情报学报 2019 年 4 月 第 38 卷 第 4 期
Journal of the China Society for Scientific and Technical Information, Apr. 2019, 38(4): 384-390

DOI: 10.3772/j.issn.1000-0135.2019.04.006

基于机器视觉的PDF学术文献结构识别

于丰畅，陆 伟

（武汉大学信息管理学院，武汉 430072）

摘 要 PDF格式在电子学术文献出版发行领域占有极其重要的地位，但因其复杂的技术规则，使得PDF无法直接被机器阅读，给针对学术文献的研究工作造成了诸多不便。本文提出了一种基于机器视觉的PDF文档结构识别方法，该方法针对常见的PDF学术论文，将PDF文件中的视觉对象和文本对象进行映射，获得内容对象的几何属性和文本属性，并辅以启发式算法对内容对象进行类型判断，得到PDF文档的物理结构和逻辑结构。该方法以直观的方式克服了其他PDF解析方法需要大量人工特征构建或大规模语料训练、难以识别公式表格等缺点，并成功地对ACL（Association for Computational Linguistics）的论文集进行了结构识别和全文抽取。

关键词 PDF；学术文献；机器视觉；结构识别

Structural Recognition of PDF Academic Literature Based on Computer Vision

Yu Fengchang, and Lu Wei

(*School of Information Management, Wuhan University, Wuhan* 430072)

Abstract: Portable Document Format (PDF) documents play an important role in the publication of academic electronic literature. However, owing to the technical and structural complexities of PDF documents, they cannot be directly read by digital devices, which in turn can hinder research studies based on academic electronic literature. Hence, this paper proposes a method based on computer vision for the structural recognition of PDF documents. The proposed method, supplemented by a heuristic algorithm, maps graphic objects and text objects present in the PDF files of academic documents and thereby obtains geometric and text attributes of the file objects. The proposed algorithm can identify the category of a PDF object for determining the physical and logical structures of a PDF document. Conventional PDF analysis methods require a significant amount of artificial feature construction and large-scale lexical corpus training and cannot identify formulae and tables. The proposed method can overcome the aforementioned shortcomings and can successfully perform full-text extraction and structural recognition of ACL data collections.

Key words: Portable Document Format (PDF); academic literature; computer vision; structural recognition

1 引 言

PDF是美国Adobe公司于1983推出的电子文档交换格式，它以PostScript语言图像模型为基础，设计初衷是能够在不同应用程序、不同平台、不同硬件上以完全相同的样式输出内容。正因为PDF具有这样的优点，它也是当今互联网常用的电子文档交换格式之一。特别地，在学术研究领域，PDF是论文的主要保存和传播格式，在世界范围内除了少数国家的少数出版机构提供HTML、CAJ、EPUB等

收稿日期：2018-09-26；**修回日期**：2018-10-07

作者简介：于丰畅，男，1990年生，博士研究生，主要研究领域为信息抽取、机器学习；陆伟，男，1974年生，教授，博士生导师，主要研究领域为信息检索、知识挖掘与可视化、竞争情报方法与技术、知识管理等，E-mail：weilu@whu.edu.cn。

格式的学术论文外，PDF格式是学术界首选的论文电子交换格式。

但是因为PDF推出时间久远，虽然有多次更新和改进，其设计目的主要关注其渲染结果，忽略了内容的结构信息，使得PDF文档内容间的逻辑结构或者语义结构均无法直接获得。导致PDF格式不能被机器直接阅读，需要经过复杂地转化处理过程才能获得文档内容的数据，并且这些数据的准确性往往难以得到保障，丢失内容、顺序错乱的情况时有发生。特别是对于学术论文中大量出现的公式、表格等内容，这些内容对于学者研究文献的重要性不言而喻，但多数现有通用PDF转化解析工具不会加以区别，依然当作普通文字处理，造成了内容大量的丢失，给研究学者的工作带了不利的影响。

本文针对学术文献中常见的双栏PDF文件提出了一套基于机器视觉的文献结构解析方法。该方法将PDF文献从视觉和文本两个视角进行识别，模拟了人类阅读文献的直观方法，即首先识别文献的物理结构分块再判断内容类型最后进行语义理解，并且针对文字、标题、图片、表格、公式等不同类型的内容进行了相应的处理，为图情学者对文献的进一步研究提供了便利。以下是本文的结构：第2节是PDF文档结构识别与内容抽取解析的相关研究，第3节是本文的研究方法，第4节是实验内容，第5节是总结与展望。

2　相关研究

数字文档相对于纸质文档的一个主要优点是：数字文档有清晰的物理结构（比如：页、栏、段落、正文、表格、图片，等等）和逻辑结构（比如：作者、标题、机构、摘要、章节，等等），但是PDF格式在设计之初是以渲染结果跨设备跨平台一致性为其优先考量，并未针对以上数字文档应有的特点进行特别的设计，导致了无法直接从PDF文件中读取版面的物理结构和内容的逻辑结构。因此对于PDF的格式识别问题主要也是从以上两个角度进行研究，多个学科的学者对于该问题进行了多个层次的研究。

早期对于PDF文件的物理结构的分析，主要是针对PDF文件图片或者扫描型PDF进行，其目的是从一张完成的页面图片中，定位不同类型的结构的位置或者得到相应的图片区域。较为传统的方法主要有三种方式：自上而下的方式、自下而上的方式和混合方式[1]。

自上而下的方式从整个页面出发，通过迭代将图片划分为不同类型的小图片，每个小图片对应于不同类型的结构。例如，Nagy等[2]通过页面元素的横纵坐标，将二维平面分割问题转化为一维字符串解析问题，然后是用规则将相应的元素进行有效区分；Baird等[3]介绍了一种按照形状运算的分割算法，该算法在白底的文稿中逐步剥离空白的矩形区域，最终得到含有内容的区域，并且内容的分割不以矩形为边界，而是形成内容实际的边界。

自下而上的方式从图片的单个像素出发，通过迭代聚合的方式形成对应于文字、线段、图片等类型的结构，如以下几种经典算法：文献[4]使用最近邻聚类算法定位行内、行间的文字、段落和空白；Kise等[5]采用基于泰森多边形的算法将文档各元素从图片中分离，并且该方法还可以处理带有一定角度的文档照片；Wahl等[6]使用约束游程算法将文档切分为文字区域、文字线段、黑色实线、矩形区域和图表区域。

而混合型的方式则从页面和像素两者同时出发，当迭代满足某些条件时停止，形成不同类型的结构图片。如文献[7]中使用连续空白作为判定文档行列的手段，并通过设定白色区域和黑色区域的长宽比阈值来进行图片分割，当黑色区域中白色空白小于阈值预设的比例时将黑色部分进行合并，由此完成自上而下和自下而上两个部分的工作。

虽然这些传统的分析方法在识别物理结构方面能够达到一定的准确率，考虑到本文研究对象是学术论文，其有着规范的格式限制，上述算法或上述算法的改进方法需要细致复杂的人工特征抽取，并且以上方法不涉及文档的逻辑结构识别，故以上方法不能完全适合本文的任务需求。也是因为相同的原因，近年的文档物理结构分析主要采用机器学习的方式进行。Chen等[8]介绍了一种无监督学习历史手写文档特征的方法，利用这些特征配合SVM可以将文档图片的每一个像素点分类为背景、文字区域等。Chen等[9]使用了一种仅用单个卷积层的应用于文档分割的卷积神经网络，该网络用于将文档照片的每一个超像素进行分类，取得了较好的效果。

随着PDF文件的日益盛行，扫描型PDF文件的数量逐渐减少，文字型PDF或者文字图片混合型PDF日益增多，这一类PDF文件的数据流中包含了文字、图片、线段等元素的信息，这也给学者分析文档结构提供了另一种方式。但是考虑到PDF格式的复杂性，学者们通常利用第三方工具将PDF转换

为诸如 XML、HTML 等机器可读的格式，然后通过分析转换后的格式结合一定的规则来解析文档的物理结构，同样的方法也适用于解析文档的逻辑结构，如文献[10]。

其他一些研究针对 PDF 文档中的部分物理结构进行识别，如文献[11-15]使用人工构建的特征或者机器学习的方式识别文档中的图表。以上的文献主要采用人工特征构建或者机器学习的方法，前者人工代价较大，而后者需要构建带有标签的数据集，也需要较大的人工成本，虽然都可以达到一定的识别准确率，但对于本文的目的并不适合。

PDF 的逻辑结构分析相对来说方法更加统一，一般采用规则式的分析方法或者正则表达式的分析方式，并按照分析结果将各个逻辑结构之间的关系表示为树状结构。Tsujimoto 等[16]将文档的物理结构和逻辑结构表示成树状图，并提出了一种从物理结构树向逻辑结构树的转换方法。Yamashita 等[17]提出了树形模型，该模型输入文档元素的少量位置信息和内容信息，根据人工构建的规则输出文档元素的逻辑分类。Ramesh 等[18]采用多层条件随机场以及多个监督学习模型对文字内容和排版信息进行分析，得到了文档中的作者、机构、引用等信息；Fauconnier 等[19]利用文档的逻辑结构树，辅助简单的规则对逻辑结构进行分类。

上述文献都在一定程度上识别了文档的物理结果或者逻辑结构，但是也存在着诸如需要复杂人工构建特征或者需要大量的训练预料的不足。更重要的是未能够较为全面地提取论文中绝大多数内容（如公式、表格等）的方法，而这些内容往往对论文的表达起到重要的作用。

对于文档的物理结构的定位，考虑到本文的研究对象具有较为规整的排版，本文将使用机器视觉方法对文档的图片进行内容定位。该方法的优点在于不需要人工构建大量特征，也不需要收集大量语料进行训练。而对于逻辑结构的分类，也因为研究对象的特点，本文将采用一种启发式算法进行分类。

3 研究方法

本文针对 PDF 格式的学术论文作为研究对象来介绍本研究方法，学术论文一般具有统一的排版规则：白色背景、内容单栏、双栏排列，各级标题字号大于正文等。需要指出的是符合以上排版方式的学术论文均适合本方法，本文以难度较大的双栏排布为例介绍该方法，单栏的情况只需要变化对象的映射方式即可。正因为具有这样的排版，本文将首先利用机器视觉的物体识别技术对文档的物理结构进行定位，然后使用一套启发式算法对各个物理结果进行逻辑结构的分类。主要工作流程如下：

（1）对于一个 PDF 文件，首先将其按照页码分断，然后将每一页单独渲染成一张图片。

（2）对于每一张图片识别其中的内容块，得到未经分类的物理结构的对象。

（3）读取 PDF 文件的数据流，将前一步中得到的内容块与数据流按照坐标的重叠关系进行映射，得到内容块对象，其中包含了物理结构的图片和内容信息。

（4）遍历所有内容对象依次使用启发式算法判断其物理结构类型（文字、标题、图片、表格、公式）和逻辑结构类型（标题、作者、段落等）。

（5）最后按照阅读顺序、按照类型输出。

以下内容将会具体介绍每一个流程的实现方式。

3.1 物理结构识别

将 PDF 学术文档按照其物理结构分割成各个部分需要两个步骤：①将 PDF 文件按页转化为图片格式；②从每一页的图片中定各物理结构的位置。

将 PDF 转化为图片格式有多种方法，本文中采用的是使用 Ghostscript 工具包来进行此项操作。Ghostscript 是一套基于 Adobe PDF 标准、PostScript 页面描述语言等而编译成的自由软件。其可以在多个主流软件平台运行，具有很好的跨平台性和可移植性。本文中我们将 PDF 文件作为 Ghostscript 的输入，输出设备设置为 “png16m”，并为了保证后续的机器视觉处理步骤达到最佳效果，PDF 渲染成图片的分辨率设置为 300 dpi，得到每一页的 PDF 文件对应的渲染之后的图片。

学术论文一般采用简单背景（纯白背景），和比较规整的排版（一般均为矩形内容框），非常适合使用机器视觉方法提取其中的内容。故第二步使用 OpenCV 机器视觉库对上述图片进行内容块识别，具体处理步骤如下：

（1）将图片转换为黑白模式。

（2）将黑白模式的图片进行二值化处理。

（3）适当地进行腐蚀和膨胀处理，使得内容的形状尽可能规整。

（4）查找图片中的边缘信息，边缘指定为所有

图形外部的边缘，包括线段、曲线、角等。

（5）找到这些边缘的包络线，且该包络线设置为矩形。

图 1 为利用 OpenCV 机器视觉库从 PDF 文件中识别论文的物理结构的示意图。图 1a 为输入的文档图片，该图片由 PDF 文件的其中一页转换而来，并经过了黑白转化、二值化、腐蚀和膨胀等处理；图 1b 为查找边缘的中间结果，图中曲线即为物体的外边缘，可以看到程序正确地找到了正文、标题、图表等内容块的外边缘；图 1c 为根据内容块的外边缘做出的矩形包络线。值得说明的是，为了便于理解，以上示意图的结果均标注于原图之上。经过该步骤之后，各个物理结构对象将会包含一张自身的图片和自身的位置坐标。

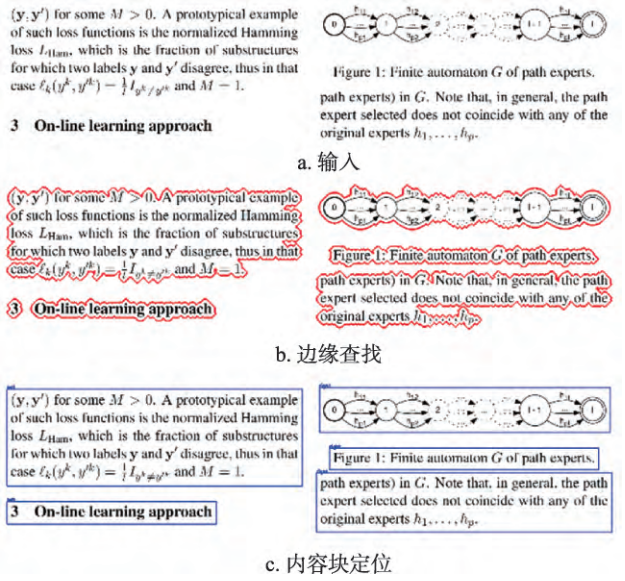


图 1　利用机器视觉处理学术论文的过程

各种物理结构在被识别之后，还需将按照阅读顺序进行分类和排序，以 ACL 论文集为例，该论文集为左右双栏排版，论文标题和部分图表跨越左右双栏。故各物理结构的图片按照在原文档的位置分为三类，即左、右以及跨越三类，各类中按照从上到下进行排列。

3.2　对象映射

定位到物理结构所在的图片后，还需要添加其对应的内容，即从 PDF 文件的数据流中读取该区域的文本内容。得益于 PDF 的设计模式，PDF 数据流中的所有内容信息都附带有位置坐标，这也是本文设计这套方法的出发点之一。

本文采用 PDFminer 作为 PDF 的解析框架，该框架采用 Python 语言编写，专注于获取和分析 PDF 文件中的文本、图形、几何数据等。该框架也提供了 PDF 文档的物理结构解析，并能够将这些结构按照树形进行组织，如图 2 所示。PDF 每一页可以分为文字框、表格、直线、矩形和图片。其中文字框又可以分为文字行，文字行包含单词，对于英文还可以分解为字符，如果是中文便不包含字符一项。

PDFminer 对于 PDF 文件的解析原理为：将文字之间的间距和人为设定的阈值进行比较，按照比较的结果将文字进行聚类，从而将单词聚类为句子，句子组合成段落等。该方法虽然简单易用，也能达到一定的准确率，但人为设定的阈值需要经过多次实验，且切换不同的 PDF 文档仍需要重新设定，导致程序的鲁棒性不高。

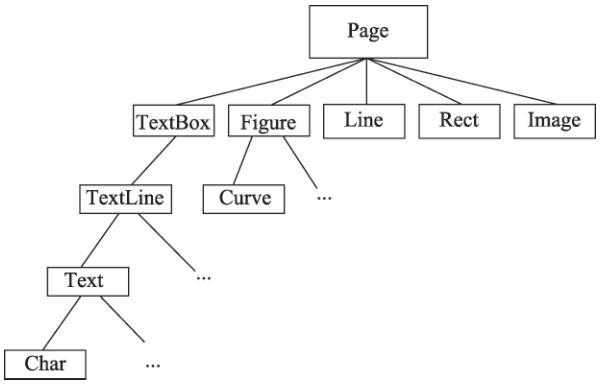


图 2　PDFminer 解析 PDF 的物理结构树

相反第 3.1 节中介绍的机器视觉方法无需对物理结构之间的距离进行参数设置，便能够很好地区分不同的物理结构。基于以上的原因，为保证本方法的准确性，本文中仅对 PDFminer 解析出来的最小单位，即字符（英文字母或中文汉字）进行操作。与第 3.1 节中将物理结构的图片分为左、右、跨越三类一样，将所有字符也分为这三类，并在每一类型中按照从上到下的阅读进行排序。PDFminer 中解析的每一个字符均包含了字符的位置坐标、字体、字号等信息，遍历字符与第 3.1 节中的物理结构，对比两者的位置坐标，按照坐标包含与否建立映射关系。考虑到第 3.1 节和第 3.2 节中解析方式的不同，位置坐标可能存在一定的误差，包含关系可以拓展为位置的重叠关系，即当字符与物理结构的位置重叠比例大于一定阈值时视为字符包含于物理结构之内。重叠比例的具体计算法如公式(1)~公式(4)所示：

$$X_{\text{overlap}} = \max(0, \min(x_{12}, x_{22}) - \max(x_{11}, x_{21})) \quad (1)$$

$$Y_{\text{overlap}} = \max(0, \min(y_{12}, y_{22}) - \max(y_{11}, y_{21})) \quad (2)$$

$$S_{\text{overlap}} = X_{\text{overlap}} \times Y_{\text{overlap}} \quad (3)$$

$$\text{Ratio}_{\text{overlap}} = \frac{S_{\text{overlap}}}{(x_{12} - x_{11}) \times (y_{12} - y_{11})} \quad (4)$$

其中，(x_{11}, y_{11})、(x_{12}, y_{12})分别为字符的左上角、右下角坐标，(x_{21}, y_{21})、(x_{22}, y_{22})分别为物理结构的左上角、右下角坐标。因为单个字符的面积一般远小于一个物理结构所占的面积，因此利用面积的重叠比例判断某一字符所属的物理结构的方法，相较于PDFminer提供的判断方法有更好的鲁棒性。

如此获得对应关系之后，便形成了新的物理结构对象，其中每一个对象包含了一张图片、内容文本、文本的字体信息以及坐标信息等，该结果将用于后续的物理类型和逻辑类型的判断。此时得到的结果是按照阅读顺序排列的内容对象，下一步为对其物理类型和逻辑类型进行分类。

3.3 对象类型判断

本文中采用一种启发式算法来对内容对象进行物理结构和逻辑结构的分类。考虑到第3.1节中使用的机器视觉的物体查找方法能够对PDF页面上全部内容进行记录，而第3.2节中的PDFminer解析器只会解析PDF文档中的文字和矢量几何图形，不会包含图片内容，故不包含任何文字信息的内容对象自然能够判断为图片。对于含有大量线段、数字的内容对象被认定为表格。对于包含符号比例较高的内容对象被认定公式。标题的字号往往较正文更大，故在统计了全文的字体字号分布之后，占主要的字体字号便是正文内容，大于该字号的表示标题内容。在标题对象中，匹配Abstract、Reference等关键字，便可定位到摘要、参考文献等标题的位置。该算法能够以较为简单的方式较准确地对物理、逻辑对象进行分类。

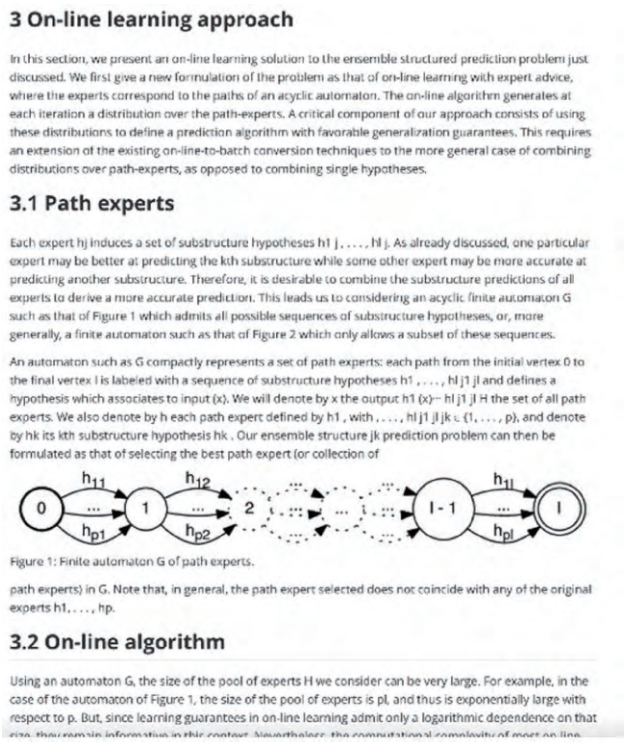
3.4 内容输出

在经过前几步的处理之后，还需将各个已分类的内容对象按照阅读顺序进行输出，本文采用的输出方式是：正文和标题以文字方式输出，并且标题采用更大的字号和粗体。图片、公式和表格均采用图片格式进行输出，最大程序地还原文档原本的排版。

4 实 验

本文选取ACL（Association for Computational Linguistics）论文集的论文作为实验对象。ACL论文的典型特征是：内容左右双栏排布，标题和部分图片跨越左右两栏，正文使用一致的字体和字号，各级标题的字号大于正文，字体异于正文字体。我们从ACL论文集中随机选取100篇论文，使用本文的格式识别方法，识别文档中的物理结构和逻辑结构后，将内容对象按照阅读顺序依次输出，生成对应的Markdown文件，人工检测每一篇论文输出的结果。其中Markdown是一种轻量级的标记语言，使用Markdown文件作为输出的原因有几点：①该文件的编写语言相较HTML、XML更简洁，更易于程序中输出部分代码的编写；②Markdown文件有着良好的可视化层级关系，有利于人工快速检测结果的正确性。一种典型的Markdown文件的结构如图3所示（该图显示的部分与图1属原PDF的同一部分），其中章节标题为加粗且具有较大的字号，由图表、公式转化而来的图片位于居中显示，正是因为Markdown格式具有这种较好的可视化特性，这种输出方式将有效地提高后续的人工检验的效率。

相应的评价指标有：标题检测准确率、正文检



3 On-line learning approach

In this section, we present an on-line learning solution to the ensemble structured prediction problem just discussed. We first give a new formulation of the problem as that of on-line learning with expert advice, where the experts correspond to the paths of an acyclic automaton. The on-line algorithm generates at each iteration a distribution over the path-experts. A critical component of our approach consists of using these distributions to define a prediction algorithm with favorable generalization guarantees. This requires an extension of the existing on-line-to-batch conversion techniques to the more general case of combining distributions over path-experts, as opposed to combining single hypotheses.

3.1 Path experts

Each expert hj induces a set of substructure hypotheses h1 j , . . . , hl j. As already discussed, one particular expert may be better at predicting the kth substructure while some other expert may be more accurate at predicting another substructure. Therefore, it is desirable to combine the substructure predictions of all experts to derive a more accurate prediction. This leads us to considering an acyclic finite automaton G such as that of Figure 1 which admits all possible sequences of substructure hypotheses, or, more generally, a finite automaton such as that of Figure 2 which only allows a subset of these sequences.

An automaton such as G compactly represents a set of path experts: each path from the initial vertex 0 to the final vertex l is labeled with a sequence of substructure hypotheses h1 , . . . , hl j1 jl and defines a hypothesis which associates to input (x). We will denote by x the output h1 (x)··· hl j1 jl H the set of all path experts. We also denote by h each path expert defined by h1 , with , . . . , hl j1 jl jk ∈ {1, . . . , p}, and denote by hk its kth substructure hypothesis hk . Our ensemble structure jk prediction problem can then be formulated as that of selecting the best path expert (or collection of

Figure 1: Finite automaton G of path experts.

path experts) in G. Note that, in general, the path expert selected does not coincide with any of the original experts h1, . . . , hp.

3.2 On-line algorithm

Using an automaton G, the size of the pool of experts H we consider can be very large. For example, in the case of the automaton of Figure 1, the size of the pool of experts is pl, and thus is exponentially large with respect to p. But, since learning guarantees in on-line learning admit only a logarithmic dependence on that size, they remain informative in this context. Nevertheless, the computational complexity of most on-line

图3 实验输出示例

测准确率、图表检测准确率和公式检测准确率。每一种准确率的计算方法均为正确检测到类型的文档数目除以总文档数目，其中正确检测的含义为包含且仅包含相应的完整对象内容。针对随机选择的100篇ACL论文，通过本文介绍的格式识别方法输出相应的Markdown格式文件，经过人工的检测和对比，标题检测准确率96%，正文检测准确率92%，图表检测准确率94%，公式检测准确率93%。结果如表1所示。

表1　针对ACL论文集的实验结果

检测类型	标题	正文	图表	公式
准确率/%	96	92	94	93

其中正文检测的准确率最低的原因是在机器视觉处理阶段，部分正文和其后的图表相邻过于紧密，被程序识别为一个内容对象所致。

在不需要复杂的人工特征构建和大规模语料库训练的情况下，针对ACL论文集总体上实验结果表明该方法以较小的人工代价和计算代价达到较为满意的准确率。

本文提供的方法有如下两个主要贡献：①利用本文的方法可以生成某一期刊或者会议的语料库，节省情报学学者人工构建语料库的时间成本；②利用该方法生成的语料库辅以人工检验之后，可以成为机器学习的训练资料，帮助训练以机器学习为核心的文档结构识别方案，达到扩充语料训练的目的。

5　小　结

本文针对现有的文档识别的主流方法的不足之处，提出了一种基于机器视觉的PDF文档结构识别的新方法，该方法无需复杂的人工特征构建或者大规模语料训练，以较小的计算代价达到了较好的准确率，并且能够识别包括公式、表格在内的论文的主要内容。本文的方法对语料库的构建和机器学习的训练资料生成都有一定的帮助。当然该方法还有一些不足之处需要改进，比如，机器视觉的识别准确率和启发式算法的分类标准均有提高和优化的空间，这也是笔者今后研究的重点方向。

参　考　文　献

[1] Mao S, Rosenfeld A, Kanungo T. Document structure analysis algorithms: a literature survey[C]// Document Recognition and Retrieval X. International Society for Optics and Photonics, 2003, 5010: 197-208.

[2] Nagy G, Seth S, Viswanathan M. A prototype document image analysis system for technical journals[J]. Computer, 1992, 25(7): 10-22.

[3] Baird H S, Jones S E, Fortune S J. Image segmentation by shape-directed covers[C]//Proceedings of the 10th International Conference on Pattern Recognition. IEEE, 1990: 820-825.

[4] O'Gorman L. The document spectrum for page layout analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15(11): 1162-1173.

[5] Kise K, Sato A, Iwata M. Segmentation of page images using the area Voronoi diagram[J]. Computer Vision and Image Understanding, 1998, 70(3): 370-382.

[6] Wahl F M, Wong K Y, Casey R G. Block segmentation and text extraction in mixed text/image documents[J]. Computer Graphics and Image Processing, 1982, 20(4): 375-390.

[7] Pavlidis T, Zhou J Y. Page segmentation and classification[J]. CVGIP: Graphical Models and Image Processing, 1992, 54(6): 484-496.

[8] Chen K, Seuret M, Liwicki M, et al. Page segmentation of historical document images with convolutional autoencoders[C]// Proceedings of the 13th International Conference on Document Analysis and Recognition. IEEE, 2015: 1011-1015.

[9] Chen K, Seuret M, Hennebert J, et al. Convolutional neural networks for page segmentation of historical document images[C]// Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. IEEE, 2017, 1: 965-970.

[10] Constantin A, Pettifer S, Voronkov A. PDFX: fully-automated PDF-to-XML conversion of scientific literature[C]// Proceedings of the 2013 ACM Symposium on Document Engineering. New York: ACM Press, 2013: 177-180.

[11] Yildiz B, Kaiser K, Miksch S. pdf2table: A method to extract table information from PDF files[OL]. http://citeseerx.ist.psu. edu/viewdoc/summary?doi=10.1.1.94.9382.

[12] Clark C, Divvala S. PDFFigures 2.0: Mining figures from research papers[C]//Proceedings of the 16th ACM / IEEE-CS on Joint Conference on Digital Libraries. New York: ACM Press, 2016: 143-152.

[13] Al-Zaidy R A, Giles C L. A machine learning approach for semantic structuring of scientific charts in scholarly documents[C]// Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications. Palo Alto: AAAI Press, 2017: 4644-4649.

[14] Siegel N, Lourie N, Power R, et al. Extracting scientific figures with distantly supervised neural networks[C]// Proceedings of the 18th ACM / IEEE Joint Conference on Digital Libraries. New York: ACM Press, 2018: 223-232.

[15] 王津涛, 康晓东, 李玫, 等. PDF 文件中可识别图像的提取[J].

计算机工程与设计, 2006, 27(9): 1539-1541.

[16] Tsujimoto S, Asada H. Understanding multi-articled documents [C]// Proceedings on 10th International Conference on Pattern Recognition. IEEE, 1990, 1: 551-556.

[17] Yamashita A, Amano T, Takahashi I, et al. A model based layout understanding method for the document recognition system[C]// Proceedings of the International Conference on Document Analysis and Recognition, Saint-Malo, France, 1991: 130-138.

[18] Ramesh S H, Dhar A, Kumar R R, et al. Automatically identify and label sections in scientific journals using conditional random fields[C]// Proceedings of Conference on Semantic Web Evaluation Challenge. Cham: Springer, 2016, 641: 269-280.

[19] Fauconnier J P, Kamel M. Discovering hypernymy relations using text layout[C]// Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. Stroudsburg: The Association for Computational Linguistics, 2015: 249-258.

（责任编辑 魏瑞斌）