arXiv:2601.01088v1 [cs.CV] 3 Jan 2026

600K-KS-OCR：面向克什米尔文字光学字符识别的大规模合成数据集

Haq NawazMalik独立研究者
orcid.org/0009-0003-1994-7640
huggingface.co/Omarrran
x.com/HAQ_NAWAZ_MALIK

2026年1月6日

摘要

本技术报告介绍了**600K-KS-OCR 数据集**，这是一个大规模合成语料库，包含约 602,000 张按词语分割的图像，旨在训练和评估面向克什米尔文字的光学字符识别系统。该数据集填补了克什米尔语领域的一项关键资源空缺。克什米尔语是一种濒危的达尔德语支语言，采用经过改良的波斯-阿拉伯书写系统，约有七百万人使用。每张图像均以 256×64 像素的分辨率渲染，并提供多种格式的对应真实标注转写，可兼容 CRNN、TrOCR 及通用机器学习流程。生成方法采用三种传统克什米尔字体，并通过全面的数据增强模拟真实文档退化，同时加入多样的背景纹理，以增强模型的鲁棒性。该数据集分为十个存档，总计约 10.6 吉字节，并依据 CC-BY-4.0 许可协议发布，以促进低资源语言光学字符识别研究。

关键词 光学字符识别 · 克什米尔文字 · 低资源语言 · 合成数据集 · 深度学习 · 波斯-阿拉伯文字

1 引言

光学字符识别（OCR）系统在拥有大量带标注训练数据的高资源语言上已经取得了显著的准确率。然而，对于标注数据集稀缺或不存在的低资源语言，其性能会大幅下降。对于采用复杂书法文字且现有计算基础设施有限的语言而言，这种差距尤其明显。

克什米尔语是克什米尔山谷的主要语言，集中体现了这些挑战。作为印度-雅利安语族中的一种达尔德语，克什米尔语使用一种经过改造的波斯-阿拉伯文字，其中包含额外字符，用于表示阿拉伯语和波斯语中不存在的音位 [1]。尽管克什米尔语的使用者约有七百万人，但在自然语言处理和计算机视觉研究中，它仍然严重缺乏代表性。

开发适用于克什米尔文字的高效 OCR 系统对于多种应用至关重要：将历史手稿和行政记录数字化、保护文学遗产，以及实现基于文本的无障碍技术。然而，大规模带标注数据集的缺失阻碍了这一领域的发展。

本技术报告介绍了 600K-KS-OCR 数据集，这是一个合成语料库，包含约 602,000 张按词级切分的图像及其对应的真实标注转写。该数据集旨在实现以下目标：

1. **规模**：为需要大量样本才能实现稳健泛化的深度学习架构提供充足的训练数据。

2. **真实性**：采用传统克什米尔字体，准确呈现该文字的书法特征。 3. **鲁棒性**：加入全面的数据增强，以模拟真实文档条件，包括噪声、模糊、几何畸变和纸张劣化。 4. **易用性**：以多种格式分发数据集，使其兼容常用的 OCR 训练框架。

本报告其余部分安排如下。第2 节介绍克什米尔文字的特点及相关研究背景。第3节详细说明数据集规格和架构。第4 节描述生成方法。第5 节说明数据格式约定。第6节讨论预期应用。第8节最后介绍数据集的获取信息。

2 背景与相关研究

2.1 克什米尔文字的特点

克什米尔语书写系统在计算机识别方面面临若干挑战：

波斯-阿拉伯文字基础。克什米尔文字源自波斯-阿拉伯字母表，但为表示该语言特有的发音，增加了额外字符。这些字符包括现有字母的变体，以及带有独特附加符号的全新字素。

从右到左书写方向。 文本从右向左流动，因此需要在图像渲染和序列建模架构中进行适当处理。

字母上下文形式。 字符会根据其在单词中的位置（词首、词中、词尾或独立形式）呈现不同形状，这是阿拉伯语和波斯语文字所共有的特征。

复杂附加符号。 克什米尔语采用了复杂的附加符号系统（元音符号、沙达符号、静音符号），用于改变发音和含义。这些符号位于基本字符的上方或下方，必须得到准确渲染和识别。

连字与连接。 相邻字符在书法传统中会彼此连接，形成连字，从而改变字母序列的视觉外观。

2.2 相关工作

波斯-阿拉伯文字的OCR研究主要集中于阿拉伯语、波斯语和乌尔都语。值得注意的数据集包括用于阿拉伯语手写体的IFN/ENIT数据库 [2], 、用于乌尔都语印刷文本的UPTI数据集 [3], 以及用于阿拉伯语手写文本的KHATT数据库 [4]。

对于克什米尔语而言，计算资源仍极为有限。此前的研究已涉及孤立字符识别 [5]，但一直缺乏适用于端到端 OCR 训练的大规模分词数据集。

合成数据生成已证明能够有效扩充 OCR 训练语料库。SynthText [6] 展示了渲染文本图像在场景文本识别中的实用价值。MJSynth 数据集 [7] 提供了数百万张合成词图像，助力基准数据集上的性能取得突破。此项工作将合成数据范式拓展至克什米尔文字。

3 数据集规格

3.1 概述

表1概述了 600K-KS-OCR 数据集的主要特征。

3.2 分区结构

该数据集分布在十个分区存档中，便于下载和模块化使用。表2列出了各分区的规格。

表1：数据集概览

属性	规格
样本总数	~602,000 个单词
档案分发	10个ZIP文件
总大小	~10.6 吉字节
图像尺寸	256×64 像素
图像格式	PNG（无损）
文本方向	从右到左（RTL）
文字系统	克什米尔语（波斯-阿拉伯字母）
输出格式	CRNN、TrOCR、CSV、JSONL
许可证	CC-BY-4.0

表2：数据集分区

分区	样本	Size
P1_OCR_数据集	50,000	~904 MB
P2_OCR_数据集	53,815	~953 MB
P3_OCR_数据集	68,741	~1.2 吉字节
P4_OCR_数据集	69,886	~1.2 吉字节
P5_OCR_数据集	69,637	~1.2 吉字节
P6_OCR_数据集	69,506	~1.2 吉字节
P7_OCR_数据集	58,228	~1.1 吉字节
P8_OCR_数据集	35,720	~620 MB
P9_OCR_数据集	86,635	~1.5 吉字节
P10_OCR_数据集	41,401	~732 MB
总计	**~602,000**	**~10.6 吉字节**

3.3 归档内容

每个分区归档均采用一致的内部结构：

清单 1：归档目录结构

```
P*_光学字符识别_数据集_ */
+-- 图像 /                  # 按词分割的 PNG 图像
+-- 数据 .csv               # 表格格式（文件名，                文本 ）
+-- data. jsonl             # 用于 TrOCR 的 JSON Lines   管线
+-- labels .txt             # 制表符分隔的 CRNN 格式
+-- 元数据.json             # 生成配置
```

这种组织方式使研究人员能够选择偏好的标签格式，同时在不同格式之间保持一致的图像引用。

4 生成方法

4.1 渲染配置

所有图像均使用标准化参数进行渲染，这些参数针对卷积循环神经网络（CRNN）和基于 Transformer 的 OCR 架构进行了优化：

该 256×64 像素尺寸是在分辨率质量与计算效率之间权衡后选定的，代表了词级 OCR 系统的一种常见配置 [8]。

4.2 排版

文本渲染　g 使用了三种字体，分别体现传统克什米尔语和纳斯塔利克书法风格：

表3：渲染参数

参数	值
图像尺寸	256×64 像素
背景颜色	#FFFFFF（白色）
文本颜色	#000000（黑色）
文本方向	从右到左
背景模式	混合（多种纹理）
色彩空间	RGB

Afan Koshur Naksh。 一种克什米尔语本土纳斯赫体字体，字形清晰易读，适用于印刷材料。这种字体在保持克什米尔语字形特色的同时，着重提升了易读性。

纳斯塔利克体。 一种经典的书法字体，呈现出正式的克什米尔文学和宗教文本所特有的对角基线和流畅连接效果。

Nakash (Narqalam)。 一种融入传统手写特征的字体，能够模拟手稿和非正式文档的条件，为纯排版字体所缺乏的自然变化提供支持。
使用多种字体能够增强模型在真实世界的克什米尔语文档所呈现的风格多样性中的泛化能力。

4.3 数据增强流程

为增强模型应对真实世界图像退化的鲁棒性，我们对生成的样本中 60% 应用了全面的数据增强流程，其余 40% 则保留为干净的基线图像。数据增强类别包括：

几何变换。 旋转（±5°）、透视畸变、倾斜和倾角变化，用于模拟扫描错位和文档定位伪影。

模糊效果。 高斯模糊和运动模糊，用于模拟图像采集过程中的对焦变化和移动。

噪声注入。 高斯噪声和椒盐噪声，用于模拟低质量数字化设备和传感器噪声的典型特征。

光度变化。 通过调整亮度和对比度、加入 JPEG 压缩伪影以及降低分辨率，模拟多种采集条件。

文档特定效果。 纸张纹理叠加、阴影和光照渐变，以及老旧或劣化源材料中常见的墨水渗化现象。

4.4 背景合成

混合背景模式融入多种纹理，以提升模型在洁白背景之外的泛化能力：

- **干净背景**：纯白色，以获得最佳基线条件
- **老旧文档风格**：旧纸、古籍书页、羊皮纸
- **文档变体**：笔记本纸、印刷书页、报纸纹理
- **纸张色调**：象牙色、奶油色、再生纸外观
- **做旧效果**：咖啡渍、水渍损坏、风化表面
- **自定义纹理**：另外十四种背景变体

这种背景多样性缓解了合成训练数据与异构真实世界文档之间的领域偏移。

4.5 生成基础设施

数据集生成采用跨四个计算核心的GPU加速并行处理，相较于顺序生成，性能提升了3-10×倍。基于GPU的数据增强能够高效地大规模应用复杂变换。

5 数据格式规范

数据集　提供四种格式的标签，以适配多种训练框架和工作流程。

5.1 CRNN格式（labels.txt）

遵循基于连接时序分类（CTC）的序列识别标准约定的制表符分隔格式：

列表 2：CRNN标签格式

```
图像_001.png    克什米尔语文本
图像_002.png    克什米尔语文本
图像_003.png    克什米尔语文本
```

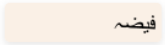
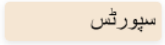
图像	
The following table presents representative samples from the dataset, demonstrating the variety of word forms and rendering styles:	
Image	Transcript
	پہلگام
	چھ
	فُٹ
	ہال
	فیضہ
	خانَن
	کوٗر
	سپورٹس
	گراونڈس
	رِیاَستی

表4：克什米尔语光学字符识别数据集样本

5.2 TrOCRFormat（data.jsonl）

JSON Lines 格式可直接与 Hugging Face Transformers 管线集成，用于 TrOCR 及类似的编码器-解码器架构：

清单 3：TrOCR 标签格式

```
{" file_name ":  " 图像_001.png ",  " 文本 ":  "克什米尔语文本 "}
{" file_name ":  " image_002.png ",  " text ":  "克什米尔语文本 "}
{" file_name ":  " 图像_003.png ",  " 文本 ":  "克什米尔语文本 "}
```

5.3 表格格式 (data.csv)

标准 CSV 格式与通用数据处理工具兼容：

清单 4：CSV标签格式

```
文件名 ,text
图像_001.png，克什米尔语文本
图像_002.png，克什米尔语文本
图像_003.png，克什米尔语文本
```

5.4 元数据模式

每个档案都包含详尽的元数据，用于记录生成参数：

清单 5：元数据 JSON 架构

```
{
  " 生成_时间 ":  "2025 -12 -26 T16 :24:34.803 Z",
  " 配置 ":  {
    " 图像_大小 ":  "256 x64 ",
    "数据增强_已启用": true ,
    " 数据增强_百分比": 60,
    " 字体 _使用":  [
      " 阿凡语 _ 克什米尔语 _ 纳克什体 " ,
      " 纳斯塔利克体 ",
      " Nakash "
    ] ,
    " 输出_格式 ":  [" CRNN ", " TrOCR ", "csv", " JSONL "]
  },
  " 样本 ":  50000,
  "干净_样本 ":  20000,
  "增强_样本 ": 30000
}
```

6 应用与影响

6.1 主要应用

600K-KS-OCR 数据集支持多种研究与开发应用：

OCR 模型训练。 该数据集规模充足，可用于从头训练 CRNN 等深度学习架构[8], TrOCR [9], 和基于注意力的序列到序列模型，也可用于对这些模型进行微调。

基准测试。 标准化数据划分和格式规范可实现对克什米尔文字 OCR 系统的一致评估，有助于开展可复现的研究比较。

迁移学习。 在该数据集上预训练的模型可迁移至相关的波斯-阿拉伯文字，也可作为面向特定领域的克什米尔语光学字符识别应用的初始化模型。

文献数字化。 经过训练的模型投入实际应用后，可将目前无法以机器可读形式获取的历史手稿、报纸、行政记录和文学档案数字化。

6.2 更广泛的影响

开发强大的克什米尔语光学字符识别能力，有助于保护这种濒危语言。文本遗产的数字化可实现：

- 对逐渐 deteriorating 的纸本文献进行档案保存
- 对数字化馆藏进行全文检索
- 为视障读者提供无障碍技术
- 用于自然语言处理研究的语料库构建

6.3 使用示例

该数据集可直接与 Hugging Face 生态系统集成：

清单 6：通过 Hugging Face Datasets 加载

```
从 datasets 导入 load_dataset

数据集 = load_dataset (
    " Omarrran /600 k_KS_光学字符识别_词语_切分_数据集 "
)
```

手动加载：

代码清单 7：手动加载数据

```
import pandas as pd
from PIL import Image

标签_df = pd . read_csv (" data . csv ")
img =图像 . open (" images / image _001. png ")
文本 =标签_df . loc [
    标签_df [' 文件名 '] == ' 图像 _001. png ',
    ' 文本 '
]。值 [0]
```

7 局限性

应当承认，该数据集存在若干局限性：

合成性质。 作为合成生成的语料库，该数据集可能无法充分呈现真实手写文档或劣化历史文档的多样性。在可获得真实文档样本时，应通过领域自适应或微调来验证模型在真实数据上的表现。

词级分割。 该数据集提供了预先分割的词图像。需要进行行级或页级识别的系统还需要额外的数据集，或针对这些粒度进行合成生成。

字体覆盖范围。 虽然三种字体提供了风格多样性，但增加字体，尤其是能够体现地域书法变体的字体，将进一步提升泛化能力。

词汇分布。 词汇分布反映了用于渲染的源文本语料库。技术、医学和法律等专业领域可能需要补充数据。

8 结论

本技术报告介绍了 600K-KS-OCR 数据集，这是一个用于克什米尔语光学字符识别的大规模合成语料库，包含约 602,000 张带有真实标注转写的词分割图像。该数据集填补了克什米尔语在计算机视觉研究中作为一种代表性不足的语言所面临的关键资源缺口。

该数据集的主要特点包括：使用传统克什米尔字体渲染的标准化 256×64 像素图像、模拟真实文档劣化情况的全面数据增强、多样化的背景纹理，以及以多种格式发布，可兼容主流光学字符识别训练框架。

该数据集采用 CC-BY-4.0 许可协议发布，以促进低资源语言光学字符识别研究。未来可通过增加字体、扩大词汇覆盖范围，以及进行行级或页面级合成生成来扩展这一资源。

数据可用性

600K-KS-OCR 数据集可通过 Hugging Face 数据集中心获取：

https://huggingface.co/datasets/Omarrran/600k_KS_光学字符识别_单词_分割_数据集

该数据集依据 CC-BY-4.0 许可协议发布，允许在注明出处的情况下使用 .

致谢

作者感谢克什米尔语社区提供语言学指导，并感谢本数据集中使用的开源字体的开发者。

参考文献

[1] Omkar N. Koul 和 Kashi Wali。《现代克什米尔语语法》。Dunwoody Press，2006 年。[2] Mario Pechwitz、S. Snoussi Maddouri、Volker Märgner、Noureddine Ellouze 和 Haikal Amiri。IFN/ENIT——阿拉伯语手写词数据库。载于*CIFED*论文集，第 2 卷，第 127–136 页，2002 年。[3] Nazly Sabbour 和 Faisal Shafait。乌尔都语印刷文本识别数据库。载于文档分析与识别（*ICDAR*），第 1339–1343 页。IEEE，2013 年。[4] Sabri A. Mahmoud、Irfan Ahmad、Wasfi G. Al-Khatib、Mohammad Alshayeb、Mohammad Tanvir Parvez、Volker Märgner 和 Gernot A. Fink。KHATT：开放式阿拉伯语离线手写文本数据库。模式识别，47(3)：1096–1112，2014 年。[5] Sheeraz Ahmad 和 Syed Arsalan。孤立克什米尔语字符识别。载于国际计算与通信技术会议，第 1–5 页，2017 年。[6] Ankush Gupta、Andrea Vedaldi 和 Andrew Zisserman。用于自然图像文本定位的合成数据。载于*IEEE*计算机视觉与模式识别会议论文集，第 2315–2324 页，2016 年。[7] Max Jaderberg、Karen Simonyan、Andrea Vedaldi 和 Andrew Zisserman。用于自然场景文本识别的合成数据与人工神经网络。载于*NIPS*深度学习研讨会，2014 年。[8] Baoguang Shi、Xiang Bai 和 Cong Yao。一种用于基于图像的序列识别及其在场景文本识别中应用的端到端可训练神经网络。*IEEE*模式分析与机器智能汇刊，39(11)：2298–2304，2017 年。[9] Minghao Li、Tengchao Lv、Jingye Chen、Lei Cui、Yijuan Lu、Dinei Florencio、Cha Zhang、Zhoujun Li 和 Furu Wei。TrOCR：基于预训练模型的基于Transformer的光学字符识别。载于*AAAI*人工智能会议论文集，第 37 卷，第 13094–13102 页，2023 年。