中图法分类号:TP391.43　　**文献标识码**:A　　**文章编号**:1006-8961(2023)06-1767-25

论文引用格式:Yang C, Liu C, Fang Z Y, Han Z, Liu C L and Yin X C. 2023. Open set text recognition technology. Journal of Image and Graphics, 28(06):1767-1791(杨春, 刘畅, 方治屿, 韩铮, 刘成林, 殷绪成. 2023. 开放集文字识别技术. 中国图象图形学报, 28(06):1767-1791)[DOI:10.11834/jig.230018]

开放集文字识别技术

杨春[1,2],刘畅[1],方治屿[1],韩铮[1],刘成林[3],殷绪成[1,2*]

1. 北京科技大学计算机与通信工程学院,北京　100083; 2. 北京科技大学模式识别与人工智能技术创新实验室,北京　100083; 3. 中国科学院自动化研究所,北京　100190

摘　要:开放环境下的模式识别与文字识别应用中,新数据、新模式和新类别不断涌现,要求算法具备应对新类别模式的能力。针对这一问题,研究者们开始聚焦开放集文字识别(open-set text recognition,OSTR)任务。该任务要求,算法在测试(推断)阶段,既能识别训练集见过的文字类别,还能够识别、拒识或发现训练集未见过的新文字。开放集文字识别逐步成为文字识别领域的研究热点之一。本文首先对开放集模式识别技术进行简要总结,然后重点介绍开放集文字识别的研究背景、任务定义、基本概念、研究重点和技术难点。同时,针对开放集文字识别三大问题(未知样本发现、新类别识别和上下文信息偏差),从方法的模型结构、特点优势和应用场景的角度对相关工作进行了综述。最后,对开放集文字识别技术的发展趋势和研究方向进行了分析展望。

关键词:文字识别;开放集模式识别;开放集文字识别(OSTR);封闭集文字识别;零样本文字识别

Open set text recognition technology

Yang Chun[1,2], Liu Chang[1], Fang Zhiyu[1], Han Zheng[1], Liu Chenglin[3], Yin Xucheng[1,2*]

1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China;

2. University of Science and Technology Beijing, Pattern Recognition and Artificial Intelligence Lab, Beijing 100083, China;

3. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Text recognition is focused on text transcription-based image processing modeling in relevance to such domains like document digitization, content moderation, scene text translation, automation driving, scene understanding, and other related contexts. Conventional text recognition techniques are often concerned about characters-seen recognition more. However, two factors in the training set of these methods are yet to be well covered, which are novel character categories and out-of-vocabulary (OOV) samples. Newly characters-related samples are often linked with OOV-based samples. However, it may pay attention to seen characters without novel combinations or contexts. For novel character categories, internet-based environments can be mainly used to face unseen ligatures like 1) emoticons and unperceived languages, 2) scene-text recognition environments, and 3) characters from foreign and region-specific languages. For digitization profiling, the undiscovered characters may not be involved in as well. Since the heterogeneity of language format to be balanced, the linguistic statistic data (e. g., *n*-gram, context, etc.) can be biased the training data gradually, which is

收稿日期:2023-01-11;**修回日期**:2023-02-28;**预印本日期**:2023-03-06

*通信作者:殷绪成　xuchengyin@ustb.edu.cn

基金项目:国家新一代人工智能(2030)重大项目(2020AAA0109701);国家杰出青年科学基金项目(62125601);国家自然科学基金项目(62076024,62006018)

Supported by: National Key R&D Program of China (2020AAA0109701); National Science Fund for Distinguished Young Scholars (62125601); National Natural Science Foundation of China (62076024, 62006018)

challenged for vocabulary-high-correlated text recognition methods. The two factors are required to yield three key scientific problems that affect the costs or efficiency in open-world applications. The novel characters are oriented for the novel spotting capability, whereas characters-unseen are rejected to replace silent seen characters. Furthermore, as the popular open-set recognition problem, three scientific problems can be leaked out as mentioned below. First, the emergence of novel characters is not efficient in many cases, in which re-training upon each occurrence is costly, and an incremental learning capability need to be strengthened after that. Second, an amount of attention is received as the generalized zero-shot learning text recognition task. Third, Linguistic bias robustness is yielded by the OOV samples. Due to the character-based nature prediction, more popular methods can be used to possess the capability to handle characters-seen OOV samples to some extent. However, such capabilities are constrained to demonstrate strong vocabulary reliance because of the capacity of language models, the open-set text recognition (OSTR) task is feasible since existing tasks like zero-shot text recognition and OOV recognition can be used to model individual aspects of the problems only. This task aims to spot and recognize the novel characters, which is robust to linguistic skews. As an extension of the conventional text recognition task, the OSTR task is used to retain a decent recognition capability on seen contents. In recent years, the OSTR task has been developing intensively in the context of character recognition. The literature review is carried out on the open-set text recognition task and its related domains. It consists of such five aspects of the background, genericity, the concept, implementation, and summary. For the background, we introduce the application background of the OSTR task and analyze the specific OSTR-derived cases. For genericity, the generic open-set recognition is introduced in brief as a preliminary of the OSTR task that is less familiar to some researchers in the text recognition field. For concept, the definition of the OSTR task is introduced, followed by a discussion on its relationship with existing text recognition tasks, e. g. , conventional close-set text recognition task and the zero-shot text recognition task. Its implementation-wise, common text recognition frameworks are first introduced. For implementation, it can be recognized as derivations of such frameworks, where the derivation is based on the three key scientific problems as following: new category spotting, incremental recognition of novel classes, and linguistic bias robustness. Specifically, the new category spotting problem refers to rejecting samples that come from an absent class of a given label set. Slightly different from the generic open-set text recognition task, the given label-set is challenged in related to the training data straightfoward. Incremental recognition refers to new categories recognition in terms of the non-retrained side information of the corresponding categories. The definition is slightly different from the common zero-shot learning definition, it can be excluded some generative adversarial network (GAN)-based transductive approaches. The linguistic bias robustness holds its original definition beyond more stressed unseen characters. For each scientific problem, its solution can be covered in text recognition and other modeling-similar related fields. The evaluation is carried out and it can mainly cover the datasets and protocols used in the OSTR task and its contexts as listed: 1) multiple protocols based public available datasets, 2) commonly used metric to measure model performance, and 3) several of popular protocols, typical methods, and the performance. Here, a protocol refers to the compositions of training sets, testing sets, and evaluation metrics. For summary, the comparative analysis of the growth and technical preference are demonstrated. Finally, the potnetialss of the trends and future research directions are predicted further.

Key words: character recognition; open set recognition; open-set text recognition (OSTR); close-set text recognition; zero-set text recognition

0 引 言

随着人工智能技术飞速发展,模式识别广泛应用在各大领域,为人们生活和生产提供了极大便利。然而,现有研究工作大多数是以封闭集假设(close-set assumption)为前提,即假设测试集中出现的所有类别在训练集中均出现过。虽然在封闭场景中这类方法取得了很好的应用效果(刘崇宇 等,2021),但在更多开放环境的真实应用(如无人驾驶、故障诊断、医疗诊断和文字识别等)中,测试集通常更加开放,类别也动态多变。在这些应用中,新数据、新模式和新类别不断涌现,传统封闭集识别方法不能应对,极大地限制了模式识别技术的应用。

为了解决这个问题,研究者引入迁移学习(transfer learning)(Shao 等,2015;Weiss 等,2016)、领

域自适应(domain adaptation)(Patel 等,2015)、零样本学习(zero-shot learning)(Fu 等,2018)和少样本学习(few-shot learning)(Bertinetto 等,2016;Snell 等,2017;Chen 等,2019)等技术,开展开放集模式识别任务的研究,目的是使分类器识别训练集中已见类别模式的同时,能拒识或发现、识别未见类别模式。近年来,开放集模式识别逐步成为模式识别领域的研究热点之一。

文字识别作为模式识别与人工智能中一个重要的研究方向,其相关技术已广泛应用于文档信息化、视频信息检索和智能阅卷等场景。目前,主流的文字识别方法(Borisyuk 等,2018;Baek 等,2019;Shi 等,2019;Wang 等,2020b)已经在封闭集的性能测试中取得了很好的效果。然而在实际应用中,大量真实应用面对的是开放环境(Zhou,2022)。例如,在网络图像/视频文字识别任务中,文本行中出现生僻字、小语种文字和 emoji 字符等;在古籍识别任务中,随考古进程不断发现,常常遇见需要定义的新字符等。因此,如何处理这些新字符和新文字,是模式识别与文字识别技术研究与应用面临的一个重要挑战。

具体来说,在开放集文字识别任务中,目前大多数面向封闭集文字识别的方法(Borisyuk 等,2018;Shi 等,2019;Wang 等,2020b)主要面临以下 3 个问题:

1)封闭集文字识别方法缺少拒识机制。当文本中存在新字符时,这些方法不但无法主动发现新字符,还会将新字符错误地识别为某类在训练集中见过的字符。在实际应用场景中,基于这些方法的文字识别系统无法主动地发现新字符。直到用户进行识别结果错误的反馈后,系统才会进行模型的更新与迭代。在模型完成更新前,相关的误识现象会持续发生。这种模型更新的处理方式极大限制了文字识别系统的应用场景。在开放环境下的应用场景(如古籍文字识别)中,当文本中普遍可能存在新字符时,用户需要对每个文本的识别结果进行人为校正。模型需要针对新字符进行更新,随后才能正确识别新字符。

2)封闭集文字识别方法无法直接识别新字符。具体来说,这些方法通常将每种字符类按照其类中心进行建模,或采用线性分类器进行字符分类。每个字符类与权重矩阵中的一组权重隐式关联。在处理新文字的识别任务时,这些方法通常只能采用重新训练模型(集合所有已见类别和新类别的训练样本重新训练)的方式来构造新字符所对应的权重。受此所限,在开放环境下,传统封闭集文字识别方法受制于较高的数据采集、人工标注和模型训练成本。

3)封闭集文字识别方法往往依赖较强的上下文信息。在开放环境中,由于新文字的上下文在训练集是未知,因此测试集和训练集语言模型存在较大偏差。在训练时,现有方法通常隐式建模了训练集的语义信息。在开放环境下,这些信息可能会对识别产生不利影响(Wan 等,2020)。当文本中存在新字符时,文字识别系统容易基于已有的上下文信息,对识别结果进行“校正”,将新字符错误识别为训练集中见过的字符。

近年针对第 1 个和第 3 个问题的研究工作较少,研究者主要聚焦于第 2 个问题。同时,大部分的研究工作集中于零样本文字识别方法。零样本文字识别方法可以分为两类。一类方法利用部件构成信息来表示单个字符(Wang 等,2019;Zhang 等,2020b;Chen 等,2021b)或整词(Chanda 等,2018,2021;Rai 等,2021),并基于该表示实现新字符的识别。由于这些部件和结构往往只适用于特定语言,因而限制了这类方法的跨语言泛化能力。另一类方法基于字符的视觉匹配(Qi 等,2018;Ao 等,2019;Zhang 等,2020a;Souibgui 等,2022),即将字符标准形状编码为分类器权重,并通过匹配度量实现新字符的识别。尽管这类方法不再受到语言的限制,但是由于计算复杂度高,难以扩展到具有较大字符集的语言的文字识别任务中。

此外,除少量方法(Zhang 等,2020a;Zhang 等,2020b;Huang 等,2021)外,上述大多数零样本文字识别方法不具有整行识别能力,无法在封闭集文字识别任务中进行评测,极大限制了这些方法的实用性(Baek 等,2019)。同时,这些方法不能解决开放环境下文字识别中的第 1 个问题(拒识机制问题),缺少对新文字的拒识能力(Fei 和 Liu,2016;Geng 等,2021)。

针对上述问题,Liu 等人(2023)给出开放集文字识别的形式化定义,并将对新文字的处理归纳为拒识和识别两种能力。具体来说,拒识能力能够帮助用户快速地发现数据流中没有发现的新字符;识别能力则帮助用户通过不重新训练模型的方式,快速

调整模型来识别发现的新字符。结合上述两种能力，Liu等人(2023)提出了一种可随数据演化的文字识别系统，将该系统形式化定义为开放集文字识别系统，并给出了评测开放集文字识别系统性能的数据集和指标。

1 开放集模式识别

1.1 定义

在具体介绍开放集模式识别的技术方法前，首先明确开放集模式识别任务的定义。为了区分不同类型的样本，Naylor(2010)根据是否可知将测试数据集划分为“已知—已知”(known-known classes，KKC)、“已知—未知”(known-unknown classes，KUC)、“未知—已知”(unknown-known classes，UKC)和“未知—未知”(unknown- unknown Classes，UUC)4个类别。

基于这种划分，Scheirer等人(2013)给出开放集识别任务的定义：在测试集中存在训练集中未见过的新类别(UUC)，方法既要正确分类见过的类别(KKC)，又要有效处理新类别(UUC)。Geng等人(2021)在此基础上，进一步明确KKC，KUC，UKC，UUC的定义，并将开放集识别任务与封闭集识别任务、少样本识别任务以及零样本识别任务等进行对比。

Zhang等人(2020c)对开放环境下的鲁棒模式识别技术进行总结，指出传统模式识别方法大多遵循封闭集假设(close-set assumption)(即假设测试集中出现的所有类别在训练集中均出现过)，并将开放环境鲁棒模式识别的工作按突破传统机器学习3个假设的角度分成突破封闭集假设、突破独立同分布假设、突破大数据假设3大类。其中，开放集识别和类别增量学习属于突破闭合集假设的问题。零样本识别从样本缺乏的角度属于突破大数据假设的问题，而从类别变化的角度属于开放集问题。

特别地，Zhang等人(2020c)提出一种基于类别增量学习(class-incremental learning)的开放集模式识别流程。该流程分为3个阶段，1)识别已知类别，并将未知类别样本放入缓冲区内；2)对未知类别样本进行人工或机器标注；3)使用新数据和新模式对模型进行类别增量学习，然后利用更新后的模型对缓冲区内的样本进行拒识或识别。

1.2 开放集模式识别技术

近年来，开放集模式识别逐步成为模式识别领域的研究热点之一。在识别新类别时，现有的开放集模式识别技术大致可分为判别式方法(discriminative methods)和产生式方法(generative methods)两大类。在拒识新类别时，通常采用分布外检测(out-of-distribution detection)的方法。

1.2.1 基于判别式方法的开放集模式识别技术

在深度学习技术出现之前，大多数判别方法都是以传统机器学习模型，如支持向量机(support vector machines，SVM)(Scheirer等，2014；Scherreik和Rigling，2016)、最近邻分类器(nearest neighbors)(Bendale和Boult，2015；Mendes Júnior等，2017)和稀疏表示(sparse representation)(Zhang和Patel，2017)等作为基线模型，并扩展到开放集模式识别方法，这类方法的效果依赖于特征的设计与挑选。

随着深度学习技术的快速发展，基于深度学习的判别方法(Chen等，2020；Shu等，2020；Yoshihashi等，2019)处理开放集模式识别任务时体现出更好的效果。Yoshihashi等人(2019)提出了一种面向开放集识别的分类—重构学习算法CROSR(classification-reconstruction learning algorithm for open set recognition)，利用潜在表示进行重构，并能够在不损害已知分类精度的前提下，鲁棒检测未知类别。Chen等人(2020)提出互换点学习框架。互换点是每个已知类别对应的类外空间的潜在表示。通过互换点引入未知信息，神经网络可以学习到更加紧凑和稳健的特征空间，并有效地分离已知类别和未知类别。Shu等人(2020)将原型学习引入开放集识别任务中，通过同时学习原型表示和原型半径，指导深度模型得到更具区分性的特征，并通过特征和原型之间的距离度量来检测未知类别。

1.2.2 基于产生式方法的开放集模式识别技术

产生式方法大多是基于深度学习的方法。按照是否依赖训练样本，产生式方法可分为基于实例的方法(Ge等，2017；Yu等，2017；Neal等，2018)和基于非实例(Geng和Chen，2022)的方法。

基于实例的产生式方法通常利用生成对抗网络(generative adversarial network，GAN)生成未知类别样本，从而帮助模型学习已知类别(KKC)和未知类别(UUC)之间的边界。Ge等人(2017)提出

G-OpenMax(generative openmax)算法，利用条件生成对抗网络生成未知类别样本。Yu 等人(2017)提出了面向开放集模式识别的对抗样本生成框架(adversarial sample generation，ASG)。

基于非实例的方法大多基于 Dirichlet 过程。Dirichlet 过程不过度依赖训练样本，可以在数据变化时实现自适应变化，使其自然适应开放集模式识别场景。通过修改分级 Dirichlet 过程，Geng 和 Chen(2022)提出基于集体决策的方法 CD-OSR(collective decision-based open set recognition model)。该方法不需要通过定义阈值的方式来区分已知类别和未知类别。

1.2.3 分布外检测技术

上述方法更多关注的是对于未知类别的识别。当未知类别是对抗样本或无关数据时，开放集模式识别任务要求方法能够对于这类未知样本进行拒识。

在这种情况下，开放集模式识别任务与一些其他任务有关，如分布外检测、异常检测(anomaly detection)、离群检测(outlier detection)和新颖性检测(novelty detection)等。Yang 等人(2021)提出了广义分布外检测(generalized out-of-distribution detection)统一框架，并在这一框架下对上述任务进行了对比和总结。

2 开放集文字识别任务

2.1 定义

为了描述开放环境下，文字识别应用对新字符发现和新字符识别的新需求，Liu 等人(2023)对开放集文字识别(open-set text recognition，OSTR)进行了形式化定义，基本含义如图 1 所示。该任务要求模型能够有效地识别已知字符(训练集见过的字符，如图 1 中的“断”、“国”、“金”等)，而且能够发现新字符(训练集未见过的字符，如图 1 中的“を”)，并且能通过不重新训练模型的方式快速调整模型来识别发现的新字符。

为了给出开放集文字识别的形式化定义，首先需要对已知字符和新字符进行明确的定义。对测试集 $\boldsymbol{C}_{\text{test}}$ 进行划分，以此给出新字符的定义。具体来说，记训练字符集 $\boldsymbol{C}_{\text{train}}$ 为所有训练样本中出现的字符集合，测试字符集 $\boldsymbol{C}_{\text{test}}$ 为所有测试样本中出现的字符集合，字符集 $\boldsymbol{C}_{\text{set}}=\boldsymbol{C}_{\text{train}}\cup\boldsymbol{C}_{\text{test}}$。字符 $c\in\boldsymbol{C}_{\text{set}}$ 可以根据是否见过和是否提供辅助信息分为如图 1 所示的 4 类，分别为“已知集内”(seen in-set character，SIC)，“已知集外”(seen out-of-set character，SOC)，“新—集内”(novel in-set character，NIC)和“新—集外”(novel out-of-set character，NOC)。缩写中的第 1 个字母(S 或 N)代表这个字符是否在训练字符集 $\boldsymbol{C}_{\text{train}}$ 中，即是否在训练集中见过。如果一个字符被包含在某个训练样本中，则 $c\in\boldsymbol{C}_{\text{train}}$，称为“已知”的字符；反之，$c\notin\boldsymbol{C}_{\text{train}}$，称为“新”字符。

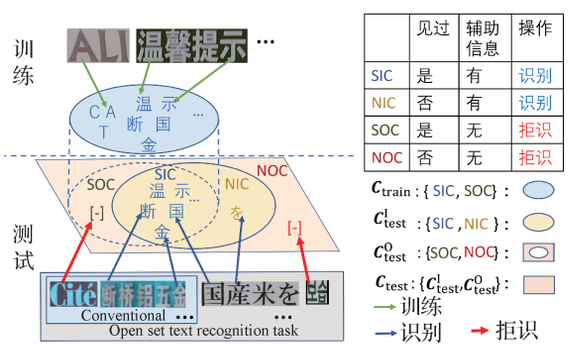

图 1 开放集文字识别任务示意图

Fig. 1 Sketch map of the open set text recognition task

缩写中的第 2 个字母(I 或 O)描述这个字符在测试时是否提供辅助信息。依据是否提供辅助信息(Pourpanah 等，2022)，可以将测试字符集 $\boldsymbol{C}_{\text{test}}$ 划分为两个互不相交子集 $\boldsymbol{C}_{\text{test}}^{k}$ 和 $\boldsymbol{C}_{\text{test}}^{u}$。辅助信息是对于单个字符的完整描述，例如字符部件、笔画序列或者字符标准形状等。已知测试字符集 $\boldsymbol{C}_{\text{test}}^{k}$ 中的字符在测试过程中提供辅助信息，而未知测试字符集 $\boldsymbol{C}_{\text{test}}^{u}$ 中的字符不提供辅助信息。如果一个字符类提供辅助信息，则称该字符类为“集内”字符($c\in\boldsymbol{C}_{\text{test}}^{k}$)，反之为“集外”字符($c\in\boldsymbol{C}_{\text{test}}^{u}$)。这里的“集”指的是 $\boldsymbol{C}_{\text{test}}^{k}$。模型需要识别集内字符，并对集外字符进行拒识。在实际应用中，用户可以通过添加或移除辅助信息的方法将字符类转换为集内或集外字符。

综上所述，开放集文字识别任务要求模型既可以识别已知文字(SIC)，又能够处理新文字(NIC、NOC)。开放集文字识别任务的输入为文字图像 $\boldsymbol{x}$(其内容为文字标签 $\boldsymbol{y}^{*}$)和任一给定字符集 $\boldsymbol{C}$，输出为预测标签 $\boldsymbol{y}$。此时，开放集文字识别函数 f 的形式化定义可表述为

$$\boldsymbol{y}=f(\boldsymbol{x},\boldsymbol{C}) \tag{1}$$

式中，$\boldsymbol{x}$为输入图像，$\boldsymbol{C}$为有辅助信息的字符集，$\boldsymbol{y}=(y_0, \cdots, y_l, \cdots, y_L)$为预测的字符序列。对于每个字符的预测结果$y_l \in \boldsymbol{C}$，如果图像中的字符在$\boldsymbol{C}$里，则模型应当输出对应的标签，否则应当输出[-]进行拒识。即

$$y_l = f(x) = \begin{cases} y_l^* & y_l^* \in \boldsymbol{C} \\ [-] & \text{其他} \end{cases} \tag{2}$$

在实际应用中，测试字符集$\boldsymbol{C}_{\text{test}}$表示待测数据流中所有可能出现的字符，这些字符既包含训练集见过的字符(SIC)和测试场景可预期的字符(NIC)，也包含不可预期的字符(NOC)。因此，Liu等人(2023)将已知的部分字符所属集合记做$\boldsymbol{C}_{\text{test}}^k$，其他未知字符类被自动归入$\boldsymbol{C}_{\text{test}}^u$。即在测试时，式(2)中的$\boldsymbol{C}=\boldsymbol{C}_{\text{test}}^k$。

此外，用户可以根据输入数据、负载和需求来调整$\boldsymbol{C}_{\text{test}}^k$。具体来说，文字识别系统与识别模型在刚上线运行时，由于无法预知数据流里有哪些新字符，因此可将所有新字符看做都属于NOC。模型上线运行后，用户可以通过人工审计拒识的样本来发现新字符，并通过提供辅助信息的方式向$\boldsymbol{C}_{\text{test}}^k$中添加这些字符，将它们变成NIC。同时管理员也可以通过移除罕见字辅助信息的方式，使模型暂时“遗忘”这些字符，实现高负载情形下的加速处理。该操作会将相应的SIC和NIC转化为SOC和NOC。

2.2 与其他文字识别任务的关联

Liu等人(2023)阐述了开放集文字识别任务与其他文字识别任务的关联，具体对比如表1所示。可以看出，1)与仅关注已知文字(SIC)的封闭集文字识别任务(Baek等，2019)相比，开放集文字识别任务不要求$\boldsymbol{C}_{\text{test}} \subseteq \boldsymbol{C}_{\text{train}}$，即允许测试集中存在训练集未出现的新字符(NIC、NOC)。2)零样本文字识别任务(Zhang等，2020b；Huang等，2021)可视做开放集文字识别任务的一类特殊情况，即满足$\boldsymbol{C}_{\text{test}} \cap \boldsymbol{C}_{\text{train}} = \varnothing$和$\boldsymbol{C}_{\text{test}}^u = \varnothing$。零样本字符识别任务则是所有文字标签的长度均为1的一个更特殊的情形。零样本识别只对新类别样本(NIC)进行识别。将新类别样本和已见类别样本(SIC)一起识别的方式称为广义零样本识别。3)开放集文字识别任务继承了封闭集文字识别对已知文字(SIC)的识别要求，以及零样本文字识别对新字符(SIC)的适应要求。

与其他任务相比，开放集文字识别任务额外引入对未知字符$c \in \boldsymbol{C}_{\text{test}}^u$的拒识任务，从而发现数据流中的新字符(对应开放集文字识别的第1个问题)。当数据流中出现新字符时，首先，模型拒识该样本，并通知用户发现疑似新字符；然后，用户人工检视拒识样本，确认新字符；最后，用户提供辅助信息，将新字符类别加入$\boldsymbol{C}_{\text{test}}^k$，使模型具有识别新字符的能力。同时，拒识功能也可以用于实验缓存机制，从而加速推理。即模型暂时遗忘不常用的字符，仅在拒识发生时，再对这些遗忘的字符进行识别。在开放集文字识别任务中，由于新字符的出现通常导致训练集的上下文语义和语言模型不再适用，所以开放集文字识别任务隐式地要求模型对训练集与测试集的上下文信息偏差有较强的健壮性。

总体来说，开放集文字识别任务可以看做是封闭集文字识别任务、小样本文字识别任务和零样本文字识别任务的超集。由于封闭集文字识别技术较为成熟，因此大多数现有开放集文字识别技术通常在某种主流封闭集文字识别框架进行扩展。

表1 开放集文字识别与其他文字识别任务的关联

Table 1 Association between open set character recognition with other character recognition tasks

任务	输入	SIC	SOC	NIC	NOC	语言
封闭集文字识别(Borisyuk等，2018；Baek等，2019；Shi等，2019；Wang等，2020b)	整词	识别	-	-	-	没有限制
零样本文字识别(Ao等，2019；Wang等，2019；Cao等，2020)	字符	-	-	识别	-	CJK
开放集文字识别(Liu等，2023)	整词	识别	拒识	识别	拒识	没有限制

注：“-”表示该识别任务不包含该类型字符集。

3 常见文字识别框架

开放集文字识别及相关研究工作主要聚焦于开放环境下新出现的3个科学问题，即未知样本发现、新类别识别和上下文信息偏差。未知类别发现问题是指开放环境下模型将数据中含有未知字符的样本（没有事先预见的）以拒识的方式发现并通知管理员的能力（Huang等，2019）；新类别识别是指数据中的未知字符被发现后，模型在不经过重新训练的情况下，只依赖辅助信息或少量样本对新发现字符类别进行识别的能力（Wang等，2018）；上下文信息偏差是指新类别文字与模型训练得到的语言模型具有较大的语义偏差，这要求开放环境下的文字识别方法对语义偏差具有鲁棒性（Wan等，2020）。由于这些问题并非开放集文字识别的特有问题，本文对相关问题综述时也涵盖了该问题在其他实际任务中的应用。

在具体介绍开放集文字识别任务的关键问题及解决方法前，本文先对文字识别框架进行总结。目前主流文字识别框架流程主要是通过预测器（predictor）对输入图像（和辅助信息）进行转录，得到预测文本结果。按照预测器的信息处理粒度，文字识别框架主要分为整词识别、基于特征归集的字符序列识别、基于标签归集的字符序列识别3种类型。文字识别整体流程和3类文字识别框架的示意图如图2所示。

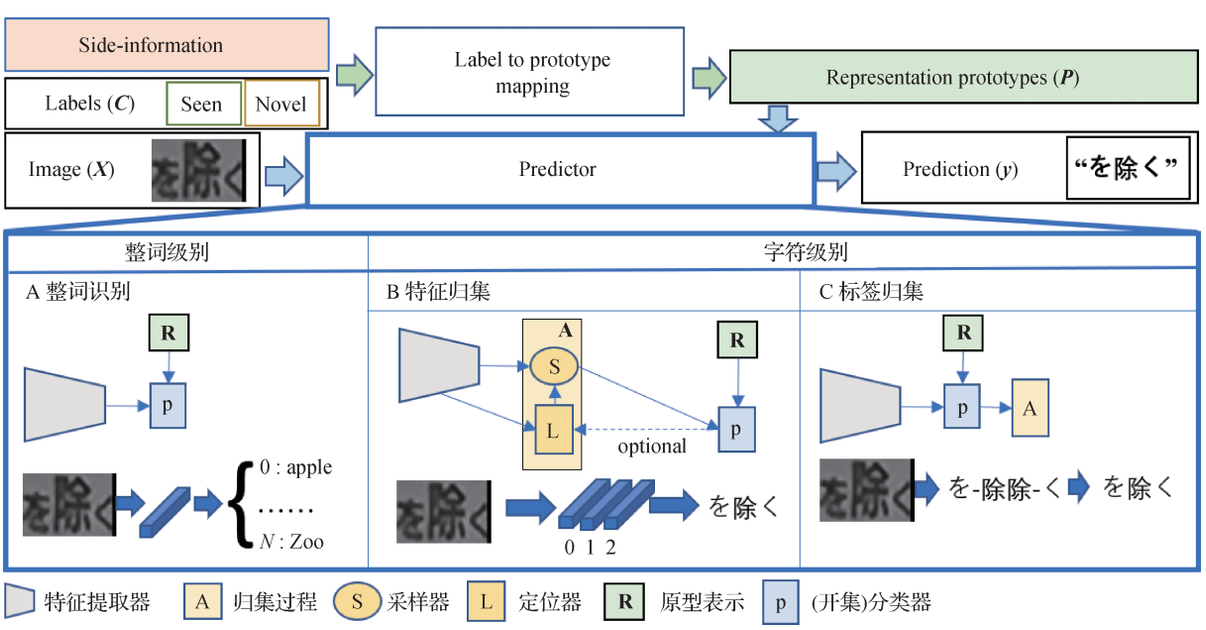

图2 主要文字识别框架

Fig. 2 Main framework for text recognition technology

3.1 整词识别

受“整体大于部分之和”的启发，面向整词识别的方法将文字识别任务定义为基于图像的词匹配（word spotting）问题或一个多类别分类（multi-class recognition）问题。这类方法对输入图像整体进行表征，并将该表征转录为文本。

在封闭集文字识别任务中，Manmatha等人（1996）提出词匹配（word spotting）方法，对历史手稿文档中相同的词文本图像进行检索。Wang和Belongie（2010）将词匹配技术应用于自然场景文本识别中，在Google的街景文本（street view text，SVT）数据集上，整词识别精度比传统文字识别模型提高了16%，证明词匹配技术也可以应用于复杂背景的自然场景文本识别。Almazán等人（2014）将识别和匹配任务解释为最近邻问题，将单词图像和文本字符串嵌入到一个公共的向量子空间或欧氏空间中，基于表示同一单词的图像特征和字符串特征应该相似的假设，将标签嵌入与属性学习相结合，实现文本的识别。Jaderberg等人（2016）将文字识别定义为一个多类别分类问题，使用深度卷积神经网络预测文本图像的字符序列和二元关系（bigram），并使用动态搜索获取的词文本图像识别结果。

在开放环境下的文本识别任务中，Chanda等人(2018)引入零样本学习范式，解决历史手写文档数字化中未知类文字识别难的问题，利用深度学习框架提取鉴别能力强的图像特征，预测自定义的手写笔画属性，进而完成单词识别。同样是针对历史文档文字识别，Rai等人(2021)基于整图特征预测11种主形状属性，提出混合模型Pho(SC)(pyramidal histogram of shapes and characters)Net，实现对零样本手写单词识别。Chanda等人(2021)基于整图特征预测13个基本形状/笔画属性，并应用零样本学习框架实现对手写孟加拉文字的识别。

由于整词识别方法通常将文本标签视为一个整体而非字符序列，因而这类方法的模型和语言信息具有较强的关联性，难以识别词典外的词。由于这种局限性，研究者开始关注基于字符序列的识别方法，即基于特征归集的字符序列识别和基于标签归集的字符序列识别。

3.2 基于特征归集的字符序列识别

基于特征归集的字符序列识别方法首先通过对整幅图像的特征进行划分和归集，得到每个字符对应的特征表示，其次对每个字符类别进行独立预测，最后将所有字符预测结果归并，得到字符序列的预测结果。这种方法不需要基于规则的后处理步骤，而且容易对字符特征进行约束(Li等，2020；Zhang等，2020b)。

按照解码框架，基于特征归集的字符序列识别方法可以划分为以ASTER(attentional scene text recognizer)(Shi等，2019)为代表的串行解码框架和以ABINet(autonomous, bidirectional and iterative language modeling network)(Fang等，2021)为代表的并行解码框架。

3.2.1 串行解码

由于文本是字符串序列，因此一种朴素的想法是按顺序对每个字符进行独立预测，并将预测结果顺序排列，从而得到文本预测结果。针对不规则形状文本的文字识别任务，Shi等人(2019)提出了ASTER网络。ASTER网络由一个校正网络和一个识别网络组成，校正网络自适应地对输入图像中的文本进行校正，然后将校正后的文本图像输入到一个基于注意力机制的序列模型，按顺序提取每个字符的感兴趣区域，并预测当前字符，直到预测结果为文本结束(end of sentence，EoS)为止。ASTER在不规则形状文本数据集上具有较好的识别性能。

为了解决循环神经网络(recurrent neural network，RNN)训练速度较慢以及层叠卷积层计算复杂度高的问题，Sheng等人(2019)提出一种非循环—序列到序列文字识别器(no-recurrence sequence-to-sequence text recognition，NRTR)，使用完整的Transformer结构对输入图像进行编码，并利用编码特征和前序预测结果对当前字符进行预测。NRTR训练的并行性更高，复杂度更低。

在现有开放集文字识别方法中，Zhang等人(2020b)充分利用汉字固有的层次结构将一个汉字解构为一个独立的树。其中，以笔画作为叶子节点，以层级式的笔画间空间结构作为根节点，并使用结合注意力机制的编解码网络实现对中文文本的串行转录。

3.2.2 并行解码

由于串行解码需要按顺序预测每个字符，所以运行速度较慢。研究者提出并行解码策略，加速解码过程。

Fang等人(2021)指出当前引入语言模型的自然场景文本识别方法具有隐式语言建模、无向特征表达和含噪输入等缺陷，没有充分挖掘语言模型的能力，因此提出了一个自主、双向和迭代的自然场景文本识别模型ABINet。ABINet的视觉模型由主干网和位置注意力模块组成。该方法使用位置注意力模块，基于查询范式将视觉特征并行转录为字符概率。

针对基于自回归结构的注意力方法容易出现注意力对齐问题，Wang等人(2020b)提出了解耦注意力网络(decoupled attention network，DAN)。该模型主要由特征采集器、卷积对齐模块和基于解耦的文本解码器构成。其中，卷积对齐模块根据编码器的输出进行特征对齐，文本解码器通过联合使用特征图和注意图对每个字符进行并行预测，可以有效解决注意力对齐问题。

Yu等人(2020)认为现有场景文本识别模型多是采用RNN结构对语义信息进行解码，解码过程难以并行，限制了计算效率，同时传播方向单一、信息无法有效传递，且会引起无用语义信息以及错误语义信息的传播。因此提出了SRN(semantic reasoning network)网络，网络中采用GSRM(global semantic reasoning module)模块实现全局语义信息的多路并行传播，从而对上下文语义信息进行充分挖掘。

Baek 等人(2019)认为当前很多场景文本识别模型在评测时使用不同的训练集和测试集,在领域内缺少一个全面公平的比较。因此提出了一个文字识别统一框架,通过对字符进行并行转录,得到文本预测结果。同时,在训练集和测试集设定下,从准确性、速度和内存需求方面分析各个模块对模型性能的贡献。

在开放集文字识别任务中,Liu 等人(2022a,2023)采用并行解码框架。这些方法首先采用注意力机制定位每个字符,并提取特征表示。然后,通过与字符标准形状匹配的方式,对每个字符进行分类。最后将结果归并为文本预测结果。

3.3 基于标签归集的字符序列识别

与特征归集方法不同,标签归集方法先对特征进行分类,然后根据分类结果进行归集,进而得到最终文本预测结果。

Shi 等人(2017a)提出了一个文字识别端到端训练模型 CRNN (convolutional recurrent neural network),将特征提取、序列建模和转录过程进行融合,输入文本行图像,预测文本序列。在不涉及字符分割和水平尺度归一化的条件下,能够处理任意长度文本。同时,该模型不受限于任何预定义的词库,在现实应用中具有很强的实用性。

Borisyuk 等人(2018)提出一个大规模图像的文本提取和文字识别方法 Rosetta。Rosetta 遵循当前主流的文字识别系统架构,分为文本检测和文本识别两部分。文本检测部分采用 Faster-RCNN 模型,检测图像中包含文本的区域。文本识别部分采用全卷积字符识别模型,将检测到的文本行图像转录为文本内容。同时,Borisyuk 等人(2018)给出了工业级文字识别应用如何对性能和效果进行折中,对完成高性能的文字识别任务有实用的参考价值。

针对不规则形状文本的文字识别任务,Cheng 等人(2018)提出一种全新的识别方法 AON(arbitrary orientation network),将文本方向分为从左到右(left to right)、从右到左(right to left)、从下到上(bottom to top)和从上到下(top to bottom)4 个方向,首先对这 4 个方向分别提取字符序列特征和权重向量,随后将 4 个字符序列特征和权重进行组合,形成最终的字符序列,最后输入带有注意力机制的解码器,得到最终文本预测结果。

目前大多数研究认为场景文字识别是一个 1 维序列预测问题。但是图像中的文本实际上分布在 2 维空间,原则上直接将文本特征压缩成 1 维形式可能会失去有用信息和引入额外噪声。Liao 等人(2019)从 2 维角度来处理场景文本识别,提出了字符注意力全卷积网路(character attention fully convolutional network, CA-FCN),用于识别不同形态的文本。场景文字识别使用语义分割网络实现,并设计了针对字符的注意力模块。通过添加构词模块,CA-FCN 能同时识别文本并预测每个字符的位置,在规则和不规则文本数据集上都能达到很好的效果。

4 开放集文字识别技术

开放集文字识别相较封闭集文字识别面临新类别和新数据的挑战。新类别指数据流中可能含有无法实现预测的字符(比如 emoji 字符)。这些新类别引入了 3 个主要问题。1)模型应该能够主动发现没有见过的新字符,并通过拒识方式通知用户,而非将其识别成某种已知的文字,即未知样本发现问题。该问题同时是开放集识别任务(Scheirer 等,2013)面临的主要问题之一。2)在发现新字符后,模型应该能通过快速调整,用于识别这些新字符类别,即新类别的识别问题。该问题同时是小样本方法和零样本方法关注的主要内容(Pourpanah 等,2022)。3)开放环境下,伴随新字符、新文字的出现,文本内容受语言演化影响,与训练的语料逐渐产生偏差,进而影响识别效能(Wan 等,2020),即上下文信息偏差问题。本节针对上述 3 个问题及其处理方法进行综述。

4.1 未知样本发现技术

拒识任务在文字识别任务中尚未得到广泛应用,目前仍然处于起步阶段(Liu 等,2023)。该问题本质上属于对于未知类别的发现任务,在开放集识别任务研究中具有一定研究基础(Fei 和 Liu,2016;Zhang 等,2021;Bao 等,2022;Ding 等,2022;Han 等,2022;Huang 等,2022b)。该任务面临两个关键挑战,即如何建模未知类别和如何利用已知类的数据在训练中模拟未知类。

4.1.1 未知类别建模

从未知类别建模的角度,现有方法可以分为两类。一类主流方法是将未知类别作为分类器的一个额外类别(Zhang 和 Ding,2021;Gupta 等,2022;Han

等，2022；Huang 等，2022b）。其中，Zhang 和 Ding（2021）使用一个可训练的类中心来建模未知类别。Han 等人（2022）提出 CFL（contrastive feature learner）模块以及将所有非正确类别分数推到未知类别之下的损失 L_{UP}，并依照不确定性对该损失加权。Huang 等人（2022b）提出适应性的未知类中心生成方法，即通过 Negative Generator 总结所有已知类别，得到未知类别的类中心。另一类方法（Du 等，2022；Liu 等，2023）根据未知样本与所有已知类别的相似程度进行判定。这些方法认为未知类实际上并不是隶属于同一个类别。同时，由于未知类别的数量并不明确，因而并不应该进行类中心的预设。Du 等人（2022）将未知类别的分类概率真值设定为在所有类别上的均匀分布，由于已知类别的数量较大，故对未知类样本的预测概率会很低。Liu 等人（2023）则用一个可训练的参数作为未知类别的置信度，省去了人工设定阈值的步骤。

以上两类方法可以看成是判别式的开放集识别方法。另一些方法则通过分布外检测的方式进行新类别的发现。例如，Ding 等人（2022）对已知模式进行采样，提取特征后求均值，再与待测特征做差，然后使用分类器判断是否属于新的类别。

4.1.2 未知类别模拟

对于未知类别的模拟，一类方法使用生成模型来生成不属于已知类别的合成样本，也称做产生式的开放集识别方法。Scheirer 等人（2013）、Ge 等人（2017）、Zhang 和 Ding（2021）都是使用生成对抗网络来生成未知样本。Zhou 等人（2021）则通过对已知样本进行混合（mix-up）的方式得到未知类的特征，该方法通过混合两个类别样本的中间层特征作为未知样本的中间层特征，解决了生成对抗网络的复杂度问题。

另一类方法（Huang 等，2022b；Liu 等，2022a，2023）通过对已知类别进行划分的方式来模拟类别。Liu 等人（2023）根据每个批次（batch）的样本将标签集合动态划分为已知类别和未知类别。通过约束 batch 中的已知类别的识别损失，使未知类别的置信度小于正确标签的置信度，并高于其他错误标签。特别地，由于未知类别标签的对应权重不参与训练，所以不会出现将未知类的正确标签推向未知类置信度之下的情况。Han 等人（2022）和 Zhou 等人（2021）的方法可以看做是动态划分的极端情况，即对每个样本进行排除正确类标签的划分。

还有一类方法（Joseph 等，2021；Du 等，2022；Gupta 等，2022）通过伪标签的方式，试图发现训练样本中没有进行标注但又与目标相似的区域，将其作为未知类。在部分方法（Du 等，2022）中这类区域也称为背景。Gupta 等人（2022）选取训练过程中前 5 个检出的非真值样本作为未知类别，选取 RPN（region proposal network）预测的背景类中置信度最高的一个（$k = 1$）候选作为未知类样本。

值得注意的是，上述 3 类方法并不排他，即同一个工作中可能出现对不同类型方法的组合应用（Zhou 等，2021）。

4.2 新类别识别技术

在零样本文字识别和开放集文字识别任务中，大部分现有方法采用修改模型的字符分类器的方式，实现对未知样本的识别。除具体使用的辅助信息外，这些分类器在处理新类别时，其方式与小样本、零样本或域迁移学习任务中的大部分方法对新类别的处理方式并无本质区别。按照新类别的处理方式，这些分类器可以分为基于属性和基于视觉匹配两大类。

4.2.1 基于属性的方法

基于属性的分类方法源自于零样本学习。此处，属性是指在所有类别上的共有特征，这些特征在经过标量化处理之后被表示成具有一定维度的属性向量。其中，每一个类别对应着唯一的属性向量。利用该属性向量作为分类原型建立已知类与未知类的迁移桥梁，应用度量学习的方法实现对类别的划分。开放集文字识别模型沿用了这一思想，通过将能够描述文字特征的诸多信息作为属性，实现对集外新类别文字的识别。基于不同的属性设计方案，现有开放集文字识别方法可大致分为基于文字先验属性的方法和基于可学习编码属性的方法。

基于文字先验属性（Chanda 等，2018；Wang 等，2018；Wang 等，2019；Zhang 等，2019；Zhang 等，2020b）的方法利用文字间共有的笔画、结构作为属性，将每一个文字解析成一个包含笔画和结构的集合（如表意描述序列），再通过与标准库对比，从而确定该文字所属类别。由于笔画、结构是所有文字的共有属性，所以此类方法可以自适应到开放集文字识别应用中。Wang 等人（2018）使用汉字的笔画和 2D 文字结构作为属性，提出针对手写汉字的笔画分

析网络 DenseRAN（radical analysis network with densely connected architecture）。该网络首先利用基于 DenseNet 的编码器将输入的手写文字图像编码为高层视觉特征，然后再利用基于 RNN 的解码器将视觉特征解耦成相应的笔画和 2D 结构，并以此作为输入图像的字符标注，实现对手写中文的离线识别。Wang 等人（2019）认为已有基于笔画的手写中文识别方法缺乏灵活的解码算法，对笔画特征表示能力弱，因此提出包含了笔画映射编码器 RME（radical mapping encoder）、笔画回归模块 RAM（radical aggregation module）和字符分析解码器 CAD（character analysis decoder）的笔画回归网络 RAN（radical analysis network），用于进一步提升笔画表征的有效性和鲁棒性。Huang 等人（2022a）提出基于海马回启发式的中文字符识别网络 HCRN（hippocampus-heuristic character recognition network），可以在只学习部分文字笔画结构的前提下完成对未知文字的识别。Lin 等人（2022）以笔画数为属性，汉字为已知类，白族文字为未知类构建数据集，并提出基于 VAE（variational autoencoder）的生成模型自动捕获字符结构信息。Diao 等人（2022）根据正字法分解和重构字符，提出基于部首提取的零样本字符识别框架 REZCR（zero-shot character recognition framework via radical extraction），并在甲骨文、中文和韩文等数据集上完成模型有效性验证。

在真实场景中，笔画和结构无法完全覆盖所有的文字。此外，以笔画和结构构造复杂文字的属性，得到的属性向量维度将会很长且无法为所有文字学习统一的表征。因此，研究者提出基于可学习编码属性的方法（Cao 等，2020；Du 等，2022；Guo 等，2022；Kumar 等，2022）。这类方法利用深度学习模型将类别自动编码为维度统一的嵌入表征，然后与图像特征进行相似度度量，从而确定字符所属类别。Cao 等人（2020）提出层级解耦嵌入 HDE（hierarchical decomposition embedding），将输入的中文字符图像编码到一个可解释的向量空间，将相应的类别标签编码为类别嵌入向量，并在向量空间内计算向量间相似度，完成输入图像到类别的匹配。Fu 和 Sigal（2016）、Fu 等人（2020）提出半监督词汇学习模型，将图像编码到语义嵌入空间，利用基于语义流形的最大边距识别框架解决开放集识别问题，显著提升了单词分类任务在零样本学习和开放集场景下的性能。

4.2.2 基于视觉匹配的方法

视觉匹配方法在开放集文字识别中的应用主要包括 Doan 和 Kalita（2017）、Prakhya 等人（2017）、Shu 等人（2017）、Zhang 等人（2020a）、Xia 等人（2021）、Song 等人（2022）和 Souibgui 等人（2022）的方法。Zhang 等人（2020a）提出通过中间表征解耦合视觉解码和语言建模两个阶段，以达到利用语言中字符重复性特征的目的，中间表征则是以相似性图的形式表达。这样，就将文本识别问题转化为视觉匹配问题，从而达到了更优的泛化性和灵活性。Souibgui 等人（2022）提出基于小样本学习的手写体文字识别方法来减少人工标注的工作量，首先检测文本行图像中给定字母的所有标志，并将获得的相似性得分译码至提取文本的最终序列。这样，也将文本识别问题转化为了视觉特征匹配问题。Shu 等人（2017）首次提出了用于开放环境下文档识别的深度学习模型。Doan 和 Kalita（2017）提出了基于增量学习的类中心最近邻分类模型，用于开放环境下的文本分类任务。He 和 Schomaker（2018）提出了基于多种类型属性的小样本开放集环境下的中文字符识别模型。多种属性类型共同结合的方法可以提供更丰富的特征，更有利于属性空间中的度量学习。Prakhya 等人（2017）提出了基于深度卷积神经网络的开放集文本分类模型。Song 等人（2022）提出利用图像的背景特征作为“伪不可见类”来促进开集分类器的学习，同时为未知类在表征空间中预留相应的位置。Xia 等人（2021）提出通过增量学习的思想动态学习文本中样本数量很少的新类别。

在其他小样本学习应用场景下，类似的基于视觉匹配的分类器也广泛用于解决新类别识别这一问题。具体来说，这些方法主要基于度量学习，并集中应用于小样本分类任务（Vinyals 等，2016；Snell 等，2017；Hou 等，2019；Kim 等，2019；Li 等，2019b；Li 等，2019c；Qiao 等，2019；Simon 等，2020；Ye 等，2020；Chen 等，2021a）。由于解决的是同一个问题，这些方法同样可以应用于开放集文字识别。这些方法大致可以分为基于最近邻的方法（Snell 等，2017；Li 等，2019a；Li 等，2019c；Qiao 等，2019；Simon 等，2020；Ye 等，2020）、基于注意力机制的方法（Vinyals 等，2016；Hou 等，2019）和基于图网络的方法（Kim 等，2019；Ma 等，2020；Chen 等，2021a）。

基于最近邻的方法主要是在表征空间中训练最

近邻分类器，使用欧氏距离或余弦距离度量样本特征间的相似度。Snell 等人（2017）最先提出使用特征提取网络将支持集和查询集中的样本映射至同一个表征空间中，并在此表征空间中训练最近邻分类器，最小化每个查询集样本到其对应的类原型的欧氏距离，每个类原型则是由该类所有支持集样本的特征平均值确定。Li 等人（2019c）提出使用图像的局部描述算子对图像级特征进行替换，并在图像的局部描述算子基础上训练最近邻分类器，从而进行图像到类别的度量。Simon 等人（2020）提出将每类的原型表示替换为由该类的样本作为基底扩展而成的子空间，对查询集样本的距离度量也就变成了样本特征点到各个类所表征的子空间的投影距离，以此投影距离作为度量准则训练最近邻分类器，与在图像级特征上训练的最近邻分类器相比，基于子空间投影距离的最近邻分类器能够保留更多的类别特征。Qiao 等人（2019）提出将针对查询集样本的基于距离度量的最近邻分类问题转化为带约束条件的凸优化问题，通过在每个特定的小样本分类任务上求得所对应的闭式解，将图像级特征从任务无关的共享嵌入空间映射至更具有鉴别力的任务级度量空间，在此空间上训练的最近邻分类器具有更高的分类精度。Ye 等人（2020）提出使用集合对集合的函数学习更具判别力的任务相关表征，这些表征更多地保留了任务级的信息，在这些任务级表征上学习的最近邻分类器具有更高的判别能力。

基于注意力的方法主要是通过注意力机制对特征图进行加权，并且一般会使用额外的辅助信息进行指导，例如任务级的上下文信息，从而达到特征增强的目的。Vinyals 等人（2016）提出使用查询集样本与支持集样本的距离度量作为注意力核，对样本标签真值进行加权求和，从而平衡各个支持集样本的权重，并且使用长短期记忆网络综合考虑任务级上下文关系。Hou 等人（2019）提出综合考虑每个查询集样本与支持集样本特征图的相似性关系，生成更贴近类真值标签的注意力图，从而改善由于样本数少导致的注意力关注区域偏移的问题。Li 等人（2019b）提出在每个特定的小样本任务中，通过遍历一次整个支持集，同时结合类内的共同性和类间的独特性，用掩码对样本特征进行增强，得到任务相关的特征。

基于图网络的方法主要是结合图神经网络发掘类内相似性和类间独特性，并且通过图网络的消息传递机制能够结合查询集共同考虑整个任务级相关的信息，同时可以引入样本标签的语义特征，进行跨模态的推理。Kim 等人（2019）提出通过在图网络中引入边标记，同时对类内相似性和类间不相似性进行显式的建模，与基于点特征隐式建模的图网络方法相比，具有更强的泛化性。Ma 等人（2020）提出用图神经网络显式地建模每个支持集—查询集样本对之间的关联，并通过消息传递算法对这些关联进行传播。与传统图网络建模不同的是，该方法中图网络的每个节点表示的是一个支持集—查询集样本对的关联，这样就可以更充分地利用不同查询集样本之间的关联信息，同时保持了类内共同性和类间独特性。Chen 等人（2021a）提出通过计算对级的样本关联来学习更丰富的样本表征，并将学到的视觉知识压缩至类级图网络中，同时引入了样本类标签的语义特征来辅助显式地学习类与类之间的关联，经过图网络的消息传播机制后可获得更具鉴别力的类级知识表征。

4.3 上下文信息偏差处理技术

上下文信息偏差（Egglin 和 Feinstein，1996）是广泛存在与机器学习不同领域中的常见问题之一（Wang 等，2020a；Yue 等，2020；Su 等，2021；Liu 等，2022b）。由于文本的特殊性，上下文信息在文字识别中起到相对更大的作用，这也导致模型对新字符的预测结果受上下文偏差影响更突出。自 2020 年，文字识别领域的研究者（Wan 等，2020；Garcia-Bordils 等，2023）开始关注这个问题。

在开放集文字识别任务中，由于训练数据的局限性，训练语料的语义信息难以充分覆盖开放的测试环境（Liu 等，2022a），导致上下文信息偏差问题相较于封闭集文字识别任务影响更大。

针对这一问题，Hu 等人（2022）通过门控的方式动态集成视觉预测与上下文信息，并使用相对较小的感受野减轻上下文偏差对网络造成的影响。Wan 等人（2020）通过相互学习，利用 KL（Kullback-Leibler）散度对齐对上下文相对不敏感的分割模型分支和对上下文相对敏感的 RNN-Attention 分支的预测概率，从而提升每个分支的性能。Liu 等人（2022a）通过显式解耦上下文信息和视觉特征对预测结果的影响，从而得到不受上下文影响的视觉特征。在其他领域，类似的因果模型（Wang 等，2020a；Yue 等，2020；Liu 等，2022b）也经常用于控制上下文

的偏差对预测结果的影响。Su 等人(2021)则通过随机合成训练数据,以此削弱目标间的耦合关系。

5 公开数据集和评测协议

为了尽可能公平地评价一个文字识别方法,已提出了一些评测协议,以方便研究者训练、测试和评价其方法。一个评测协议包括数据(训练、测试)和评价指标两部分。本节从数据、指标和协议3方面进行综述。

5.1 公开数据集

一些常见的公开数据集的具体细节如表2所示。其中,标签信息度表示标注信息中是否含有字符位置信息、词条的内容和词条的语言。字典指该数据集中是否附有每个词条预设的候选真值及其数量。对于私有数据(Huang 等,2021)本文不进行讨论。下面按照真实数据集和合成数据集对现有公开集的规模、语言和特点进行分类介绍。

常见的真实数据集按照语言可以分为英文数据集和其他语言数据集,英文数据集的使用最多。其中,常用于训练的拉丁语言数据集包括 MJ(MJSynth)和ST(SynthText)两个合成数据集,而真实集常用于测试,包括 IIIT5k(Mishra 等,2012)、CUTE(curved text dataset)(Risnumawan 等,2014)、SVT(street view text)(Wang 等,2011)、SVTP(street view text-perspective)(Phan 等,2013)、IC03(ICDAR2003)(Lucas 等,2005)、IC13(Karatzas 等,2013)和 IC15(Karatzas等,2015)等。此外还有google book集(Zhang等,2020a),Coco-text(Veit等,2016)也偶见使用。

近年来,研究者也提出了其他语言的数据集,其中,很大一部分为中英混合的数据集,涵盖自然场景、网络图像和脱机/在线手写等场景,其属性详见表2。其中,MLT(multi-lingual scene text)为一个较大的包含9种语言的多语言数据集。

表2 常见的文字识别数据集(Chen等,2022)

Table 2 Common character recognition datasets (Chen et al., 2022)

数据集	来源	语言	样本量			标签信息			字典
			总数	训练集	测试集	字符	词	语言	
IIIT5K (Mishra等,2012)	真实, 场景	英文	5 000	2 000	3 000	√	√	×	50/1k
SVT (Wang等,2011)	真实, 场景	英文	725	211	514	×	√	×	50
IC03 (Lucas等,2005)	真实, 场景	英文	2 268	1 157	1 111	√	√	×	50/full/50k
IC13 (Karatzas等,2013)	真实, 场景	英文	5 003	3 564	1 439	√	√	×	×
SVT-P (Phan等,2013)	真实, 场景	英文	639	0	639	×	√	×	50/full
CUTE80 (Risnumawan等,2014)	真实, 场景	英文	288	0	288	×	√	×	×
IC15 (Karatzas等,2015)	真实, 场景	英文	6 545	4 468	2 077	×	√	×	×
Synth90k (Jaderberg等,2014)	人工合成	英文	~9 M	–	–	×	√	×	×
SynthText (Gupta等,2016)	人工合成	英文	~6 M	–	–	√	√	×	×
MTWI (He等,2018)	真实, 网络	中文, 英文	290 206	141 476	148 730	×	√	×	×
RCTW-17 (Shi等,2017b)	真实, 场景	中文, 英文	–	–	–	×	√	×	×
CTW (Yuan等,2019)	真实, 场景	中文, 英文	1 018 402	812 872	103 519	√	√	×	×
SCUT-CTW1500 (Liu等,2017)	真实, 场景	中文, 英文	10 751	7 683	3 068	×	√	×	×
LSVT (Sun等,2019)	真实, 场景	中文, 英文	–	–	–	×	√	×	×
ArT (Chng等,2019)	真实, 场景	中文, 英文	98 455	50 029	48 426	×	√	×	×
ReCTS-25k (Liu等,2017)	真实, 场景	中文, 英文	119 713	108 924	10 789	√	√	×	×
MLT (Nayef等,2019)	真实, 场景	多语言	191 639	89 177	102 462	×	√	√	×
OLHWDB1 (Liu等,2011)	真实, 手写	中文	3.9 M	–	–	√	×	×	×

注:“–”表示数据缺失;“√”表示数据集提供种类型的标注;“×”表示数据集不提供种类型的标注;字典中,50,1k,full,50k描述的是字典大小。

5.2 评价指标

5.2.1 识别任务

文字识别任务的常见指标有行准确率(line accuracy, LA)和字符准确率(character accuracy, CA)。其中,行准确率在场景文字识别中也称为词准确率,或简称准确率,其定义为

$$LA = \frac{\sum_{i=1}^{N} Same(PR_i, GT_i)}{N} \tag{3}$$

式中,$Same$为两个字符串相同的示性函数,N为数据集中的总词条数,PR_i和GT_i分别表示第i个文本行样本的预测结果和真值。

字符准确率通常用1-NED进行度量。NED是标准化的编辑距离(normalized edit distance)的缩写,通常用编辑距(edit distance, ED)计算。字符准确率的两个常见定义具体为

$$CA = 1 - \frac{\sum_{i}^{N} ED(GT_i, PR_i)}{\sum_{i}^{N} Len(GT_i)} \tag{4}$$

或

$$CA = 1 - \frac{1}{N}\sum_{i}^{N} \frac{ED(GT_i, PR_i)}{\max(Len(GT_i), Len(PR_i))} \tag{5}$$

注意这两种定义的分母不同。式(4)先对每个词条求平均编辑距离,再对数据集求平均(Wang等,2020b)。式(5)则将整个数据集的所有字符串起来求编辑距离(Chng等,2019),因此该指标更突出长文本的识别性能。字符准确率往往不作为衡量模型好坏的主要指标,且有一定二义性。对于字符识别任务来说,由于单个样本中只包含一个字符,此时行准确率与字符准确率计算结果相同。

5.2.2 拒识任务

对于新字符拒识任务,OSTR使用词级别的召回率(Recall)、准确率(Precision)和F值(F-measure)对拒识性能进行衡量。具体为

$$R = \frac{\sum_{i}^{N} Rej(PR_i) \cdot Rej(GT_i)}{\sum_{i}^{N} Rej(GT_i)}$$

$$P = \frac{\sum_{i}^{N} Rej(PR_i) \cdot Rej(GT_i)}{\sum_{i}^{N} Rej(PR_i)} \tag{6}$$

式中,Rej为样本中含有集外字符(SOC和NOC)的示性函数。PR_i和GT_i分别表示第i个文本行样本的预测结果和真值。OSTR使用行级别的衡量标准而非字符级别的衡量标准,从而简化指标计算,并避免时序对齐过程中产生的歧义。这是由于无论含有集外字符的文本行样本中是否含有其他集内字符,该样本都需要人为检视。同时,提供文本行中的集外字符的具体位置对检视工作量并没有明显帮助。

具体到每一个指标,召回率描述的是对于一个含有集外字符的样本,模型对该样本进行拒识并提示管理员对当前字符集进行修订的概率。准确率则度量在所有模型发出的集外字符警报中,管理员需要手动忽略的假警报的比例。按照惯例,使用F值表示召回率和准确率的综合指标,其计算式为

$$F = \frac{2RP}{R + P} \tag{7}$$

5.3 评测协议

评测协议包括使用的训练集、测试集和评价指标。本文总结了一些典型任务对应的评测协议,如表3和表4所示,分别对应训练集和测试集。对于封闭集文字识别,最常见的协议是在MJ(Jaderberg等,2014)和ST(Gupta等,2016)这两个合成数据集上进行训练,在IIIT5k(Mishra等,2012)、CUTE(Risnumawan等,2014)、SVT(Wang等,2011)、SVTP(Phan等,2013)、IC03(Lucas等,2005)、IC13(Karatzas等,2013)和IC15(Karatzas等,2015)等数据集上进行测试,并统计行准确率,这种评测协议称为www(what is wrong with scene text recognition model comparisons)(Baek等,2019)。该协议在不同方法中存在若干变种,例如追加字符集别训练标注、只用一个训练集训练或使用部分数据集进行测试等。

特别地,针对中文封闭集文字识别,Yu等人(2021)给出了一种4种中文场景下的评测协议Fudan。在该协议下,对典型文字识别方法进行了测试,使用行准确率和字符准确率来度量识别效果。注意协议在评测某些方法时会做出一些变化,如引入字符定位的标注(Liao等,2019)、引入外来的语言模型(Qiao等,2020;Fang等,2021)或加入新的训练数据(Li等,2019a)等。在技术选型时,要注意这些变更可能造成的成本和性能影响。

www和Fudan两种评测协议可以反映模型应对集内字符的识别能力。在大多数开放环境下,集内文字构成的样本仍然占多数,因此选取开放集文字

表 3　常见文字识别评测协议（训练集）

Table 3　Common evaluation protocol for text recognition task (training set)

任务类型	名称	变种	训练集		
			数据集	字符集	额外标注
封闭集	www（Baek 等，2019）	Common	MJ, ST	英文，数字	–
		Semantic			语言模型
		Character			字符位置
	Fudan（Yu 等，2021）	Scene	RCTW, ReCTS, LSVT, ArT, CTW	7 938 个中文字符	部件组成
		Web	MTWI		
		Document	FudanVI		
		Handwriting	SCUT-HCCDoc		
零样本	HDE（Cao 等，2020）	HWDB	HWDB subsets	中文字符集 A[1]	部件组成
		CTW	CTW subsets	中文字符集 B[2]	
	Fudan（Chen 等，2021b）	HWDB	HWDB subsets	中文字符集 A	部件组成
		CTW	CTW subsets	中文字符集 B	
开放集	OS-OCR（Liu 等，2023）	GZSL-JAP	RCTW, LSVT, ArT, CTW, MLT（Latin, Chinese）	中文，英文，数字	–
		GZSL-KR			–
		OSR			–
		GOSR			–
		OSTR			–

注：[1]中文字符训练集大小分为 5 组，每组字符集大小分别为 500，1 000，1 500，2 000，2 755。[2]中文字符训练集大小分为 5 组，每组字符集大小分别为 500，1 000，1 500，2 000，3 150。“–”表示无额外标注。

识别方法时应当兼顾封闭集文字识别性能。

对于零样本文字识别，在字符级别训练和测试目前形成了两种评测协议，称为 HDE（Cao 等，2020）和 Fudan（Chen 等，2021b）。

需要注意的是，虽然这两个协议数据集与划分规模相同，但是由于划分规则的区别造成了显著的难度区别。两种协议分别用 HWDB 和 CTW 两个数据集进行训练和测试。这两种协议都是首先划定测试集，在剩下的字符中划定不同规模的训练集。不同的是，HDE 是随机进行划分，Fudan 是按照某种特定规则进行划分。

目前，行级别的零样本文字识别和开放集文字识别尚未形成共识，这些方法（Zhang 等，2020a；Zhang 等，2020b；Huang 等，2021；Liu 等，2022a；Souibgui 等，2022；Liu 等，2023）使用不同的训练集、测试集和评价指标进行评测。这里，本文列出与开放集文字识别任务相关性较强的 OSOCR（Liu 等，2023）和 OpenCCD（Liu 等，2022a）方法使用的协议。该协议在中英文上进行模型训练，并报告韩文和日文上的行准确率，该组协议同时也报告字符准确率作为参考。

对于开放集文字识别，OSOCR 给出了完整的对新旧字符的识别能力和拒识能力的评测协议 OSTR。在问题层面，OSTR 协议是涵盖封闭集文字识别和零样本文字识别的更一般的情形。同时，该工作给出了另外两种特殊情形。1）聚焦拒识的 OSR（open-set reconition）协议，对应用户只希望通过重新训练来获得对新字符识别的能力。2）忽略对已知字符（seen out-of-set character，SOC）拒识需求的 GOSR（generalized open-set recognition）协议，即用户处理的字符集大小可控，不需要通过临时遗忘不常出现的字符进行提速。

5.4　主流方法及性能

对于上述协议和评价指标，表 5 和表 6 列举了一些主流方法及使用 www 和 Fudan 两种评测协议的性能。表中，标签（tag）列给出对实现的补充说明，*代

表4 常见文字识别评测协议(测试集)

Table 4 Common evaluation protocol for text recognition task (test set)

任务类型	名称	变种	测试集					指标
			数据集	字符集				
				SIC	SOC	NIC	NOC	
封闭集	www(Baek等,2019)	Common	IIIT5k, CUTE80, SVT, SVTP, IC03, IC13, IC15	英文,数字	–	–	–	LA
		Semantic			–	–	–	
		Character			–	–	–	
	Fudan(Yu等,2021)	Scene	RCTW, ReCTS, LSVT, ArT, CTW	7 938个中文字符	–	–	–	LA/ CA
		Web	MTWI		–	–	–	
		Document	FudanVI		–	–	–	
		Handwriting	SCUT-HCCDoc		–	–	–	
零样本	HDE(Cao等,2020)	HWDB	HWDB子集(剩余1 000未见字符)	–	–	1 000新字符[1]	–	LA
		CTW	CTW子集(剩余500未见字符)	–	–	500新字符		
	Fudan(Chen等,2021b)	HWDB	HWDB subsets(剩余1 000未见字符)	–	–	1 000新字符	–	LA
		CTW	CTW subsets(剩余500未见字符)	–	–	500新字符	–	LA
开放集	OS-OCR(Liu等,2023)	GZSL-JAP	MLT-Japanese	–	–	–	–	LA/ CA
		GZSL-KR	MLT-Korean	–	–	–	–	LA/ CA
		OSR	MLT-Japanese[2]	SK,英文,数字	–	–	UK, Ka	LA/ R/P/F
		GOSR	MLT-Japanese	SK,英文,数字	–	UK	Ka	LA/ R/P/F
		OSTR	MLT-Japanese	SK	英文,数字	UK	Ka	LA/ R/P/F

注:[1]Fudan协议中,训练和测试的字符集并非随机划分。该协议的划分首先将字符按照某种规则排序后,将最末尾500或1 000个字符作为测试类别。这导致该协议明显难于HDE协议。[2]此处SK指日文中与一类中文简体汉字重合的部分,UK指日文中不在一类中文简体汉字的日文汉字。Ka指日文中的平假名和片假名。“–”表示不包含该类型字符集。

表协议变种,+/–代表使用更多或更少的训练数据,R代表第三方复现版本。这里对复现的定义包括使用原作释出的代码在自己的训练集上训练得到的模型。考虑到调参和环境对性能影响也比较大,并非所有方法都在全部的协议或测试集上进行了测试。总体而言,对于www协议(Baek等,2019),在英文数据集上的性能已经整体较为成熟,方法以特征归集为主。另一方面,特征归集类方法对中文文字识别适应性稍差,可能与训练数据量、字符数量和词条长度有关。

零样本字符识别广泛使用HDE(Cao等,2020)和Fudan协议(Chen等,2021b)进行评价,不同方法的性能分别如表7和表8所示。由于字符可以看做“独字成行”的文本行,所以此处字符准确率等同于行准确率。Fudan协议的训练字符集和测试字符集是按照GB18030-2005划分的,但没明确HWDB(handwriting database)的排序原则。该划分方式会导致测试字符集选取更多的生僻字,造成Fudan协议难度显著增加。

目前,对于完整功能的开放集文字识别方法的协议和方法较少,OSOCR和OpenCCD方法的性能如表9所示。该评测协议尚处于起步阶段,有很大提升空间,尤其是对新字符的拒识性能尚有较大欠缺。

6 国内外研究进展

目前,开放集文字识别工作大多是国内研究团

表 5　典型的封闭集识别方法及性能（www 协议）

Table 5　Typical closed set identification methods and performance（www protocol）

类型	方法	www 协议，结果为行准确率（LA）							
		Tag	IIIT5k	CUTE	IC03	IC13	IC15	SVT	SVTP
标签归集	CRNN（Shi 等，2017a）	R（Zhang 等，2020a）	82.9	65.5	92.6	89.2	64.2	81.6	70.0
	Rosetta（Borisyuk 等，2018）	R（Zhang 等，2020a）	84.3	69.2	92.9	89.0	66.0	84.7	73.8
	CA-FCN（Liao 等，2019）	−*	92.0	78.1	/	91.4	/	82.1	/
特征归集	ASTER（Shi 等，2019）	/	93.4	79.5	94.5	91.8	76.1	93.6	78.5
	MORAN（Luo 等，2019）	/	91.2	77.4	95.0	92.4	68.8	88.3	76.1
	DAN（Wang 等，2020b）	/	94.3	84.4	95.0	93.9	74.5	89.2	80.0
	SEED（Qiao 等，2020）	*	93.4	84.0	/	93.5	75.8	88.4	82.0
	ABINet（Fang 等，2021）	*	96.2	89.2	/	97.4	86.0	93.5	88.5
	SAR（Li 等，2019a）	+	95.0	89.6	/	94.0	78.8	91.2	86.4
	TransOCR（Chen 等，2021c）	+	/	/	/	/	/	/	/

注：“/”表示该方法未在此数据集上进行评测，表中空白处代表标准协议，R（）代表非方法原作者报告的结果，括号中数值为该结果复现的来源论文，“+”代表额外数据，“*”代表额外模态的数据（例如字符 mask、语料等），“−”代表使用了比标准协议更少的训练数据。

表 6　典型的封闭集识别方法及性能（Fudan 协议）

Table 6　Typical closed set identification methods and performance（Fudan protocol）

类型	方法	Fudan 协议，结果为 LA（CA）				
		Tag	Scene	Web	Document	Hand-writing
标签归集	CRNN（Shi 等，2017a）	R（Yu 等，2021）	54.94（0.742）	56.21（0.745）	97.41（0.995）	48.04（0.843）
	Rosetta（Borisyuk 等，2018）	/	/	/	/	/
	CA-FCN（Liao 等，2019）	/	/	/	/	/
特征归集	ASTER（Shi 等，2019）	R（Yu 等，2021）	59.37（0.801）	57.83（0.782）	97.59（0.995）	45.90（0.819）
	MORAN（Luo 等，2019）	R（Yu 等，2021）	54.68（0.710）	49.64（0.679）	91.66（0.984）	30.24（0.651）
	DAN（Wang 等，2020b）	/	/	/	/	/
	SEED（Qiao 等，2020）	R（Yu 等，2021）	45.37（0.708）	31.35（0.571）	96.08（0.992）	21.10（0.555）
	ABINet（Fang 等，2021）	/	/	/	/	/
	SAR（Li 等，2019a）	R（Yu 等，2021）	53.80（0.738）	50.49（0.705）	96.23（0.993）	30.95（0.732）
	TransOCR（Chen 等，2021c）	/	67.81（0.817）	62.74（0.782）	97.86（0.996）	51.67（0.835）

注：“/”表示该方法未在此数据集上进行评测，表中空白处代表标准协议，R（）代表非方法原作者报告的结果，括号中数值为该结果复现的来源论文，“+”代表额外数据，“*”代表额外模态的数据（例如字符 mask、语料等）。

表 7　零样本字符识别典型方法及性能（HDE 协议）

Table 7　Typical methods and performance of zero-sample character recognition（HDE protocol）　/%

方法	HWDB（结果为准确率 LA）					CTW（结果为准确率 LA）				
	500	1 000	1 500	2 000	2 755	500	1 000	1 500	2 000	3 150
DenseRAN（Wang 等，2018）	1.70	8.44	14.71	19.51	30.68	0.12	1.50	4.95	10.08	15.95
FewRAN（Wang 等，2019）	33.60	41.50	63.80	70.60	77.20	2.36	10.49	16.59	22.03	28.45
HDE（Cao 等，2020）	33.70	53.91	66.27	73.42	80.95	23.53	38.47	44.17	49.79	57.42

表8　零样本字符识别典型方法及性能(Fudan协议)

Table 8　Typical methods and performance of zero sample character recognition (Fudan protocol)

方法	HWDB(结果为LA)					CTW(结果为LA)				
	500	1 000	1 500	2 000	2 755	500	1 000	1 500	2 000	3 150
DenseRAN(Wang等,2018)	1.70	8.44	14.71	19.51	30.68	0.15	0.54	1.60	1.95	5.39
HDE(Cao等,2020)	4.90	12.77	19.25	25.13	33.49	0.82	2.11	3.11	6.96	7.75
Stroke(Chen等,2021b)	5.60	13.85	22.88	25.73	37.91	1.54	2.54	4.32	6.82	8.61
ACPM(Zu等,2022)	9.72	18.50	27.74	34.00	42.43	3.44	6.18	10.65	15.40	21.29

表9　开放集文字识别典型方法及性能

Table 9　Typical methods and performance of open set character recognition

名称	GZSL-JAP (LA)	GZSL-KR (LA)	OSR (LA/R/P/F)	GOSR(LA/R/P/F)	OSTR(LA/R/P/F)
OpenCCD-L (Liu等,2023)	41.31	19.16	–/–/–/–	–/–/–/–	–/–/–/–
OSOCR-L (Liu等,2022a)	30.83	1.35	74.35/11.27/98.28/20.23	56.03/3.03/63.52/5.78	58.57/24.46/93.78/38.80

注:"–"表示在对应数据集上没有官方测试结果。

队开展的。总的来说,国内外研究目标和应用场景具体差别不大。这些研究都是为了解决快速适应多语言(Zhang等,2020a;Liu等,2022a,2023;Souibgui等,2022)或古籍文字识别(Chanda等,2018,2021;Huang等,2021;Rai等,2021)中遇到的新字符处理问题。

从识别语言的角度,国内的研究以CJK字符为主(Cao等,2020;Huang等,2021),国外则对不同语系的语言进行研究。He和Schomaker(2018)以中文为研究对象,Chanda等人(2018,2021)研究中世纪拉丁文、孟加拉语和古韩文,Zhang等人(2020a)研究拉丁语系,Souibgui等人(2022)研究一些其他的小语种。

从文本识别粒度的角度,字符级别和行级别在国内外均有研究。国内外在研究字符级识别时,都是将该任务视为开放集分类问题,更宽泛地说,是视为细粒度长尾分类问题。而行级别则在字符级别基础上引入了序列信息的处理问题。相应地,国内外也都有使用字符识别对通用开放集、少(零)样本或长尾识别方法进行验证的示例(Bertinetto等,2016;Mishra等,2022)。国内外对于开放集识别的任务定义没有明显差别。

从整体技术路线的角度,国内外研究没有明确的分界。在开放集分类技术上,国内外均有基于部件和基于匹配的做法。在框架上,国内研究团队所提出的方法大多是针对某种封闭集识别框架的扩展;国外研究团队较少采用主流基于字符序列的封闭集文本识别框架,而是沿用早期的整词分类框架的方法(Chanda等,2018,2021);也有部分国外研究团队提出了新的封闭集文字识别框架并将其扩展为开放集文字识别模型(Zhang等,2020a;Souibgui等,2022)。

从开集分类器实现的角度,国内外的工作主要集中在基于属性和基于视觉匹配的两类方法上。一方面,国内外团队研究基于部件的方法时,大体思路均是通过部件的复用获得对未知类别的泛化能力。在研究工作中,不同的识别语言通常导致不同的部件设计与实现。另一方面,国内(Liu等,2022a,2023)和国外(Zhang等,2020a;Souibgui等,2022)针对基于视觉匹配的方法的研究整体差别不大。对于拒识任务,由于完整特性的开放集文字识别任务仍然是个新概念,在文字识别领域,目前只有国内的少数团队在做(Liu等,2023)。在文本以外目标的开放集识别任务中,国内外研究团队都有相关工作。

从实现的角度,国内研究工作的实现方式相对多样。对于基于属性匹配的方法(Chanda等,2018,2021;He和Schomaker,2018;Rai等,2021),国外研究团队(Chanda等,2018,2021;Rai等,2021)对属性进行无顺序的计数,在训练意义上很像ACE(aggregation cross-entropy)损失(Xie等,2019);国内研究团

队(Zhang等,2018,2020b;Chen等,2021b)采用直接属性比较的方法,使用部件序列作为比较对象;部分国内研究工作(Cao等,2020;Huang等,2021)也将属性信息编码为特征向量,并将之与图像特征进行直接比较,这类方法在国外文字识别相关研究工作中不太常见,但在国外通用零样本的研究中比较普遍。对于基于视觉匹配的方法,国内外研究工作均有涉及。特别地,针对CJK字符集较大的现象,在进行基于视觉匹配的开放集文字识别方法训练时,国内研究团队常常会引入对标签的采样机制,以此减少计算资源开销。

7 发展趋势与技术展望

虽然零样本文字识别(Zhang等,2020a;Huang等,2021;Souibgui等,2022)和开放集文字识别方法(Liu等,2022a,2023)取得了一定的研究进展,但现有方法在性能上还存在较大局限。开放集文字识别技术发展的主要趋势包括以下几个方面:1)跨语言文字识别技术。现有工作对不同语言文字形状差异处理及跨语言文字识别的鲁棒性不强。主要体现在开放集分类器难以进行跨语系字符迁移,同时对注意力机制的定位能力也有较大影响。这类问题属于领域鸿沟(domain gap),是近年来模式识别与机器学习的热门研究方向之一。在开放集文字识别任务中,一方面可以增加训练的语言数来缓解这个问题;另一方面可以从上游的域泛化(domain generalization)(Wang等,2021)或者无监督域适应(unsupervised domain adaption)(Zhang,2021)方法中进行借鉴,以此解决对跨语言甚至跨语系新字符识别问题。2)字符细粒度分析技术。现有模型性能受字符集$\boldsymbol{C}_{\text{test}}^{k}$的大小影响较大,这意味着模型对字符细节区别不够敏感。该问题可以视为字符的细粒度分类问题(fine-grained classification)(Wei等,2022)。未来需考虑应用细粒度分类的方法来提升开放集文字识别方法的可扩展性。3)新字符归纳发现技术。现有方法无法自动对数据中的新—集外(NOC)字符进行归纳。在古籍识别的过程中,归纳能力可以用来对新字符进行半自动或自动的字形分析,从而减少专家的工作量。现有开放集文字识别方法只能做到初步的拒识,尚不能做到对新字符的自动归纳。目前,新样本归纳(Huang等,2019)属于较新的研究方向,尚没有成熟的研究,有待不同领域的共同探讨。4)语言模型增量演化技术。目前大多数开放集文字识别方法(Wan等,2020;Liu等,2022a)试图排除语言模型的影响。事实上,在很多应用环境中,语言模型是渐近增量式演化的,因此可以对语言模型进行无监督训练(Devlin等,2019),建立基于增量学习的语义模型,并使之随测试数据演化,从而提高开放集文字识别性能。

另外,与主流封闭集文字识别算法相比,现有的零样本和开放集文字识别方法在处理已知字符的性能上还有一定差距。除了语言模型增量演化外,这一问题还可以通过更换特征提取网络(Atienza,2021)或者使用校正模块(Jaderberg等,2015;Luo等,2019;Yang等,2019)等方式进行改进。

致　谢　本文由中国图象图形学学会文档图像分析与识别专委会组织撰写,该专委会链接为http://www.csig.org.cn/detail/2551。

参考文献(References)

Almazán J, Gordo A, Fornés A and Valveny E. 2014. Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(12): 2552-2566[DOI: 10.1109/TPAMI.2014.2339814]

Ao X, Zhang X Y, Yang H M, Yin F and Liu C L. 2019. Cross-modal prototype learning for zero-shot handwriting recognition//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 589-594[DOI: 10.1109/ICDAR.2019.00100]

Atienza R. 2021. Vision transformer for fast and efficient scene text recognition//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 319-334[DOI: 10.1007/978-3-030-86549-8_21]

Baek J, Kim G, Lee J, Park S, Han D, Yun S, Oh S J and Lee H. 2019. What is wrong with scene text recognition model comparisons? Dataset and model analysis//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4714-4722[DOI: 10.1109/ICCV.2019.00481]

Bao W T, Yu Q and Kong Y. 2022. OpenTAL: towards open set temporal action localization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 2969-2979[DOI: 10.1109/CVPR52688.2022.00299]

Bendale A and Boult T. 2015. Towards open world recognition//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1893-1902[DOI: 10.1109/CVPR.2015.7298799]

Bertinetto L, Henriques J F, Valmadre J, Torr P H S and Vedaldi A. 2016. Learning feed-forward one-shot learners//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 523-531

Borisyuk F, Gordo A and Sivakumar V. 2018. Rosetta: large scale system for text detection and recognition in images//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. London, UK: ACM: 71-79 [DOI: 10.1145/3219819.3219861]

Cao Z, Lu J, Cui S and Zhang C S. 2020. Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding. Pattern Recognition, 107: #107488 [DOI: 10.1016/j.patcog.2020.107488]

Chanda S, Baas J, Haitink D, Hamel S, Stutzmann D and Schomaker L. 2018. Zero-shot learning based approach for medieval word recognition using deep-learned features//Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition. Niagara Falls, USA: IEEE: 345-350 [DOI: 10.1109/ICFHR-2018.2018.00067]

Chanda S, Haitink D, Prasad P K, Baas J, Pal U and Schomaker L. 2021. Recognizing bengali word images——A zero-shot learning perspective//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 5603-5610 [DOI: 10.1109/ICPR48806.2021.9412607]

Chen C F, Yang X S, Xu C S, Huang X H and Ma Z. 2021a. ECKPN: explicit class knowledge propagation network for transductive few-shot learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6596-6605 [DOI: 10.1109/cvpr46437.2021.00653]

Chen G Y, Qiao L M, Shi Y M, Peng P X, Li J, Huang T J, Pu S L and Tian Y H. 2020. Learning open set network with discriminative reciprocal points//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 507-522 [DOI: 10.1007/978-3-030-58580-8_30]

Chen J Y, Li B and Xue X Y. 2021b. Zero-shot Chinese character recognition with stroke-level decomposition//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI.org: 615-621 [DOI: 10.24963/ijcai.2021/85]

Chen J Y, Li B and Xue X Y. 2021c. Scene text telescope: text-focused scene image super-resolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12021-12030 [DOI: 10.1109/cvpr46437.2021.01185]

Chen X X, Jin L W, Zhu Y Z, Luo C J and Wang T W. 2022. Text recognition in the wild: a survey. ACM Computing Surveys, 54(2): #42 [DOI: 10.1145/3440756]

Chen Z T, Fu Y W, Zhang Y D, Jiang Y G, Xue X Y and Sigal L. 2019. Multi-level semantic feature augmentation for one-shot learning. IEEE Transactions on Image Processing, 28(9): 4594-4605 [DOI: 10.1109/TIP.2019.2910052]

Cheng Z Z, Xu Y L, Bai F, Niu Y, Pu S L and Zhou S G. 2018. AON: towards arbitrarily-oriented text recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5571-5579 [DOI: 10.1109/cvpr.2018.00584]

Chng C K, Liu Y L, Sun Y P, Ng C C, Luo C J, Ni Z H, Fang C M, Zhang S T, Han J Y, Ding E R, Liu J T, Karatzas D, Chan C S and Jin L W. 2019. ICDAR2019 robust reading challenge on arbitrary-shaped text——RRC-ArT//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 1571-1576 [DOI: 10.1109/icdar.2019.00252]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA: ACL: 4171-4186 [DOI: 10.18653/v1/n19-1423]

Diao X L, Shi D Q, Tang H, Wu L, Li Y Z and Xu H. 2022. REZCR: a zero-shot character recognition method via radical extraction [EB/OL]. [2022-08-17]. https://arxiv.org/pdf/2207.05842.pdf

Ding C B, Pang G S and Shen C H. 2022. Catching both gray and black swans: open-set supervised anomaly detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 7378-7388 [DOI: 10.1109/CVPR52688.2022.00724]

Doan T and Kalita J. 2017. Overcoming the challenge for text classification in the open world//Proceedings of the 7th IEEE Annual Computing and Communication Workshop and Conference (CCWC). Las Vegas, USA: IEEE: 1-7 [DOI: 10.1109/CCWC.2017.7868366]

Du Y, Wei F Y, Zhang Z H, Shi M J, Gao Y and Li G Q. 2022. Learning to prompt for open-vocabulary object detection with vision-language model//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 14064-14073 [DOI: 10.1109/CVPR52688.2022.01369]

Egglin T K and Feinstein A R. 1996. Context bias. A problem in diagnostic radiology. JAMA, 276(21): 1752-1755 [DOI: 10.1001/jama.276.21.1752]

Fang S C, Xie H T, Wang Y X, Mao Z D and Zhang Y D. 2021. Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 7094-7103 [DOI: 10.1109/cvpr46437.2021.00702]

Fei G L and Liu B. 2016. Breaking the closed world assumption in text classification//Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: ACL: 506-514 [DOI: 10.18653/v1/n16-1061]

Fu Y W and Sigal L. 2016. Semi-supervised vocabulary-informed learning//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5337-5346 [DOI: 10.1109/CVPR.2016.576]

Fu Y W, Wang X M, Dong H Z, Jiang Y G, Wang M, Xue X Y and Sigal L. 2020. Vocabulary-informed zero-shot and open-set learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42 (12): 3136-3152 [DOI: 10.1109/TPAMI. 2019. 2922175]

Fu Y W, Xiang T, Jiang Y G, Xue X Y, Sigal L and Gong S G. 2018. Recent advances in zero-shot recognition: toward data-efficient understanding of visual content. IEEE Signal Processing Magazine, 35(1): 112-125 [DOI: 10.1109/msp.2017.2763441]

Garcia-Bordils S, Mafla A, Biten A F, Nuriel O, Aberdam A, Mazor S, Litman R and Karatzas D. 2023. Out-of-vocabulary challenge report//Proceedings of Computer Vision —— ECCV 2022 Workshops. Tel Aviv, Israel: Springer: 359-375

Ge Z Y, Demyanov S and Garnavi R. 2017. Generative openmax for multi-class open set classification//Proceedings of 2017 British Machine Vision Conference. London, UK: BMVA Press: #42 [DOI: 10.5244/c.31.42]

Geng C X and Chen S C. 2022. Collective decision for open set recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1): 192-204 [DOI: 10.1109/TKDE.2020.2978199]

Geng C X, Huang S J and Chen S C. 2021. Recent advances in open set recognition: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43 (10): 3614-3631 [DOI: 10.1109/TPAMI.2020.2981604]

Guo X Q, Liu J, Liu T L and Yuan Y X. 2022. SimT: handling open-set noise for domain adaptive semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7022-7031 [DOI: 10.1109/CVPR52688.2022.00690]

Gupta A, Narayan S, Joseph K J, Khan S, Khan F S and Shah M. 2022. OW-DETR: open-world detection transformer//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9225-9234 [DOI: 10.1109/CVPR52688.2022.00902]

Gupta A, Vedaldi A and Zisserman A. 2016. Synthetic data for text localisation in natural images//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2315-2324 [DOI: 10.1109/cvpr.2016.254]

Han J M, Ren Y Q, Ding J, Pan X J, Yan K and Xia G S. 2022. Expanding low-density latent regions for open-set object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9581-9590 [DOI: 10.1109/CVPR52688.2022.00937]

He M C, Liu Y L, Yang Z B, Zhang S, Luo C J, Gao F Y, Zheng Q, Wang Y P, Zhang X and Jin L W. 2018. ICPR2018 contest on robust reading for multi-type web images//Proceedings of the 24th International Conference on Pattern Recognition. Beijing, China: IEEE: 7-12 [DOI: 10.1109/ICPR.2018.8546143]

He S and Schomaker L. 2018. Open set Chinese character recognition using multi-typed attributes [EB/OL]. [2023-01-11]. https://arxiv.org/pdf/1808.08993.pdf

Hou R B, Chang H, Ma B P, Shan S G and Chen X L. 2019. Cross attention network for few-shot classification//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #360

Hu J S, Liu C Y, Yan Q D, Zhu X Y, Yu F L, Wu J J and Yin B. 2022. Vision-language adaptive mutual decoder for OOV-STR [EB/OL]. [2022-09-02]. https://arxiv.org/pdf/2209.00859.pdf

Huang G J, Luo X Y, Wang S W, Gu T L and Su K L. 2022a. Hippocampus-heuristic character recognition network for zero-shot learning in Chinese character recognition. Pattern Recognition, 130: #108818 [DOI: 10.1016/j.patcog.2022.108818]

Huang S P, Wang H B, Liu Y G, Shi X S and Jin L W. 2019. OBC306: a large-scale oracle bone character recognition dataset//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 681-688 [DOI: 10.1109/icdar.2019.00114]

Huang S Y, Ma J W, Han G X and Chang S F. 2022b. Task-adaptive negative envision for few-shot open-set recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7161-7170 [DOI: 10.1109/CVPR52688.2022.00703]

Huang Y H, Jin L W and Peng D Z. 2021. Zero-shot Chinese text recognition via matching class embedding//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 127-141 [DOI: 10.1007/978-3-030-86334-0_9]

Jaderberg M, Simonyan K, Vedaldi A and Zisserman A. 2014. Synthetic data and artificial neural networks for natural scene text recognition [EB/OL]. [2022-12-09]. https://arxiv.org/pdf/1406.2227.pdf

Jaderberg M, Simonyan K, Vedaldi A and Zisserman A. 2016. Reading text in the wild with convolutional neural networks. International Journal of Computer Vision, 116 (1): 1-20 [DOI: 10.1007/s11263-015-0823-z]

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. 2015. Spatial transformer networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2017-2025

Joseph K J, Khan S, Khan F S and Balasubramanian V N. 2021. Towards open world object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5826-5836 [DOI: 10.1109/cvpr46437.2021.00577]

Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar V R, Lu S J, Shafait F, Uchida S and Valveny E. 2015. ICDAR 2015 competi-

tion on robust reading//Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia: IEEE: 1156-1160 [DOI: 10.1109/icdar.2015.7333942]

Karatzas D, Shafait F, Uchida S, Iwamura M, i Bigorda L G, Mestre S R, Mas J, Mota D F, Almazàn J A and de las Heras L P. 2013. ICDAR 2013 robust reading competition//Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, USA: IEEE: 1484-1493 [DOI: 10.1109/icdar.2013.221]

Kim J, Kim T, Kim S and Yoo C D. 2019. Edge-labeling graph neural network for few-shot learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11-20 [DOI: 10.1109/cvpr.2019.00010]

Kumar P, Pathania K and Raman B. 2022. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. Applied Intelligence: #6 [DOI: 10.1007/s10489-022-04046-6]

Li B C, Tang X, Qi X B, Chen Y H and Xiao R. 2020. Hamming OCR: a locality sensitive hashing neural network for scene text recognition [EB/OL]. [2020-09-23]. https://arxiv.org/pdf/2209.10874.pdf

Li H, Wang P, Shen C H and Zhang G Y. 2019a. Show, attend and read: a simple and strong baseline for irregular text recognition//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, the 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. Honolulu, USA: AAAI: 8610-8617 [DOI: 10.1609/aaai.v33i01.33018610]

Li H Y, Eigen D, Dodge S, Zeiler M and Wang X G. 2019b. Finding task-relevant features for few-shot learning by category traversal//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1-10 [DOI: 10.1109/cvpr.2019.00009]

Li W B, Wang L, Xu J L, Huo J, Gao Y and Luo J B. 2019c. Revisiting local descriptor based image-to-class measure for few-shot learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7253-7260 [DOI: 10.1109/cvpr.2019.00743]

Liao M H, Zhang J, Wan Z Y, Xie F M, Liang J J, Lyu P Y, Yao C and Bai X. 2019. Scene text recognition from two-dimensional perspective//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, the 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. Honolulu, USA: AAAI: 8714-8721 [DOI: 10.1609/aaai.v33i01.33018714]

Lin W W, Ma T, Zhang Z Q, Li X F and Xue X S. 2022. Variational autoencoder for zero-shot recognition of bai characters. Wireless Communications and Mobile Computing, 2022: #2717322 [DOI: 10.1155/2022/2717322]

Liu C, Yang C and Yin X C. 2022a. Open-set text recognition via character-context decoupling//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE/CVF: 4513-4522 [DOI: 10.1109/cvpr52688.2022.00448]

Liu C, Yang C, Qin H B, Zhu X B, Liu C L and Yin X C. 2023. Towards open-set text recognition via label-to-prototype learning. Pattern Recognition, 134: #109109 [DOI: 10.1016/j.patcog.2022.109109]

Liu C L, Yin F, Wang D H and Wang Q F. 2011. CASIA online and offline Chinese handwriting databases//Proceedings of 2011 International Conference on Document Analysis and Recognition. Beijing, China: IEEE: 37-41 [DOI: 10.1109/icdar.2011.17]

Liu C Y, Chen X X, Luo C J, Jin L W, Xue Y and Liu Y L. 2021. Deep learning methods for scene text detection and recognition. Journal of Image and Graphics, 26(6): 1330-1367 (刘崇宇, 陈晓雪, 罗灿杰, 金连文, 薛洋, 刘禹良. 2021. 自然场景文本检测与识别的深度学习方法. 中国图象图形学报, 26(6): 1330-1367) [DOI: 10.11834/jig.210044]

Liu R Y, Liu H, Li G, Hou H D, Yu T H and Yang T. 2022b. Contextual debiasing for visual recognition with causal mechanisms//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12745-12755 [DOI: 10.1109/CVPR52688.2022.01242]

Liu Y L, Jin L W, Zhang S T and Zhang S. 2017. Detecting curve text in the wild: new dataset and new solution [EB/OL]. [2017-12-06]. https://arxiv.org/pdf/1712.02170.pdf

Lucas S M, Panaretos A, Sosa L, Tang A, Wong S, Young R, Ashida K, Nagai H, Okamoto M, Yamamoto H, Miyao H, Zhu J M, Ou W W, Wolf C, Jolion J M, Todoran L, Worring M and Lin X F. 2005. ICDAR 2003 robust reading competitions: entries, results, and future directions. International Journal of Document Analysis and Recognition (IJDAR), 7(2/3): 105-122 [DOI: 10.1007/s10032-004-0134-3]

Luo C J, Jin L W and Sun Z H. 2019. MORAN: a multi-object rectified attention network for scene text recognition. Pattern Recognition, 90: 109-118 [DOI: 10.1016/j.patcog.2019.01.020]

Ma Y Q, Bai S H, An S, Liu W, Liu A S, Zhen X T and Liu X L. 2020. Transductive relation-propagation network for few-shot learning//Proceedings of the 29th International Joint Conference on Artificial Intelligence. [s.l.]: IJCAI.org: 804-810 [DOI: 10.24963/ijcai.2020/112]

Manmatha R, Han C F and Riseman E M. 1996. Word spotting: a new approach to indexing handwriting//Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE: 631-637 [DOI: 10.1109/CVPR.1996.517139]

Mendes Júnior P R, de Souza R M, de O. Werneck R, Stein B V, Pazinato D V, de Almeida W R, Penatti O A B, da S. Torres R and Rocha A. 2017. Nearest neighbors distance ratio open-set classifier.

Machine Learning, 106(3): 359-386 [DOI: 10.1007/s10994-016-5610-8]

Mishra A, Alahari K and Jawahar C. 2012. Scene text recognition using higher order language priors//Proceedings of 2012 British Machine Vision Conference. Surrey, UK: BMVA Press: 127.1-127.11 [DOI: 10.5244/C.26.127]

Mishra S, Zhu P and Saligrama V. 2022. Learning compositional representations for effective low-shot generalization [EB/OL]. [2022-04-17]. https://arxiv.org/pdf/2204.08090.pdf

Nayef N, Patel Y, Busta M, Chowdhury P N, Karatzas D, Khlif W, Matas J, Pal U, Burie J C, Liu C L and Ogier J M. 2019. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition——RRC-MLT-2019//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 1582-1587 [DOI: 10.1109/ICDAR.2019.00254]

Naylor A R. 2010. Known knowns, known unknowns and unknown unknowns: a 2010 update on carotid artery disease. The Surgeon, 8(2): 79-86 [DOI: 10.1016/j.surge.2010.01.006]

Neal L, Olson M, Fern X, Wong W K and Li F X. 2018. Open set learning with counterfactual images//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 620-635 [DOI: 10.1007/978-3-030-01231-1_38]

Patel V M, Gopalan R, Li R N and Chellappa R. 2015. Visual domain adaptation: a survey of recent advances. IEEE Signal Processing Magazine, 32(3): 53-69 [DOI: 10.1109/msp.2014.2347059]

Phan T Q, Shivakumara P, Tian S X and Tan C L. 2013. Recognizing text with perspective distortion in natural scenes//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 569-576 [DOI: 10.1109/iccv.2013.76]

Pourpanah F, Abdar M, Luo Y X, Zhou X L, Wang R, Lim C P, Wang X Z and Wu Q M J. 2022. A review of generalized zero-shot learning methods. IEEE Transactions on Pattern Analysis and Machine Intelligence:#3191696 [DOI: 10.1109/TPAMI.2022.3191696]

Prakhya S, Venkataram V and Kalita J. 2017. Open set text classification using CNNs//Proceedings of the 14th International Conference on Natural Language Processing. Kolkata, India: NLP Association of India: 466-475

Qi H, Brown M and Lowe D G. 2018. Low-shot learning with imprinted weights//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5822-5830 [DOI: 10.1109/cvpr.2018.00610]

Qiao L M, Shi Y M, Li J, Wang Y H, Huang T J and Wang Y W. 2019. Transductive episodic-wise adaptive metric for few-shot learning//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3603-3612 [DOI: 10.1109/iccv.2019.00370]

Qiao Z, Zhou Y, Yang D B, Zhou Y C and Wang W P. 2020. SEED: semantics enhanced encoder-decoder framework for scene text recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13525-13534 [DOI: 10.1109/cvpr42600.2020.01354]

Rai A, Krishnan N C and Chanda S. 2021. Pho(SC)Net: an approach towards zero-shot word image recognition in historical documents//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 19-33 [DOI: 10.1007/978-3-030-86549-8_2]

Risnumawan A, Shivakumara P, Chan C S and Tan C L. 2014. A robust arbitrary text detection system for natural scene images. Expert Systems with Applications, 41(18): 8027-8048 [DOI: 10.1016/j.eswa.2014.07.008]

Scheirer W J, de Rezende Rocha A, Sapkota A and Boult T E. 2013. Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(7): 1757-1772 [DOI: 10.1109/TPAMI.2012.256]

Scheirer W J, Jain L P and Boult T E. 2014. probability models for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(11): 2317-2324 [DOI: 10.1109/TPAMI.2014.2321392]

Scherreik M D and Rigling B D. 2016. Open set recognition for automatic target classification with rejection. IEEE Transactions on Aerospace and Electronic Systems, 52(2): 632-642 [DOI: 10.1109/taes.2015.150027]

Shao L, Zhu F and Li X L. 2015. Transfer learning for visual categorization: a survey. IEEE Transactions on Neural Networks and Learning Systems, 26(5): 1019-1034 [DOI: 10.1109/TNNLS.2014.2330900]

Sheng F F, Chen Z N and Xu B. 2019. NRTR: a no-recurrence sequence-to-sequence model for scene text recognition//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 781-786 [DOI: 10.1109/icdar.2019.00130]

Shi B G, Bai X and Yao C. 2017a. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11): 2298-2304 [DOI: 10.1109/TPAMI.2016.2646371]

Shi B G, Yang M K, Wang X G, Lyu P Y, Yao C and Bai X. 2019. ASTER: an attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9): 2035-2048 [DOI: 10.1109/TPAMI.2018.2848939]

Shi B G, Yao C, Liao M H, Yang M K, Xu P, Cui L Y, Belongie S J, Lu S J and Bai X. 2017b. ICDAR2017 competition on reading Chinese text in the wild (RCTW-17)//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 1429-1434 [DOI: 10.1109/icdar.2017.233]

Shu L, Xu H and Liu B. 2017. DOC: deep open classification of text documents//Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: ACL: 2911-2916 [DOI: 10.18653/v1/d17-1314]

Shu Y, Shi Y M, Wang Y W, Huang T J and Tian Y H. 2020. P-ODN: prototype-based open deep network for open set recognition. Scientific Reports, 10(1): #7146 [DOI: 10.1038/s41598-020-63649-6]

Simon C, Koniusz P, Nock R and Harandi M. 2020. Adaptive subspaces for few-shot learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4136-4145 [DOI: 10.1109/cvpr42600.2020.00419]

Snell J, Swersky K and Zemel R. 2017. Prototypical networks for few-shot learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 4080-4090

Song N, Zhang C and Lin G S. 2022. Few-shot open-set recognition using background as unknowns//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 5970-5979 [DOI: 10.1145/3503161.3547933]

Souibgui M A, Fornés A, Kessentini Y and Megyesi B. 2022. Few shots are all you need: a progressive learning approach for low resource handwritten text recognition. Pattern Recognition Letters, 160: 43-49 [DOI: 10.1016/j.patrec.2022.06.003]

Su Y K, Sun R Z, Lin G S and Wu Q Y. 2021. Context decoupling augmentation for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6984-6994 [DOI: 10.1109/iccv48922.2021.00692]

Sun Y P, Ni Z H, Chng C K, Liu Y L, Luo C J, Ng C C, Han J Y, Ding E R, Liu J T, Karatzas D, Chan C S and Jin L W. 2019. ICDAR 2019 competition on large-scale street view text with partial labeling —— RRC-LSVT//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 1557-1562 [DOI: 10.1109/icdar.2019.00250]

Veit A, Matera T, Neumann L, Matas J and Belongie S. 2016. COCO-text: dataset and benchmark for text detection and recognition in natural images [EB/OL]. [2023-01-11]. https://arxiv.org/pdf/1601.07140.pdf

Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K and Wierstra D. 2016. Matching networks for one shot learning//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 3637-3645

Wan Z Y, Zhang J L, Zhang L, Luo J B and Yao C. 2020. On vocabulary reliance in scene text recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11422-11431 [DOI: 10.1109/cvpr42600.2020.01144]

Wang J D, Lan C L, Liu C, Ouyang Y D and Qin T. 2021. Generalizing to unseen domains: a survey on domain generalization//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI.org: 4627-4635 [DOI: 10.24963/ijcai.2021/628]

Wang K and Belongie S. 2010. Word spotting in the wild//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer: 591-604 [DOI: 10.1007/978-3-642-15549-9_43]

Wang K, Babenko B and Belongie S. 2011. End-to-end scene text recognition//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 1457-1464 [DOI: 10.1109/iccv.2011.6126402]

Wang T, Huang J Q, Zhang H W and Sun Q R. 2020a. Visual commonsense R-CNN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10757-10767 [DOI: 10.1109/cvpr42600.2020.01077]

Wang T W, Xie Z C, Li Z, Jin L W and Chen X L. 2019. Radical aggregation network for few-shot offline hand written Chinese character recognition. Pattern Recognition Letters, 125: 821-827 [DOI: 10.1016/j.patrec.2019.08.005]

Wang T W, Zhu Y Z, Jin L W, Luo C J, Chen X X, Wu Y Q, Wang Q Y and Cai M X. 2020b. Decoupled attention network for text recognition//The 34th AAAI Conference on Artificial Intelligence, AAAI 2020, the 32nd Inn+ovative Applications of Artificial Intelligence Conference, IAAI 2020, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020. New York, USA: AAAI: 12216-12224 [DOI: 10.1609/aaai.v34i07.6903]

Wang W C, Zhang J S, Du J, Wang Z R and Zhu Y X. 2018. DenseRAN for offline handwritten Chinese character recognition//Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition. Niagara Falls, USA: IEEE: 104-109 [DOI: 10.1109/icfhr-2018.2018.00027]

Wei X S, Song Y Z, Mac Aodha O, Wu J X, Peng Y X, Tang J H, Yang J and Belongie S. 2022. Fine-grained image analysis with deep learning: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 8927-8948 [DOI: 10.1109/TPAMI.2021.3126648]

Weiss K, Khoshgoftaar T M and Wang D D. 2016. A survey of transfer learning. Journal of Big Data, 3(1): #9 [DOI: 10.1186/s40537-016-0043-6]

Xia C Y, Yin W P, Feng Y H and Yu P. 2021. Incremental few-shot text classification with multi-round new classes: formulation, dataset and system//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [s.l.]: ACL: 1351-1360 [DOI: 10.18653/v1/2021.naacl-main.106]

Xie Z C, Huang Y X, Zhu Y Z, Jin L W, Liu Y L and Xie L L. 2019. Aggregation cross-entropy for sequence recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6538-6547 [DOI: 10.1109/cvpr.2019.00670]

Yang J K, Zhou K Y, Li Y X and Liu Z W. 2021. Generalized out-of-distribution detection: a survey [EB/OL]. [2023-01-11]. https://arxiv.org/pdf/2110.11334.pdf

Yang M K, Guan Y S, Liao M H, He X, Bian K G, Bai S, Yao C and Bai X. 2019. Symmetry-constrained rectification network for scene text recognition//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9146-

9155［DOI：10.1109/iccv.2019.00924］

Ye H J，Hu H X，Zhan D C and Sha F. 2020. Few-shot learning via embedding adaptation with set-to-set functions//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：8805-8814［DOI：10.1109/cvpr42600.2020.00883］

Yoshihashi R，Shao W，Kawakami R，You S，Iida M，and Naemura T. 2019. Classification-reconstruction learning for open-set recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：4016－4025［DOI：10.1109/CVPR.2019.00414］

Yu D L，Li X，Zhang C Q，Liu T，Han J Y，Liu J T and Ding E R. 2020. Towards accurate scene text recognition with semantic reasoning networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：12110-12119［DOI：10.1109/cvpr42600.2020.01213］

Yu H Y，Chen J Y，Li B，Ma J Q，Guan M N，Xu X X，Wang X C，Qu S B and Xue X Y. 2021. Benchmarking Chinese text recognition：datasets，baselines，and an empirical study［EB/OL］.［2021-12-30］. https://arxiv.org/pdf/2112.15093.pdf

Yu Y，Qu W Y，Li N and Guo Z M. 2017. Open category classification by adversarial sample generation//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne，Australia：IJCAI.org：3357-3363［DOI：10.24963/ijcai.2017/469］

Yuan T L，Zhu Z，Xu K，Li C J，Mu T J and Hu S M. 2019. A large Chinese text dataset in the wild. Journal of Computer Science and Technology，34(3)：509-521［DOI：10.1007/s11390-019-1923-y］

Yue Z Q，Zhang H W，Sun Q R and Hua X S. 2020. Interventional few-shot learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver，Canada：Curran Associates Inc.：#230

Zhang C H，Gupta A and Zisserman A. 2020a. Adaptive text recognition through visual matching//Proceedings of the 16th European Conference on Computer Vision. Glasgow，UK：Springer：51-67［DOI：10.1007/978-3-030-58517-4_4］

Zhang H and Ding H H. 2021. Prototypical matching and open set rejection for zero-shot semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：6954-6963［DOI：10.1109/ICCV48922.2021.00689］

Zhang H and Patel V M. 2017. Sparse representation-based open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence，39(8)：1690-1696［DOI：10.1109/TPAMI.2016.2613924］

Zhang H L，Xu H and Lin T E. 2021. Deep open intent classification with adaptive decision boundary//Proceedings of the 35th AAAI Conference on Artificial Intelligence，AAAI 2021，the 33rd Conference on Innovative Applications of Artificial Intelligence，IAAI 2021，the 11th Symposium on Educational Advances in Artificial Intelligence，EAAI 2021.［s.l.］：AAAI：14374-14382［DOI：10.1609/aaai.v35i16.17690］

Zhang J Q，Lertvittayakumjorn P and Guo Y K. 2019. Integrating semantic knowledge to tackle zero-shot text classification//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Minneapolis，USA：ACL：1031-1040［DOI：10.18653/v1/n19-1108］

Zhang J S，Du J and Dai L R. 2020b. Radical analysis network for learning hierarchies of Chinese characters. Pattern Recognition，103：#107305［DOI：10.1016/j.patcog.2020.107305］

Zhang J S，Zhu Y X，Du J and Dai L R. 2018. Radical analysis network for zero-shot learning in printed Chinese character recognition//Proceedings of 2018 IEEE International Conference on Multimedia and Expo. San Diego，USA：IEEE：1-6［DOI：10.1109/ICME.2018.8486456］

Zhang X Y，Liu C L and Suen C Y. 2020c. Towards robust pattern recognition：a review. Proceedings of the IEEE，108(6)：894-922［DOI：10.1109/jproc.2020.2989782］

Zhang Y S. 2021. A survey of unsupervised domain adaptation for visual recognition［EB/OL］.［2021-12-13］. https://arxiv.org/pdf/2112.06745.pdf

Zhou D W，Ye H J and Zhan D C. 2021. Learning placeholders for open-set recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：4401-4410［DOI：10.1109/cvpr46437.2021.00438］

Zhou Z H. 2022. Open-environment machine learning. National Science Review，9(8)：#123［DOI：10.1093/nsr/nwac123］

Zu X Y，Yu H Y，Li B and Xue X Y. 2022. Chinese character recognition with augmented character profile matching//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa，Portugal：ACM：6094-6102［DOI：10.1145/3503161.3547827］

作者简介

杨春，男，讲师，主要研究方向为模式识别、计算机视觉、文档分析与识别。E-mail：chunyang@ustb.edu.cn

殷绪成，通信作者，男，教授，主要研究方向为模式识别与计算机视觉、文字识别（文档图像分析与识别）、信息检索与自然语言处理、人工智能芯片技术及应用。

E-mail：xuchengyin@ustb.edu.cn

刘畅，男，博士研究生，主要研究方向为小样本学习、文本识别和文本检测。E-mail：lasercat@gmx.us

方治屿，男，博士研究生，主要研究方向为零样本学习和知识图谱。E-mail：mr.fangzy@foxmail.com

韩铮，男，博士研究生，主要研究方向为小样本学习、域适应和强化学习。E-mail：han970421@163.com

刘成林，男，研究员，主要研究方向为模式识别理论与方法、文字识别与文档分析。E-mail：chenglin.liu@ia.ac.cn