# 人工智能时代的文字识别技术

XX

Feb 2023

**摘要**

文字普遍存在于人们的日常生活中。OCR 技术可以将图像中的文字转换成便于计算机所理解的机器编码的形式，可以提高计算机在图文处理中的泛用性和效率。OCR 有着悠久的历史，随着时间的推移和技术的进步，现在已经有多种方法来实现 OCR 技术。本论文综述中收集了各种与 OCR 领域有关的文献。介绍了当前 OCR 领域中的具有代表性的数据库、流行的研究方法以及研究趋势等。

**关键词：** OCR、分类、人工智能、机器学习、深度学习

## 1 简介

计算机有着高速的检索速度，使用计算机对文章进行检索与人工查看各种图像文本相比，不仅效率更高，而且成本更低。然而，人们在生活中常见的容易理解的图像文本形式却不容易被计算机所理解。

光学字符识别（Optical Character Recognition, OCR）是一种将输入的图像文本转换成计算机容易理解的机器编码的技术 [1]。OCR 能够处理不同的场景的图像，如打印的纸质材料，手写的纸质材料，视频中显示的字幕，道路上的广告标语等等。

OCR 历史悠久，早在 1965 年，在纽约世博会展示的机器 “IBM 1287”，是有史以来第一个光学阅读器，能够读取手写数字，被视作 OCR 领域的先驱。在 20 世纪 70 年代，人们主要专注于提高 OCR 的响应时间和性能 [2] 。21 世纪后，引入了二值化技术，人们主要使用传统的机器学习方法设计 OCR ，如支持向量机、随机森林、K 近邻算法、决策树 [3] 等。近年来，随着深度学习技术的不断成熟和计算机硬件能力的不断提升，对 OCR 的研究逐渐转向深度学习方法，人们不再强调手工的特征提取技术，使用的深度

学习方法包括卷积神经网络、循环神经网络、长短期记忆神经网络、自注意力机制 [4] 等。自然语言处理技术也开始与 OCR 相结合，赋予 OCR “理解”文字内容的能力，进一步提高了 OCR 对文字的识别率 [5]。到目前为止，虽然机器识别文字的性能得到了显著的提升，但是与人理解图像文字的能力仍存在一定的差距，提升 OCR 的鲁棒性、效率和智能性仍是 OCR 发展的主要趋势 [6]。

图 1: IBM 1287

在 OCR 中，根据输入数据的形式可进一步分为脱机识别和联机识别。脱机识别的输入数据是以扫描图像的形式出现的，图像中的文字形式又可以进一步分为印刷体文字和手写体文字。而联机识别的输入数据性质是动态的，输入的数据具有一定的速度，位置和定位点的笔尖运动，联机识别往往比脱机识别更加复杂 [7]。

本文献综述的组织结构如下：第 2 章介绍 OCR 领域常用到的各种数据集（中文、英文、数字）。第 3 章介绍 OCR 领域使用的方法（传统机器学习方法、深度学习方法及性能指标）。第 4 章介绍 OCR 领域的研究趋势。第 5 章对本次系统文献综述做出总结。

## 2 数据集

衡量不同的 OCR 的好坏，需要使用标准化的数据集来进行比较。在机器学习领域中，也需要包含足够数量的数据的数据集对模型进行训练。对于

文字的形式，分为印刷体文字和手写体文字。不同语言的文字也有着不同的特征，OCR 在不同语言中的文字识别性能也会有差异。以下各小节介绍了中文、英文、数字中公开可用的数据集。

## 2.1 MNIST

MNIST（Mixed National Institute of Standards and Technology database）被认为是机器学习领域最常用的数据集之一。它是 NIST 数据集的一个子集。该数据集包含 60000 张图像的训练集和 10000 张图像的测试集。训练集由来自 250 个不同人手写的数字构成，其中 50% 是高中学生，50% 来自人口普查局的工作人员，测试集也是同样比例的手写数字数据。图像均经过大小归一化处理成 $28 \times 28$ 像素的灰度图像 [8]。

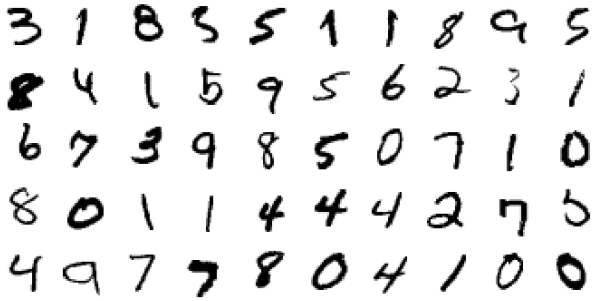


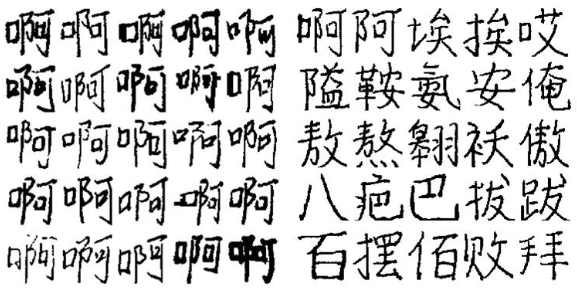
图 2: MNIST 数据集

## 2.2 HCL 2000

HCL 2000 是一个脱机手写汉字数据集。该数据库中共有 3755 个常用汉字，由 1000 个不同的笔者书写。笔者的背景（如年龄、职业、性别和受教育程度等）亦被纳入数据集中 [9]。

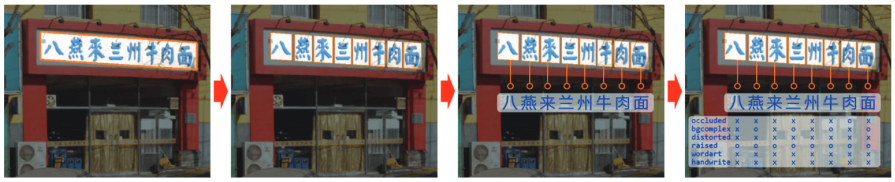
## 2.3 Chinese Text in the Wild

Chinese Text in the Wild 包含了来自腾讯街景从中国几十个不同城市拍摄的 32285 幅图像，共有 1018402 个汉字，3850 种汉字。每个图像中，所有中文文本都有其字符、边界框、是否被遮挡、背景复杂、失真、3D 突起、是否为艺术字体、是否为手写体等注释 [10]。



(a) (b)

图 3: HCL2000 数据集




(a) (b) (c) (d)

图 4: Chinese Text in the Wild 数据集

## 2.4 IAM

IAM 是基于的 Lancaster-Oslo/Bergen (LOB) 语料库的英文手写数据库。由 400 位笔者书写，共有 10055 种英文文本，其中包含 82227 个单词 [11]。

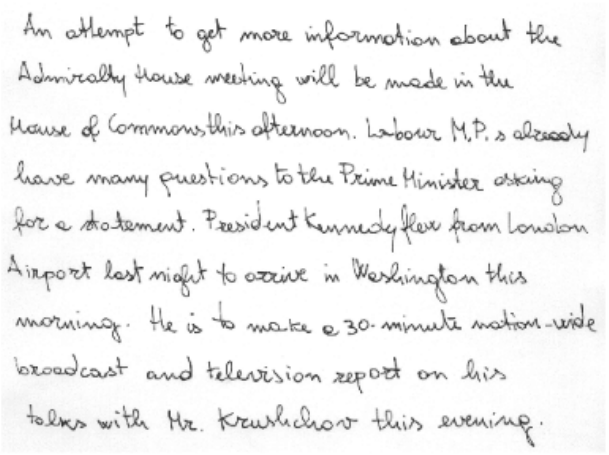



图 5: IAM 数据集

## 2.5 ICDAR 2013

ICDAR 是国际文档分析与识别领域的顶级会议之一。ICDAR 2013 是一个英文标注的自然场景图片的数据集。训练集包含 229 张图像，测试集包含 233 张图像 [12]。


图 6: ICDAR 2013 数据集

### 2.6 Total-text

Total-Text 是一个英文标注的弯曲文本的自然场景图片的数据集。训练集包含 1255 张图像，测试集包含 300 张图像。该数据集可以评估 OCR 对弯曲文本的鲁棒性 [13]。


图 7: Total-Text 数据集

## 3 研究方法

21 世纪后，人们主要使用机器学习方法来设计 OCR。OCR 在机器学习领域中一般被视为是一个分类问题。分类是一个在给定的输入数据上学习模型的过程，并将其映射或标记为预定义的类别或类 [3]。本章主要从传统机器学习方法和深度学习方法两个方面介绍 OCR 的研究方法。

### 3.1 传统机器学习方法

一般来说，使用传统机器学习方法的 OCR 需要对输入的图像进行图像预处理、特征提取和文字识别等操作。

#### 3.1.1 图像二值化

图像二值化是图像预处理的常用操作。它将整幅图像仅用两种颜色来表示，以丢弃大多数无用的信息，便于后续的处理。将灰度大于阈值 T 的像素点的灰度设为 1，将灰度小于阈值 T 的像素点的灰度设为 0。

$$g(x,y) = \{\,1\,, f(x,y) < T0, f(x,y) \geq T$$

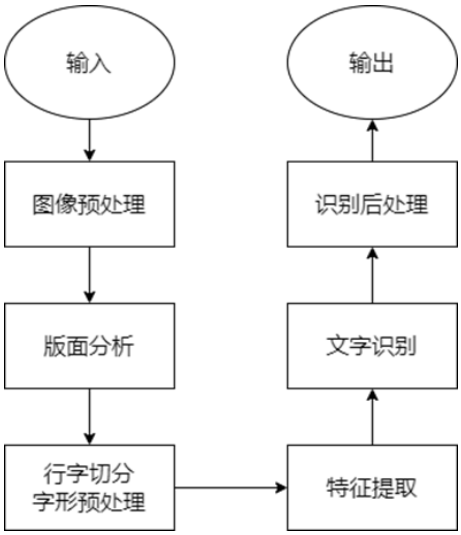



图 8: 传统机器学习方法流程图

阈值 T 既可由人为设定一个定值，也可使用 OSTU 等算法由计算机通过图像特征计算得到 [14]。

图 9: 莱娜图（左）与图像二值化处理后的莱娜图（右）

### 3.1.2 图像预处理

#### 3.1.2.1 图像规范化

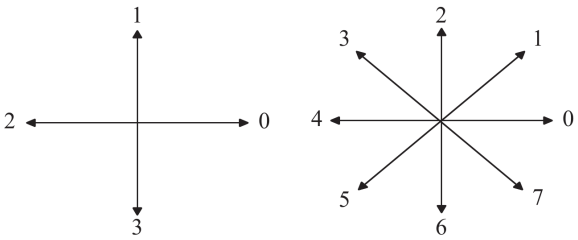
字符图像的多样性是 OCR 的一个主要难题，使用图像规范化可以减少字符图像的形状变化。常见的规范化技术有尺寸归一化、矩归一化、剪切变化、透视变换、非线性规范化等 [15]。

#### 3.1.2.2 图像降噪

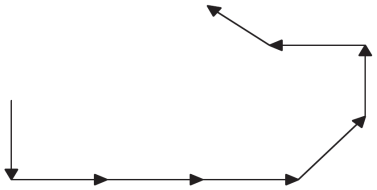
图像降噪可以使得图像保留大部分的主要特征的同时，又能够去除图像中的无用信息。常用的传统图像算法有高斯滤波、均值滤波、中值滤波、NLM 算法等 [16]。

### 3.1.3 链码直方图

在传统机器学习方法中，文字识别的关键就是区别字符图像中的不同特征。对于文字，最常用的特征提取技术之一是链码直方图（CCH）。链码直方图可以有效地描述图像/字符边界/曲线等特征，从而帮助对字符进行分类。还有其他类型的方向特征，如 NCFE 特征和梯度特征等。将这些特征结合成一个特征向量以进行之后的文字识别操作 [17]。




(a) 4-邻接链码　　(b) 8-邻接链码



链码：6->0->0->0->0->
1->2->4->3

(c) 某段轮廓的8-邻接链码

图 10: 链码直方图

### 3.1.4 文字识别算法

#### 3.1.4.1 k-近邻算法（kNN）

k-近邻算法是一种用于分类的非参数机器学习算法，既简单又有效。给定一个数据集，对于数据中集的每一个输入，在训练数据集中找到与该实例中最

邻近的 k 个实例，这 k 个实例的多数属于某个类，就把该输入实例分为这个类 [18]。

#### 3.1.4.2 支持向量机（SVM）

在深度学习方法普及之前，支持向量机是手写数字识别、图像分类、人脸检测、对象检测和文本分类最强大的技术之一 [19]。支持向量机是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。[18] 在数据集中数据数量较少的情况下，支持向量机有时的表现比深度学习方法还要好。

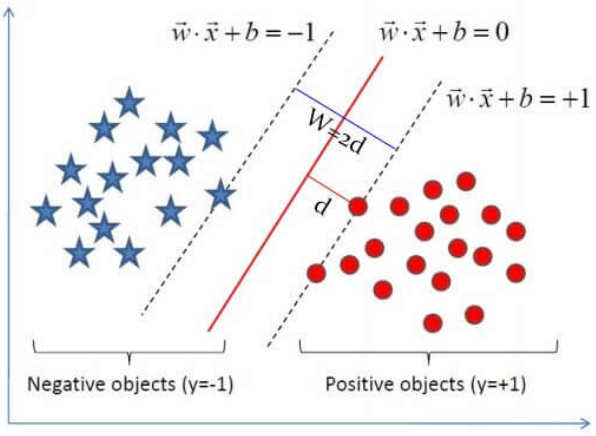



图 11: 支持向量机示意图

## 3.2 深度学习方法

传统的机器学习方法受限于手工设计特征的表达能力和处理流程的复杂性，在复杂场景下很难达到理想的文字识别效果，深度学习技术的出现很好地弥补了这一不足 [6]。

### 3.2.1 神经网络（NN）

神经网络由生物神经元的结构启发得来。通过数学可以证明，只要神经元的个数足够多，神经网络能拟合任意函数。神经网络对给定的输入数据进行建模，并将其映射到预定的类别中。最典型的神经网络结构是全连接层神

经网络结构。随着深度学习技术的出现，神经网络层数越来越多。仅使用全连接层神经网络结构处理计算机视觉类的问题容易导致过拟合的现象。因而出现简化参数量的循环神经网络和卷积神经网络等结构。目前，神经网络被视作 OCR 所使用的最佳分类技术之一。

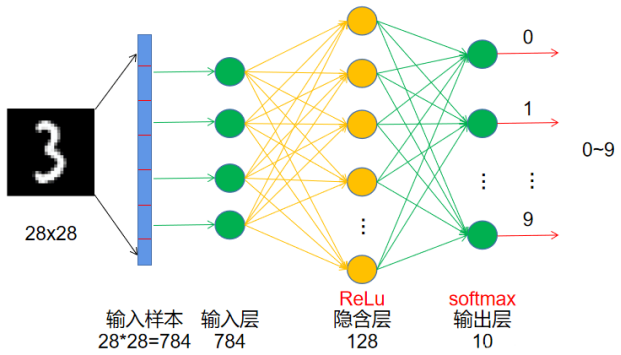



图 12: 全连接层神经网络示意图

### 3.2.2 卷积神经网络（CNN）

卷积神经网络在 OCR 领域中取得巨大成效 [20] 因而应用广泛。一般完整的卷积神经网络流程包括：读入输入图像、进行多次卷积、池化操作、Flatten 操作、多层全连接层操作、softmax 处理、输出分类结果。

Melnyk 等人提出了一种基于卷积神经网络的离线手写汉字识别的架构，并且可以解决神经网络中的可解释性问题 [21]。

卷积神经网络训练出的模型并不能识别图像的缩放和旋转，空间变换网络（STNs）是 Jaderberg 等人提出的一种卷积神经网络架构模型，通过变换输入的图片，降低受到数据在空间上多样性的影响，来提高卷积网络模型的分类准确率，而不是通过改变网络结构 [22]。

### 3.2.3 循环神经网络（RNN）

循环神经网络是一类以序列数据为输入，在序列的演进方向进行递归且所有节点按链式连接的神经网络。如果将输入的图像视作一组序列，也可将循环神经网络应用于 OCR 领域，Farisa Benta Safir 等人就使用了两种不

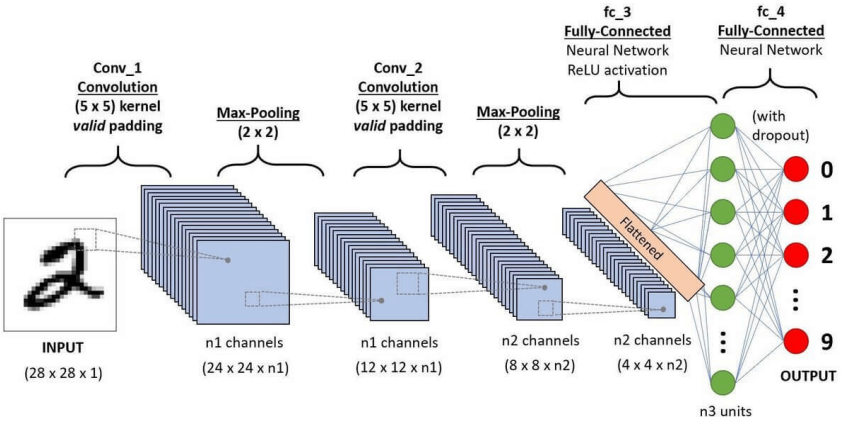



图 13: 卷积神经网络示意图

同的循环神经网络：长短期记忆网络（LSTM）和门控循环单元（GRU）构建了一个用于孟加拉语的端到端的 OCR 系统 [23]。

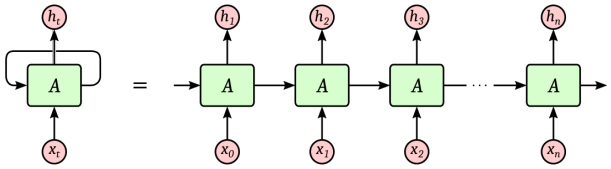



图 14: 循环神经网络示意图

### 3.2.4 自注意力机制（Self-Attention）

自注意力机制最早由 Google 的 Transformer 架构中所提出并使用，取代了以往自然语言处理任务中的循环神经网络结构 [24]。此后，在文本识别领域中，Transformer 模型被频繁采用，其结构的优势带来了显著的效率提升。TrOCR 利用了图像 Transformer 作文文本编码器，仅使用一个简单的编码器-解码器模型的情况下，TrOCR 在打印文本和手写文本识别中均取得了当时最先进的准确率 [25]。

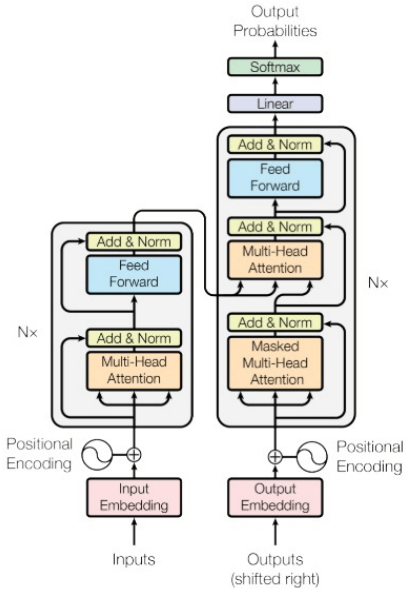



图 15: 自注意力机制示意图

#### 3.2.5 OCR 与自然语言处理技术的结合

近年来，OCR 中开始引入自然语言处理技术，增强了对内容的“理解”能力，通过对上下文语义信息的关联，OCR 在复杂场景下的文字识别能力得到增强。Javier Ferrando 等人将 OCR 识别的结果与 BERT 模型产生的预测相结合，提高了 OCR 的准确性 [5]。

### 3.3 评价指标

OCR 以字符为单位进行统计和分析，评价的指标有：

1. 字段召回率，指被完全正确识别字段（测试输出结果与字段的所有字符完全匹配）数量与总字段数比值。

2. 字段准确率，指被完全正确识别字段（测试输出结果与字段的所有字符完全匹配）数量与测试返回识别结果的字段数量比值。

3. 字符召回率（recall），指被完全正确识别字符数量与真实字符总数的比值，可以反应识别错和漏识别的情况。

4. 字符准确率（precision），指被完全正确识别字符数量与测试返回的字符数的比值，可以反应识别错和多识别的情况。

5. $F_\beta - Score$，可以综合反映字符识别召回效果和字符识别准确效果，计算公式如下：

$$F_\beta - Score = (1 + \beta^2) \cdot \frac{recall \cdot precision}{\beta^2 \cdot (recall + precision)}$$

当 $\beta = 1$ 时，被称为 F1 分数，又称为平衡 F 分数，其值为精确率和召回率的调和平均数 [6]。

# 4 研究趋势

OCR 领域的技术迭代十分迅速。近年来，OCR 领域的研究已从传统机器学习方法逐渐转为深度学习方法，很少强调手工提取图像特征。卷积神经网络在 OCR 领域的研究中被广泛使用。在自注意力机制被提出以后，又逐渐取代了卷积神经网络在 OCR 中的地位。传统机器学习方法也在和深度学习方法混合使用。自然语言处理技术也开始和 OCR 混合使用，通过对上下文的推断进一步提高了 OCR 识别的准确率。

”text in the wild” 的识别仍是 OCR 的一个难点。由于这类图像的场景复杂，OCR 需要考虑背景噪声、透视、光照、不同字体等因素的影响。还需要一个足够全面的数据集以尽可能包含日常生活中出现的文字变化 [26]。

# 5 结论

OCR 有着悠久的历史。人工智能技术的发展使着 OCR 的性能越来越好。本文献综述中主要介绍了 OCR 领域中一些具有代表性的数据集。传统机器学习方法往往需要对数据集进行图像预处理和手工提取特征。k-近邻算法、支持向量机都是传统机器学习方法的代表性算法之一。而传统机器学习方法在复杂场景下很难达到理想的文字识别效果，本文献综述还介绍了计算机视觉的一些深度学习架构（全连接层神经网络、卷积神经网络、循环神经网络、自注意力机制等）在 OCR 领域中的应用。

# 参考文献


[1] C.C. Tappert, C.Y. Suen, and T. Wakahara. "The State of the Art in Online Handwriting Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.8 (Aug./1990), pp. 787–808. ISSN: 01628828. DOI: 10.1109/34.57669. URL: http://ieeexplore.ieee.org/document/57669/.

[2] S. Mori, C.Y. Suen, and K. Yamamoto. "Historical Review of OCR Research and Development". In: *Proceedings of the IEEE* 80.7 (July 1992), pp. 1029–1058. ISSN: 00189219. DOI: 10.1109/5.156468. URL: http://ieeexplore.ieee.org/document/156468/.

[3] 周志华. 机器学习. 清华大学出版社, Jan. 1, 2016. 425 pp. ISBN: 978-7-302-42328-7.

[4] 伊恩 · 古德费洛, 约书亚 · 本吉奥, and 亚伦 · 库维尔. 深度学习. Trans. by 赵申剑 et al. 人民邮电出版社, July 1, 2017. 500 pp. ISBN: 978-7-115-46147-6.

[5] Javier Ferrando et al. "Improving Accuracy and Speeding up Document Image Classification through Parallel Systems". In: vol. 12138. 2020, pp. 387–400. DOI: 10.1007/978-3-030-50417-5_29. arXiv: 2006.09141 [cs]. URL: http://arxiv.org/abs/2006.09141.

[6] 中国信息通信研究院云计算与大数据研究所, 中国人工智能产业发展联盟, and 深圳市腾讯计算机系统有限公司. 智能文字识别（*OCR*）能力评测与应用白皮书. Sept. 2020. URL: http://aiiaorg.cn/index.php?m=content&c=index&a=show&catid=14&id=186.

[7] Scott D. Connell and Anil K. Jain. "Template-Based Online Character Recognition". In: *Pattern Recognition* 34.1 (Jan. 2001), pp. 1–14. ISSN: 00313203. DOI: 10.1016/S0031-3203(99)00197-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S0031320399001971.

[8] Cheng-Lin Liu et al. "Handwritten Digit Recognition: Benchmarking of State-of-the-Art Techniques". In: *Pattern Recognition* 36.10 (Oct. 2003), pp. 2271–2285. ISSN: 00313203. DOI: 10.1016/S0031-3203(03)

00085-2. URL: https://linkinghub.elsevier.com/retrieve/pii/S0031320303000852.

[9] Honggang Zhang et al. “HCL2000 - A Large-scale Handwritten Chinese Character Database for Handwritten Character Recognition”. In: *2009 10th International Conference on Document Analysis and Recognition*. 2009 10th International Conference on Document Analysis and Recognition. Barcelona, Spain: IEEE, 2009, pp. 286–290. ISBN: 978-1-4244-4500-4. DOI: 10.1109/ICDAR.2009.15. URL: http://ieeexplore.ieee.org/document/5277700/.

[10] Tai-Ling Yuan et al. “A Large Chinese Text Dataset in the Wild”. In: *Journal of Computer Science and Technology* 34.3 (May 2019), pp. 509–521. ISSN: 1000-9000, 1860-4749. DOI: 10.1007/s11390-019-1923-y. URL: http://link.springer.com/10.1007/s11390-019-1923-y.

[11] U.-V. Marti and H. Bunke. “The IAM-database: An English Sentence Database for Offline Handwriting Recognition”. In: *International Journal on Document Analysis and Recognition* 5.1 (Nov. 1, 2002), pp. 39–46. ISSN: 1433-2833, 1433-2825. DOI: 10.1007/s100320200071. URL: http://link.springer.com/10.1007/s100320200071.

[12] Fei Yin et al. “ICDAR 2013 Chinese Handwriting Recognition Competition”. In: *2013 12th International Conference on Document Analysis and Recognition*. 2013 12th International Conference on Document Analysis and Recognition (ICDAR). Washington, DC, USA: IEEE, Aug. 2013, pp. 1464–1470. ISBN: 978-0-7695-4999-6. DOI: 10.1109/ICDAR.2013.218. URL: http://ieeexplore.ieee.org/document/6628856/.

[13] Chee Kheng Chng and Chee Seng Chan. *Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition*. Oct. 28, 2017. arXiv: 1710.10400 [cs]. URL: http://arxiv.org/abs/1710.10400.

[14] Rafael C. Gonzalez and Richard E. Woods. 数字图像处理 *(第四版)*. Trans. by 阮秋琦 and 阮宇智. 电子工业出版社, May 1, 2020. 748 pp. ISBN: 978-7-121-37747-1.

[15] Cheng-Lin Liu et al. “Aspect Ratio Adaptive Normalization for Handwritten Character Recognition”. In: *Advances in Multimodal Interfaces —ICMI 2000*. Ed. by Tieniu Tan, Yuanchun Shi, and Wen Gao. Red. by Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen. Vol. 1948. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 418–425. ISBN: 978-3-540-41180-2 978-3-540-40063-9. DOI: 10.1007/3-540-40063-X_55. URL: http://link.springer.com/10.1007/3-540-40063-X_55.

[16] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. “Non-Local Means Denoising”. In: *Image Processing On Line* 1 (Sept. 13, 2011), pp. 208–212. ISSN: 2105-1232. DOI: 10.5201/ipol.2011.bcm_nlm. URL: https://www.ipol.im/pub/art/2011/bcm_nlm/?utm_source=doi.

[17] Cheng-Lin Liu et al. “Handwritten Digit Recognition: Investigation of Normalization and Feature Extraction Techniques”. In: *Pattern Recognition* 37.2 (Feb. 2004), pp. 265–279. ISSN: 00313203. DOI: 10.1016/S0031-3203(03)00224-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S0031320303002243.

[18] 李航. 统计学习方法 (第 2 版). 清华大学出版社, May 1, 2019. 464 pp. ISBN: 978-7-302-51727-6.

[19] Abdelhak Boukharouba and Abdelhak Bennia. “Novel Feature Extraction Technique for the Recognition of Handwritten Digits”. In: *Applied Computing and Informatics* 13.1 (Jan. 2017), pp. 19–26. ISSN: 22108327. DOI: 10.1016/j.aci.2015.05.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S221083271500006X.

[20] Cheng-Lin Liu and Hiromichi Fujisawa. “Classification and Learning Methods for Character Recognition: Advances and Remaining Problems”. In: *Machine Learning in Document Analysis and Recognition*. Ed. by Simone Marinai and Hiromichi Fujisawa. Red. by Janusz Kacprzyk. Vol. 90. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 139–161. ISBN: 978-3-540-76279-9 978-3-540-76280-5. DOI:

10.1007/978-3-540-76280-5_6. URL: http://link.springer.com/10.1007/978-3-540-76280-5_6.

[21] Pavlo Melnyk, Zhiqiang You, and Keqin Li. "A High-Performance CNN Method for Offline Handwritten Chinese Character Recognition and Visualization". In: *Soft Computing* 24.11 (June 2020), pp. 7977–7987. ISSN: 1432-7643, 1433-7479. DOI: 10.1007/s00500-019-04083-3. URL: http://link.springer.com/10.1007/s00500-019-04083-3.

[22] Max Jaderberg et al. *Spatial Transformer Networks*. Feb. 4, 2016. arXiv: 1506.02025 [cs]. URL: http://arxiv.org/abs/1506.02025.

[23] Farisa Benta Safir et al. *End-to-End Optical Character Recognition for Bengali Handwritten Words*. May 9, 2021. arXiv: 2105.04020 [cs]. URL: http://arxiv.org/abs/2105.04020.

[24] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 5, 2017. arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762.

[25] Minghao Li et al. *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*. Sept. 6, 2022. arXiv: 2109.10282 [cs]. URL: http://arxiv.org/abs/2109.10282.

[26] Jamshed Memon, Maira Sami, and Rizwan Ahmed Khan. *Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)*. Dec. 31, 2019. arXiv: 2001.00139 [cs]. URL: http://arxiv.org/abs/2001.00139.