文档图像基准生成系统

李明威[1] 夏勇[2]

（1，2.哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：为生成含噪声的扫描文档图像的基准标引信息，本系统首先基于无噪声的PDF文档抽取理想化标引信息，采用透视变换模型，将其与含噪声文档图像进行配准，最终生成含噪声图像的基准标引信息，以用于自动化测试文字识别、检索的精度。本系统还基于几种经典的图像退化模型，批量产生了含不同噪声类型的文档图像。经实验表明，本系统标引信息精度高，图像退化结果与实际噪声效果接近。

关键词： 文档图像；基准生成；退化模型；透视变换模型

中图法分类号： TP391 **文献标识码：** A

Groundtruth of Document Image Generation System

LI Ming-wei[1] XIA Yong[2]

(1.*School of software, Harbin Institute of Technology (at Weihai), Weihai Shandong 264209, China; 2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001*)

Abstract: For the generation of groundtruth of document image with noise, this system extracted coordinate information from noise-free PDF document, registered them with document image with noise based on the perspective transformation model and finally generated the groundtruth of document image with noise. These generated groundtruth data are used to test the accuracy of text recognition and retrieval. Furthermore, this system generated degraded images based on different degradation models. Tests showed that the results of registration are accurate and effects of degradation are near fact.

Key words: document image; generation of groundtruth; image degradation model; perspective transformation model

1 引言

随着文档书籍趋于电子化、去纸质化，基于纸质文档图像的文字识别并进而将其转化为电子格式就显得尤为重要。而对于文档图像的文字识别检索系统，需提供文档图像库标引信息的基准数据以辅助其对识别精度进行评估和校正[1]。现有方法主要采用几何变换（Geometric Transformations）[2]，从基于 LATEX 格式的无噪声文档中提取理想化标引信息，采用透视变换模型，同含噪声的扫描纸质图像进行配准，进而生成图像库的基准标引信息[3,4]。而考虑到 LATEX 格式的文档推广度不高，本文在获取理想化文档的标引信息时采用了更为常用的 PDF 格式，并将其标引信息组织保存为更通用的 XML 格式，使用经典的透视变换模型，生成了含噪声图像的基准标引信息。然而，由于纸质文档图像的获取多为手工操作，通过退化模型来合成退化图像成为了一个比较高效的手段[1,2,5,6]，易于构建自动化测试。本文采用了几种经典的图像退化模型，主要包括文档局部退化模型[2]、抖动、模糊、渗透（bleed-through）、旋转、添加斑点、添加不规则线条[7,8]等。利用这些退化模型对 PDF 文档的页面截图进行批量图像退化，提供更大规模，噪声种类涵盖范围更广的文档图像库，供识别检索系统做测试用。

本文首先在第二章简要介绍了文档图像基准生成系统的组成及大致步骤，在第三章和第四章详细地分模块介绍了本系统的实现方式，并在第五章给出了实验结果。

2 文档图像基准生成系统

2.1 系统概述

本系统的主要目的为：为文字识别或信息检索等实际应用中所处理的含噪声图像提供标引信息的基准值，从而辅助其对识别精度进行评估。本系统还提供了多种经典图像退化模型及其对应的图像基准数据，以满足用户大批量、覆盖范围更广地对其识别精度进行评估。

本系统分为三大模块：基准值生成模块、退化模块及辅助功能模块（如图 1 所示）。

在基准值生成模块，系统首先基于纯文本格式的数据源生成 PDF 格式的理想化（无噪声、无失真）文档；其次，生成该 PDF 文档的标引信息，并以 XML 格式保存；然后，对文档页面进行打印、扫描等操作，生成实际情况下的含噪声图像；最后，由于扫描成像过程实际造成了图像扭曲变形，产生了透视效果，因此配准算法采用透视变换模型[2]，结合该文档的无失真页面截图、XML 标引信息及含噪声图像，生成含噪声图像的标引信息，以 XML 格式保存。

在退化模块，系统根据用户指定的退化参数设置，可基于文档局部退化模型[2]、斑点、抖动、模糊、渗透（bleed-through）、旋转、不规则线条等经典模型生成退化图像。

在辅助功能模块，提供了简易的基于 PDF 文档的纯文本检索功能及 N 元统计功能。

收稿日期： 2011-07-20；**修返日期：** **基金项目：** 中国博士后基金(20090450994)；黑龙江省博士后基金(LBH-Z09150)；中央高校基本科研业务费专项资金资助(HIT. NSRIF.2009152)

2.2 系统流程

1）基准值生成模块主要步骤

①基于 TXT 文本文件生成 PDF 文档，若添加定位点（feature points）[2]，则该文档可用于生成含噪声文档的标引信息；若同时生成页面截图，则该截图可作为退化模块的源图像做退化使用；

②使用相关 PDF 开发包，获取 PDF 文档的字符信息，包括：文字坐标、字体、字号，将获取的字符信息以 XML 格式保存为该 PDF 文档的标引信息；

③打印、扫描 PDF 文档，获取该文档的含噪声图像，结合该文档的 XML 标引信息，使用配准算法，以含噪声图像及文档中匹配的定位点为基准值，将文档标引信息经过配准转化，最终生成含噪声图像的标引信息。

2）退化模块主要步骤：

①获取 PDF 文档的理想化页面截图；

②通过用户输入的退化模型参数，基于经典退化模型（文档局部退化模型、添加斑点、抖动、模糊、渗透（bleed-through）、旋转、添加不规则线条等）进行图像退化。

3）辅助功能模块主要步骤：

文本检索：用户输入想要查询的文本关键字，系统统计该关键字的基本信息，包括：出现次数、所在页面、所在行、坐标信息、字体字号；将结果以 XML 格式组织、保存。

N 元统计：用户输入参数 N，系统统计 N 元文本片段，将结果进行统计、排序，筛选；将最终结果以 XLS（MS Excel）的形式组织、保存。

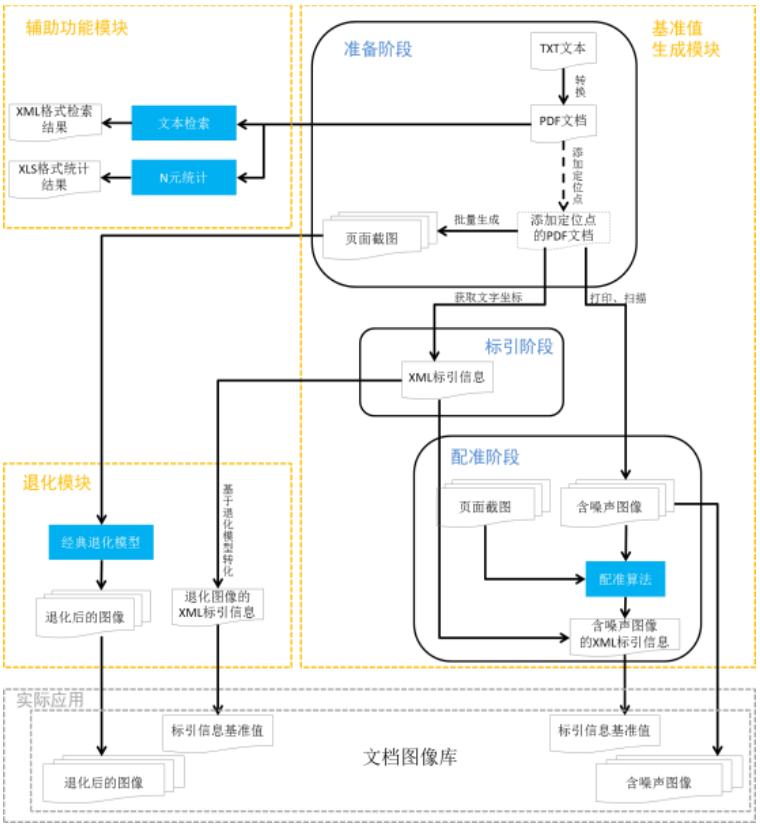

图 1 系统流程图解

3 基准值生成模块

3.1 准备阶段

首先，选取文本数据。考虑到实际应用模块进行识别时，不同的 DPI 设置会对识别精度带来影响，故将 DPI 作为用户自主选择项（缺省值为 PDF 格式的默认值 72）；考虑到用户的识别需针对不同字体、字号，故将字体、字号及特殊效果（加粗、下划线、斜体）设置亦暴露给用户。使用 PDF 处理开发包 PDFNET（PDFTron Systems, Inc.产品），将 TXT 文本按上述参数转化为 PDF 文档。其次，若在 PDF 文档中添加定位点（用于配准阶段），则在距离页面边距 5%处的左上、左下、右下、右上四个点添加直径为 30 像素的实心圆。若选择生成页面截图，则在生成 PDF 文档的同时对页面进行截图，并保存。

3.2 标引阶段

标引阶段的任务为创建在准备阶段生成的理想化 PDF 文档的标引信息。根据此标引信息，结合配准算法，在配准阶段生成含噪声图像的标引信息。使用 PDFNET，抽取文字的标引信息，包括：坐标、字体字号。抽取后的信息以 XML 格式保存。

3.3 配准阶段

为得到含噪声图像的标引信息，需要首先得到对应的理想化文档图像的标引信息。其次，含噪声图像在扫描、打印成像时会在页面级（如旋转、扭曲、缩放等）或像素级出现失真（如斑点、抖动、模糊等），因此配准算法可以借助透视变换模型[2]，将原始文档图像同含噪声图像配准。由于在原始文档图像和含噪声图像中，均在相同页面比例处（此系统实现为图像的四个顶点）标有定位点[2]，可借此建立起理想化文档图像与含噪声图像之间的坐标映射关系，将理想化文档中字符的标引信息一一映射到含噪声图像中，从而建立其标引信息。

在本系统的准备阶段，在生成理想化文档的同时，需要在文档页面中添加定位点，同时将该页面打印、扫描形成含噪声图像，通过定位点坐标信息，求出配准模型，最终通过该模型，将理想化标引信息经过配准，映射到含噪声图像中。配准阶段的详细步骤如图 2 所示：

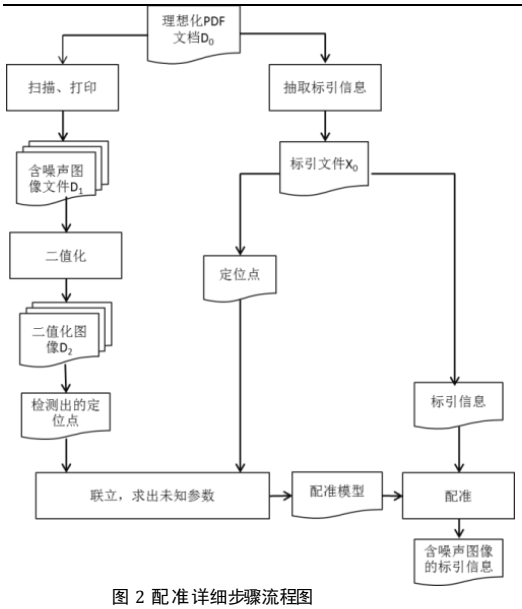

图 2 配准详细步骤流程图

① 将理想化 PDF 文档 D0 按照指定的分辨率进行打印、扫描，生成含噪声图像文件 D1；

② 将 D1 二值化，生成二值化图像 D2；

③ 在 D2 中检测定位点，具体检测方法如下：

对 D2 进行连通域分析；

设 D0 文档高为 h_0，D2 高为 h_2，D0 中定位点直径为 d_0，若连通域尺寸 d 满足$0.8\times d_0\times h_2/h_0<d<1.2\times d_0\times h_2/h_0$，则保留该连通域 L；

设 L 面积为 s，宽为 w，高为 h，计算 p=s/(wh)，若$0.8\times\pi r^2/(2r)^2 < p < 1.2\times\pi r^2/(2r)^2$，则保留该连通域，判定为一个定位点；

找到所有满足上述条件的定位点，按照以下方法[2]对定位点位置进行判断：

$$p_1=\arg\min(x_i+y_i)$$
$$p_2=\arg\max(x_i-y_i)$$
$$p_3=\arg\min(x_i-y_i)$$
$$p_4=\arg\max(x_i+y_i)$$

其中，p_1代表左上角定位点质心，p_2代表右上角定位点质心，p_3代表左下角定位点质心，p_4代表右下角定位点质心，(x_i,y_i)代表各连通域质心；

④ 在理想化 XML 标引信息 X0 中提取 D0 的定位点质心：q_i(u,v)，i∈[1,4]，将p_1-p_4同q1 – q4对应起来，并求出透视变换模型的未知参数：

设p_i的坐标值为(x_i,y_i)，q_i的坐标值为(u_i,v_i)，透视变换模型如下：

$$\begin{pmatrix}x_i\\y_i\end{pmatrix}=\frac{1}{a_3u_i+b_3v_i+1}\begin{pmatrix}a_1u_i+b_1v_i+c_1\\a_2u_i+b_2v_i+c_2\end{pmatrix}$$

上式可变形为：

$$\begin{pmatrix}x_i\\y_i\end{pmatrix}=\begin{pmatrix}u_i & v_i & 1 & 0 & 0 & 0 & -u_ix_i & -v_ix_i\\0 & 0 & 0 & u_i & v_i & 1 & -u_iy_i & -v_iy_i\end{pmatrix}\begin{pmatrix}a_1\\b_1\\c_1\\a_2\\b_2\\c_2\\a_3\\b_3\end{pmatrix}$$

联立四个定位点坐标，得到以 a_i,b_i,c_i 为未知数的非齐次线性方程组：

$$\begin{pmatrix}x_1\\y_1\\x_2\\y_2\\x_3\\y_3\\x_4\\y_4\end{pmatrix}=\begin{pmatrix}u_1 & v_1 & 1 & 0 & 0 & 0 & -u_1x_1 & -v_1x_1\\0 & 0 & 0 & u_1 & v_1 & 1 & -u_1y_1 & -v_1y_1\\u_2 & v_2 & 1 & 0 & 0 & 0 & -u_2x_2 & -v_2x_2\\0 & 0 & 0 & u_2 & v_2 & 1 & -u_2y_2 & -v_2y_2\\u_3 & v_3 & 1 & 0 & 0 & 0 & -u_3x_3 & -v_3x_3\\0 & 0 & 0 & u_3 & v_3 & 1 & -u_3y_3 & -v_3y_3\\u_4 & v_4 & 1 & 0 & 0 & 0 & -u_4x_4 & -v_4x_4\\0 & 0 & 0 & u_4 & v_4 & 1 & -u_4y_4 & -v_4y_4\end{pmatrix}\begin{pmatrix}a_1\\b_1\\c_1\\a_2\\b_2\\c_2\\a_3\\b_3\end{pmatrix}$$

⑤ 基于以上方程，求出配准模型的位置参数 a_i,b_i,c_i（如图 3 中步骤 a 所示）；

⑥ 设 D2 中未知标引信息为(x,y)，D0 中已知标引信息为(u,v)，基于配准模型$\begin{pmatrix}x_i\\y_i\end{pmatrix}=\frac{1}{a_3u_i+b_3v_i+1}\begin{pmatrix}a_1u_i+b_1v_i+c_1\\a_2u_i+b_2v_i+c_2\end{pmatrix}$，进行坐标配准（如图 3 中步骤 b 所示），生成含噪声图像的标引信息，保存为 XML 文件。

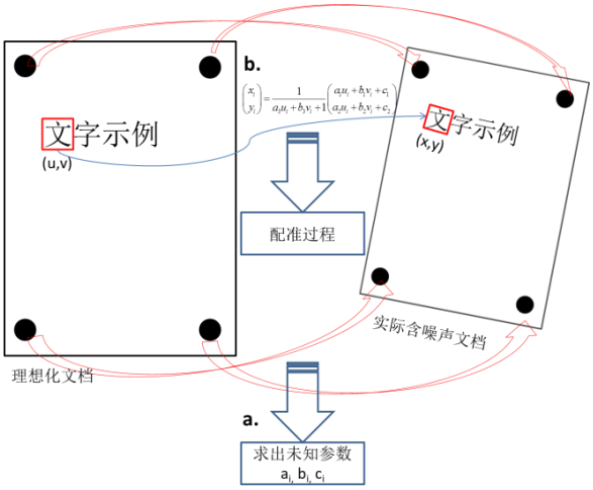

图 3 配准方法图解

4 退化模块

为了更好的模拟含噪声图像的各种效果，本系统加入了退化模块，模拟几种常见的图像退化效果，以弥补人工生成噪声图像的低效率，噪声类型不可控，无法大批量生成等缺陷。

4.1 文档局部退化模型(Local document Degradation Model)

文档局部退化模型[2]针对二值化图像，模拟了图像模糊抖动的效果，具体步骤如下：

① 依据退化模型：

$$\begin{cases}p(0|1)=\alpha_0e^{-\alpha d^2}+\eta\\p(1|0)=\beta_0e^{-\beta d^2}+\eta\end{cases}$$

该公式指出，距离前景（文字）边缘越近，越容易发生像素偏转，造成图像失真。

② 将像素翻转后的图像以参数 k 为结构元素尺寸进行闭运算。

图像退化结果如图 4 所示：

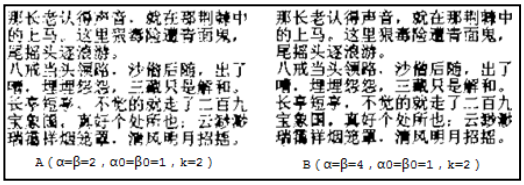

图 4 文档局部模型退化结果

4.2 旋转

基于图像变换矩阵，取原图像（PDF 页面截图），以其中心为原点进行旋转，并同时根据旋转角度，调整相应标引信息基准值。

4.3 斑点

在含噪声图像中，可能出现不规则分布的斑点，会在一定程度上影响识别精度。本系统中，依据[3]中的方法，斑点的生成的采用随机分布的思想。

4.4 模糊

由于模糊对识别精度有很大影响[8]，本系统采用了高斯滤波方法对原始理想化图像进行卷积处理，模拟模糊效果。

模糊、旋转、添加斑点的结果如图 5 所示：

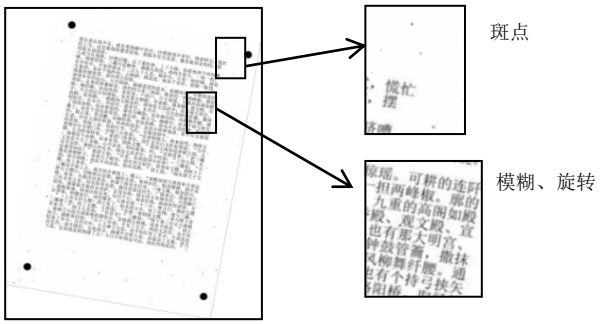

图 5 模糊、旋转、添加斑点退化效果

4.5 不规则线条

由于打印机老化等原因，会在页面中造成一定程度上的不规则分布的横竖线条纹。本系统基于此，模拟了图像中的横竖线退化模型，其中，线条数、线条长宽、黑色像素密度等由用户控制。

4.6 抖动

抖动用于模拟在打印、扫描过程中出现的干扰。本系统依据[3]中的方法，设定像素抖动半径 R，即使用 $X(u,v)=X(u\pm r_1, v\pm r_2)$。$(r_1, r_2 \in [0,R])$。该模型即表示在 R 尺寸的窗口内，某像素点随机选取临近点作为新的像素值，模拟抖动效果。

抖动、添加不规则线条的结果如图 6 所示：

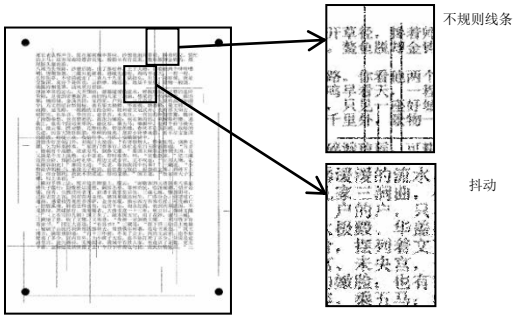

图 6 抖动、添加不规则线条退化结果

4.7 渗透（bleed-through）

在进行复印、扫描工作时，经常会由于纸张反面的油墨透到另一面而影响了另一面的画面清晰度。在本系统中，将两张图像在经过一定模糊处理之后进行反向叠加，模拟渗透[3]的效果。详细模型如下：

$$M_{u,v} = \begin{cases} p_{u,v} - \alpha(p_{u,v} - b_{u,v}) & \text{若 } p_{u,v} - b_{u,v} > threshold \\ p_{u,v} \end{cases}$$

其中$p_{u,v}$指正面图像在(u,v)点的像素值，$b_{u,v}$为反面进行 X 轴方向翻转、模糊操作后的图像在(u,v)点的像素值。

渗透退化结果如图 7 所示：

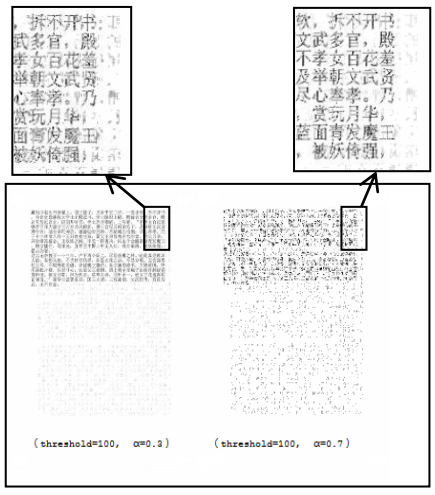

图 7 渗透退化效果

5 实验结果

为了测试基准生成平台的性能，我们采用了中国的四大经典名著小说《红楼梦》、《西游记》、《三国演义》、《水浒传》进行了测试。

我们首先得到了这些小说的纯文本，然后基于我们的系统平台转化为 PDF 格式的文档，并利用 PDF 的 SDK 接口提取了文字的坐标信息。将 PDF 文档经过打印、扫描得到了实际的扫描文档图像。经过配准后，得到实际扫描文档的基准信息。此外，还基于各种退化模型对 PDF 的截图（理想化的图像）进行了各种退化效果的测试。结果显示，实际扫描图像的配准精度很高，误差为 1 个像素点。一个实际的扫描文档配准的例子如图 8 所示。图 8 为实际的扫描文档图像及相应的坐标信息。从该图可见，实际扫描文档的字符信息的坐标非常准确，且有很好的可视化效果和交互方式。在该平台中，文字的标引信息部分采用 XML 格式，其标引信息与图像内容进行了关联，可以非常便捷的查看配准的效果。

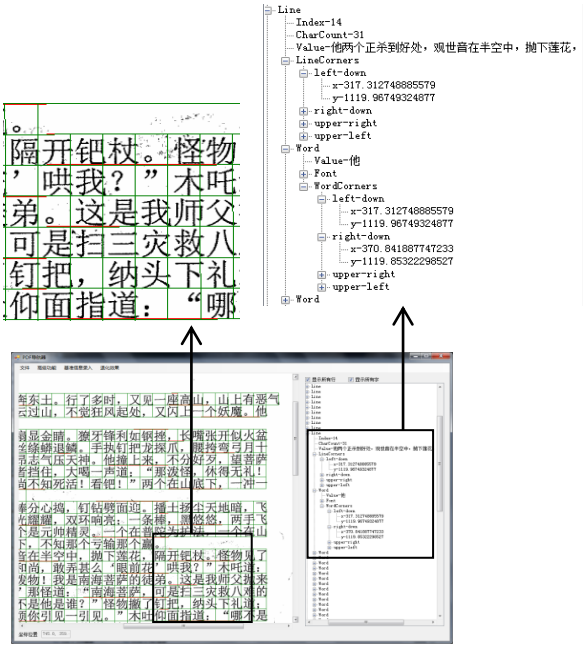

图 8 实际扫描文档图像基准信息标引结果

结束语

本系统基于 Visual C#平台进行开发，利用了 PDF 的 SDK，基于 PDF 格式的无噪声文档，采用透视变换模型，创建了文档图像库及其标引信息，经实验验证，有很高的精度。

为了满足大规模文档识别或检索的自动化性能评估，本论文提供的文档基准生成系统可以根据用户的设定，生成各种不同类型的文档图像及基准信息。由于实际扫描文档难以做到大规模级，本系统还提供了几种常用的退化模型及退化的文档和基准信息，从而为进行大规模的识别性能或检索性能的评测提供了丰富可靠的测试数据。下一步，还将继续针对各种退化模型，研究相应的退化复原方法。

参考文献:

[1] H.S. BAIRD. The State of the Art of Document Image Degradation Modeling, 2000[C]: Riode Janeiro, Brazil: Proc. of 4th IAPR International Workshop on Document Analysis Systems, 2000: 1-16.

[2] Tapas KANUNGO. Document Degradation and Methodology for Degradation Model Validation [D]. Seattle: University of Washington, 1996:

56-68
[3] ZI Gang. GROUNDTRUTH GENERATION AND DOCUMENT IMAGE DEGRADATION [D]. Maryland: University of Maryland, 2005: 49-57
[4] T. KANUNGO and R. M. HARALICK. An Automatic Closed-Loop Methodology for Generating Character Groundtruth for Scanned Documents[J], IEEE Transactions on Pattern Analysis and Machine Intelligence , February 1999, vol. 21, no. 2: pp. 179-183.
[5] Carlos A. B. MELLO, Rafael D. LINS. Generation of images of historical documents by composition, 2002[C]. McLean, Virginia, USA: ACM Symposium on Document Engineering, 2002:1-7
[6] G. BAL, G. AGAM, O. FRIEDER, G. FRIEDER. Interactive degraded document enhancement and ground truth generation, 2008[C]: San Jose, CA, USA: Document Recognition and Retrieval XV, 2008: 1-9.
[7] J. ZHAI, W. LIU etc. A line drawing degradation model for performance characterization, 2003[C]. Edinburgh, Scotland: Seventh International Conference on Document Analysis and Recognition, 2003: 618-622.
[8] T.K. HO, Henry S. BAIRD. Evaluation of OCR Accuracy Using Synthetic Data, 1995[C]. Las Vegas, U.S: Symposium on Document Analysis and Information Retrieval, 1995: 413-422