第 33 卷第 5 期　　河 北 科 技 大 学 学 报　　Vol. 33, No. 5
2012 年 10 月　　Journal of Hebei University of Science and Technology　　Oct. 2012

文章编号：1008-1542(2012)05-0434-05

# 基于马尔科夫链模型的论文格式审查系统

唐心亮[1]，王　靖[2]，王震洲[3]

(1. 河北科技大学人事处，河北石家庄　050018；2. 唐山师范学院计算机科学系，河北唐山 063000；3. 河北科技大学信息科学与工程学院，河北石家庄　050018)

**摘　要**：提出一种基于图像处理审查论文格式的方法，该方法依据论文电子版文档图像像素点间的相关性，应用马尔科夫链模型分割文档图像为正文、标题和图片部分，在人工设定的论文格式规则基础上，对论文的每页图像进行分类审查，可有效提高论文格式审查效率。

**关键词**：论文格式审查；图像分割；马尔科夫链；格式规则

**中图分类号**：TP392　　**文献标志码**：A

# Examination of paper format based on Markov-Chain model

TANG Xin-liang[1], WANG Jing[2], WANG Zhen-zhou[3]

(1. Personnel Department, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. Department of Computer Science, Tangshan Normal University, Tangshan Hebei 063000, China; 3. College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

**Abstract**: A method of examining paper format based on image segmentation was proposed. According to the relevance of pixels in the image of electronic paper, paper image was segmented into title, text and pictures by using Markov-Chain model. And every paper image can be examined in accordance with segmentation followed by manually setting the paper format rules. The test results indicate that it is helpful to improve the efficiency of examining paper format.

**Key words**: examination of paper format; image segmentation; Markov-Chain; format rules

随着中国高等教育进入大众化阶段，普通本科生、研究生以及各类学位的攻读人数逐年增加，论文格式审查工作也逐年繁重。格式审查费时费力，然而，目前国内外尚无具体的论文格式审查系统的研究和应用成果。近年来，图像处理技术在各领域得到广泛应用，笔者结合该技术提出一种审查论文格式的方法。

由于不同学校的论文格式要求各异，论文不同部分的格式要求也不尽相同，因此论文审查难度较大。为了提高论文检测的速度和准确性，同时满足论文审查技术的适应性，笔者提出的审查论文格式的技术主要包括 2 部分。1)图像分割：用马尔科夫链模型对被测论文电子版文档进行分割，分割出每页图像的正文、标题、图片部分。2)格式审查：手工设定的论文格式规则，在此基础上提取不同分割区域中文档图像对应的特征值，不同的区域使用不同的格式审查方法，分割出来被测论文进行分类审查，若不满足其对应的审查方法，表示未通过审查，并使用红色标记。该方法不仅能满足不同论文规则的要求，而且有效地提高了论文审查效率和准确性。论文格式审查流程如图 1 所示。

收稿日期：2012-05-28；修回日期：2012-09-06；责任编辑：李　穆
基金项目：河北省自然科学基金资助项目(F2012208004)；河北科技大学校立基金资助项目(XL201027)
作者简介：唐心亮(1977-)，男，河北成安人，讲师，博士研究生，主要从事计算机应用方面的研究。

## 1　基于马尔科夫链图像分割

### 1.1　图像预处理

论文原始图像在采集过程中会引入噪声，减弱了论文图像中的信息，影响图像分割和格式审查效果，针对该问题笔者采用均值滤波和中值滤波的方法对图像进行去噪处理，提高图像信噪比。但是在图像的去噪过程中会平滑原始图像的边缘，文献[1]中提出了基于二维小波变换的图像增强算法，笔者结合该算法实现对去噪后的论文文档图像的增强，为后续图像分割和格式审查奠定基础。

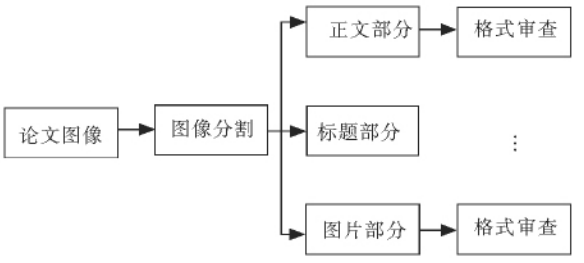



图 1　论文格式审查流程图

Fig. 1　Flow chat of paper format examination

### 1.2　马尔科夫链

经过图像预处理之后的论文文档图像，应用马尔科夫链分类器实现图像的分割。马尔科夫链(Markov-Chain)[2]是指具备系统在将来发生某件事的条件概率与其过去发生的事件无关，只与系统的当前状态相关的随机过程。如果随机过程$\{(X_t),t\in T\}$，其中，时间集合 $T=0,1,2,\cdots$。设定 $i$ 对应 $t$ 时刻随机过程 $X_t$ 的状态，即 $X_t=i$，此时 $X_{t+1}$ 在时刻 $t+1$ 的状态 $j$ 的概率分布 $P_{ij}$ 只与 $X_t$ 在前一时刻 $t$ 的状态 $i$ 有关，即有

$$P\{X_{t+1}=j,|X_0=k_0,X_1=k_1,\cdots,X_{t-1}=k_{t-1},X_t=i\}=P\{X_{t+1}=j|X_t=i\},\tag{1}$$

对于 $t=0,1,2,\cdots$和每一序列 $i,j,k_0,\cdots,k_{t-1}$均成立，则此随机过程$\{X_t\}(t=0,1,2,\cdots)$被称为马尔科夫链。马尔科夫链有稳定的转移概率，即

$$P_{ij}=P\{X_{t+1}=j|X_t=i\};\tag{2}$$

$$P_{ij}^{(n)}=P\{X_{t+n}=j|X_t=i\}。\tag{3}$$

概率转移矩阵为

$$\boldsymbol{P}=\begin{vmatrix}P_{00} & P_{01} & \cdots & P_{0i}\\ P_{10} & P_{11} & \cdots & P_{1i}\\ \vdots & \vdots & \vdots & \vdots\\ P_{i0} & P_{i1} & \cdots & P_{ii}\end{vmatrix},\tag{4}$$

满足 $P_{ij}\geqslant 0,n=0,1,2,\cdots$，且 $\sum_{j=0}^{\infty}P_{ij}^{(n)}=1,n=0,1,2,\cdots$。

### 1.3　论文图像分割

将采集到的论文电子版文档图像视为是 1 个向量的集合，每个像素点将对应某个向量的分量，2 个像素点间的相关性则可使用某种条件概率来描述，1 页论文图像就可视作多个满足条件概率的连续状态的集合，因此采用马尔科夫链模型(Markov-Chain 模型)[3] 进行图像分割的方法是可行的。

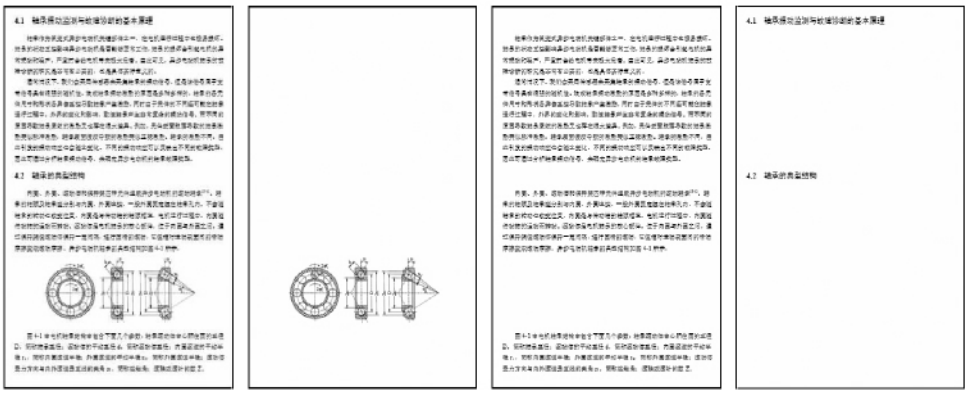
由于每页论文文档图像的标题、正文和图片等的位置不固定，计算对论文文档图像进行蛇形扫描，得到论文文档图像向量 $\boldsymbol{Y}$，$\boldsymbol{Y}=(y_1,y_2,\cdots,y_i,\cdots,y_n)$，其中 $y_i,y_{i+1}$ 代表相邻的 2 个像素点，使用 $n_{ij}(Y)$ 来表征相邻的像素点从值 $i$ 到 $j$ 跳变的次数，则可得到其跳变概率为 $P(y_i\rightarrow y_j)=P(n_{ij})$，令 $P_{ij}=P(n_{ij})$[4]，即可计算出文档图像像素点跳变的概率分布矩阵 $\boldsymbol{P}$。

$$\boldsymbol{P}=\begin{vmatrix}P_{00} & P_{01} & \cdots & P_{0n}\\ P_{10} & P_{11} & \cdots & P_{1n}\\ \vdots & \vdots & \vdots & \vdots\\ P_{n0} & P_{n1} & \cdots & P_{nn}\end{vmatrix}。\tag{5}$$

利用监督学习的方法，通过大量的论文文档图像对马尔科夫链模型进行训练，分割出图像中正文、标题和图片几部分。

### 1.4　论文图像分割结果

根据上述对论文文档图像分割方法，将一篇待检测文档图像进行分割，分割结果如图 2 所示，其中图 2a)为待测文档的原图像，图 2b)、图 2c)、图 2d)分别为使用马尔科夫链分类器分割原图像后对应的图片部分、正文区域、标题部分。图像分割的正确结果为下一步格式审查奠定了基础。



a) 原图 b) 图片部分 c) 正文区域 d)标题部分

图2 论文图像分割结果

Fig. 2 Results of paper image segmentation

## 2 格式审查技术

分别将前面分割出来的结果进行审查。论文格式可以根据要求手动输入到系统中，系统根据不同的格式规则要求，对相应的待测文档进行审查，这里以前面分割出来的正文图像部分为例，进行格式审查。

### 2.1 图像二值化

由于笔者采用的格式审查算法针对灰度图像，在特征提取前，首先将正文图像进行图像二值化[5]，采用最大熵法将正文图像按照灰度级划分为前景和背景2个不同区域。计算正文图像中不同灰度级的概率 $p(x)$，灰度级的熵为 $H=-\sum_{i=1}^{n}P_i\ln P_i$。图像二值化就是以灰度级 $T$ 分割图像，像素点的灰度级低于 $T$ 灰度级视为目标物体O，像素点的灰度级高于 $T$ 视为背景B，则灰度级在本区的分布概念如下。

目标物体区 O：$P_i/P_t, i=1,2,\cdots,t$； (6)

背景区 B：$P_i/(1-P_t), i=t+1,t+2,\cdots,L-1$。 (7)

正文图像中目标物体区的熵为

$$H_{\mathrm{O}}=-\sum_{i=1}(P_i/P_t)\ln(P_i/P_t),\quad i=1,2,\cdots,t。 \tag{8}$$

正文图像中背景区的熵为

$$H_{\mathrm{B}}=-\sum_{i}[P_i/(1-P_t)]\ln[P_i/(1-P_t)],\quad \text{其中 } i=t+1,t+2,\cdots,L-1。 \tag{9}$$

对于图像中每个灰度级分别求取 $w=H_{\mathrm{O}}+H_{\mathrm{B}}$，选取使 $w$ 最大的灰度级作为阈值。

### 2.2 字符切分

论文正文格式设置不同必然影响字的宽度和高度，因此将处理后正文图像进行再次分割，切分出正文图像中的行信息，进而分割出每行字符，然后比对该正文格式规则下的字符高度和宽度，实现对正文文本格式的审查。

1) 行切分

对正文图像水平积分投影，利用文本行间存在的空白区域，实现行切分[6]，假设文档图像中第 $i$ 行、第 $j$ 列的像素值为 $g(i,j)$，则第 $i$ 行水平方向各项上的积分投影为 $\sum_{j=1}^{L}g(i,j)$，其中 $L$ 为行的长度。行间由于噪声的影响，水平积分投影不为0，因此，文字与行空隙间的水平积分将出现一个锐减或锐增，利用该性质可切分行。

各行的积分投影：$F(i)=\sum_{j=1}^{L}g(i,j)$，各行积分的平均值 $\mathrm{Avg}=\sum_{i=1}^{n}F(i)/N$，其中 $N$ 为文字图像的行数。

从下往上依次搜索尚未切分的文本，直到第1个满足下列2个条件的像素行 $i$ 时，则第 $i$ 行为文本行的下界。

有连续 $n$ 行满足

$$(F(i)>\mathrm{Avg}/p)\cap(F(i+1)>\mathrm{Avg}/p)\cap\cdots\cap(F(i+n-1)>\mathrm{Avg}/p),\tag{10}$$

从第 $i-n+1$ 行到第 $i$ 行中至少有 1 行满足 $F(k)<\mathrm{Avg}/p$，其中 $i-n+1\leqslant k\leqslant i$ 对尚未切分的文本按自上而下的顺序依次搜索，当搜索到第 1 个满足下列 2 个条件的像素行 $i$ 时，第 $i$ 行为文本行的上界。

有连续 $m$ 行满足

$$(F(i)>r)\cap(F(i+1)>r)\cap\cdots\cap(F(i+m-1)>r),\tag{11}$$

从第 $i-m+1$ 行到第 $i$ 行中至少有 1 行满足 $F(k)>t$，其中 $i-m+1\leqslant k\leqslant i$。

2) 字切分

上述行切分分出文本的行信息，以分割出的行文本的垂直投影为基础，自左向右搜索汉字，确定其两边边界，汉字、标点符号使用最大度回溯字切分算法切分，根据汉字的方块性特点，切分单字，该方法不仅可以实现，而且还具有自适应性[7]。设定行数为 $I$ 的正文文本，字宽 $W=\frac{1}{I}\sum_{i=1}^{I}(i_b-i_a+1)$，其中 $i_b$ 为第 $i$ 行的上界，$i_a$ 为该行的下界值。设定 $W_m$ 表征文字的最大宽度，$d$ 表征回溯范围，$j_a$ 表征第 $j$ 个字的左边界，切分步骤如下。

步骤 1：在 $j_a\leqslant j\leqslant j_b+W_m$ 范围内计算第 1 个 $\sum_{j=1}^{L}g(i,j)=0$ 的点。

步骤 2：在 $j_a+W_m-d\leqslant j\leqslant j_a+W_m$ 范围内求使得 $\sum_{j=1}^{L}g(i,j)$ 取最小值得点 $j_b$，则 $j_b$ 就是第 $j$ 字的右边界。

步骤 3：从 $j_b$ 处向右，计算 $\sum_{j=1}^{L}g(i,j)$，当 $\sum_{j=1}^{L}g(i,j)\neq 0$ 时，设为 $j_a'$，若满足 $j_a'>j_b$，则 $j_a'$ 就是第 $j+1$ 个文字的左边界，重复以上步骤即切分出来每行文字。

### 2.3　正文格式审查

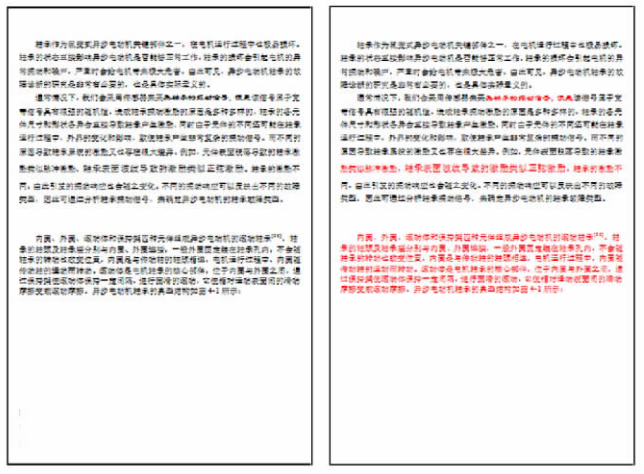
同一个字在不同正文格式设置时的高度是不一样的，例如字体格式的设置，即使是同一字号设置，字形不一致，该字的宽度、高度也不尽相同，因此可以通过比较每个字符的高度和宽度，来判定该字是否符合正文格式要求。经过训练得到该字正文格式下的平均高度和宽度的参考值 $H,G$。系统需要对该格式下正确的文本进行训练，得到该格式下字符高度、宽度的波动范围值 $\Delta H,\Delta G$。设文字的高度的极值分别为 $H_{\max}=H+\Delta H,H_{\min}=H-\Delta H$；设文字的宽度的极值分别为 $G_{\max}=G+\Delta G,G_{\min}=G-\Delta G$。设定第 $i$ 行第 $j$ 个字的左边界为 $j_a$，右边界为 $j_b$，上边界为 $h_a$，下边界为 $h_b$，若该字符满足 $(G_{\min}\leqslant j_b-j_a\leqslant G_{\max})\,\&\&\,(H_{\min}\leqslant h_b-h_a\leqslant H_{\max})$，则认定该字符满足格式要求，可通过格式审查；若不满足上述条件，则将该正文部分字体颜色设为红色输入，表示未通过该规则下的格式审查。

## 3　实验结果

以河北科技大学学位论文正文格式要求为例，在系统中手工输入格式要求，字体为小四号宋体，首行缩进 2 个字符，行距最小值为 20 磅。设定格式之后，对该规则下的正确文档图像进行训练，得到特征值 $G_{\min}$，$G_{\max}$，$H_{\min}$，$H_{\max}$，然后对待测文档进行格式审查，审查结果如图 3 所示，其中图 3a)为待测文档图像，文档中部分正文格式不符合该格式要求，图 3b)为格式审查输出结果，不符合格式要求的文档部分使用红色(图中虚字)进行标记。

## 4　结　论

提出一种基于图像处理的论文格式审查技术，将电子版的论文图像采集到系统中，利用文档图像相邻像素点之间的相关性，结合马尔科夫链分类器对每页论文文档图像进行分割，得到图像中的图片、正文、标题部分。手工设定论文中的格式要求，提取此格式下论文中的特征值，审查被测论文格式，不仅提高了论文检测的速度和正确性，而且具有适应性。



a) 待测文档图像 b) 审查结果

图 3 正文文档图像格式审查结果

Fig. 3 Results of examination of text format

## 参考文献：


[1] 杨 静. 基于小波变换的低对比度图像增强方法[J]. 计算机时代(Computer Era),2011(1):10-12.

[2] 刘绍辉,孙建超,姚鸿勋. 一种改进的基于马尔科夫链的扩频图像隐写分析方法[J]. 中国科学院研究生院学报(Journal of the Graduate School of the Chinese Academy of Science),2011, 28(5):690-695.

[3] 宋锦萍,侯玉华,杨晓艺,等. 基于小波域多状态隐马尔科夫树模型的自适应正文图像分割算法[J]. 电子学报(Chinese Journal of Electronices),2007,35(1):118-122.

[4] 杜新宇,刘光耀. 基于马尔科夫链的光侧图像自动判读方法[J]. 计算机工程与应用(Computer Engineering and Applications),2008,44(28):246-248.

[5] 常丹华,何耘娴,苗 丹. 中英混排文档图像粘连字符分割方法的研究[J]. 激光与红外(Laser & Infrared),2010,40(12):1 369-1 373.

[6] 许伦辉,陈衍平,修科鼎. 基于图像处理的静态车牌识别技术[J]. 江西理工大学学报(Journal of Jiangxi University of Science and Technology),2011,23(1):47-50.

[7] 杨 霈. 基于小波分析的字符图像分割技术[J]. 太原科技大学学报(Journal of Taiyuan University of Science and Technology),2007,28(4):288-290.

[8] 韩立华,王学军,王晓芬. 多特征融合及 SVM 相关反馈技术在教育资源图像检索中的应用[J]. 河北科技大学学报(Journal of Hebei University of Science and Technology),2010,31(3):240-244.

[9] 杨丽娟,刘教民,王震洲,等. 基于分块帧差的视频图像运动检测[J]. 河北科技大学学报(Journal of Hebei University of Science and Technology),2006,27(1):89-92.


# 向本期载文的审稿专家致谢

**本期《河北科技大学学报》共发表论文 21 篇。这些论文的发表是与有关专家的认真审读、细查资料、推敲分析、中肯评价分不开的。对此，本编辑部特向这些专家表示敬意，对他们的辛勤劳动表示感谢。**

**本期载文的审稿专家名单如下(按姓名的汉语拼音顺序排列)：**

郭宪民 郭彦平 李国庭 李新福 李艳艳
李永刚 刘 英 孟范平 史兰香 苏连青
王志新 魏世泽 杨志荣 翟学良 张 宁

**(本刊编辑部)**