(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111783416 A
(43)申请公布日 2020.10.16

(21)申请号 202010511448.5

(22)申请日 2020.06.08

(71)申请人 青岛科技大学
地址 266000 山东省青岛市崂山区松岭路99号

(72)发明人 许灿辉 史操 孙春奇 陶冶 刘国柱 程远志

(74)专利代理机构 青岛中天汇智知识产权代理有限公司 37241
代理人 王丹丹 刘晓

(51)Int.Cl.
G06F 40/189(2020.01)

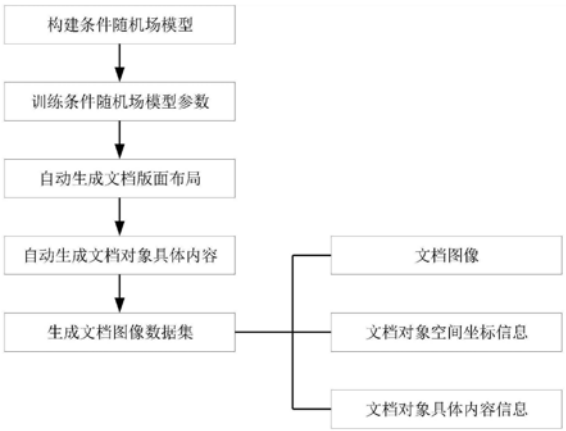
权利要求书2页 说明书10页 附图4页

(54)发明名称

一种利用先验知识构建文档图像数据集的方法

(57)摘要

本发明公开一种利用先验知识构建文档图像数据集的方法，首先采用条件随机场对文档版面进行建模；然后训练条件随机场模型参数，获取文档版面先验知识；使用训练好的模型自动生成文档版面布局；进而根据生成的版面布局，生成文档对象的具体内容；最终将文档转换成文档图像，实现对文档图像数据集的构建。本方案基于条件随机场对文档版面结构进行建模，获取文档版面的先验知识，并利用先验知识自动生成文档图像数据集，节省时间和人力成本，避免由于人工标注带来的无效标注；通过条件随机场及先验知识指导生成文档图像的版面结构，使生成的版面更贴近出版物、符合阅读习惯，提高数据集的利用率及深度学习精度；并且，生成文档图像集的同时，能够提供文档图像中文本对象的文字编码信息。



CN 111783416 A

1.一种利用先验知识构建文档图像数据集的方法,其特征在于,包括以步骤:

步骤A、将文档版面信息刻画为文档对象空间、彩色直方图空间和文档对象类型空间,且文档对象空间元素与彩色直方图空间元素一一对应,彩色直方图空间元素与文档对象类型空间元素一一对应;

基于条件随机场对文档版面结构进行建模,得到文档对象彩色直方图序列X与文档对象类型标签序列Y的线性链条随机场;

步骤B、采集各种已有文档页面数据,训练条件随机场模型参数,对条件随机场权值进行求解;

步骤C、基于高斯混合模型随机生成彩色直方图序列X,根据训练好的模型求解文档对象类型标签序列Y,自动生成文档版面布局;

步骤D、采集各种已有文档对象数据,根据生成的文档版面布局,生成文档对象的具体内容;

步骤E、将步骤D生成的文档转换成文档图像,构建文档图像数据集,所述文档图像数据集包含文档图像、文档对象空间坐标信息和文档对象具体内容信息。

2.根据权利要求1所述的利用先验知识构建文档图像数据集的方法,其特征在于:所述步骤A中在对文档版面结构进行建模时,具体采用以下方式:

(1)确定文档页面中的对象序列DO_i,计算每一个文档对象对应的彩色直方图,确定与彩色直方图序列X对应的文档对象类型标签序列Y;

(2)将$X=[x_1,x_2,x_3,...x_N]$和$Y=[y_1,y_2,y_3,...y_N]$均视为随机变量序列,在给定随机变量序列X的情况下,随机变量序列Y的条件概率分布P(Y|X)便构成条件随机场,进而可以得到:

$$P(Y|X)\propto \exp(g(Y|X)) \qquad (15)$$

且:

$$\sum_{\mathbf{Y}'=1}^{N\times M} P(\mathbf{Y}|\mathbf{X})=1 \qquad (14)$$

其中,N为文档对象的数量,M为标签类型的数量;

$$g(Y|X)=\sum_i\sum_{j,j'}\lambda_{j,j'}f_{j,j'}(X,y_{i-1},i)=\sum_i\sum_{j,j'}\lambda_{j,j'}f_{j,j'}(y_i) \qquad (12)$$

$$f_{j,j'}(y_i)=f_{j,j'}(X,y_{i-1},i) \qquad (9)$$

其中,特征函数族$f_{j,j'}(y_i)$第一个脚标j表示当前节点y_i所属的类型,第二个脚标j′表示前一个节点y_{i-1}所属的类型,$f_{j,j'}(y_i)$表示节点y_i的特征函数,$\lambda_{j,j'}$表示权值,特征函数的个数以及特征函数值即为先验知识。

3.根据权利要求2所述的利用先验知识构建文档图像数据集的方法,其特征在于:所述步骤B中,将$\lambda_{j,j'}$假定为相互独立,并定义对数似然函数:

$$L(\lambda)=\log\prod_{\mathbf{X},\mathbf{Y}}\exp\left(g(\mathbf{Y}|\mathbf{X})\right) \qquad (16)$$

式(16)对待求权值求导,寻找驻点:

$$\frac{\partial L(\lambda)}{\partial\lambda}=0 \qquad (17)$$

对条件随机场权值$\lambda_{j,j'}$进行求解时,具体采用以下方式:

(1)基于文档解析工具解析采集的文档页面数据,获得X序列和Y序列;

(2)根据Y序列样本,设置文档对象类型标签序列的标签类型;

(3)设定函数族$f_{j,j'}(y_i)$中的函数特征函数值;

(4)基于优化算法求解式(17),进而得到条件随机场权值$\lambda_{j,j'}$。

4.根据权利要求1所述的利用先验知识构建文档图像数据集的方法,其特征在于:所述步骤C自动生成文档版面布局的方式如下:

步骤C1、基于高斯混合高斯模型

$$GMM(\alpha)=\sum_{k=1}^{S}\phi_k\frac{1}{\sqrt{2\sigma_k^2}}\exp\left[-(\alpha-\mu_k)/2\sigma_k^2\right] \qquad (18)$$

$$\sum_{k=1}^{S}\phi_k=1 \qquad (19)$$

随机生成序列X中各节点的红色、绿色、蓝色三个颜色通道的直方图,以表征文档对象,其中μ_k和σ_k^2采用随机数自动生成,进而得到X序列;

步骤C2、基于彩色直方图空间元素与文档对象类型空间元素一一对应,对Y序列进行求解,即自动生成文档版面布局。

5.根据权利要求4所述的利用先验知识构建文档图像数据集的方法,其特征在于:所述步骤C1中生成X序列的过程具体如下:

(1)设定序列X的节点数量N;

(2)为序列X的每个节点x_i的每个颜色通道的颜色直方图分别设定参数S;

(3)基于随机数生成算法设置参数μ_k和σ_k^2;

(4)基于高斯混合模型生成序列X。

6.根据权利要求1所述的利用先验知识构建文档图像数据集的方法,其特征在于:所述步骤D中,具体采用以下方式:

步骤D1、收集数据集:

Set_j,j=1,2,3...M (20)

其中,Set_1,Set_2,...Set_M=文本集,公式集,...页码集;

步骤D2、基于步骤C生成的Y序列采用TeX标记语言及数据集生成文档对象具体内容。

7.根据权利要求6所述的利用先验知识构建文档图像数据集的方法,其特征在于:所述步骤D2中,文档对象具体内容的生成过程采用自顶向下的方法:

(1)首先生成页眉;

(2)生成页面中的栏目数;

(3)从第一栏开始,根据Y序列、以及定义的数据集生成页面对象DO_i,同时记录DO_i的空间坐标信息以及DO_i的具体内容信息;

(4)若文档不止一栏,则当第一栏结束后继续生成第二栏,直至最后一栏结束;

(5)当文档中所有栏目均生成后,生成页脚、页码。

8.根据权利要求1所述的利用先验知识构建文档图像数据集的方法,其特征在于:所述步骤A中,所述文档对象类型标签序列所包含的标签类型包括但不限于文本、公式、图、图名、表、表名、页眉、页脚和页码。

一种利用先验知识构建文档图像数据集的方法

技术领域

[0001] 本发明涉及一种图像生成方法，属于图像数据集构建领域，具体涉及一种利用先验知识构建文档图像数据集的方法。

背景技术

[0002] 在文档图像处理的诸多领域中，如分割、分类、检索等领域，带标记的文档图像集是机器学习过程中不可或缺的数据基础，尤其是近年来，深度学习在基于大数据的人工智能领域中成为了重要研究方法，与传统的机器学习相比，深度学习需要更多训练数据。

[0003] 目前，文档图像数据集一方面来自人工标注，例如：牛津大学机器人研究组(Robotics Research Group)设计的图像标注工具VIA(“Abhishek Dutta and Andrew Zisserman.2019.The VIA Annotation Software for Images,Audio and Video.In Proceedings of the 27th ACM International Conference on Multimedia(MM'19),October 21-25,2019,Nice,France.ACM,New York,NY,USA.”，使用VIA工具可以使用不同形状(矩形、圆、椭圆、多边形，等等)对图像区域进行手工标注。

[0004] 另外，也有采用自动生成的方法获取文档图像及标注信息，如2017年文档分析与识别国际会议(International Conference on Document Analysis and Recognition,ICDAR)上的论文(【2】D.He,S.Cohen,B.Price,D.Kifer and C.L.Giles,"Multi-Scale Multi-Task FCN for Semantic Page Segmentation and Table Detection")中将段落、图、表格、标题、段落标题、列表等等元素进行随机排列生成文档图像数据集，用于深度学习训练。同样，申请公布号为【CN 108898188 A】的发明专利也公开一种图像数据集辅助标记系统及方法，利用神经网络训练的思想对神经网络训练所需的图像进行初步特征提取训练，对图像进行识别标记获得神经网络所需的标签文档格式，在大量的图像信息中获得某一类的标签文档。

[0005] 对于人工标注而言，虽然其具有很强灵活性，标注过程中可以弹性更改标注策略，标注结果能够较好地契合预期，但是，其缺点也是显然的，即标注过程费时、人力成本高昂，而且标注质量与标注人员的熟练程度成正比；相对于人工标注，文档图像数据集自动生成方法可以较好地克服人工标注的不足，但是也存在不可避免的问题，比如，出版业具有自身的行业规范，不同出版物的版面设计也遵循特定的规律，通过这些规律更好地展示文档内容，若随机生成的文档图像不能很好地契合出版物的排版规律，使得训练出来的模型应用于真实出版物文档图像时，不能体现模型的最佳性能。

[0006] 因此，本发明旨在对出版物版面的客观规律进行建模，从而更有利于机器学习中模型的训练，在机器学习领域，这种已经存在的客观规律，也被称作“先验知识”。

发明内容

[0007] 本发明针对现有获得文档图像数据集存在的缺陷，提出一种利用先验知识构建文档图像数据集的方法，基于条件随机场对文档版面结构进行建模，并利用先验知识自动生

成文档图像数据集，可有效节省时间和人力成本，避免由于人工标注带来的无效标注。

[0008] 本发明是采用以下的技术方案实现的：一种利用先验知识构建文档图像数据集的方法，包括以步骤：

[0009] 步骤A、将文档版面信息刻画为文档对象空间、彩色直方图空间和文档对象类型空间，且文档对象空间元素与彩色直方图空间元素一一对应，彩色直方图空间元素与文档对象类型空间元素一一对应；

[0010] 基于条件随机场对文档版面结构进行建模，得到文档对象彩色直方图序列X与文档对象类型标签序列Y的线性链条随机场；

[0011] 步骤B、采集各种已有文档页面数据，训练条件随机场模型参数，对条件随机场权值进行求解；

[0012] 步骤C、基于高斯混合模型随机生成彩色直方图序列X，根据训练好的模型求解文档对象类型标签序列Y，自动生成文档版面布局；

[0013] 步骤D、采集各种已有文档对象数据，根据生成的文档版面布局，生成文档对象的具体内容；

[0014] 步骤E、将步骤D生成的文档转换成文档图像，构建文档图像数据集，所述文档图像数据集包含文档图像、文档对象空间坐标信息和文档对象具体内容信息。

[0015] 进一步的，所述步骤A中在对文档版面结构进行建模时，具体采用以下方式：

[0016] (1)确定文档页面中的对象序列DO_i，计算每一个文档对象对应的彩色直方图，确定与彩色直方图序列X对应的文档对象类型标签序列Y；

[0017] (2)将$X=[x_1,x_2,x_3,...x_N]$和$Y=[y_1,y_2,y_3,...y_N]$均视为随机变量序列，在给定随机变量序列X的情况下，随机变量序列Y的条件概率分布$P(Y|X)$便构成条件随机场，进而可以得到：

[0018] $$P(Y|X)\propto\exp(g(Y|X)) \quad (15)$$

[0019] 且：

[0020] $$\sum_{\mathbf{Y}'=1}^{N\times M}P(\mathbf{Y}|\mathbf{X})=1 \quad (14)$$

[0021] 其中，N为文档对象的数量，M为标签类型的数量；

[0022] $$g(Y|X)=\sum_i\sum_{j,j'}\lambda_{j,j'}f_{j,j'}(X,y_{i-1},i)=\sum_i\sum_{j,j'}\lambda_{j,j'}f_{j,j'}(y_i) \quad (12)$$

[0023] $$f_{j,j'}(y_i)=f_{j,j'}(X,y_{i-1},i) \quad (9)$$

[0024] 其中，特征函数族$f_{j,j'}(y_i)$第一个脚标j表示当前节点y_i所属的类型，第二个脚标j′表示前一个节点y_{i-1}所属的类型，$f_{j,j'}(y_i)$表示节点y_i的特征函数，$\lambda_{j,j'}$表示权值，特征函数的个数以及特征函数值即为先验知识。

[0025] 进一步的，所述步骤B中，将$\lambda_{j,j'}$假定为相互独立，并定义对数似然函数：

[0026] $$L(\lambda)=\log\prod_{\mathbf{X},\mathbf{Y}}\exp\left(g(\mathbf{Y}|\mathbf{X})\right) \quad (16)$$

[0027] 式(16)对待求权值求导，寻找驻点：

[0028] $$\frac{\partial L(\lambda)}{\partial\lambda}=0 \quad (17)$$

[0029] 对条件随机场权值$\lambda_{j,j'}$进行求解时，具体采用以下方式：

[0030] (1)基于文档解析工具解析采集的文档页面数据,获得X序列和Y序列;

[0031] (2)根据Y序列样本,设置文档对象类型标签序列的标签类型;

[0032] (3)设定函数族$f_{j,j'}(y_i)$中的函数特征函数值;

[0033] (4)基于优化算法求解式(17),进而得到条件随机场权值$\lambda_{j,j'}$。

[0034] 进一步的,所述步骤C自动生成文档版面布局的方式如下:

[0035] 步骤C1、基于高斯混合高斯模型

[0036] $$GMM(\alpha)=\sum_{k=1}^{S}\phi_k\frac{1}{\sqrt{2\sigma_k^2}}\exp\left[-(\alpha-\mu_k)/2\sigma_k^2\right] \quad (18)$$

[0037] $$\sum_{k=1}^{S}\phi_k=1 \quad (19)$$

[0038] 随机生成序列X中各节点的红色、绿色、蓝色三个颜色通道的直方图,以表征文档对象,其中μ_k和σ_k^2采用随机数自动生成,进而得到X序列;

[0039] 步骤C2、基于彩色直方图空间元素与文档对象类型空间元素一一对应,对Y序列进行求解,即自动生成文档版面布局。

[0040] 进一步的,所述步骤C1中生成X序列的过程具体如下:

[0041] (1)设定序列X的节点数量N;

[0042] (2)为序列X的每个节点x_i的每个颜色通道的颜色直方图分别设定参数S;

[0043] (3)基于随机数生成算法设置参数μ_k和σ_k^2;

[0044] (4)基于高斯混合模型生成序列X。

[0045] 进一步的,所述步骤D中,具体采用以下方式:

[0046] 步骤D1、收集数据集:

[0047] Set_j,j=1,2,3...M (20)

[0048] 其中,Set_1,Set_2,...Set_M=文本集,公式集,...页码集;

[0049] 步骤D2、基于步骤C生成的Y序列采用TeX标记语言及数据集生成文档对象具体内容。

[0050] 进一步的,所述步骤D2中,文档对象具体内容的生成过程采用自顶向下的方法:

[0051] (1)首先生成页眉;

[0052] (2)生成页面中的栏目数;

[0053] (3)从第一栏开始,根据Y序列、以及定义的数据集生成页面对象DO_i,同时记录DO_i的空间坐标信息以及DO_i的具体内容信息;

[0054] (4)若文档不止一栏,则当第一栏结束后继续生成第二栏,直至最后一栏结束;

[0055] (5)当文档中所有栏目均生成后,生成页脚、页码。

[0056] 进一步的,所述步骤A中,所述文档对象类型标签序列所包含的标签类型包括但不限于文本、公式、图、图名、表、表名、页眉、页脚和页码。

[0057] 与现有技术相比,本发明的优点和积极效果在于:

[0058] 本方案基于条件随机场(CRF)对文档版面结构进行建模,获取文档版面的先验知识,并利用先验知识自动生成文档图像数据集,节省时间和人力成本,避免由于人工标注带来的无效标注;通过条件随机场及先验知识指导生成文档图像的版面结构,使生成的版面

更贴近出版物、符合阅读习惯，提高数据集的利用率及深度学习精度；并且，生成文档图像集的同时，能够提供文档图像中文本对象的文字编码信息(ASCII、Unicode等)。

附图说明

[0059] 图1为本发明实施例构建文档图像数据集的流程示意图；

[0060] 图2为本发明实施例文档对象序列示意图；

[0061] 图3为本发明实施例条件随机场样本序列示意图；

[0062] 图4为本发明实施例自动生成文档图像示意图；

[0063] 图5为本发明实施例生成的文档图像数据集结构示意图。

具体实施方式

[0064] 为了能够更加清楚地理解本发明的上述目的、特征和优点，下面结合附图及实施例对本发明做进一步说明。在下面的描述中阐述了很多具体细节以便于充分理解本发明，但是，本发明还可以采用不同于在此描述的其他方式来实施，因此，本发明并不限于下面公开的具体实施例。

[0065] 本实施例提供了一种利用先验知识构建文档图像数据集的方法，首先采用条件随机场对版面结构进行建模，并对模型进行训练，然后根据模型随机生成文档对象序列，最终生成文档图像数据集，同时，在数据集中保留了文档对象的空间坐标信息和具体内容信息，如图1所示，具体包括以下步骤：

[0066] 第一步、将文档版面信息刻画为文档对象空间、彩色直方图空间和文档对象类型空间，且文档对象空间元素与彩色直方图空间元素一一对应，彩色直方图空间元素与文档对象类型空间元素一一对应；

[0067] 基于条件随机场对文档版面结构进行建模，得到文档对象彩色直方图序列X与文档对象类型标签序列Y的线性链条随机场；

[0068] 第二步、训练条件随机场模型参数，对条件随机场权值进行求解；

[0069] 第三步、基于高斯混合模型随机生成彩色直方图序列X，根据训练好的模型求解文档对象类型标签序列Y，自动生成文档版面布局；

[0070] 第四步、采集各种已有文档对象数据，根据生成的文档版面布局，生成文档对象的具体内容；

[0071] 第五步、将文档转换成文档图像，构建文档图像数据集，所述文档图像数据集包含文档图像、文档对象空间坐标信息和文档对象具体内容信息。

[0072] 本实施例中，将文档版面信息刻画为空间映射关系，如图2和图3所示，将文档版面信息抽象为三个空间，即文档对象空间、彩色直方图空间和文档对象类型空间，三个空间之间存在两种映射关系：文档对象空间←→彩色直方图空间，彩色直方图空间←→文档对象类型空间，利用这两种映射关系，即可以对文档版面信息采用条件随机场进行建模，也可以指导文档图像的自动生成。

[0073] 具体的，下面结合具体的实施例对本发明方案进行详细的介绍：

[0074] 第一步、采用条件随机场对文档版面进行建模；

[0075] 将采集的PDF文档页面中的对象看作一个序列，记作：

[0076] $DO_i, i=1,2,3...N$ (1)

[0077] 其中，DO_i表示第i个文档对象，比如图2中的文档对象序列共包含8个对象：DO_1，DO_2，DO_3，...DO_8，这8个对象如图3第一行所示；

[0078] 计算每一个对象的彩色直方图：

[0079] $x_i=[Hist_R(DO_i), Hist_G(DO_i), Hist_B(DO_i)]$ (2)

[0080] 上式中$Hist_R(DO_i)$，$Hist_G(DO_i)$，$Hist_B(DO_i)$分别代表文档对象DO_i的红色、绿色、蓝色三个颜色通道的直方图，根据对象序列DO_1，DO_2，DO_3，...DO_8计算得到彩色直方图x序列：x_1，x_2，x_3，...x_N如图3第二行所示；

[0081] 确定与彩色直方图x_i.序列所对应的类型标签序列，如图3中第三行所示：

[0082] $y_i, i=1,2,3...N$ (3)

[0083] $y_i \in \{Type_j | j=1,2,3...M\}$ (4)

[0084] 其中，$Type_j$为标签类型，所述标签类型包括文本、公式、图、图名、表、表名、页眉、页脚和页码等标签。

[0085] 将彩色直方图序列、标签类型序列分别定义为：

[0086] $X=[x_1, x_2, x_3, ...x_N]$ (5)

[0087] $Y=[y_1, y_2, y_3, ...y_N]$ (6)

[0088] 该序列中每一个节点元素均对应一个文档对象类别，如图、文本、页码等。

[0089] 由于x_i与y_i具有相同的序列结构(即元素个数均为N，且元素之间一一对应，从图3中也显而易见)，则X和Y可以构成一个线性链条随机场(linear chain conditional random fields)。具体而言，将$X=[x_1, x_2, x_3, ...x_N]$和$Y=[y_1, y_2, y_3, ...y_N]$均看作随机变量序列，在给定随机变量序列X的情况下，随机变量序列Y的条件概率分布$P(Y|X)$便构成了条件随机场，若满足马尔可夫性(无后效性)：

[0090] $P(y_{i+1}|X, y_1, y_2, y_3, ...y_N)=P(y_{i+1}|X, y_i)$ (7)

[0091] 则称$P(Y|X)$为线性链条随机场。

[0092] 为了计算式(7)所表示的条件概率，需要进一步定义序列Y中节点y_i的特征函数族，考虑到更好地描述特征函数族，首先需要定义式(4)中的标签类型，本实施例中以9个类型为例具体说明：

[0093] $\{Type_1, Type_2, ...Type_9\}$

[0094] =｛文本，公式，图，图名，表，表名，页眉，页脚，页码｝ (8)

[0095] 即式(4)中M=9，j=1,2,...9；那么特征函数族就可以定义为：

[0096] $f_{j,j'}(y_i)=f_{j,j'}(X, y_{i-1}, i)$ (9)

[0097] 其中，函数$f_{j,j'}(y_i)$第一个脚标j表示当前节点y_i所属的类型，第二个脚标j′表示前一个节点y_{i-1}所属的类型，$f_{j,j'}(y_i)$表示节点y_i的特征函数，等式右边$f_{j,j'}(X, y_{i-1}, i)$表示在给定序列X的情况下，序列Y的第i节点y_i的特征函数值仅与前一节点y_{i-1}有关，这与式(7)所描述一致。

[0098] 根据式(8)的定义，当前节点y_i可能的类型数为M=9且前一节点y_{i-1}可能的类型数亦为M=9，所以函数族$f_{j,j'}(y_i)$中的函数个数为$M\times M=9\times 9=81$其函数值如下：

[0099] $$\left[f_{j,j'}(y_i)\right]_{M\times M}=\begin{bmatrix}1&1&0&1&1&0&1&0&0\\1&1&0&1&1&0&1&0&0\\1&1&0&1&1&0&0&0&0\\0&0&1&0&0&0&0&0&0\\0&0&0&0&0&1&0&0&0\\1&1&0&1&1&0&1&0&0\\0&0&0&0&0&0&0&0&0\\1&1&0&1&1&0&0&0&0\\1&1&0&1&1&0&0&1&0\end{bmatrix}_{M\times M} \quad (10)$$

[0100] 其中，j和j′亦分别表示矩阵的行、列坐标，j＝1,2,...9，j′＝1,2,...9。以第7行为例，此行所有元素均为0，因为式(8)中Type$_7$为“页眉”，而$f_{7,j'}(y_i)=0$则表示：在文档页面中的当前对象标签y_i若是“页眉”，则“页眉”前不能有任何其他元素。同时，为函数族的每一个函数定义一个权值：

[0101] $[\lambda_{j,j'}]_{M\times M}$ (11)

[0102] 这里M×M个权值为模型的待求参数。当给定特定的两个序列：X和Y可以通过特征函数族的加权求和用以评估当前Y与X的契合程度：

[0103] $g(Y|X)=\sum_i\sum_{j,j'}\lambda_{j,j'}f_{j,j'}(X,y_{i-1},i)=\sum_i\sum_{j,j'}\lambda_{j,j'}f_{j,j'}(y_i)$ (12)

[0104] 显然，对于特定的X，可能的类别序列Y一共有N×M种，那么P(Y|X)可以定义为：

[0105] $$P(\mathbf{Y}|\mathbf{X})=\frac{\exp\left(g(\mathbf{Y}|\mathbf{X})\right)}{\sum_{\mathbf{Y}'=1}^{N\times M}\exp\left(g(\mathbf{Y}'|\mathbf{X})\right)} \quad (13)$$

[0106] 且：

[0107] $$\sum_{\mathbf{Y}'=1}^{N\times M}P(\mathbf{Y}|\mathbf{X})=1 \quad (14)$$

[0108] 其中，特征函数的个数以及特征函数值即为先验知识，函数族表征了文档版面的先验知识，其将用于指导文档版面结构的自动生成，定义了特征函数族之后，便可以计算式(13)，式(14)说明式(13)满足概率的基本性质，但是通常为了计算方便，使用式(13)的非规范化概率形式：

[0109] $P(Y|X)\propto\exp(g(Y|X))$ (15)

[0110] 最终，式(15)定义了图3中第二层彩色直方图序列X(式(5))与第三层类型标签序列Y(式(6))的线性链条随机场，其中，式(10)中的特征函数值根据实际情况进行设置即可，式(11)定义的权值为系统待求参数。

[0111] 第二步，训练条件随机场模型参数，得到条件随机场中的待求解权值$\lambda_{j,j'}$；

[0112] 本实施例中，通过设计条件随机场模型中的特征函数值，并且使用优化算法(如：梯度下降法、牛顿法、拟牛顿法等等)求解特征函数权值，最终，完成模型的求解过程，得到文档版面的权值$\lambda_{j,j'}$；

[0113] 具体的，在求解式(11)定义的权值时，将$\lambda_{j,j'}$假定为相互独立，定义对数似然函数：

[0114] $$L(\lambda)=\log\prod_{\mathbf{X},\mathbf{Y}}\exp\left(g\left(\mathbf{Y}\mid\mathbf{X}\right)\right) \quad (16)$$

[0115] 然后,式(16)对待求权值求导,寻找驻点:

[0116] $$\frac{\partial L(\lambda)}{\partial\lambda}=0 \quad (17)$$

[0117] 针对式(17),可使用梯度下降法、牛顿法、拟牛顿法,等一些列优化算法进行求解。

[0118] 整个求解$\lambda_{j,j'}$的过程概括如下:

算法 1 条件随机场权值求解算法

<1> 使用 PDF 解析工具，解析 PDF 文档，获取 X 序列、Y 序列;

<2> 根据 Y 序列样本，设置式（8）中的文档对象类别;

[0119] <3> 设定式（10）中的特征函数值;

<4> 使用优化算法（如：梯度下降法、牛顿法、拟牛顿法，等等）求解式（17），最终得到式（11）定义的权值;

<5> 算法结束。

[0120] 第三步,使用训练好的模型自动生成文档版面布局

[0121] 基于混合高斯模型随机生成X序列,根据第一步、第二步所获得的条件随机场模型,采用维特比算法求解Y序列,即Y序列表征了自动生成的文档版面布局;

[0122] 自动生成文档版面布局就是生成图3中第三行的Y序列,即:式(6)。为了自动获得Y序列,可使用图3中第三行的X序列通过式(15)求解得到,利用空间映射关系:“文档对象”空间←→“彩色直方图”空间,使用混合高斯模型实现。

[0123] 首先需要获得X序列,使用混合高斯模型(Gaussian Mixture Model):

[0124] $$GMM(\alpha)=\sum_{k=1}^{S}\phi_k\frac{1}{\sqrt{2\sigma_k^2}}\exp\left[-(\alpha-\mu_k)/2\sigma_k^2\right] \quad (18)$$

[0125] $$\sum_{k=1}^{S}\phi_k=1 \quad (19)$$

[0126] 随机生成$X=[x_1,x_2,...x_8]$序列中节点$x_i=[Hist_R(DO_i),Hist_G(DO_i),Hist_B(DO_i)]$的红色、绿色、蓝色三个颜色通道的直方图,用以表征文档对象,其中μ_k和σ_k^2采用随机数自动生成,由此得到X序列,然后利用空间映射关系:“彩色直方图”空间←→“文档对象类型”空间,使用维特比算法(Viterbi algorithm)求解Y序列,即:自动生成文档版面信息。

[0127] 整个过程归纳为算法2:

算法 2 文档版面自动生成算法

<1> 设定序列 $\mathbf{X}$ 的节点数量 N，见式（1）和式（2）

<2> 为序列 $\mathbf{X}$ 的每个节点 x_i 的每个颜色通道的颜色直方图：

$Hist_R(DO_i), Hist_G(DO_i), Hist_B(DO_i)$ 分别设定式（18）中的参数 S；

[0128]

<3> 基于随机数生成算法设置式（18）中的参数 μ_k 和 σ_k^2；

<4> 基于式（18）生成序列 $\mathbf{X}$；

<5> 根据维特比算法求解 $\mathbf{Y}$ 序列；

<6> 算法结束。

[0129] 第四步，根据生成的版面布局，生成文档对象的具体内容：

[0130] 首先采集各种文档对象数据，然后使用第三步生成的Y序列生成文档中对象的具体内容；为了生成文档对象的具体内容，需要根据式(4)收集数据集，本实施例采用式(8)的定义收集数据集：

[0131] $Set_j, j=1,2,3...M$ (20)

[0132] Set_j对应于式(4)中的$Type_j$，具体而言，根据式(8)有：

[0133] $Set_1, Set_2, ...Set_9$＝文本集，公式集，...页码集 (21)

[0134] 接着基于第三步生成的Y序列采用TeX标记语言及式(21)的数据集生成文档对象具体内容，生成过程采用“自顶向下”的方法：页面→栏→页面对象。

[0135] <1>首先生成页眉；

[0136] <2>接着生成页面中的栏目数；

[0137] <3>从第一栏开始根据Y序列、式(21)中的数据集，生成页面对象，即：式(1)中的DO_i，同时记录DO_i的空间坐标信息(对象边框信息)：

[0138] DO_i-Coors (22)

[0139] 以及DO_i的具体内容信息(文字编码、公式、图、表，等等)：

[0140] DO_i-Content (23)

[0141] <4>若文档不止一栏，则当第一栏结束后继续生成第二栏，直至最后一栏结束；

[0142] <5>当文档中所有栏目均生成后，生成页脚、页码；

[0143] <6>以上过程均采用TeX标记语言实现，根据TeX标记语言，采用PDF引擎自动生成PDF文档。

[0144] 注：并非式(8)中所有类型的文档对象都需要出现在页面上，例如：一个文档页面允许没有页眉、页脚、页码，也可以只包含文本，主要由算法2的输出Y序列决定。当然，也可以通过人工设置要求页面必须包含特定文档对象。

[0145] 将以上过程归纳为算法3：

算法 3 文档对象具体内容生成算法

<1> 使用 PDF 解析工具，解析 PDF 文档，获取如式（21）所描述数据集；并根据数据集、算法 2 生成的 $\mathbf{Y}$ 序列，采用 TeX 标记语言生成文档具体内容；

<2> 若页面包含页眉，则生成页眉；

<3> 随机生成页面栏目数；

<4> 生成文档第一栏中的文档对象 DO_i（式（1））及对应的空间坐标信息 $DO_i - Coors$（式（22））、具体内容信息 $DO_i - Content$（式（23））；

<5> 若文档不止一栏，则继续生成第二栏的文档对象，直至最后一栏结束；

<6> 若页面包含页脚、页码，则生成页脚、页码；

<7> 根据 TeX 标记语言，使用 PDF 引擎生成 PDF 文档；

<6> 算法结束。

[0146]

[0147] 第五步，将文档转换成文档图像，构建文档图像数据集，所述文档图像数据集包括文档图像、文档对象空间坐标信息和文档对象具体内容信息；

[0148] 根据算法3生成的PDF文档，每一页都转换成文档图像，如图4给出一张自动生成的图像，将每一张生成的文档图像定义为：

[0149] $\mathrm{DocImage}_c, c=1,2,...\mathrm{Num}$ (24)

[0150] Num表示文档图像数据集的图像数量，同时将式(22)所表示的文档对象空间坐标映射至文档图像中，得到：

[0151] $\mathrm{DO}_i\text{-Coors}'$ (25)

[0152] 那么，文档图像数据集可表示为：

[0153] $\mathrm{DocImageSet}=\{\mathrm{ele}_c\}, c=1,2,...\mathrm{Num}$ (26)

[0154] $\mathrm{ele}_c=\{\mathrm{DocImage}_c, \mathrm{DO}_{i,c}\text{-Coors}', \mathrm{DO}_{i,c}\text{-Content}\}$ (27)

[0155] 式(26)定义了文档图像数据集，其中ele_c如图5虚线框所示，包含了一张图像中的N个文档对象空间坐标信息(式(27)中$\mathrm{DO}_{i,c}\text{-Coors}'$)，其与N个文档对象具体内容信息一一对应(式(27)中$\mathrm{DO}_{i,c}\text{-Content}$)。

[0156] 可见，本方案基于条件随机场对文档版面结构进行建模，可有效节省时间和人力成本，避免由于人工标注带来的无效标注；而且使用条件随机场对版面结构进行建模，用以指导生成文档图像的版面结构，使生成的版面更贴近出版物、符合阅读习惯，并且生成文档图像集的同时，能够提供文档图像中文本对象的文字编码信息(ASCII、Unicode等)，提高数据集的利用率及深度学习精度。

[0157] 以上所述，仅是本发明的较佳实施例而已，并非是对本发明作其它形式的限制，任何熟悉本专业的技术人员可能利用上述揭示的技术内容加以变更或改型为等同变化的等效实施例应用于其它领域，但是凡是未脱离本发明技术方案内容，依据本发明的技术实质对以上实施例所作的任何简单修改、等同变化与改型，仍属于本发明技术方案的保护范围。

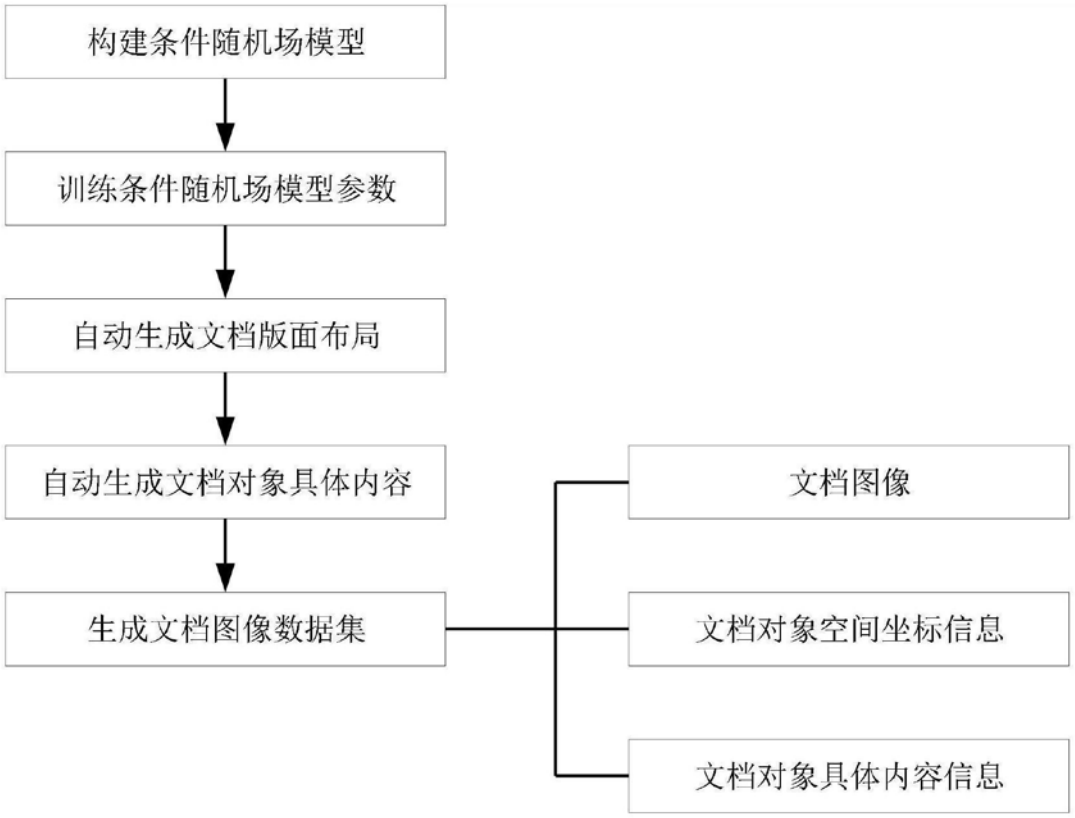

构建条件随机场模型
训练条件随机场模型参数
自动生成文档版面布局
自动生成文档对象具体内容
生成文档图像数据集
文档图像
文档对象空间坐标信息
文档对象具体内容信息

图1

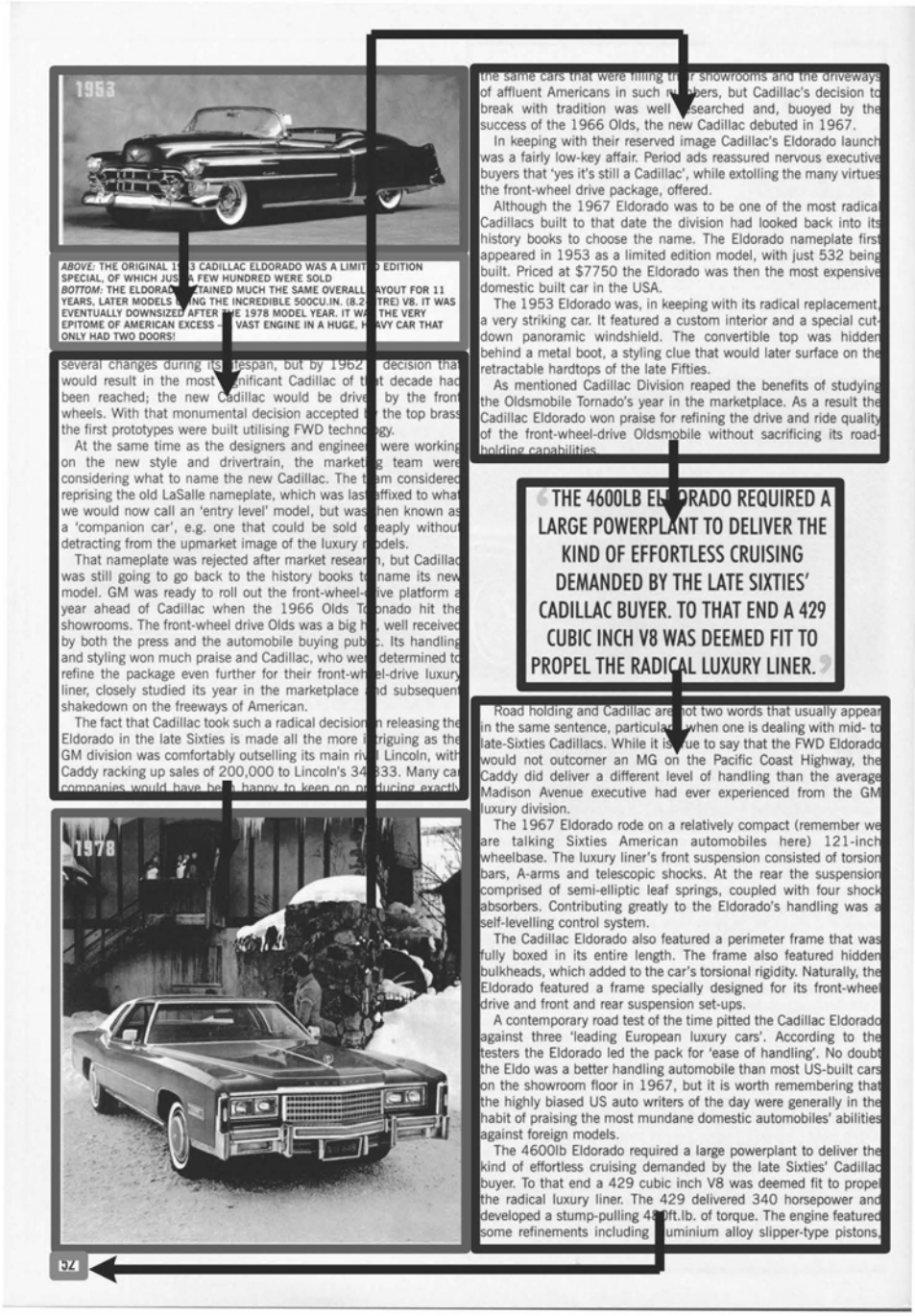

1953
ABOVE: THE ORIGINAL 1953 CADILLAC ELDORADO WAS A LIMITED EDITION SPECIAL, OF WHICH JUST A FEW HUNDRED WERE SOLD
BOTTOM: THE ELDORADO RETAINED MUCH THE SAME OVERALL LAYOUT FOR 11 YEARS, LATER MODELS USING THE INCREDIBLE 500CU.IN. (8.2 LITRE) V8. IT WAS EVENTUALLY DOWNSIZED AFTER THE 1978 MODEL YEAR. IT WAS THE VERY EPITOME OF AMERICAN EXCESS – A VAST ENGINE IN A HUGE, HEAVY CAR THAT ONLY HAD TWO DOORS!
several changes during its lifespan, but by 1962 a decision that would result in the most significant Cadillac of that decade had been reached; the new Cadillac would be driven by the front wheels. With that monumental decision accepted by the top brass the first prototypes were built utilising FWD technology.
At the same time as the designers and engineers were working on the new style and drivertrain, the marketing team were considering what to name the new Cadillac. The team considered reprising the old LaSalle nameplate, which was last affixed to what we would now call an 'entry level' model, but was then known as a 'companion car', e.g. one that could be sold cheaply without detracting from the upmarket image of the luxury models.
That nameplate was rejected after market research, but Cadillac was still going to go back to the history books to name its new model. GM was ready to roll out the front-wheel-drive platform a year ahead of Cadillac when the 1966 Olds Toronado hit the showrooms. The front-wheel drive Olds was a big hit, well received by both the press and the automobile buying public. Its handling and styling won much praise and Cadillac, who were determined to refine the package even further for their front-wheel-drive luxury liner, closely studied its year in the marketplace and subsequent shakedown on the freeways of American.
The fact that Cadillac took such a radical decision in releasing the Eldorado in the late Sixties is made all the more intriguing as the GM division was comfortably outselling its main rival Lincoln, with Caddy racking up sales of 200,000 to Lincoln's 34,833. Many car companies would have been happy to keep on producing exactly
1978
the same cars that were filling their showrooms and the driveways of affluent Americans in such numbers, but Cadillac's decision to break with tradition was well researched and, buoyed by the success of the 1966 Olds, the new Cadillac debuted in 1967.
In keeping with their reserved image Cadillac's Eldorado launch was a fairly low-key affair. Period ads reassured nervous executive buyers that 'yes it's still a Cadillac', while extolling the many virtues the front-wheel drive package, offered.
Although the 1967 Eldorado was to be one of the most radical Cadillacs built to that date the division had looked back into its history books to choose the name. The Eldorado nameplate first appeared in 1953 as a limited edition model, with just 532 being built. Priced at $7750 the Eldorado was then the most expensive domestic built car in the USA.
The 1953 Eldorado was, in keeping with its radical replacement, a very striking car. It featured a custom interior and a special cut-down panoramic windshield. The convertible top was hidden behind a metal boot, a styling clue that would later surface on the retractable hardtops of the late Fifties.
As mentioned Cadillac Division reaped the benefits of studying the Oldsmobile Tornado's year in the marketplace. As a result the Cadillac Eldorado won praise for refining the drive and ride quality of the front-wheel-drive Oldsmobile without sacrificing its road-holding capabilities.
"THE 4600LB ELDORADO REQUIRED A LARGE POWERPLANT TO DELIVER THE KIND OF EFFORTLESS CRUISING DEMANDED BY THE LATE SIXTIES' CADILLAC BUYER. TO THAT END A 429 CUBIC INCH V8 WAS DEEMED FIT TO PROPEL THE RADICAL LUXURY LINER."
Road holding and Cadillac are not two words that usually appear in the same sentence, particularly when one is dealing with mid- to late-Sixties Cadillacs. While it is true to say that the FWD Eldorado would not outcorner an MG on the Pacific Coast Highway, the Caddy did deliver a different level of handling than the average Madison Avenue executive had ever experienced from the GM luxury division.
The 1967 Eldorado rode on a relatively compact (remember we are talking Sixties American automobiles here) 121-inch wheelbase. The luxury liner's front suspension consisted of torsion bars, A-arms and telescopic shocks. At the rear the suspension comprised of semi-elliptic leaf springs, coupled with four shock absorbers. Contributing greatly to the Eldorado's handling was a self-levelling control system.
The Cadillac Eldorado also featured a perimeter frame that was fully boxed in its entire length. The frame also featured hidden bulkheads, which added to the car's torsional rigidity. Naturally, the Eldorado featured a frame specially designed for its front-wheel drive and front and rear suspension set-ups.
A contemporary road test of the time pitted the Cadillac Eldorado against three 'leading European luxury cars'. According to the testers the Eldorado led the pack for 'ease of handling'. No doubt the Eldo was a better handling automobile than most US-built cars on the showroom floor in 1967, but it is worth remembering that the highly biased US auto writers of the day were generally in the habit of praising the most mundane domestic automobiles' abilities against foreign models.
The 4600lb Eldorado required a large powerplant to deliver the kind of effortless cruising demanded by the late Sixties' Cadillac buyer. To that end a 429 cubic inch V8 was deemed fit to propel the radical luxury liner. The 429 delivered 340 horsepower and developed a stump-pulling 480ft.lb. of torque. The engine featured some refinements including aluminium alloy slipper-type pistons,
74

图2

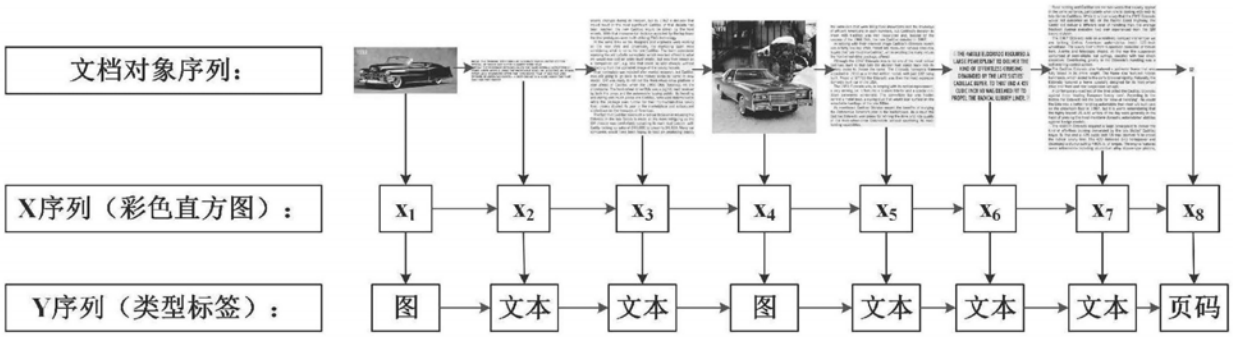
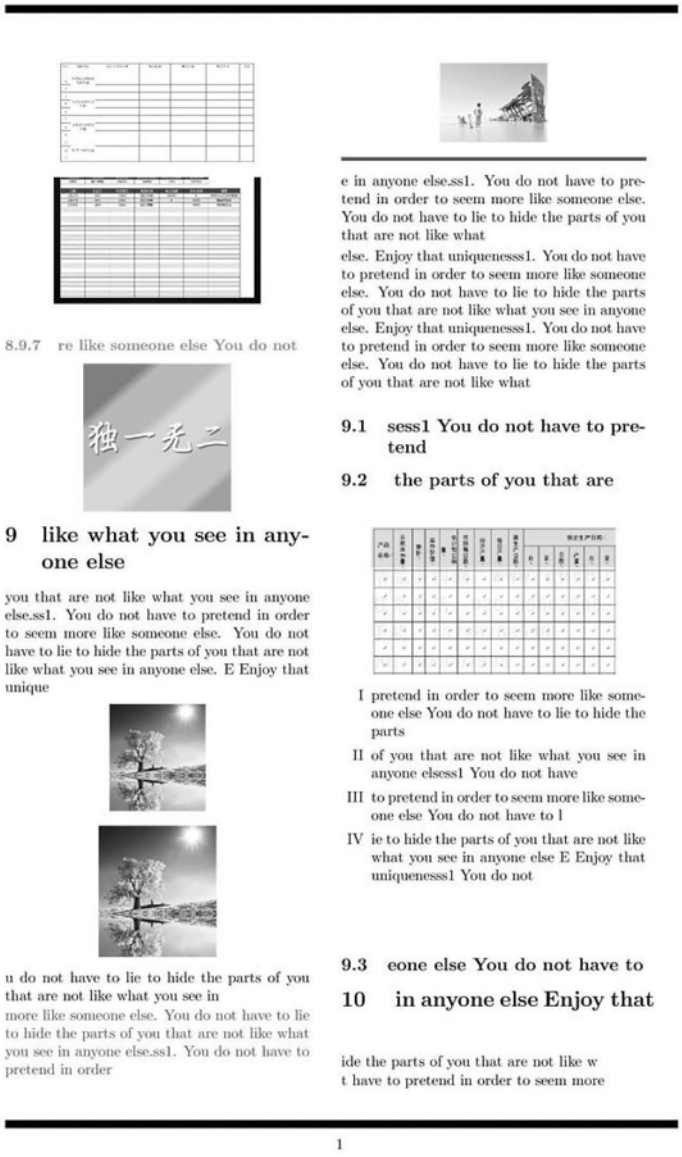


图3



8.9.7 re like someone else You do not

9 like what you see in anyone else

you that are not like what you see in anyone else.ss1. You do not have to pretend in order to seem more like someone else. You do not have to lie to hide the parts of you that are not like what you see in anyone else. E Enjoy that unique

u do not have to lie to hide the parts of you that are not like what you see in more like someone else. You do not have to lie to hide the parts of you that are not like what you see in anyone else.ss1. You do not have to pretend in order

e in anyone else.ss1. You do not have to pretend in order to seem more like someone else. You do not have to lie to hide the parts of you that are not like what else. Enjoy that uniquenesss1. You do not have to pretend in order to seem more like someone else. You do not have to lie to hide the parts of you that are not like what you see in anyone else. Enjoy that uniquenesss1. You do not have to pretend in order to seem more like someone else. You do not have to lie to hide the parts of you that are not like what

9.1 sess1 You do not have to pretend

9.2 the parts of you that are

I pretend in order to seem more like someone else You do not have to lie to hide the parts

II of you that are not like what you see in anyone elsess1 You do not have

III to pretend in order to seem more like someone else You do not have to l

IV ie to hide the parts of you that are not like what you see in anyone else E Enjoy that uniquenesss1 You do not

9.3 eone else You do not have to

10 in anyone else Enjoy that

ide the parts of you that are not like w
t have to pretend in order to seem more

1

图4

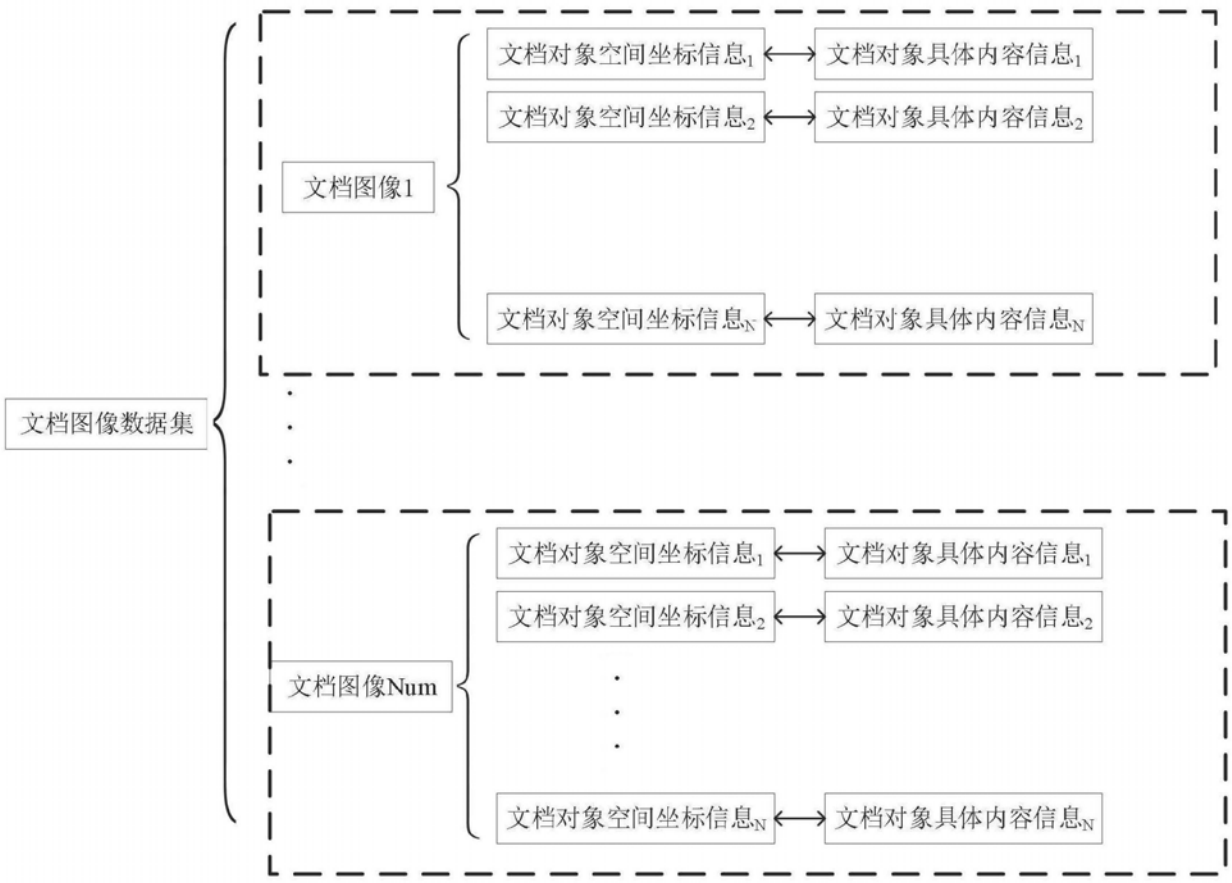

文档图像数据集
文档图像1
文档对象空间坐标信息1
文档对象具体内容信息1
文档对象空间坐标信息2
文档对象具体内容信息2
文档对象空间坐标信息N
文档对象具体内容信息N
文档图像Num
文档对象空间坐标信息1
文档对象具体内容信息1
文档对象空间坐标信息2
文档对象具体内容信息2
文档对象空间坐标信息N
文档对象具体内容信息N

图5