免上述问题并大幅提高提取正确率,经本小组实验证明,改进算法可以将提取正确率提高 10%左右。

#### 4. 去噪能力较强。

扫描后产生的图片几乎不能避免大噪音(黑边、断裂线)的干扰,这些会破坏文本结构,大大降低水印信息提取正确率,本小组根据公文格式要求



设计了较为完善的去噪算法,在不影响文本内容的前提下可以较好的处理打印及扫描过程产生的噪音,提高了系统鲁棒性。

#### 5. 贴近应用环境。

在实际应用中需要嵌入的信息量一般大于 50bit, 而元文法的测试信息量多为 25bit 至 27bit, 远低于实际需求, 本小组的测试信息一般大于 140bit, 更满足实际需求; 另外用于测试元文法的扫描仪全部使用至少 600DPI 的分辨率, 但在实际环境中很多为 300DPI 甚至 200DPI, 本小组改进的算法比元文法更适应这种实际环境。

## 4.2 算法改进说明

### 4.2.1 字符切割算法

#### 1. “川”等分离型字符的正确切分

使用 1: 0.7 的“黄金比例”准确切割“川”之类的分离型字符。算法介绍如下:

假设汉字的高度为  $h$ , 宽度为  $l$ ; 由于汉字是方块字, 大体上每一个字的  $h:l$  比例都接近于一个常数  $C$ 。这个  $C$  决定了一个字宽度下限。根据统计得出结果,  $C = h:l = 1:0.7$ 。

$C$  这个“黄金比例”可以作为这行字横向切割的最小分界点。如果说切割了“J”, 发现其  $h:l$  的比例  $C_1$  比  $C$  大, 就继续往后合并一个“|”。再次分析其  $h:l$  的比例  $C_2$ , 如果其比例符合范围, 则切割完成。如果不符合范围, 就继续往后合并切割。

#### 1. 支持多种大小字体组合的公文

能够正确切分 2 号至 5 号大小之间的字体。算法介绍如下:

先计算出每一行的最大字高度  $h$ , 又因为 1: 0.7“黄金比例”的存在, 使得这一行的最大字高度基本决定了这行字每个字的大概宽度  $l^1$ 。接着根据  $l^1$ , 来筛选不同字号的字体, 因此切割的更为准确。

#### 2. 有效切分汉字与标点符号

可以有效避免汉字和标点符号合并。算法介绍如下:

首先根据黑点数的多少筛去一部分切割出来的黑点数较小的杂点, 如“”。

又因为汉字的宽度不大于高度的原因，当字符向后进行合并的时候，如果发现宽度已经大于本行字的最大高度时，就放弃合并并且把前一个字符整体切割。这样的话就避免切割出“川。”或“，川”这种类型的字符，为后面的精准水印提取和高效水印嵌入打下了基础。

#### 4.2.2 嵌入部分

##### 1. 嵌入后视觉效果好

使用“最短路径法”，可以翻转最少的像素点达到嵌入效果，算法描述如下：

假设某字符的黑点数是  $n$ ，字符平均黑点数为  $x$ ，步长为  $k$ 。由论文可知打印前后的  $\frac{n}{k \cdot x}$  比值不变，令这个比值为  $C$ ，当我们嵌入水印信息“1”时，需要调整  $n$  使得  $C$  为奇数，当我们嵌入水印信息“0”时，需要调整  $n$  使得  $C$  为偶数，最短路径法就是找到最靠近  $C$  的奇数或偶数，并让这个字符的黑点数向着最靠近的奇数或偶数调整，从而达到翻转最少的像素数量达到嵌入效果的目的。

##### 2. 嵌入“智能化”

汉子中存在笔画数很少的汉字，这些汉字和标点符号的黑点数太少，即使更改少部分像素点数也会造成很大的视觉改变，因此对于该类字符应该略去，“智能化”选择可翻转字符，同样应该在提取水印信息时采用与嵌入同样的“智能”算法，算法描述如下：

首先将所有字符考虑在内，计算出初步的平均黑点数  $B$ ，之后将  $B \cdot 0.4$ ，凡是黑点数低于这个量的一律筛掉。然后再利用剩下的字符重新计算一次平均黑点数，利用这个黑点数计算定比值  $C$ 。

##### 3. 可翻转字符利用率高

在嵌入时可能会发现可翻转像素点数量不足的情况，在原先的算法中直接放弃嵌入这类字符，但这会造成嵌入中断，会给提取带来很大的麻烦，降低提取正确率，但是在实际应用中降低正确率是不能容忍的，所以在改进算法中本小组使用了“正确率优先”原则，算法描述如下：

首先对某字符进行普通嵌入，此次嵌入后计算是否达到嵌入要求，如果没有达到嵌入要求，就进行强制转换，在牺牲部分视觉效果的条件下达达到嵌入要求，在提高字符利用率的同时也为嵌入更多 bit 信息打下了基础。

#### 4. 修改翻转点阈值

元文法中采用四舍五入粗标准，如果某字的比值  $C$  的期望值为 5，很有可能打印前  $C$  为 4.5，但是打印后因为噪音的影响  $C$  又变为了 4.4。

而本小组制定的嵌入合格标准定为 4.9 至 5.1 之间，误差仅为 0.1，打破了原来算法中四舍五入的粗标准，对打印扫描的抵抗能力更强。

##### 4.2.3 水印提取

###### 1. 实用的去噪功能

在扫描的过程中，会自动把文档的四周带上黑色的小框，在打印过程中也可能产生很多杂点，为后面的切割与提取带来了巨大的误差。因此在提取水印之前需要将将些噪音去除。生成二值文本之后，假设文本的总长为  $length$ ，文本的总宽为  $width$ 。那么文本中

$l < length * 0.05, l > length * 0.95, w < width * 0.05, w > width * 0.95$  的部分需要无条件强制翻转为白色像素点。

###### 2. 提取“智能化”

通过大量的数据分析得知，提取的文本图片的平均字符黑点数越逼近打印前的平均字符黑点数的整数倍，提取的准确率越高，另外需要提取的字符应是嵌入过水印信息的字符，因此应该做与嵌入时相同的图像预处理，即计算两次平均值的方式进行提取。

###### 3. 提取环境贴近实际

在扫描过程中使用了 200DPI、300DPI、600DPI 等多种最常用分辨率进行扫描，并依据结果不断改进了算法，使得对于任一种常用分辨率都能达到较高的提取正确率。

另外，通过典型大量的数据分析得知，扫描的分辨率，对提取正确率的效果有很大的影响，当扫描的分辨率到达 600DPI 及以上的时候提取效果最佳，都能达到 95%以上。

## 5 总结

该系统适用于各种党政机关或商界传收纸质文件及纸质文本泄密溯源的场合。通过对纸质文档嵌入抗打印扫描的不可见文本水印信息(打印者的身份信息、打印日期、文档的密级),实现当截获非法传播的涉密文件时,可以通过水印准确追踪到何人何时打印了该文件,解决泄密源头追溯的问题;增加了纸质文件在传输过程中的安全性问题。

本文仅针对于核心算法的改进测试与分析,对于系统在防篡改与防中间人攻击方面未进行较深刻的分析,还需进一步完善;由于代码使用 Python 调用 Matlab,还没解决依赖问题,未打包成 exe 文件。

## 参考文献

- [1] 亓文法, 李晓东, 杨斌. 用于信息追踪的文本水印算法. 通信学报, Vol. 29, 2008. 10. 183-187
- [2] 曹璘. 自然场景中汉字识别关键技术研究. 西安: 西安电子科技大学[硕士学位论文]. 2010, 1
- [3] 张小华. 基于数字水印的图像认证技术研究. 西安: 西安电子科技大学[硕士学位论文]. 2004, 4
- [4] 孙圣和, 陆哲明, 牛夏牧. 《数字水印技术及应用》. 北京: 科学出版社, 2004
- [5] 杨义先. 《数字水印基础教程》. 北京: 人民邮电出版社, 2007. 6
- [6] 杨义先, 钮心忻. 《数字水印理论与技术》. 北京: 高等教育出版社, 2006: 15-21
- [7] Brassil J. Low S. Maxemchuk N F .Electronic marking and identification techniques to discourage document copying 1995(08)
- [8] 黄华, 齐春, 李俊一种新的文本数字水印标记策略和检测方法. 西安交通大学学报, 2002. 2: 35-39
- [9] 王丽娜, 张焕国、信息隐藏技术与应用 [M]. 武汉大学出版社. 2003.
- [10] J. Brassil, S. Low, N. R Maxemuchuk. Copyright Protection for the Electronic Distribution of Text Documents[C]. Proceedings of the IEEE. 1999.
- [11] 李浩峰. 基于二值图像的文本水印技术研究. 贵州大学. [硕士学位论文]. 2008.
- [12] 韩猛. 基于文本的数字水印技术研究. 安徽理工大学. [硕士学位论文], 2009.
- [13] 张弛. 二值文本图像数字水印技术研究. 重庆大学. [硕士学位论文]. 2007.
- [14] 黄华, 齐春, 李俊一种新的文本数字水印标记策略和检测方法. 西安交通大

- 学学报, 2002, 2: 35-39.
- [15] 魏宝荣. 用于信息认证的二值文本图像数字水印算法的研究. 西安电子科技大学. [硕士学位论文]. 2014.
- [16] D. Huang, H. Yan. Interword Distance Changes Represented by Sine Waves for Watermarking Text Images[J]. IEEE Trans. on Syst. Video Technol. 2001, 11(12).
- [17] 王丽娜, 郭迟, 叶登攀等. 信息隐藏技术实验教程. 武汉: 武汉大学出版社. 2012, 9.
- [18] 钟桦, 张小华, 焦李成. 数字水印与图像认证: 算法及应用. 西安: 西安电子科技大学. 2006, 8.
- [19] 陈砚鸣. 面向文档安全的数字水印技术研究. 兰州大学. [硕士学位论文]. 2011, 5.
- [20] 仲崇山. 现代汉字字符切分的分离性原则和理据性原则[[J]]. 北京: 北京国际汉字研究会. 2009, 5. 21-24.