杨华美，刘乐平，李文伟，等. 水工混凝土材料非结构化文本解析与表格数据库构建[J]. 水利水电技术(中英文)，2025，56(S2)：66-70.
YANG Huamei，LIU Leping，LI Wenwei，et al. Unstructured text analysis and table database construction of hydraulic concrete materials [J]. Water Resources and Hydropower Engineering，2025，56(S2)：66-70.

水工混凝土材料非结构化文本解析与表格数据库构建

杨华美[1]，刘乐平[2]，李文伟[1]，邓旭方[3]，李曙光[1]，陈正虎[3]，邓　伦[3]
(1. 中国长江三峡集团有限公司，北京　100038；2. 天津大学 水利工程智能建设与运维全国重点实验室，天津　300350；3. 中国长江电力股份有限公司，湖北 武汉)

摘　要：在水利工程历史建设过程中，受到文本信息化水平的限制，积累了大量以纸质文本和扫描图像形式保存的水工混凝土材料不可编辑文档，难以直接有效利用材料数据，极大增加了材料知识应用的难度。提出一种基于机器视觉和深度学习的文档解析方法，准确高效地将水工混凝土材料文本信息和表格数据转化为可编辑形式。进一步，基于已解译的表格信息，构建了水工混凝土材料表格数据库，实现了混凝土材料数据的高效查询和统一管理。以实际工程的水工混凝土材料文档为例验证新方法的可行性，结果表明，文档解析方法各项子任务的准确率均达90%以上，有助于混凝土材料不可编辑资源的自动化再利用。

关键词：水工混凝土材料；版面结构划分；文本检测与识别；表格数据库
DOI：10.13928/j.cnki.wrahe.2025.S2.016
中图分类号：TV431；TP391.1　　**文献标志码**：A　　**文章编号**：1000-0860(2025)S2-0066-05

Unstructured text analysis and table database construction of hydraulic concrete materials

YANG Huamei[1], LIU Leping[2], LI Wenwei[1], DENG Xufang[3], LI Shuguang[1], CHEN Zhenghu[3], DENG Lun[3]
(1. China Three Gorges Corporation, Beijing　100038; 2. State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, Tianjin　300350; 3. China Yangtze Power Co., Ltd., Wuhan　430014, Hubei, China)

Abstract: In the process of historical construction of water conservancy projects, limited by the level of text informatization, a large number of non-editable documents of hydraulic concrete materials have been accumulated in the form of paper texts and scanned images, making it difficult to directly and effectively utilize material data, greatly increasing the difficulty of applying material knowledge. A document parsing method was proposed based on machine vision and deep learning, which accurately and efficiently converts the text information and table data of hydraulic concrete materials into editable form. Furthermore, based on the interpreted table information, a database of hydraulic concrete material tables was constructed, achieving efficient querying and unified management of concrete material data. Taking the actual engineering hydraulic concrete material document as an example to verify the feasibility of new method, the result show that the accuracy of each subtask of the document parsing method is over 90%, which is helpful for the automated reuse of non-editable resources of concrete materials and improves the data

收稿日期：2024-10-15
基金项目：中国长江电力股份有限公司科研项目资助(Z212302036)
作者简介：杨华美(1986—)，女，副教授，博士，主要从事水工混凝土性能研究。E-mail：yang_ huamei@ ctg. com. cn
通信作者：刘乐平(2004—)，男，博士研究生，主要从事水利工程文本智能分析研究。E-mail：liuleping@ tju. edu. cn

service capability in the field of water conservancy engineering.

Keywords: hydraulic concrete materials; layout structure division; text detection and recognition; table database

0 引 言

从水利工程组织设计到运营投产的长期过程中，涉及许多不同种类混凝土的生产试验，产生了大量水工混凝土材料数据，清晰展示了工程建设中材料的使用情况。由于过往信息存储水平低，此类数据多是以纸质文本或扫描图片等形式保存的不可编辑文本，无法从中获取相应的文本信息，增加了水工混凝土数据管理与分析的难度[1]。目前工程领域多是采用人工解译的方式将不可编辑文档转化为电子格式文本，然而，人工转化费时费力且准确率不可控，难以适应基元类别多样的文档，准确高效地处理不可编辑文档数据成为了水利工程领域新的需求。

为获取不可编辑文档中的文本信息，各个领域的学者针对不可编辑文档转化任务开展了一系列研究，依据任务流程可分为版面结构划分、文本检测与识别等子任务[2]，大致经历了传统算法、基于机器学习的算法以及基于机器视觉的算法等三个阶段。为解决试验中纸质材料数据遗失问题，目前各大机构都在大力开发各种材料数据库，材料数据库系统已经逐渐成为材料基因工程的支柱。然而，水利工程的混凝土材料文本数据仍处于孤岛状态，各项研究尚未建立起其显、隐性数据之间的联系，难以从文本信息中挖掘潜在的有用信息，大量水工混凝土材料的文本数据缺乏统计与管理，不易被快速检索与高效利用[3]。因此，构建水工混凝土表格数据库可以提高信息存储、查询以及利用的能力，为后续智能分析工程文档打好坚实的基础。

基于上述需求，本文基于机器视觉和深度学习方法，提出了水工混凝土材料不可编辑文档解析方法，实现了快速高效获取大量可编辑的混凝土材料数据。进一步，建立了水工混凝土材料表格数据库，便于混凝土材料信息的快速查找和获取，实现了解译信息的智能管理和再利用，对水利水电工程管理、数据应用等方面具有重要意义。

1 研究框架

本文围绕水工混凝土材料不可编辑文档，通过图像处理、机器视觉以及深度学习等方法构建了水工混凝土材料表格数据库，研究方法总体结构如图 1 所示，主要研究内容如下。(1)划分水工混凝土材料文档版面结构。针对水工混凝土材料不可编辑文档，基于 OpenCV 机器视觉库构建基于机器视觉的版面结构划分模型，实现不可编辑 PDF 文档的智能分类任务。(2)提取水工混凝土材料多元文本数据内容。基于划分后的文字及表格图片，构建文档智能检测与识别模型，分别运用 CTPN + CRNN + CTC 模型以及 TableOCR2 模型处理非结构化数据以及结构化数据，保证材料数据提取的准确性。(3)构建水工混凝土材料多元数据库。以原始多元文本数据为导向，以文本、表格、图片为基础，建立水工混凝土材料表格数据库，智能存储多元混凝土材料数据，有助于进一步的材料数据分析与应用。

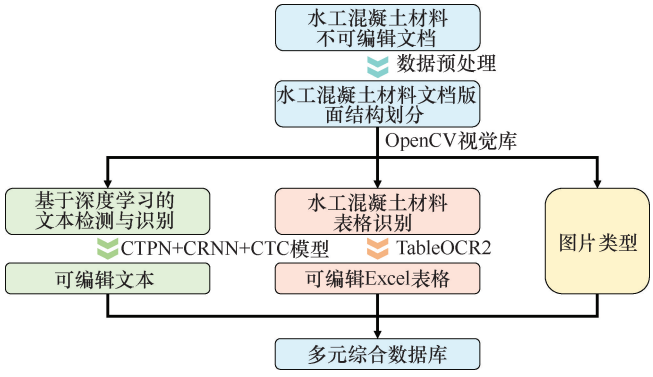


图 1 研究方法总体结构

2 基于机器视觉和深度学习的文档解析方法

2.1 水工混凝土材料文档版面结构划分

版面结构划分是不可编辑文档图像分析和处理的前提，旨在将不可编辑文档中的数据按照文字区域、表格区域、图像区域等类别进行划分。本文基于 OpenCV 机器视觉库构建版面结构划分方法，实现了水工混凝土材料文档中文字、表格、图像等区域的划分，其包含图像识别预处理、表格线位置检测、表格内容块位置信息获取以及文字内容块位置信息获取。

2.2 基于深度学习的文本检测与识别

基于对各类文本检测模型的分析，采用 CTPN 作为检测本文文本的算法。CTPN 是基于 Faster R-CNN 改进的一种文本检测模型，包括卷积层、循环层以及全连接层三部分，主要由 VGG16 模型，Bi-LSTM 模型、FC 层以及 RPN 网络组成。采用基于 CTC 的文本

识别模型识别本文文档，即采用 CRNN+CTC 的网络结构。其中，CRNN 是一种卷积神经网络，主要负责获取文本序列特征并对文本进行识别；CTC 是一种计算损失函数的方法，负责将预测结果与实际文本所在区域对齐。

2.3 混凝土材料表格识别与数据库构建

表格识别旨在从图像等不可编辑的媒介中提取出结构化数据，将表格中的数据信息识别为文本字符并将表格数据按照原有格式存放。本文采用 TableOCR 表格识别模型完成水利工程混凝土材料文档的表格识别，该模型支持图片内常规表格、无线表格以及多表格的检测和识别，且可以将识别结果保存为 Excel 格式。

水工混凝土材料信息的存储与运用是本研究的最终目的，从水工混凝土材料不可编辑文档中提取的文本信息需要由知识管理系统统一管理，随时检索利用才能体现其应用价值。基于识别出的 EXCEL 表格信息，选用 MySQL 数据库构建水工混凝土材料表格数据库，此数据库主要分为三个部分，分别为原料库、试件信息库和试验数据库。

3 工程实例分析

3.1 工程数据预处理

本文采用溪洛渡水电站工程的水工混凝土材料文档作为数据源，其包含 8 个章节，共 286 页，涵盖了原料组成成分、原料性能试验数据，混凝土组成成分以及混凝土性能试验数据等内容。在文档结构方面，文档格式为不可编辑的 PDF 文档，包含文字、表格以及图片等结构化及非结构化数据。因不可编辑 PDF 文档结构复杂且较难解析，无法直接从 PDF 文档中得到版面结构的元素信息，因此将 PDF 文档进行文档格式转化、图像倾斜校正后，再对图像文档进行版面结构划分。

3.2 评价指标

本文采用精确率、召回率以及 *F*1 分数三个指标评价本文版面结构划分、文本检测与识别以及表格识别等各项子任务的效果。其中，精确率指正确预测为正类的样本占所有正类样本的比例；召回率指正确预测为正类的样本占预测为正类的比例；*F*1 分数将精确率和召回率结合起来，可实现对样本进行综合评价。

3.3 试验结果与分析

3.3.1 版面结构划分

经过 OpenCV 机器视觉库对本文研究报告进行版面结构划分后，共有 540 张图片被划分出来。采用精确率、召回率以及 *F*1 分数三个指标结果评估本文方法的可行性，可知版面结构划分文本区域的准确率为 88.77%，召回率为 94.40%，*F*1 分数为 91.50%。分析已划分的研究报告，发现由于柱状图结构与表格结构相似，导致部分柱状图被错误提取出来。

为减少柱状图被提取的数量、提高表格提取的正确率，分别对版面结构划分的轮廓区域面积阈值以及区域内交点个数阈值进行调整。轮廓区域面积代表版面结构划分提取区域的大小，区域内交点个数阈值代表轮廓区域内横竖线的交点个数，若此两项指标阈值设置过大，会导致小面积或交点个数较少的区域无法被提取，造成已识别表格的数量较少；反之，则会导致不符合表格面积要求或表格交点个数要求的区域被提取，导致错误识别表格的数量较多。以精确率、召回率以及 *F*1 分数作为评价指标，分析轮廓区域面积和区域内交点个数对表格划分结果的影响。总体来讲，在区域内交点数量的阈值为 20 时，表格划分的精确率和召回率都较高，并且 *F-Score* 的值最大，代表此时表格划分的效果较好，因此区域内交点个数阈值取值为 20。

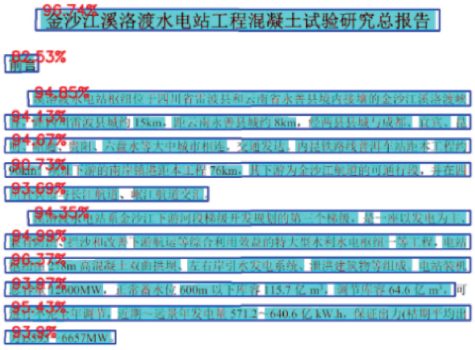
以上述确定的阈值对本文研究报告进行版面结构划分，采用精确率、召回率以及 *F*1 分数三个指标作为评价表格划分效果的指标。若划分的图片中包含表格信息，即认为该版面结构划分正确，否则视为版面结构划分错误。以上述指标为评价依据，可得表格划分的准确率为 96.34%，召回率为 97.77%，*F*1 分数为 97.05%。

3.3.2 文本检测与识别

文本检测通常使用交互比表示文本检测算法预测文本框的效果，以文本行为单位，标注每一行文本检测框的交互比，规定若 $IoU(A, B) \geqslant 0.5$，视为文本被成功检测，否则认为检测的文本候选框不合格。基于本文研究报告可得，文本检测模型的各行检测正确率均可达到 90% 以上，综合检测正确率可达到 95.8%。在文本识别任务中，采用精确率、召回率以及 *F*1 分数三个指标结果进行判断。若文本识别数据与原文本数据结果一致时，即认为该文本数据识别正确，若上下标没有按照原有样式标注、标点符号及单位未识别成功，则认为该文本数据识别错误。以上述指标和规定为依据，可得文本识别模型的精确率为 97.39%，召回率为 97.93%，*F*1 分数为 97.66%。以红色框线圈出识别错误文本位置，文本检测与识别样例如图 2 所示。

3.3.3 表格识别与数据库构建

在表格结构检测中，采用上述表格评价指标评估表格结构检测的效果，可得表格结构检测的精确率为98.02%，召回率为99.43%，*F*1分数为98.72%，表格结构检测可视化效果如图3(a)所示。表格数据识别是指将表格图片中检测到的文本、数学符号等数据识别为可编辑字符数据。当表格识别数据与原文本数据结果一致时，即认为该单元格数据识别正确，否则为识别错误。采用精确率、召回率以及*F*1分数对表格数据识别效果进行评价，可得表格数据识别的精确



金沙江溪洛渡水电站工程混凝土试验研究总报告
前言
溪洛渡水电站枢纽位于四川省雷波县和云南省永善县境内接壤的金沙江溪洛渡峡谷，距四川雷波县城约15km，距云南永善县城约8km，经两县县城与成都、宜宾、昆明、昭通、贵阳、六盘水等大中城市相连，交通发达。内昆铁路线普洪车站距本工程约90km。坝址下游的南岸镇港距本工程76km,其下游为金沙江航道的可通行段，并在四川宜宾市与长江航道、岷江航道交汇
溪洛渡水电站系金沙江下游河段梯级开发规划的第三个梯级，是一座以发电为主、兼有防洪、拦沙和改善下游航运等综合利用效益的特大型水利水电枢纽一等工程，电站枢纽由278m高混凝土双曲拱坝、左右岸引水发电系统、泄洪建筑物等组成。电站装机总容量12600MW，正常蓄水位600m以下库容115.7亿m，调节库容64.6亿m，进行不完全年调节，近期～远景年发电量571.2～640.6亿kWh，保证出力(枯期平均出力)3395～6657MW。

图2 文本检测与识别样例

表1-1水泥物理力学性能检验结果(试验中心)

试验编号	生产厂家 及品种	密度(g/cm')	比表面积(m2/kg)	稠度(%)	凝结时间(h:min) 初凝	凝结时间(h:min) 终凝	安定性	烧失量(%6)	SO; (%)	碱含量(%)	MgO(%)	抗折强度(MPa) 3d	抗折强度(MPa) 7d	抗折强度(MPa) 28d	抗压强度(MPa) 3d	抗压强度(MPa) 7d	抗压强度(MPa) 28d	水化热(Jkg) 3d	水化热(Jkg) 7d
1	重庆腾辉42.5中热	3.18	342	26.0	3:26	4:12	合格	1.02	1.69	0.50	3.51	4.2	5.8	8.5	18.8	29.7	57.9	239	281
2	广安腾辉42.5中热	3.23	277	22.8	3:01	3:52	合格	0.66	1.81	0.45	3.71	3.7	4.7	7.8	16.3	23.4	51.0	219	246
125	贵州水城42.5中热	3.23	256	23.2	3:01	3:55	合格	0.95	1.68	0.44	3.58	3.3	5.0	8.3	16.0	24.7	51.0	195	240
154	四川峨嵋42.5中热	3.21	334	23.2	2:01	2:46	合格	0.75	1.99	0.56	4.11	4.3	5.7	8.9	19.4	29.0	45.6	217	232
5	四川双马42.5中热	3.22	341	22.9	2:30	3:15	合格	0.74	1.54	0.42	3.26	3.2	5.1	8.1	15.2	25.2	54.6	218	253
6	贵州乌江42.5中热	3.12	334	23.9	5:24	6;09	合格	1.74	2.48	0.60	2.06	4.5	6.4	8.2	22.2	33.8	51.0	220	255
7	四川嘉华42.5低热	3.22	372	26.0	2:42	3;47	合格	1.10	2.66	0.47	2.97	1	5.3	8.4	—	23.2	54.0	212	243
4	贵州水城42.5低热	3.23	248	22.3	2:55	3;35	合格	0.76	1.88	0.35	3.80	1	3.7	7.2	/	16.2	45.0	199	252
155	四川峨嵋42.5低热	3.22	396	24.7	2:20	3:31	合格	1.10	3.04	0.53	3.75	/	3.4	8.3		13.8	45.6	177	193
GB200-2003	42.5中热	乡250 >60min				≤12h	合格	3.0 3.5		<0.60	5.0	≥3.0	>4.5	>6.5	≥12.0	>22.0	>42.5	<251	<293
42.5低热												>3.5 >6.5 ≥13.0 >42.5 ×230 ≤260							

(a) 表格结构检测可视化效果

表1-2水泥物理力学性能检测结果(成勘院)

序号	水泥品种	细度(%)	密度(g/cm')	标稠(%)	初凝(h:min)	终凝(h:min)	安定性	抗压强度(MPa) 3d	抗压强度(MPa) 7d	抗压强度(MPa) 28d	抗折强度(MPa) 3d	抗折强度(MPa) 7d	抗折强度(MPa) 28d	水化热(直接法)(kJ/kg) 3d	水化热(直接法)(kJ/kg) 7d
一	贵州水城低热	2.3	3.28	24.0	3:17	4:03	合格	13.2	17.6	47.2	3.3	4.6	8.4	170	221
2	水城中热(6月)	1.0	3.26	23.6	3:37	4:38	合格	15.9	22.1	45.5	3.2	5.2	7.0	221	276
3	水城中热(9月)	1.1	3.25	23.0	3:45	4:40	合格	16.3	27.3	50.0	3.5	5.3	8.3	210	268
4	贵州乌江中热	2.8	3.23	23.6	4;34	5:28	合格	25.2	34.2	51.9	5.2	6.8	8.7	220	281
5	江油中热	1.8	3.24	22.8	2:50	3;41	合格	17.1	23.0	50.1	3.9	5.7	8.4	214	271
6	重庆腾辉中热	0.9	3.24	26.4	3:20	4:06	合格	21.8	24.8	52.3	4.8	5.9	7.3	224	286
7	广安腾辉中热	1.0	3.26	25.8	3:33	4:17	合格	21.2	22.2	52.2	4.9	5.2	9.2	225	271

(b) 表格识别可视化结果

(c) 表格存储示例

图3 表格识别可视化结果与数据库存储示例

率为 94.13%，召回率为 95.51%，*F*1 分数为 94.82%，表格识别可视化结果如图 3(b)所示。由于表中存在含有下标的数据，规定若上下标没有按照原有样式标注，则认为该单元格数据识别错误，用红框以及红字进行标注。

基于已识别的水工混凝土材料表格数据，采用 MySQL 数据库建立水工混凝土材料数据库，将已识别的表格文件以列名为属性存储至数据库中。以表格名称作为数据库表名，共存放 273 张表格，实现了表格的自动存储和智能查询的功能，表格存储示例如图 3(c)所示。

4 结 论

针对上述问题，本文依托水工混凝土材料的不可编辑文档，构建了水工混凝土材料表格数据库，为混凝土材料文档提供“文档划分-信息识别-数据管理”的一站式数据服务，主要研究内容如下。

(1)本文结合 OpenCV 机器视觉库和 PP-Structure 模型库提出一种适用于水工混凝土材料不可编辑文档的版面结构划分方法，高效准确的将文档版面划分为文字、表格两类区域。

(2)考虑到文字与表格的数据结构不同，本文采用不同的高精度模型对文档信息进行检测与识别。以 CTPN+CRNN+CTC 模型作为文字检测与识别模型，以 TableOCR2 作为表格检测与识别模型，实现了对文字与表格数据的高精度提取。

(3)基于识别的表格结构及表格信息，基于 Mysql 构建了水工混凝土材料表格数据库，存储文档中每个表格的数据信息，实现了材料数据智能存储、查询、管理等操作。

参考文献：

[1] 何殷鹏，张梦溪，李文伟，等. 金沙江下游水电站数字混凝土研究与应用[J]. 水力发电学报，2022，41(10)：1-17.

[2] 王珂，杨芳，姜杉. 光学字符识别综述[J]. 计算机应用研究，2020，37(S2)：22-24.

[3] 刘海定，汤爱涛，潘复生，等. 材料科学数据库的研究现状及其发展趋势[J]. 材料导报，2004，18(9)：5-7.

（责任编辑 王 璐）