2012 年以来 OCR 前沿论文进展

2018.11.04 方建勇，苏镌，邹博（问题来自两位讨论）

提示：采用手机 safari 微软翻译技术

1. **在对工作：了解工人在人群工作中的交互作用**

作者:何建菊,尹明

摘要: 众包作为一种工具，帮助解决计算机难以解决的问题，越来越受欢迎。以前在众包方面的工作往往假定工人独立完成众包工作。本文放宽了人群工作的独立性，探讨了引入工人之间直接、同步、自由的互动对众筹的影响。特别是，在教育环境中的同伴教学概念的推动下，我们研究了同伴交流在众包环境中的作用。在具有同行交流的众包环境中，要求一对工人一起完成同样的任务，首先独立地生成任务的初始答案，然后自由地相互讨论任务并更新答案讨论结束后。我们通过实验研究了众影中的同伴交流对众包平台上各种常见任务类型的影响，包括图像标记、光学字符识别（ocr）、音频转录和营养分析。我们的实验结果表明，与工人独立完成工作的任务相比，同行沟通任务的工作质量显著提高。然而，参与同行沟通的任务对影响工人未来在同类任务中的独立表现的影响有限。少

2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

2. **利用开源引擎对 19 世纪弗拉克图尔脚本的艺术光学人物识别现状**

   作者:christian reul, uwe springmann, christoph wick, frank pupe

   摘要: 在本文中，我们评估光学字符识别（**ocr**）19 世纪的 fraktur 脚本没有书籍特定的培训使用混合模型，即模型训练识别各种字体和排版从以前看不见的来源。我们描述了导致强大混合 **ocr** 模型的培训过程，并将其与流行的开源发动机 OCRopus 和 Tesseract 的免费模型以及 abbyy 最先进系统的商业状态进行比较。为了进行评估，我们使用了来自 19 世纪的书籍、期刊和字典中的各种看不见的数据。实验表明，用真实数据训练混合模型优于合成数据训练，新型 **ocr** 发动机卡拉马里的性能大大优于其他发动机，平均降低 abbyys 字符错误率（cer)70%，导致平均核证的排减量低于 1%。少

   2018 年 10 月 8 日提交;最初宣布 2018 年 10 月。

3. **在德国弗拉克图尔和早期现代拉丁语的历史文献中训练 ocr 引擎的地面真理**

   作者:uwe springmann, christian reul , stefanie dipper, johnes baiter

摘要: 本文以打印文本行图像及其转录的形式描述了历史 **ocr** 的德语和拉丁语 (:textumtument{地面事实}) 数据集。此数据集称为 "纹理 {g4 组}", 由 313 173 行对组成, 涵盖了从 15 世纪到 19 世纪以 fraktur 类型印刷的书籍的广泛打印日期, 并在 ccby 4.0 许可证下公开提供。gt 作为线映像/转录对的特殊形式使其直接用于在 lstm 架构 (如 Tesseract 4 或 OCRopus) 中使用重复神经网络的 **ocr** 软件的应用, 从而直接使用它。我们还提供了一些预先训练的 ocropus 模型, 用于在 95% (早期打印) 和 98%--(19 世纪 fraktur 打印) 之间的数据集的数据集的字符精度在看不见的测试用例中的亚库, 以及一个 perl 脚本来协调不同测试用例生成的 gt 转录规则, 并给出如何构造 gt 的 ocr 目的的提示, 其要求可能不同于语言动机的转录。少

2018 年 9 月 14 日提交;最初宣布 2018 年 9 月。

4. **重新循环您的 ocr: 在罗马化梵文中重复使用 ocr 后文本校正 ocr**

   作 者 :amrith krishna, bodhistwa prasad majumder, rajesh shreedhar bhat, pawan goyal

   文摘: 我们提出了**一种后 ocr** 文本校正方法, 用于将罗马化梵文中的文本数字化。由于缺乏资源, 我们的方法使用用罗马语言编写的**其他语言训练的 ocr** 模型。目前, 没有关于罗马化梵文 **ocr 的**数据集。因此, 我们引导 430 张图像的数据集, 在两个不同的设

置中进行扫描，并提供相应的地面事实。对于训练，我们为这两个设置合成训练图像。我们发现，使用复制机制（gu 等人，2016 年）在字符识别速率（crr）中，比目前解决单调序列序列任务的最先进模型增加 7.69 的百分比（schnober 等人，2016 年）。我们发现，我们的系统在消除 **ocr** 易发错误方面是稳健的，因为对于其中一个数据集设置，我们的系统从 **ocr** 输出中获得了 800.01% 的 crr，crr 为 35.76%。对这些模型进行的人类判断调查显示，我们提出的模型产生的预测比其他系统更快理解，更快地改善人类。少

2018 年 9 月 6 日提交;**最初宣布** 2018 年 9 月。

5. **cg-diqa：基于字符梯度的无参考文档图像质量评估**

作者:李宏宇,范珠,邱俊华

摘要：文档图像质量评估（diqa）是实际应用中一个重要且具有挑战性的问题。为了预测文档图像的质量分数，本文提出了一种新的基于字符梯度的无参考 diqa 方法，将 **ocr** 精度作为地面真理质量度量。在基于最大稳定极值区域（mser）的方法检测到的字符补丁上计算字符梯度。字符补丁对字符识别具有重要意义，因此适用于估计文档图像质量。在基准数据集上的实验表明，该方法在估计文档图像质量评分方面优于最先进的方法。少

2018 年 7 月 11 日提交;**最初宣布** 2018 年 7 月。

6. **卡拉马里–一种基于光字符识别的高性能张力图深度学习包**

作者:christoph wikick, christian reul, frank pupe

文摘: 当代和历史数据的光学字符识别（**ocr**）仍然是许多研究人员关注的焦点。特别是历史版画需要图书特定的**培训 ocr** 模型才能实现适用的结果（springmann 和 lüdeling, 2016 年, reul 等人, 2017 年 a)。为了减少人为地注释地面真相（gt）的努力，投票和预培训等各种技术已证明是非常有效的（reul 等人, 2018a, reul 等人, 2018a)。卡拉马里是一种新的开源 **ocr** 线路识别软件，它既使用最新的深层神经网络（dnn）在 tensorflow 中实现，又为预训练和投票等技术提供本机支持。由卷积神经网络（cnns）和长短内存（lstm）层构建的可定制网络架构由格雷夫斯等人（2006 年）的所谓连接器时间分类（ctc）算法进行训练。gpu 的可选使用大大减少了训练和预测的计算时间。我们使用两个不同的数据集来比较卡拉马里与 ocrapy、OCRopus3 和 Tesseract 4 的性能。卡拉马里在用现代英语编写的 uw3 数据集上的字符错误率（cer）为 0.11，在用德语 fraktur 编写的 dta19 数据集上的字符错误率（cer）达到 0.11，这大大优于现有软件的结果。少

2018 年 8 月 6 日提交;**v1** 于 2018 年 7 月 5 日提交;**最初宣布** 2018 年 7 月。

7. **通过前安特先知不等式的最佳在线争用解决方案**

作者:euiwoong lee, sahalil sinla

抽象: 联机争用解决方案（ocrss）是由 feldman、svensson 和 zenklusen 提出的一种通用技术，用于在线方式在 matroid 多边形中舍入分数解决方案。它发现了在几个存在承诺约束的随机组合问题中的应用: 在看到一个随机元素的值时，算法必须立即和不可撤销地决定是否在始终保持一个独立设置中的 matroid。尽管 ocrs 立即导致先知的不平等，但这些先知的不平等并不是最优的。我们能否转而使用先知的不等式来设计最佳的 ocrs？我们设计了第一个最佳 1 / 2-**ocrs 的成熟，通过**减少问题，以设计一个成熟的先知不平等，我们比较了一个更强的基准，事先放松。我们还介绍和设计了最佳(1--1 / e)-随机顺序 crs 的拟阵，这是类似于 ocrs，但到达是随机的均匀选择。少

2018 年 6 月 24 日提交;**最初宣布** 2018 年 6 月。

8. **无氧学 ocr 的置信度预测**

作者:noam mor , lior wolf

摘要: 拥有可靠的精度分数对于 **ocr** 的实际应用至关重要，因为这样的系统是根据错误读数的数量来判断的。基于词汇的 **ocr** 系统处理本质上是多类分类问题，通常使用明确考虑词汇的方法，以提高准确性。但是，在无词典的方案中，筛选错误需要显式的置

信度计算。在本工作中，我们展示了两种明确的置信度测量技术，并表明它们能够显著减少标准基准和专有数据集上的误读。少

2018 年 5 月 28 日提交;最初宣布 2018 年 5 月。

9. **ocr lstm 中的内隐语言模型**

   作者:ekraam sabir, stephen rawls, prem natarajan

   摘要: 神经网络已成为 **ocr** 的首选技术，但它们如何以及为什么能够提供卓越性能的许多方面仍是未知数。使用 lstm 的当前神经网络技术与以前最先进的 hmm 系统之间的一个关键区别是，hmm 系统具有很强的独立性假设。相比之下，lstm 对解码过程中可以考虑的上下文量没有明确的约束。在本文中，我们展示了他们学习隐式 lm，并试图用等效的 n-gram 上下文来描述 lm 的强度。我们表明，这种隐式学习的语言模型与一组随机字符（即不自然发生的序列）相比，对我们的合成测试集提供了 2.4%--cer 改进，并且 lstm 学习使用多达 5 个字符的上下文（这大约是 88 帧在我们的配置）。我们相信，这是首次尝试在基于 lstm 的 **ocr** 系统中表征隐式 lm 的强度。少

   2018 年 5 月 23 日提交;最初宣布 2018 年 5 月。

10. **感知文本：一种新的具有可变形 psroi 池的感知文本模块，用于面向多方向的场景文本检测**

作者:杨强鹏,程孟丽，周文蒙,陈燕,邱明辉,林伟,朱伟

摘要: 在许多计算机视觉应用中，事件场景文本检测，尤其是对于多面向文本区域，是最具挑战性的任务之一。与常见的对象检测任务不同，场景文本在纵横比、比例和方向等方面往往存在较大的差异。为了解决这个问题，我们从实例感知分割的角度提出了一种新的端到端场景文本检测器感知文本。我们设计了一个新的感知文本模块，并引入了可变形的 psroi 池处理多面向文本检测。在 icdar2015、rctw-17 和 msra-td500 数据集上进行的广泛实验证明了我们的方法在有效性和效率方面的优越性。我们提出的方法在 icdar2015 挑战和其他数据集上的最先进性能方面取得了第一名的成果。此外，我们还发布了我们的实施作为 **ocr** 产品，可供公众访问。少

2018 年 5 月 7 日提交;**v1** 于 2018 年 5 月 3 日提交;**最初宣布** 2018 年 5 月。

11. **利用长期记忆 rnn-lstm 对递归神经网络进行快速识别的法语词汇识别**

作者:saman sarraf

文摘: 光学字符识别（**ocr**）是计算机视觉中的一个基本问题。研究表明，在使用深度学习方法和拓扑结构对印刷字符进行分类方面取得了重大进展。在目前的算法中，具有称为 mn-lstm 的长

期记忆块的递归神经网络在准确率方面提供了最高的性能。使用从互联网上收集到的包括所有符号和口音在内的最顶尖的 5,000 个法语单词，对 mn-lstm 模型进行了培训和测试，用于几个案例。使用了六种字体来生成 **ocr** 示例，并准备了一个包含这六种字体中所有示例的附加数据集，用于培训和测试。对经过训练的 mn-lstm 模型进行了测试，分别实现了 99.98798 和 99.98798 的编辑距离和序列误差的准确率。准确的预处理，然后是高度归一化（深度学习中的标准化方法），使 mn-lstm 模型能够以最有效的方式进行训练。这项机器学习工作还揭示了 mn-lstm 拓扑识别打印字符的鲁棒性。少

2018 年 4 月 10 日提交;最初宣布 2018 年 4 月。

12. **感知中国移民：移动应用如何提供对全球移民流动的洞察**

作者:薛敏辉,格里戈拉斯,希瑟·李,基思·罗斯

摘要: 如今，许多国家都有 "以国家为中心的移动应用"，这些应用是主要由特定国家的居民使用的移动应用。其中许多以国家为中心的应用还包括一项基于位置的服务，该服务利用智能手机对智能手机当前 gps 位置的 api 访问。在本文中，我们研究如何利用这种以位置为基础的服务的以国家为中心的应用程序来研究与族裔和文化群体相关的散居国外者。我们的方法结合了 gps 黑客、手机自动任务工具和 **ocr**，为散居国外者生成迁移统计数据。

作为一个案例研究，我们将我们的方法应用到微信上，微信是中国国内和世界华裔中非常受欢迎的应用。利用微信，我们收集 32 个城市的华侨数据。我们还使用 google 地点 api 收集每个城市的中国企业的数据。这些综合数据为现代华侨及其近年来的变化提供了有趣的见解。少

2018 年 3 月 22 日提交;**v1** 于 2018 年 3 月 22 日提交;**最初宣布** 2018 年 3 月。

13. **野生的中文文本**

作者:袁泰玲,朱哲,徐坤,李成军，胡世敏

摘要: 我们在野外介绍中文文本，这是一个非常大的中文文本数据集在街景图像。虽然文档图像中的光学字符识别（**ocr**）研究得很好，而且有许多商业工具，但自然图像中文本的检测和识别仍然是一个具有挑战性的问题，特别是对于更复杂的字符集如中文文本。缺乏培训数据一直是一个问题，特别是对于需要大量培训数据的深度学习方法。本文提供了一个新创建的中文文本数据集的详细情况，该数据集由专家在 3 万多张街景图像中注释了约 100 万个汉字。这是一个具有挑战性的数据集，具有良好的多样性。它包含平面文本、凸起文本、城市文本、农村文本、光照差的文本、远文本、部分遮挡文本等。对于数据集中的每个字符，批注包括其基础字符、边界框和 6 个属性。这些属性指示它是否具有复

杂的背景、是否引发、是手写的还是打印的，等等。此数据集的大尺寸和多样性使其适合于训练用于各种任务（特别是检测和识别）的健壮神经网络。我们使用几个最先进的网络给出基线结果，包括 alesnet、overfeat、google 初始和 resnet 用于字符识别，以及 YOLOv2 用于图像中的字符检测。总体而言，google 宗称在识别方面具有最佳性能，排名第一的精度为 80.5，而 YOLOv2 在检测方面的 map 为 71.0%。数据集、源代码和训练有素的模型都将在网站上公开提供。少

2018 年 2 月 28 日提交;最初宣布 2018 年 3 月。

14. **通过预培训、投票和主动学习相结合，提高早期印刷书籍的 ocr 准确性**

作 者 :christian reul, uwe springmann, christoph wick, frank pupe

文摘: 我们结合了三种方法，显著提高了早期印刷书籍**训练的 ocr** 模型的 **ocr** 精度: (1) 预训练方法利用了存储在现有模型中的信息，这些信息在各种排版中进行了训练（混合模型），而不是从头开始培训。(2) 使用单一 ocr 引擎 (ocropus) 对一组地面真相数据（线图像及其转录）进行交叉折叠训练，产生一个委员会，其成员随后还通过采用前 n 种替代品投票选出最佳结果,他们的内在信心价值。(3) 按照最大分歧的原则，我们选择选民最不同意的额外培训线路，期望他们为随后的培训（主动学习）提供最高的

信息收益。对六本早期印刷书籍的评价得出了以下结果：平均而言，预训练和投票的结合，在从同一混合模型开始训练五折时，性格准确率提高了 46%。在使用不同的预培训模式时，这一数字上升到 53%，这凸显了不同选民的重要性。纳入主动学习使获得的结果平均增加了 16%（对六本书中的三本进行了评价）。总体而言，拟议的方法仅在 60 行进行训练时，平均错误率为 2.5%。使用 1，000 行的大量地面真相池，使错误率进一步下降，平均低于 1%。少

2018 年 2 月 28 日提交;**v1** 于 2018 年 2 月 27 日提交;**最初宣布** 2018 年 2 月。

15. 利用深卷网提高早期印刷图书的 ocr 精度

作者:christoph wikick, christian reul, frank pupe

文摘：本文提出了一种卷积网络和 lstm 网络相结合的方法，以提高早期印刷图书 ocr 的精度。虽然基于线的 **ocr 的**标准模型使用单个 lstm 图层，但我们在 lstm 层之前使用 cnn 和池层组合。由于可训练参数的数量较多，网络的性能依赖于大量的训练示例来释放其功能。因此，误差减少了 44%，导致 cer 为 1% 及以下。为了进一步改进结果，我们使用投票机制来实现下面的字符错误率（cer)0。5.用于训练和预测书籍的深层模型的运行时与浅层网络非常相似。少

2018 年 2 月 27 日提交;最初宣布 2018 年 2 月。

16. 具有对抗性文本图像的愚弄 ocr 系统

作者:宋从正,维塔利·什马季科夫

摘要: 我们证明，基于深度学习的最先进的光学字符识别（**ocr**）容易受到对抗性图像的攻击。对印刷文本图像的微小修改不会改变文本对人类读者的含义，导致 **ocr** 系统 "识别" 不同的文本，在这种文本中，对手选择的某些词被其语义对立面所取代。这完全改变了 **ocr** 系统和使用 **ocr** 对其输入进行预处理的 nlp 应用程序所产生的输出的含义。少

2018 年 2 月 14 日提交;最初宣布 2018 年 2 月。

17. e2e-mlt-一种不受约束的多语言场景文本端到端方法

作者:yash patel, mashal buš ta, jiri matas

摘要: 提出了一种多语言场景文本定位、识别和脚本识别的端到端方法。该方法基于一组卷积神经网络。该方法称为 e2e-mlt，在自然图像和裁剪字脚本识别中实现了最先进的联合定位和脚本识别性能。e2e-mlt 是第一个发布的用于场景文本的多语言 **ocr** 。实验表明，获得准确的多语言多脚本注释是一个具有挑战性的问题。少

2018 年 1 月 30 日提交;最初宣布 2018 年 1 月。

18. **智能手机屏幕截图中的文本提取和检索：构建媒体中的生活存储库**

作者:agnese chiatti, mu jung cho, anupriya gagneja, xiang, miriam brinberg, katie roehrick, sagnik ray choudhury, nilam ram, byron reeves, c. 李·贾尔斯

摘要: 日常参与生活体验越来越多地与移动设备的使用交织在一起。在行为研究中使用秒级的屏幕捕捉，并实施 "及时" 的卫生干预措施。数字信息的心理广度的不断增加将继续使人们看到的实际屏幕成为生活经历的首选甚至需要的数据来源。有效和高效的信息提取和数字屏幕截图检索是成功使用屏幕数据的关键先决条件。在本文中，我们提出了实验工作流程，我们利用: (i) 预处理一个独特的屏幕截图集合, (ii) 提取嵌入在图像中的非结构化文本, (iii) 基于结构化模式组织图像文本和元数据, (iv) 索引生成的文档集合，以及 (v) 允许通过专用的垂直搜索引擎应用程序进行图像检索。所采用的程序集成了传统图像处理、光学字符识别 (**ocr**) 和图像检索的不同开源库。我们的目标是评估是否以及如何将最先进的方法应用于这一新颖的数据集。我们展示了如何将基于 opencv 的预处理模块与基于长期短期内存 (lstm) 的 Tesseract **ocr** 版本相结合，而无需经过临时培训，从而获得了 74% 的字符级精度。此外，我们还将处理后的存储库作为专用图

像检索系统的基线，用于行为和预防科学家的即时使用和应用。我们讨论了文本信息提取和检索的问题，这些问题是截图图像案例所特有的，并提出了今后的重要工作。少

2018 年 1 月 4 日提交;最初宣布 2018 年 1 月。

19. 一种新的 ocr 英语字母的排序检测与校正方法

作 者 :chinmay chinara, Mishra nath, subhajeet mishra, sangram keshari sahoo, farida ashraf ali

摘要: 光学字符识别一直是数字计算机出现时的一个具有挑战性的领域。在信息对人类和机器都是可读的情况下，是需要的。**ocr** 的过程由一组预测识别精度水平的前、后处理步骤组成。本文讨论了 **ocr** 过程中涉及的预处理步骤之一，即倾斜（倾斜）检测和校正。所提出的斜检测算法称为 cog（重力中心）法，用于倾斜校正法的方法称为子像素移位法。该算法保持了简单，并进行了优化，实现了高效的倾斜检测和校正。对该算法在测试后的性能进行了恰当的论证。少

2018 年 1 月 2 日提交;最初宣布 2018 年 1 月。

20. 早期印刷书籍中 OCRopus 模型培训的迁移学习

作者:christian reul, christoph wiick , uwe springmann, frank pupe

文摘: 提出了一种方法，在只有少量外交转录的情况下，显著降低从在早期印刷书籍上训练的 ocorpus 模型中获得的 **ocr** 文本的字符错误率。这是通过在培训期间从现有模型中构建来实现的，而不是从零开始。为了克服预训练模型的字符集和附加的地面真值之间的差异，对 OCRopus 代码进行了调整，以允许字母表的扩展或减少。现在，当加载现有模型时，字符集可以灵活地从预先训练的字母表中添加和删除字符。在我们的实验中，我们在早期的拉丁文版画上使用了一个自我训练的混合模型，在现代英语和德语的 fraktur 文本上使用了两个标准的 OCRopus 模型。对七本早期印刷书籍的评价显示，来自拉丁混合模式的训练与从零开始进行的平均误差分别减少了 43% 和 26%，分别为 60 行和 150 行地面真相。此外，研究表明，即使在与新增加的培训和测试数据无关的数据上进行混合模型的构建，也能显著提高识别结果。少

2017 年 12 月 21 日提交;**v1** 于 2017 年 12 月 15 日提交;**最初宣布** 2017 年 12 月。

21. 利用交叉折叠训练和投票提高早期印刷书籍的 ocr 精度

作者:christian reul, uwe springmann, christoph wick, frank pupe

文摘: 本文介绍了一种显著降低从早期印刷书籍训练的 OCRopus 模型中获得的 **ocr** 文本的字符错误率的方法。该方法采用交叉折叠训练和基于信心的投票相结合的方法。在不同的子

集中分配可用的地面真相后，将执行几个训练过程，每个过程都会产生一个特定**的 ocr** 模型。然后，通过考虑识别的**字符**、它们的替代方案以及分配给每个字符的置信度值，对这些模型生成的 ocr 文本进行投票，以确定最终输出。对七本早期印刷书籍的实验表明，该方法通过将误差量减少多达 50% 以上，大大优于标准方法。少

2017 年 11 月 27 日提交;最初宣布 2017 年 11 月。

22. 泰卢固语光学字符识别（ocr）：数据库、算法及应用

作 者 :konkimalla chandra prakash, y. m.srikar , gayam trishal, sourajmanal , sumohana s. channappayya

摘要: 泰卢固语是全世界 8, 000 多万人使用的德拉维迪安语。泰卢固文字的光学字符识别（**ocr**）具有广泛的应用，包括教育、保健、管理等。然而，美丽的泰卢固文字与英语和德语等日耳曼文字有很大的不同。这使得使用日耳曼 **ocr 解决**方案到泰卢固的转移学习成为一项不平凡的任务。为了应对 ocr 对泰卢固的**挑战**，我们在这项工作中做出了三个贡献: (一）泰卢固字符数据库, (二）基于深度学习的 **ocr** 算法，以及（iii）在线部署该算法的客户端服务器解决方案。为了泰卢固人和研究界的利益，我们将以 https://gayamtrishal.github.io/OCR_Telugu.github.io/ 的价格免费提供我们的代码

2017 年 11 月 20 日提交;最初宣布 2017 年 11 月。

23. aon: 面向仲裁的文本识别

作者:郑占展,徐阳柳,范白，易牛,施良浦,周水庚

摘要: 从自然图像中识别文本由于其应用的多样性，是计算机视觉中的一个热门课题。尽管几十年来对光学字符识别（**ocr**）进行了持久的研究，但从自然图像中识别文本仍然是一项具有挑战性的任务。这是因为场景文本往往是不规则的（例如弯曲的、任意的或严重扭曲的）安排，而这些安排在文献中尚未得到很好的处理。现有的文本识别方法主要适用于规则（水平和正面）文本，不能对处理不规则文本进行小加概括。本文开发了任意定向网络(aon)，直接捕获不规则文本的深层特征，并将其组合到基于注意的解码器中生成字符序列。通过只使用图像和文字级别的注释，可以对整个网络进行端到端培训。对各种基准（包括 cute80、svt 透视、iiit5k、svt 和 icdar 数据集）进行的广泛实验表明，所提出的基于 aon 的方法在不规则数据集中实现了最先进的性能，与现有的主要基准相当。常规数据集中的方法。少

2018 年 3 月 22 日提交;v1 于 2017 年 11 月 11 日提交;**最初宣布** 2017 年 11 月。

24. 光学字符识别系统的研究综述

作者:noman 回教, zeeshan 回教, nzia noor

文摘: 光学字符识别（ocr）多年来一直是人们关注的话题。它被定义为将文档图像数字化为其组成字符的过程。尽管进行了几十年的紧张研究，但开发具有与人类相当的能力的 ocr 仍然是一个开放的挑战。由于这种具有挑战性的性质，来自业界和学术界的研究人员开始关注光学字符识别。在过去几年里，参与角色识别研究的学术实验室和公司数量大幅增加。本研究旨在总结目前在 ocr 领域所做的研究。报告概述了 ocr 的不同方面，并讨论了旨在解决 ocr 问题的相应建议。少

2017 年 10 月 3 日提交;最初宣布 2017 年 10 月。

25. 卷积神经网络结合文本和视觉特征的页面流分割

作者:gregor wiedemann, gerhard heyer

摘要: 近年来，(追溯）纸质档案数字化成为私人和公共档案的一项重大工作，也是电子邮件室应用中的一项重要任务。作为第一步，工作流程涉及文档的扫描和光学字符识别（ocr)。保存单页扫描的文档上下文是此上下文中的主要要求。为了方便涉及大量纸张扫描的工作流，页面流分段（pss）的任务是将扫描的图像流自动分离到多页文档中。在一个数字化项目和德国联邦档案馆的合作中，我们开发了一种基于卷积神经网络（cnn）的新方法，将图像和文本特征结合起来，以获得最佳的文档分离结果。评估表明，

我们的 pss 架构实现了高达 93% 的精度，这可以被看作是这项任务的一项新的最新技术。少

2018 年 2 月 8 日提交;**v1** 于 2017 年 10 月 9 日提交;**最初宣布** 2017 年 10 月。

26. 基于动态形状编码的场景图像和视频帧中的文本搜索

作 者 :partha pratim roy, ayan kumar bhunia, avirup bhattacharyya, umapada pal

摘要: 从自然场景图像和视频帧中检索文本信息是一项具有挑战性的任务，因为它存在着复杂的字符形状、低分辨率、背景噪声等固有的问题。可用**的 ocr** 系统通常无法在场景/视频帧中检索此类信息。关键字发现是检索信息的另一种方式，可在此类方案中执行高效的文本搜索。然而，目前场景/视频图像中的单词发现技术是特定于脚本的，它们主要是为拉丁文脚本开发的。本文提出了一种新的基于动态形状编码的自然场景图像和视频帧文本检索的词点框架。该框架旨在利用相应脚本的动态脚本关键字生成来搜索多个脚本中的查询关键字。我们使用了使用隐马尔可夫模型（hmm）的两阶段单词发现方法，通过识别行的脚本来检测给定文本行中的翻译关键字。采用一种新的基于无监督动态形状编码的新方案对相似形状字符进行分组，以避免混淆，提高文本对齐方式。接下来，验证假设位置，以提高检索性能。为了评估从自然场景图像和视频帧中搜索关键字的拟议系统，我们考虑了两个

流行的印度脚本，如孟加拉语（孟加拉语）和 devanagari 以及英语。在印度 scripts[1] 中区域识别方法的启发下，区域化文本信息被用来提高印度语脚本中的传统单词识别性能。在我们的实验中，考虑了由不同场景的图像和英语、孟加拉语和 devanagari 脚本的视频帧组成的数据集。所得结果表明了我们提出的单词识别方法的有效性。少

2018 年 7 月 30 日提交;**v1** 于 2017 年 8 月 18 日提交;**最初宣布** 2017 年 8 月。

27. 多语言 ocr 的顺序到标签脚本标识

作者:yasuhisa fujii, karel driesen , jonathan bacash, ash hurst, ashok c. popat

文摘: 我们描述了一种新的线级脚本识别方法。以前的工作重新设计了一个基于每个字符脚本代码的 **ocr** 模型，用于计算以获得行级脚本标识。这有两个缺点。首先，作为序列到序列的模型，对于行脚本标识的顺序到标签问题来说，它比需要的要复杂得多。这使得培训变得更加困难，运行效率低下。其次，与学习模型相比，计数启发式可能是次优的。因此，我们将行脚本标识重新编码为一个从标签序列来解决的问题，并使用两个组件来解决它，即训练的端到端：编码器和汇总器。编码器将线图像转换为要素序列。汇总器聚合序列以对线条进行分类。我们测试各种总结器与相同的感受式卷积网络作为编码器。在 30 个脚本中包含 232 种语言

的扫描书籍和照片上进行的实验显示，与基线相比，脚本识别错误率降低了 16%。这种改进的脚本识别将脚本错误识别的字符错误率降低了 33%。少

2017 年 8 月 17 日提交;v1 于 2017 年 8 月 15 日提交;**最初宣布** 2017 年 8 月。

28. 用于字体分类的卷积神经网络

作者:chris tensmeyer, daniel saunders, tony martíez

摘要: 将页面或文本行分类为字体类别有助于转录，因为单字符识别（ocr）通常比全字体 ocr 更准确。我们提出了一个基于卷积神经网络（cnn）的简单框架，在这个框架中，cnn 接受了培训，将小块文本划分为预定义的字体类。为了对页面或线条图像进行分类，我们将美国有线电视新闻网的预测与密集提取的补丁进行了计算。我们表明，该方法在 40 种阿拉伯计算机字体的具有挑战性的数据集上实现了最先进的性能，线级精度为 98.8.%。同样的方法也达到了最高的报告精度为 26.6%，预测古写脚本类在页面水平上的中世纪拉丁文手稿。最后，我们分析了美国有线电视新闻网在拉丁文手稿上学到的特征，并发现证据表明美国有线电视新闻网正在学习抄写脚本类之间的定义形态差异，以及过度适应类相关的滋扰因素。我们提出了一种新的数据增强形式，提高了文本黑暗的鲁棒性，进一步提高了分类性能。少

2017 年 8 月 11 日提交;最初宣布 2017 年 8 月。

29. stn-ocr: 一种用于文本检测和文本识别的单一神经网络

作者:christian bartz, haojin yang, christoph meinel

摘要: 检测和识别自然场景图像中的文本是一项具有挑战性但尚未完全解决的任务。在过去几年里，提出了几个试图解决两个子任务中至少一个子任务（文本检测和文本识别）的新系统。本文提出了 stn-ocr，这是朝着场景文本识别的半监督神经网络迈出的一步，可以端到端进行优化。与大多数现有的作品，包括多个深度神经网络和几个预处理步骤，我们建议使用一个单一的深度神经网络，学习检测和识别文本从自然图像在半监督的方式。stn-**ocr** 是一个集成和共同学习空间转换器网络的网络，它可以学习检测图像中的文本区域，以及一个获取已识别文本区域并识别其文本内容的文本识别网络。我们研究我们的模型在一系列不同任务（字符和文本行的检测和识别）上的行为。公共基准数据集的实验结果表明，我们的模型能够处理各种不同的任务，而不会对其整体网络结构进行实质性改变。少

2017 年 7 月 27 日提交;最初宣布 2017 年 7 月。

30. 基于文本增强和形状编码的场景图像和视频帧中的数据场检索

作者:partha pratim roy, ayyan kumar bhunia, umapada pal

摘要: 由于分辨率低、模糊、背景噪声等原因，场景图像和视频帧中的文本识别难度较大。由于传统**的 ocr** 在这类图像中的表现并不理想，使用关键字进行信息检索可能是索引检索此类文本信息的另一种方式。日期是一个有用的信息，它有各种应用，包括日期的视频搜索，索引或检索。本文提出了一种基于数据识别的自然场景图像和视频帧信息检索系统，其中文本出现背景复杂。我们提出了一种基于线的日期发现方法，该方法使用隐马尔可夫模型(hmm）来检测给定文本中的日期信息。从一行中搜索不同的日期模型，而不对字符或单词进行分段。在 rgb 中给定文本行图像，我们应用有效的灰度图像转换来增强文本信息。利用小波分解和梯度子带来增强灰度文本信息。其次，从灰度图像和二值图像中提取定向梯度（phog）特征的金字塔直方图，用于日期发现框架。二元图像和灰度图像特征是由基于 mlp 的串联方法结合在一起的。最后，为了进一步提高单词的性能，采用了基于形状编码的方案，在单词识别过程中将同一类中的相似形状字符组合在一起。在我们的实验中，构建了三个不同的日期模型，以搜索具有数字日期的类似日期，这些日期包含数字值、标点符号和分号，其中包含带数字的日期以及场景/视频文本中的月份。我们已经在 1648 文本行上测试了我们的系统，结果显示了我们提出的日期发现方法的有效性。少

2017 年 7 月 21 日提交;最初宣布 2017 年 7 月。

31. 基于颜色通道选择的场景图像和视频帧中的文本识别

作者:ayan kumar bhunia, gautam kumar,partha pratim roy, r. balasubramanian, umapada pal

摘要: 近年来，自然场景图像和视频帧中的文本识别由于其复杂性和挑战性而受到研究人员的越来越多的关注。由于分辨率低、效果模糊、背景复杂、字体不同、图像和视频帧中文本的颜色和变体对齐方式等，在这种情况下很难识别文本。目前的大多数方法通常采用二值化算法将其转换为二值图像，接下来的 ocr 是用来获得识别结果的。本文提出了一种基于颜色通道选择的场景图像和视频帧文本识别新方法。在该方法中，首先自动选择颜色通道，然后考虑选择颜色通道进行文本识别。我们的文本识别框架基于隐马尔可夫模型（hmm），它使用从选定的颜色通道中提取的定向梯度特征的金字塔直方图。从彩色通道的每个滑动窗口中，我们的颜色通道选择方法从滑动窗口分析图像属性，然后使用多标签支持向量机（svm）分类器来选择能够提供最佳颜色通道的颜色通道。在滑动窗口中的识别结果。每个滑动窗口的这种颜色通道选择被发现比考虑整个单词图像的单一颜色通道更有成效。分析了基于小波变换特征优于其他特征的多标签支持向量机颜色通道选择的五种不同特征。我们的框架已在不同的公开场景/视频文本图像数据集上进行了测试。对于 devanagari 脚本，我们收集了自己的数据集。实验结果令人鼓舞，显示了该方法的优越性。少

2017 年 7 月 27 日提交;**v1** 于 2017 年 7 月 21 日提交;**最初宣布** 2017 年 7 月。

32. 基于投影的基于轮廓振幅滤波器的阿拉伯语字符分割

作者:mahmoud a. a. mousa , mohammed s. sayed , mahmoud i. abdalla

摘要: 阿拉伯语是对光学字符识别（ocr）提出特殊挑战的语言之一。阿拉伯语的主要挑战是它大多是草书。因此，必须执行一个分割过程，以确定字符的起始位置和结束位置。此步骤对于字符识别至关重要。本文提出了阿拉伯语字符分割算法。该算法使用基于投影的方法概念来分隔线条、单词和字符。这是通过使用轮廓的振幅滤波器和简单的边缘工具来查找字符分离。当应用于不同的机器打印文档和不同的阿拉伯语字体时，我们的算法显示出很有希望的性能。少

2017 年 7 月 3 日提交;**最初宣布** 2017 年 7 月。

33. 基于单一分类器的基于局部纹理特征的源打印机分类无源系统

作者:sharad joshi , nitin khanna

摘要: 检查印刷文件是否有潜在伪造和侵犯版权行为的一个重要方面是识别源打印机，因为这有助于查明泄漏情况和检测伪造文件。本文提出了一种利用所有打印字母同时对印刷文档扫描图像

进行源打印机分类的系统。该系统使用基于本地纹理模式的功能和单个分类器对所有打印的字母进行分类。从扫描图像中提取字母，使用连接的成分分析，然后进行形态过滤，而无需使用 **ocr**。每个字母被细分为一个平面区域和一个边缘区域，这两个区域的局部 tetra 模式是单独估计的。利用战略性构造的池技术提取最终特征向量。该方法已在 10 台打印机的公开数据集和以 600 dpi 分辨率扫描的 18 台打印机的新数据集上进行了测试，并以 4 种不同字体打印了 300 dpi。结果表明，该方法具有形状独立性，因为使用单个分类器的方法优于现有的基于手工制作的基于特征的方法，并且通过使用所有打印的字母，需要的训练页数要小得多。少

2017 年 6 月 22 日提交;最初宣布 2017 年 6 月。

34. 搜索：培训具有全球-本地损失的 rnn

作者:rémi leblond, jean-baptiste alayrac, anton osokin, simon lacoste-jujen

摘要: 我们提出了一种新的递归神经网络训练算法--一种新的递归神经网络训练算法，该算法是在结构化预测的 "学习搜索" (l2s) 方法的启发下进行的。rnn 在结构化预测应用（如机器翻译或解析）中已取得广泛的成功，并且通常使用最大似然估计（mle）进行训练。不幸的是，这种训练损失并不总是测试错误的适当替代因素：仅仅通过最大限度地提高地面真相概率，就无法利用结构化损失提供的大量信息。此外，它还引入了培训和预测之间的差

异（如曝光偏差），这可能会影响测试性能。相反，布什恩利用类似于测试的搜索空间探索引入更接近测试误差的全局局部损失。我们首先演示了在两个不同任务上比 mle 更高的性能: **ocr** 和拼写更正。然后，我们提出了一个子采样策略，使考察能够扩展到较大的词汇量。这使我们能够验证我们的方法在机器翻译任务上的好处。少

2018 年 3 月 4 日提交;**v1** 于 2017 年 6 月 14 日提交;**最初宣布** 2017 年 6 月。

35. 基于导数的多通道图像分割组件树

作者:tobias böttger, dominik gutermuth

摘要: 我们介绍了基于派生的基于组件树的概念，用于任意数量的通道的图像。该方法是专门用于灰度图像的经典组件树的自然扩展。类似的结构使许多基于组件树的灰度图像处理技术能够转换为高光谱和彩色图像。作为一个示例应用，我们提出了一种提取最大稳定同构区域（mshr）的图像分割方法。该方法与 mser 非常相似，但可应用于具有任意通道数的图像。与 mser 不同的是，我们的方法隐式分割区域，比灰度图像的背景更浅、更暗，并且可以在 mser 将失败的 **ocr** 应用程序中使用。我们介绍了一种基于局部洪水的沉浸式结构，用于基于派生的组件树结构，该结构在像素数量上是线性的。在实验中，我们发现，随着信道数量的

增加，运行时的规模是有利的，并可能改进基于 mser 的算法。少

2018 年 4 月 19 日提交;**v1** 于 2017 年 5 月 4 日提交;**最初宣布** 2017 年 5 月。

36. 莎士比亚第一本作品集中的自动编辑归因

作者:maria ryskina, hannah alpert-ableams, dan garrette , taylor berg-kirkpatrick

摘要: 编辑归因是由设置类型的个人对历史印刷文档中的页面进行聚类，它是一项书目任务，它依赖于对印刷页面的正交变化和视觉细节的分析。本文介绍了一种新的无监督模型，该模型共同描述了区分合成器所需的文本特征和视觉特征。应用于莎士比亚的第一个作品集的图像，我们的模型预测的属性，同意与书目的人工判断，与 87% 的准确性，即使是文本，是 **ocr** 的输出。少

2017 年 4 月 25 日提交;**最初宣布** 2017 年 4 月。

37. ocrapose ii：基于 ocrs 的使用手机图像的室内定位系统

作者:hamed sadeghi, shahrokh valaee, shahram shirani

文摘: 本文提出了一种基于 ocr（光学字符识别）的定位系统--ocrapose ii，该系统适用于许多室内场景，包括办公楼、停车场、机场、杂货店等。在这些情况下，字符（即文本或数字）可用

作本地化的合适的独特地标。该系统利用 ocr 读取查询静止图像中的这些字符，并使用平面图提供了粗略的位置估计。然后，利用 ocr 引擎提供的信息查找查询的深度和视角，以优化位置估计。利用图像线段和 ocr 框信息，导出了查询视角和深度估计的新公式。通过室内场景实验，验证了该系统的适用性和有效性。结果表明，与最先进的基准相比，该系统在位置识别率和平均定位误差方面表现出了更好的性能，特别是在稀疏数据库条件下。少

2017 年 4 月 18 日提交;最初宣布 2017 年 4 月。

38. 一种优化 fpga cnc 实现 dsp 块利用率的整体方法

作 者 :kamel abdelouahab, cedric bourrasset, maxime pelcat, françois berry, jean-charles quinton, jocelyn serot

文摘: 深神经网络正在成为图像理解的事实上的标准模型，更普遍的是计算机视觉任务的标准模型。由于它们涉及高度并行的计算, cnn 非常适合当前的细粒可编程逻辑器件。因此，在 fpga 上成功地实施了多个美国有线电视新闻网加速器。遗憾的是, fpga 资源（如逻辑元素或 dsp 单元）仍然有限。本文提出了一种基于近似计算和设计空间探索的整体方法，以优化 fpga 上 cnn 实现的 dsp 块利用率。在 altera stratix v 器件上实现可重构的 ocr 卷积神经网络，并改变数据表示和 cnn 拓扑结构，以找到 dsp 块利用率方面的最佳组合时，对该方法进行了测试。分类精度。本探索生成了 76 种具有 5 个不同定点表示的 76 个 cnn 拓

扑的数据流体系结构。最有效的实现使用 8% 的可用 dsp 块以 256 x 256 分辨率执行 883 分类。少

2017 年 3 月 21 日提交;最初宣布 2017 年 3 月。

39. 阿拉伯光学字符识别（ocr）的重要新进展

作者:maxim romanov, matthew thomas miller, sarah bowen savant, benjamin kiessling

摘要: op 光 iti 团队在 90 年代的经典阿拉伯文字文本中实现了光学字符识别（**ocr**）的准确率。这些数字是基于我们对 7 个不同质量和字体的阿拉伯语脚本文本的测试，总共超过 7000 条线。这些准确率不仅比经典阿拉伯文字文本的各种专有 **ocr 选项的**实际准确率有了显著提高，而且同样重要的是，它们是使用开源 **ocr** 生产的软件，从而使我们能够使这种阿拉伯语脚本的 **ocr** 技术免费提供给更广泛的伊斯兰，波斯语和阿拉伯研究社区。少

2017 年 3 月 28 日提交;最初宣布 2017 年 3 月。

40. 基于内容的基于内容的基于 cnn 特征融合的文档图像检索

作者:毛坦,袁思平,苏永新

摘要: 数字化文档的迅速增加对文献图像检索提出了很高的要求。传统的文档图像检索方法依赖于复杂**的基于 ocr** 的文本识别和文本相似性检测，提出了一种新的基于内容的方法，其中更多的

是对特征提取和融合的关注。在该方法中，通过不同的 cnn 模型提取文档图像的多个特征。之后，提取的美国有线电视新闻网功能被减少，并融合为加权平均功能。最后，根据特征与所提供的查询图像的相似性对文档图像进行排序。对一组从包含中英文文献的学术论文转换的文档图像进行实验，结果表明，该方法具有较好的检索文本相似的文档图像能力。内容，并融合 cnn 功能可以有效地提高检索精度。少

2017 年 8 月 31 日提交;**v1** 于 2017 年 3 月 23 日提交;**最初宣布** 2017 年 3 月。

41. **推特 100k：一个现实世界的数据集，为薄弱的监管跨媒体检索**

作者:胡玉婷,梁正,易阳,黄永峰

摘要: 本文为弱监督的跨媒体检索提供了一个新的大型数据集，名为 twitter 100k。当前的数据集，如维基百科、nus 宽数据集和 flickr30k 数据集，有两个主要限制。首先，这些数据集缺乏内容多样性，即只涵盖一些预定义的类。其次，这些数据集中的文本是用组织良好的语言编写的，从而导致与实际应用程序不一致。为了克服这些缺点，提出的 twitter 100k 数据集具有两个方面的特点: 1) 它有 100，000 个从 twitter 随机爬来爬来的图像文本对，因此在图像类别中没有约束;2) 推特 100k 中的文本由用户以非正式语言编写。由于强监督方法利用了实践中可能缺失的类标签，

本文重点研究了跨媒体检索中的弱监督学习，在训练中只利用文本图像对。我们广泛地评估了四种子空间学习方法和通信自动编码器的三个变体的性能，以及维基百科、flickr30k 和 twitter 100k 上的各种文本功能。提供了新颖的见解。作为一个小的贡献，灵感来自于 twitter 100k 的特点，我们提出了一种基于 ocr 的跨媒体检索方法。实验表明，基于 ocr 的方法提高了基线性能。少

2017 年 3 月 20 日提交;**最初宣布** 2017 年 3 月。

42. 语言独立的单文档图像超分辨率使用 cnn 提高识别能力

作者:ram krishna pandey, a g ramakrishnan

摘要: 文档图像的识别在还原古文和经典文本中有着重要的应用。这个问题涉及到质量的提高，然后再传递给一个经过适当培训的 ocr ，以获得对文本的准确识别。图像增强和质量改进是重要的步骤，因为后续识别取决于输入图像的质量。在某些情况下，高分辨率图像不可用，我们的实验表明，随着文档图像空间分辨率的降低，ocr 精度显著降低。因此，唯一的选择是提高此类文档图像的分辨率。目标是在给定单个低分辨率二值图像的情况下构造高分辨率图像，这就构成了单图像超分辨率的问题。以往在超分辨率方面的大部分工作都涉及比文档图像具有更多信息内容的自然图像。在这里，我们使用卷积神经网络来学习低分辨率和相应的高分辨率图像之间的映射。我们尝试了不同数量的层、参数设置

和非线性函数，以构建一个快速的端到端文档图像超分辨率框架。我们提出的模型显示了 75 dB 泰米尔图像约 4 db 的非常好的 psnr 改进，从而使 **ocr** 提高了 3% 的字级精度。与最近的基于稀疏的自然图像超分辨率技术相比，它所需的时间更短，因此对实时文档识别应用非常有用。少

2017 年 1 月 30 日提交;最初宣布 2017 年 1 月。

43. larex–用于早期印刷书籍布局分析和区域提取的半自动开源工具

作者:christian reul, uwe springmann, frank pupe

摘要: 介绍了一种用于早期印刷书籍布局分析的半自动开源工具。larex 使用基于规则的连接组件方法，该方法非常快速，便于用户理解，并允许在必要时进行直观的手动更正。页面 xml 格式用于支持集成到现有**的 ocr** 工作流。评价显示, larex 提供了一种高效、灵活的方式来分割早期印刷书籍的几页。少

2017 年 1 月 20 日提交;最初宣布 2017 年 1 月。

44. 高度自动化布局分析和 ocr 的案例研究: "德海利根·勒本" (1488)

作者:christian reul, marco dittrich, martin gruner

文摘: 本文首次全面记录了从印加文（1450–1500）开始应用于早期印刷书籍的高质量数字化过程。以1488年在纽伦堡印刷的 "der heiligen leben" 为例，详细说明了整个 **ocr** 相关工作流程，包括预处理、布局分析和文本识别。每一步都记录所需的时间支出。字符识别在字符（97.57%）和单词（97.57）水平上都产生了优异的效果。此外，还对高度自动化（larex）和手动（Aletheia）布局分析方法进行了比较。通过大大自动化分割，所需的人工工作量从 100 多个小时大幅减少到不到 6 小时，只导致 **ocr** 精度略有下降。从这项研究中可以得出从 incunabula 中提取全文所需的人类努力的现实估计。完整工作的打印页与 **ocr** 结果一起可在线查阅，可供检查和下载。少

2017 年 1 月 20 日提交;**最初宣布** 2017 年 1 月。

45. 重新分析 ocr 的历史文本

作者:florian fink, klaus–u。schulz, uwe springmann

摘要: 在没有地面真伪的情况下，不可能自动确定 ocr 的文本中 **ocr** 错误的确切频谱和发生。然而，对于 ocr ' ed 历史打印的交互式后校正，提供统计配置文件以提供具有关联频率的错误类的估计，并指出推测错误和可疑的令牌，是非常有用的。.在刷新（2013）中引入的方法计算这样的配置文件，结合词汇，模式集和高级匹配技术在一个专门的期望最大化（em）过程中。在这里，

我们从三个方面对该方法进行了改进：第一，刷新（2013）中的方法不是自适应的：通过实际的后校正步骤获得的用户反馈不能用于计算细化的配置文件。我们引入了一种开放的方法的变体，并考虑到用户的修正步骤。这导致在识别错误**的 ocr** 令牌方面具有更高的精度。其次，在后纠正过程中，经常会发现新的历史模式。我们表明，在语言背景资源中添加新的历史模式会带来第二种改进，通过区分 ocr 错误之外的历史拼写来实现更高的精度。第三，刷新（2013）中的方法不会主动使用无法在基础通道模型中解释的令牌。我们表明，将这些不可解释的令牌添加到一组猜想的错误中，可以显著改进错误检测的召回，同时提高精度。少

2017 年 1 月 19 日提交;最初宣布 2017 年 1 月。

46. 从历史区划挖掘工业化时空数据

作 者 :david berenbaum, dwyer deighan, thomas marlow, ashley lee, scott fr 客观

摘要: 尽管许多领域的海量数据越来越多，但由于缺乏收集、数字化和组装的自动化和可扩展方法，往往无法获得关于社会环境现象的历史数据。我们开发了一种数据挖掘方法，用于从打印目录中提取表格化、地理编码的数据。虽然扫描和光学字符识别（**ocr**）可以将打印文本数字化，但仅这些方法并不能捕获基础数据的结构。我们的管道集成了页面布局分析和 **ocr** ，可从结构化文本中

提取表格、地理编码的数据。我们通过将此方法应用于罗得岛记录了 41 年工业土地使用的扫描制造登记，证明了这种方法的效用。由此产生的时空数据可用于以前不可能达成的工业化的社会环境分析。特别是，我们发现了有力的证据，表明制造业分散在该州首府普罗维登斯的城市核心，沿着 95 号州际公路南北走廊。少

2016 年 12 月 3 日提交;最初宣布 2016 年 12 月。

47. 基于多层感知器的文本图像识别

作者:singh vijendra, nisha vasudeva, hem jyotsana parashar

摘要: 图像处理领域最大的挑战是识别印刷和手写格式的文档。光学字符识别 **ocr** 是一种文档图像分析，其中扫描的数字图像包含机器打印或手写脚本输入**到 ocr** 软件引擎，并将其转换为可编辑的机器可读的数字文本格式。神经网络是用来模拟大脑执行特定任务或感兴趣的功能的方式的: 神经网络是在数字计算机上的软件中模拟的。字符识别是指将打印的文本文档转换为翻译的 unicode 文本的过程。以书籍、报纸、杂志等形式提供的印刷文件使用标准扫描仪进行扫描，这些扫描仪可生成扫描文件的图像。线是通过一种算法来识别的，我们可以在这个算法中识别线的顶部和底部。然后用算法计算各行字符边界，然后利用这些计算将字符与图像隔离，然后通过基本反向传播对每个字符进行分类。每个图像字符由 30 * 20 像素组成。利用反向传播神经网络对多

层神经网络中的误差进行了反传播校正，并通过前馈方法对整流神经元值进行了传递。少

2016 年 12 月 2 日提交;最初宣布 2016 年 12 月。

48. ocr 文本校正的统计学习

作者:jiemei, aminul isam, yasjing wu , abisalrahman moh' d , evangelos e. milios

文摘: 光学字符识别（**ocr**）的准确性对于后续应用于文本分析管道的成功至关重要。最近的 **ocr** 后处理模型显著提高**了 ocr 生成文本**的质量，但仍容易建议从有限的观察中纠正候选人，同时没有充分考虑到这些文本的特点。**ocr** 错误。本文介绍了如何利用外部语料库，并将 **ocr** 的具体特征整合到回归方法中，以纠正 **ocr** 产生的错误，从而扩大候选建议空间。评价结果表明，在理论修正上限为 78 的情况下，我们的模型可以纠正 61.5 的 ocr 误差（考虑前 1 个建议）和 **71.5 的 ocr** 误差（考虑前 3 个建议)%.少

2016 年 11 月 21 日提交;最初宣布 2016 年 11 月。

49. 如何对芬兰大型历史报纸馆藏资源稀缺进行词汇质量评价

作者:kimo kettunen, tuula päkkönen

摘要: 芬兰国家图书馆将 1771 年至 1910 年期间在芬兰出版的历史报纸数字化。该系列包含大约 125 万页芬兰语和瑞典语。收藏的

芬兰部分包括约 24.6 亿字。国家图书馆的数字收藏是通过 digi.kansalliskirjasto.fi 的网络服务（也称为 digi）提供的。部分报纸材料（1771 年至 1874 年）也可在 finclarin 财团提供的芬兰语言银行免费下载。还可以通过在哥德堡大学的 Språkbanken 开发并由赫尔辛基大学 finclarin 团队扩展的 korp 环境访问这些集合，以提供文本资源的一致性。在坦佩雷大学的 digi 报纸材料中，还制作了克兰菲尔德式的信息检索测试集。**ocred** 馆藏的质量是数字人文学科的一个重要课题，因为它影响着馆藏的一般可用性和可搜索性。没有单一的可用方法来评估大型藏品的质量，但可以使用不同的方法来近似质量。本文讨论了不同的语料库分析风格方法，以近似的整体词汇质量的芬兰部分的 digi 集合。方法包括使用平行样本和单词错误率、使用形态分析仪、单词频率分析以及与可比较编辑的词汇数据进行比较。我们在质量分析方面的目标有两个：第一，分析词汇数据的现状，其次，建立一套评估方法，建立一个紧凑的质量评估程序，例如新的 **o 常用**或后纠正的材料。在论文的讨论部分，我们将综合我们不同分析的结果。少

2016 年 11 月 16 日提交;最初宣布 2016 年 11 月。

50. 旧内容和现代工具–在 1771–1910 芬兰 ocred 历史报纸集中搜索命名实体

作者:kimo kettunen, eetu mäkelä, teemu ruokolainen, juha kuokkala, laura löfberg

摘要: 命名实体识别 (ner)、名称和名称的搜索、分类和标记, 如文本中频繁的信息元素, 已成为文本数据的标准信息提取过程。ner 已应用于许多类型的文本和不同类型的实体: 报纸、小说、历史纪录、人、地点、化合物、蛋白质家族、动物等。一般来说, ner 系统的性能取决于类型和域, 并且使用的实体类别也各不相同 (nadeau 和 sekine, 2007 年)。最一般的一套命名实体通常是对地点、个人和组织进行三方分类的某种版本。本文用数字化芬兰历史报纸收藏的 digi 中的数据报告了首次对 ner 的大规模试验和评估。本研究的实验、结果和讨论为芬兰历史报纸网络收藏的发展服务。digi 系列包含 1771 年至 1910 年芬兰语和瑞典语的 1, 960, 921 页报纸材料。我们在评估中只使用芬兰文件的材料。**ocred** 报纸系列有很多 **ocr** 错误;其估计的字水平正确性约为 70–75% (kettunen 和 päkkönen, 2016)。我们的主要 ner 标签是芬兰芬兰的一个基于规则的标签, 由 fin-clarin 财团提供。我们还展示了利用阿尔托大学语义计算研究小组 (seco) 的工具进行有限类别语义标记的结果。还对另外三个工具进行了简要评估。这项研究报告首次在芬兰 **ocred** 报纸的历史收藏中公布了 ner 的大规模成果。研究结果以类似的噪声数据补充了其他语言的 ner 结果。少

2016 年 11 月 9 日提交;最初宣布 2016 年 11 月。

51. 汉语–英语混合字符分割作为语义分割

作者:郑华斌,王景宇，黄正杰,杨阳,潘荣

文摘: 多语言印刷文档的 **ocr** 字符分割由于不同语言字符的多样性而难以进行。以往的方法主要集中在单语文本上，不适合多语言语言的情况。在这项工作中，我们特别解决了汉语英语混合情况，将其重新定义为语义分割问题。我们利用了在语义分割领域中被称为完全卷积网络（fcn）的成功体系结构。给定足够宽的接受场，fcn 可以利用水平位置周围的必要上下文来确定这是否是一个分裂点。作为一种深层神经架构，fcn 可以自动从原始文本线图像中学习有用的功能。虽然我们对具有模拟随机扰动的合成样品进行了训练，但我们的 fcn 模型很好地推广到了真实世界的样品中。实验结果表明，我们的模型明显优于以往的方法。少

2016 年 11 月 15 日提交;**v1** 于 2016 年 11 月 7 日提交;**最初宣布** 2016 年 11 月。

52. 还没呢？将传统序列到序列模型与单声道字符串转换任务上的编码解码器解码神经网络进行比较

作者:carsten schnober, steffeneger, erek–lan do dinh , iryna gurevych

摘要: 我们分析了编码解码器神经模型的性能，并将其与已知的既定方法进行了比较。后者代表了不同类型的传统方法，这些方

法适用于单调的序列序列任务 **ocr** 后校正、拼写校正、图形到音素转换和引光。这些任务对于各种更高层次的研究领域具有重要的现实意义，包括数字人文、自动文本校正和语音识别。我们研究通用的深度学习方法如何适应这些任务，以及它们与既有和更专业的方法相比的表现，包括我们自己对修剪后的 crf 的适应.

2016 年 10 月 26 日提交;**v1** 于 2016 年 10 月 25 日提交;**最初宣布** 2016 年 10 月。

53. 车辆识别自动牌照和车牌识别系统的建议

作者:哈米德·萨海伊

摘要: 本文提出了一种利用图像处理算法提取通过给定位置的车辆牌照号码的自动机械化牌照和车牌识别（lnpr）系统。为实施拟议的系统，不需要安装 gps 或无线电频率识别（rfid）等其他设备。该系统使用特殊的摄像头，从每辆路过的车辆上拍照，并将图像转发到计算机上，由 lpr 软件进行处理。车牌识别软件采用定位、定位、归一化、分割等不同算法，最终实现光学字符识别（**ocr**）。生成的数据将用于与数据库上的记录进行比较。实验结果表明，该系统成功地检测和识别了真实图像上的车辆号牌。该系统还可用于安全和交通控制。少

2016 年 10 月 9 日提交;**最初宣布** 2016 年 10 月。

54. ocr ++: 一个可靠的从学术文章中提取信息的框架

作 者 :mayank singh, barnopriyo barua , priyank palod, manvigarg , sidharthasatapathy, samuel bushi, kumar ayush, krishna sai rohith, tulasigamidi, pawan goyal, animesh mukherjee

摘要: 本文提出了 ocr++，这是一个开源框架，旨在从学术文章(包括元数据（标题、作者姓名、隶属关系和电子邮件)、结构（章节标题和正文、表格）中提取各种信息任务。和图表标题、url 和脚注）和参考书目（引文实例和参考)。我们分析了一套不同的科学文章写在英语，以了解通用的写作模式，并制定规则来开发这个混合框架。广泛的评价表明，拟议的框架优于现有的最先进的工具，在结构信息提取方面有很大的优势，同时改进了元数据和书目提取任务的性能，这两方面的精度（约 50% 的改进）和处理时间（约 52% 的改进)。在 3 0 名研究人员的帮助下进行的用户体验研究显示，研究人员发现这个系统非常有帮助。作为另一个目标，我们讨论了两个新的用例，包括自动从会议记录中提取公共数据集的链接，这将进一步加快数字图书馆的发展。框架的结果可以作为一个整体导出到结构化的 tei 编码的文档中。我们的框架可在 http://cnergres.iitkgp.ac.in/ocr+ +/home/在线访问。少

2016 年 9 月 23 日提交;**v1** 于 2016 年 9 月 21 日提交;**最初宣布** 2016 年 9 月。

55. 图像到标记生成，并注意到粗件

作者:邓云天,安西·卡内尔维托,林杰瑞,亚历山大·拉什

摘要: 我们提出了一个神经编码器解码器模型，将图像转换为基于可扩展的粗到精细注意力机制的表示标记。我们的方法是在映像到 latex 生成的上下文中进行评估的，我们引入了一个新的真实的呈现数学表达式数据集，这些表达式与 latex 标记配对。我们表明，与使用基于 ctc 的模型的神经 **ocr 技术不同**，基于注意的方法可以解决这一非标准 **ocr** 任务。我们的方法在域内呈现的数据上的性能大大优于经典的数学 **ocr** 系统，而且通过预培训，在域外手写数据上的性能也很好。为了降低与基于注意的方法相关的推理复杂性，我们引入了一个新的粗到细关注层，在应用注意之前选择一个支持区域。少

2017 年 6 月 13 日提交;**v1** 于 2016 年 9 月 16 日提交;**最初宣布** 2016 年 9 月。

56. mt3s: 针对视力障碍者的移动土耳其场景文本到语音系统

作者:muusan bastan, hilal kandemir , busra canturk

摘要: 阅读文本是视障人士的基本需求之一。我们开发了一个移动系统，可以读取土耳其场景和书籍文本，使用快速梯度的多尺度文本检测算法进行实时操作，Tesseract **ocr** 引擎进行字符识别。

我们在我们构建的一个新的、公开的移动土耳其场景文本数据集上评估了我们系统的 ocr 准确性和运行时间，并与最先进的系统进行了比较。事实证明，我们的系统速度更快，能够在移动设备上运行，ocr 精度与最先进的系统相当。少

2016 年 8 月 17 日提交;最初宣布 2016 年 8 月。

57. **历史印刷的 ocr 及其在构建历时语料库中的应用：利用 ridges 草药语料库的案例研究**

作者:u. springmann, a. lüdeling

摘要: 本文介绍了一个案例研究的结果，该研究将基于神经网络的光学字符识别（ocr）应用于 1487 年至 1870 年间印刷的书籍的扫描图像，方法是在 ridges 草药文本语料库[@OdebrechtEtAlSubmitted]。培训特定**的 ocr** 模型是可能的，因为必要的 * 地面真相 * 可作为错误更正的外交抄录。**对 ocr** 结果进行了评估，以确定其对看不见的测试集的地面真实值的准确性。单个文档的机器可读文本的字符和单词准确性（正确识别的项目的百分比）从 94% 到 99% 以上（字符级别）和 76% 到 97%（单词级别）不等。这包括最早的印刷书籍，直到最近，ocr 方法还认为这些书籍是无法访问的。此外，在由不同打印日期和不同排版 *（混合模型）* 组成的语料库的一部分上训练的 ocr 模型已在包含其他字体的其他部分的书籍上测试其预测能力,大部

分产生的字符精度远远高于 90%。因此，似乎有可能构造在一系列字体上训练的广义模型，这些字体可以应用于各种各样的历史印刷，但仍然会产生很好的效果。然后，一些页面的适度后校正工作将使各个模型的培训具有更好的准确性。使用这种方法，包括早期打印在内的历时语料库可以比手动转录更快、更便宜。这里报告的 **ocr** 方法为我们将印刷的文本文化遗产通过主要是自动手段转化为电子文本提供了可能性，这是扫描书籍大规模转换的先决条件。少

2017 年 2 月 1 日提交;**v1** 于 2016 年 8 月 6 日提交;**最初宣布** 2016 年 8 月。

58. 车牌字符分割的基准

作 者 :gabriel resende gonçalves , sirlene pio gomes da silva, david menotti, william robson schwartz

摘要: 自动车牌识别（alpr）是近年来许多研究的热点。一般来说，alpr 分为以下几个问题: 轨道车辆的检测、车牌检测、车牌字符的分段和光学字符识别（**ocr**）。尽管有商业解决方案可用于受控的购置条件，例如停车场的入口，但在处理从道路和公路等不受控制的环境中获得的数据时，alpr 仍然是一个悬而未决的问题。成像传感器。由于摄像机捕获的车牌的多个方向和比例，alpr 的一项非常具有挑战性的任务是车牌字符分割（lpcs）步骤，该步骤需要（接近）最佳，以实现高识别由 **ocr**。为了解决 lpcs 问题，

本工作提出了一个新的基准，由一个数据集组成，专门针对评估协议中 alpr 的字符分割步骤。此外，我们提出了 jacard-centroid 系数，这是一种新的评价方法，比 jacard 系数更适合地面真相注释中边界框的位置。该数据集由 2,000 个巴西车牌组成，由 14 000 个字母数字符号及其相应的边界框注释组成。我们还提出了一种新的简单方法来有效地执行 lpcs。最后，我们基于四种 lpcs 方法对数据集进行了实验评估，并论证了字符分割对于实现准确 **ocr** 的重要性。少

2016 年 10 月 31 日提交;**v1** 于 2016 年 7 月 11 日提交;**最初宣布** 2016 年 7 月。

59. 识别字母表字符和数学符号的人工神经网络和模糊逻辑

作者:giuseppe airófarulla, Tiziana armano, anna capietto, nadir muru, rosaria rossini

摘要: 光学字符识别软件（**ocr**）是获取可访问文本的重要工具。我们建议使用人工神经网络（ann）来开发能够识别正常文本和公式的模式识别算法。我们提出了一个反向传播算法的原始改进。此外，我们还描述了一种利用模糊逻辑分离触摸字符的新的图像分割算法。少

2016 年 7 月 6 日提交;**最初宣布** 2016 年 7 月。

60. 历史打印 ocr 模型的自动质量评估和（半自动）自动改进

作者:u. springmann, f. fink , k. u. schulz

摘要: 历史印刷的良好 **ocr** 结果依赖于在外交转录方面训练的识别模型的可用性，而这既是稀缺资源，也是耗时的。我们不需要为每个历史字体训练一个单独的模型，而是提出了一种策略，从在各种字体的一组可用转录上训练的模型开始。这些 \ 强调 {混合模型} 导致同一时期的一组打印测试版的字符准确率超过 90%，但在训练数据中没有任何表示，这表明通过泛化来克服排版障碍的可能性从几个字体到一段时间内使用的更大的一组（类似）字体。然后将这些混合模型的输出作为基线，通过全自动方法和涉及少量手动转录的半自动方法进一步改进。为了在没有任何地面真相的情况下评估训练过程中生成的一系列模型中每个模型的识别质量，我们引入了两个易于观察的数量，它们与真实的准确性密切相关。这些数量是 \ 强调 {均值字符置信度 c}（由 **ocr** 引擎 ocropus 给出的）和 \并不是因为令牌词性 l}（从现代单词形式中测量 ocr 令牌的**距离**，同时考虑到历史拼写模式,可以为任何 **ocr** 引擎计算）。全自动方法能够提高混合模型的结果只有 1–2 个百分点，而已经有 100–200 手校正线导致更好的 **ocr** 结果，字符错误率只有几个百分点。此过程最大限度地减少了地面真相的产生量，并不依赖于以前特定排版模型的构造。少

2016 年 10 月 20 日提交;**v1** 于 2016 年 6 月 16 日提交;**最初宣布** 2016 年 6 月。

61. 摄像机捕获文档的一种自动地面真相生成的通用方法

作 者 :sheraz ahmed, muhammad imran malik , muhammad zeshan afzal, koichi kise, masakazu iwamura,and 列 as dengel, marcus lwamura

摘要: 本文的贡献有四个方面。第一笔贡献是一种新颖、通用的方法，用于自动生成摄像机捕获的文档图像（书籍、杂志、文章、发票等）。它使我们能够构建大规模（即数百万张图像）标记为相机捕获扫描的文档数据集，而无需任何人为干预。该方法是通用的，与语言无关，可用于生成任何草书和非草书语言（如英语、俄语、阿拉伯语、乌尔都语等）的标记文档数据集（扫描和相机化）。为了评估所提出的方法的有效性，使用该方法生成了两个不同的英文和俄文数据集。对两个数据集中的样本的评估显示，99% 的图像被正确标记。第二个贡献是一个大型的相机捕获的字符和单词图像的数据集（称为 c3wi），其中包括 100 万字图像（1 000 万个字符图像），这些图像是在基于相机的实际采集中捕获的。此数据集可用于对摄像机捕获的文档上的字符识别系统进行培训和测试。第三个贡献是一种新的方法来识别相机捕获的文档图像。该方法以长期短期存储器为基础，性能优于最先进的**基于相机的 ocr** 方法。作为第四个贡献，我们将使用所提供的 c3wi 数据集，进行各种基准测试，以揭示商业（abbyy）、开源（Tesseract）和基于相机的 **ocr 的行为**。评价结果表明，现有的 **ocr** 已经在扫描文档上获得了很高的精度，在相机捕获的文档图像上的性能有限;其中

abbyy 的精度为 75%，Tesseract 的准确率为 50.22%，而所提出的字符识别系统的精度为 95.10。少

2016 年 5 月 4 日提交;最初宣布 2016 年 5 月。

62. 空中文字书写

作者:saira beg, m. fahad khan, faisal baig

摘要: 本文提出了一种基于视频的实时指向方法，允许在移动摄像头前的空中绘制和书写英语文本。提出的方法主要有两个任务:首先跟踪视频帧中的彩色指尖，然后将英语 **ocr** 应用于绘制的图像上，以识别书面字符。此外，所提出的方法提供了一种自然的人机交互，不需要键盘、手写笔、笔或手套等进行字符输入。对于实验，我们开发了一个使用 java 语言的 opencv 的应用程序。我们在三星 galaxy3 android 手机上测试了该方法。结果表明，该算法在不同形状字母的测试中，平均精度为 92.083。在这里，使用了 3000 多个不同的磁性 3d 形状的字符 [ref:http://learnrnd.com/news.php?id=Magnetic_3D_Bio_Printing]。我们提出的系统是基于软件的方法和相关的非常简单，快速和容易。它不需要传感器或任何硬件，而不需要相机和繁文缛节。此外，拟议的方法可以适用于所有断开连接的语言，但有一个问题，即它对颜色敏感，因为在开始字符书写之前，背景中存在任何红色都可能导致错误的结果。少

2016 年 4 月 27 日提交;最初宣布 2016 年 4 月。

63. 基于字符校正和基于特征的单词分类的 ocr 纠错

作者:ido kissos, nachum dershowitz

文摘:本文探讨了学习分类器在 **ocr** 后文本校正中的应用。阿拉伯语的实验表明，该方法集成了加权混淆矩阵和浅语言模型，改善了绝大多数的分割和识别错误，这是我们数据集上最常见的错误类型。

2016 年 4 月 21 日提交;最初宣布 2016 年 4 月。

64. 电视新闻广播中的叠加文本提取

作者:raghvendra kannao, prithwijit guha

摘要: 覆盖频段中的文本数据传达了广播视频中对新闻事件的简要描述。文本提取的过程变得具有挑战性，因为叠加文本以千差万别的格式呈现，并且通常具有动画效果。我们注意到，现有的基于边缘密度的方法由于其简单性和运算速度，非常适合我们的应用。然而，这些方法对阈值敏感，并具有较高的假阳性率。本文提出了一种基于对比度增强的叠加文本检测预处理阶段和一种基于参数的无边缘密度方案，用于高效的文本波段检测。本文的第二个贡献是一种新的文本区域跟踪方法，它对所有可能的检测失败案例进行了正式识别。跟踪阶段使我们能够建立文本波段的时间

存在及其随时间的链接。第三个贡献是采用 tesseract **ocr,**用于使用网络新闻文章进行覆盖文本识别的具体任务。在从三个印度英语电视新闻频道以及基准数据集上获得的新闻视频上，对拟议的方法进行了测试，并发现该方法具有优势。少

2016 年 4 月 2 日提交;最初宣布 2016 年 4 月。

65. 利用连接网络的集合来改进基于拼接的场景文本脚本识别

作者:lluis gomez, Anguelos nicoraou, dimeosthenis karatzas

摘要: 本文重点研究了场景文本图像中的脚本识别问题。面对这个最先进的 cnn 分类器的现状并不简单，因为它们未能解决场景文本实例的一个关键特征：它们的长宽比极其可变。我们在这里提出了一个基于补丁的分类框架，以保持图像中具有其类特征的判别部分，而不是像典型的 cnn 分类器那样将输入图像调整为固定长宽比。我们描述了一种新的方法，基于使用组合的连接网络，共同学习判别式中风零件的表示及其在基于拼接的分类方案中的相对重要性。我们使用此学习过程进行的实验演示了两个公共脚本识别数据集中最先进的结果。此外，我们还提出了一个新的公共基准数据集，用于评估多语言场景文本端到端读取系统。在此数据集中进行的实验证明了脚本识别在一个完整的端到端系统中的关键作用，该系统将我们的脚本识别方法与以前发布的文本检测器和现成的 **ocr 引擎结合**在一起。少

2017 年 2 月 1 日提交;**v1** 于 2016 年 2 月 24 日提交;**最初宣布** 2016 年 2 月。

66. 利用主动学习在历史文献中进行字体识别

作 者 :anshul gupta, ricardo gutierrez-osuna, matthew christy, richard furuta, laura mandell

摘要: 识别历史文献中使用的字体类型（例如罗马字体、黑信字体）可以帮助光学字符识别（**ocr**）系统产生更准确的文本转录。为此，我们提出了一个主动学习策略，可以显著减少训练字体分类器所需的标记样本数量。我们的方法提取基于图像的功能，这些功能利用了字体在单词级别上的几何差异，并将它们合并为文档中每个页面的单词包表示形式。我们根据不确定性、差异和多样性标准评估六种采样策略，并在包含 3000 多个带有黑信、罗马和混合字体的历史文档的数据库中对其进行测试。我们的结果表明，不确定性和多样性的组合实现了最高的预测精度（89% 的测试用例正确分类），而只需要一小部分数据（17%）的标签。我们讨论了这一结果对历史文献大规模数字化项目的影响。少

2016 年 1 月 26 日提交;**最初宣布** 2016 年 1 月。

67. coco-文本：自然图像中文本检测与识别的数据集和基准

作 者 :andrias veit, tomasmatera , lucas neumann, jiri matas, serge belongie

摘要: 本文介绍了 coco-text 数据集。近年来，像 sun 和 imagenet 这样的大型数据集推动了场景理解和对象识别的进步。coco-text 的目标是在自然图像中推进最先进的文本检测和识别技术。该数据集基于 ms coco 数据集，其中包含复杂日常场景的图像。收集这些图像时没有考虑到文本，因此包含各种各样的文本实例。为了反映自然场景中文本的多样性，我们用 (a) 边界框的位置对文本进行注释，(b) 对机器印刷文本和手写文本进行细粒度分类，(c) 将文本分类为可辨认和难以辨认的文本，(d)文本和 (e) 易读文本的抄本。数据集包含超过 63k 图像中的 173k 文本批注。我们提供了注释准确性的统计分析。此外，我们还对我们数据集上三种最先进的光字符识别（**ocr**）方法进行了分析。虽然场景文本检测和识别近年来取得了长足的进步，但我们发现了推动未来工作的重大缺陷。少

2016 年 6 月 19 日提交;**v1** 于 2016 年 1 月 26 日提交;**最初宣布** 2016 年 1 月。

68. 数字和游戏：全景与艺术

作者:mathieu andro, imad 萨利赫

摘要: 本文概述了用于标记或 **ocr** 校正的数字图书馆的主要游戏化项目。本概述之后是一个最先进的功能，动机，贡献者的社会学和游戏化的范围相比，严重的游戏和明确的众包。最后，对显式众

包和游戏化的结果进行了比较。[英文标题：数字图书馆与游戏化：概述和最新之处]少

2015 年 12 月 28 日提交;最初宣布 2015 年 12 月。

69. 光学字符识别中的序列学习序列

作者:devendra kumar sahu, mohak sukhwani

摘要：针对印刷文本光学字符识别（**ocr**），提出了一种基于端到端的递归编码解码器序列学习方法。与当今使用连接式时间分类（ctc）输出层的最先进的 **ocr** 解决方案不同，我们的方法对序列的结构和长度进行了简约的假设。我们使用两步编码器解码器的方法——(a) 一个重复的编码器读取可变长度的印刷文本单词图像，并将其编码为固定的尺寸嵌入。(b) 这种固定尺寸嵌入随后由解码器结构理解，该解码器结构将其转换为可变长度的文本输出。我们的架构提供了相对于连接器时间分类（ctc）输出层的竞争性能，同时在更自然的环境中执行。从编码器中学习到的深字图像嵌入可用于基于文本的打印检索系统。任何可变长度输入的表达式固定维度嵌入都加快了检索任务的速度，使检索任务更加高效，这在其他递归神经网络体系结构中是不可能的。我们通过训练我们的网络进行预测任务的预测任务，在序列中的表达性和学习性的长期短期记忆（lstm）在序列学习的顺序，通过分割自由打印

文本 ocr 的经验性。通过对单词预测和检索这两个任务的定量和定性评价，证明了所提出的印刷文本体系结构的效用。少

2015 年 12 月 27 日提交;**v1** 于 2015 年 11 月 13 日提交;**最初宣布** 2015 年 11 月。

70. **通过一种新的预处理方法提高文档图像的 ocr 精度**

作者:abdeslam el harraj, naoufal raissouni

摘要: 数码相机和移动文档图像采集是光学字符识别和文本检测领域出现的新趋势。在某些情况下，这种过程集成了许多扭曲，并产生扫描不良的文本或文本照片图像和自然图像，导致不可靠的 ocr 数字化。本文提出了一种新的非参数和无监督的文档图像失真补偿方法，旨在优化 ocr 精度。我们的方法依赖于非常有效的文档图像增强技术堆栈来恢复整个文档图像的变形。首先，我们提出了一种局部亮度和对比度调整方法，以有效地处理照明变化和图像照明的不规则分布。其次，我们使用优化的灰度转换算法将文档图像转换为灰度级别。第三，我们使用非锐化蒙版方法锐化生成的灰度图像中的有用信息。最后，采用最优全局二值化方法对 ocr 识别的最终文档图像进行了准备。该方法能显著提高文本检测率和光学字符识别精度。为了证明我们方法的有效性，对标准数据集进行了详尽的实验。少

2015 年 9 月 11 日提交;**最初宣布** 2015 年 9 月。

71. ocr 扩展名-本地标识符、标记的 guid、文件 io 和数据块分区

作者:jiri Dokulil , siegfried benkner

摘要: 我们提出了几个延长开放社区运行时（**ocr**）规范的建议。扩展是具有本地有效性的标识符，它使用期货的概念为 **ocr** 实现提供更多的优化机会，标记带有创建者函数的 guid，这些标识基于本地标识符，并允许开发人员创建在并发创建对象的情况下不受竞争条件影响的 **ocr** 对象数组，一个简单的文件 io 接口，它建立在现有数据块概念的基础上，最后是数据块分区，这允许在多个任务希望访问数据块的不相交部分的情况下，更好地控制和灵活。少

2015 年 9 月 10 日提交;**最初宣布** 2015 年 9 月。

72. 噪声源的主题稳定性

作者:jing su, oisín boydell, derekgreene , gerard 林奇

文摘: 主题建模技术，如 lda，最近被应用到语音记录和 **ocr** 输出。这些语料库可能包含嘈杂或错误的文本，可能会破坏主题稳定性。因此，了解主题建模算法在应用于噪声数据时的性能非常重要。本文表明，不同类型的文本噪声会对不同主题模型的稳定性产生不同的影响。通过这些观察，我们提出了文本语料库生成的指导

原则，重点是自动语音转录。我们还提出了噪声语料库的主题模型选择方法。少

2015 年 8 月 5 日提交;**最初宣布** 2015 年 8 月。

73. 在线争用解析方案

作者:moran feldman, ola svensson, rico zenklusen

摘要: 我们介绍了一种新的在线优化问题舍入技术，该技术与争用解析方案有关，该技术最初是在子模块化函数最大化的背景下引入的。我们的舍入技术，我们称之为在线争用解决方案（ocrs），适用于许多在线选择问题，包括贝叶斯在线选择、忘记张贴定价机制和随机探测模型。它允许处理一组广泛的约束，并共享脱机争用解决方案的许多强大属性。特别是，可以将不同约束族的 ocrs 组合在一起，为它们的交集获得 **ocrs** 。此外，我们还可以在我们考虑的在线设置中近似地最大化子模块化功能。因此，我们得到了一个广泛适用的在线选择问题框架，该框架在可处理的约束类型、可以处理的客观功能以及对对手。此外，我们还从文献中解决了两个悬而未决的问题;即，我们提出了第一个常量因子约束的显性张贴价格机制的矩阵约束约束，以及第一个加权随机探测与最后期限的恒因子算法。少

2015 年 10 月 14 日提交;**v1** 于 2015 年 8 月 1 日提交;**最初宣布** 2015 年 8 月。

74. 增强光学字符识别：一种超分辨率方法

作者:赵东,朱锡梅,邓玉斌，陈易武,余巧

摘要: 文本图像超分辨率是计算机视觉界一个具有挑战性但开放的研究问题。特别是低分辨率图像妨碍了典型光学字符识别（**ocr**）系统的性能。在本文中，我们总结了我们进入 icdar2015 文本图像超级分辨率竞赛的内容。实验以 icdar2015 textsr 数据集和发布的 tesseract-ocr 3.02 系统为基础。我们报告说，我们赢得的文本图像超分辨率框架的输出极大地提高了低分辨率图像作为输入**的 ocr** 性能，达到了 77.19 的 **ocr** 准确率，这与使用的精度得分相当原始高分辨率图像为 11.80%。少

2015 年 6 月 6 日提交;**最初宣布** 2015 年 6 月。

75. 信德相关与阿拉伯语脚本改编语言识别的研究

作者:dil nawaz hakro, a. z. talib, zeeshan bhatti, g. n. moja

摘要: 大量出版物可用于光学字符识别（**ocr**）。大量的研究，以及文章是为拉丁文，中文和日文脚本。从 **ocr** 的角度来看，阿拉伯语脚本也是成熟的脚本之一。共享阿拉伯语脚本或其扩展字符的自适应语言;仍然缺乏 **ocr** 为他们的语言。在本文中，我们介绍了研究人员在阿拉伯语及其相关和适应语言方面的努力。本次调查分为不同的部分，其中介绍后是信德语的属性。介绍了 ocr 工艺技

术和方法，以及各研究人员使用的**方法**。最后一节专门对今后的工作进行了讨论，并对结论进行了讨论。少

2014 年 12 月 13 日提交;最初宣布 2014 年 12 月。

76. 现代光学字符识别技术综述

作者:尤金·博罗维科夫

摘要:数字文档识别领域的最新进展。以印刷文档图像为重点，讨论了光学字符识别（**ocr**）和文档图像增强在拉丁文和非拉丁文脚本中的应用中的主要进展。此外，我们还回顾和讨论了手写文档的可用技术..。更多

2014 年 12 月 12 日提交;最初宣布 2014 年 12 月。

77. 从非合作查询中进行高效媒体检索

作 者 :kevin shih, wei di , vignesh jagadeesh, robinson piramuthu

摘要: 文本在人造世界中无处不在，当涉及到书名和作者姓名时，很容易实现。利用斯坦福移动视觉搜索数据集的图书封面设置中的图像以及 open 编制. org 中的其他书籍封面和元数据，我们构建了一个大规模的图书封面检索数据集，完成了 100k 的封面和标题及作者。每个字符串。由于我们的查询图像对于干净的文本提取条件很差，我们提出了一种方法来提取匹配的**噪声和错误**

的 ocr 读数，并在标准的文档查找问题中将其与干净的作者和书名字符串进行匹配设置。最后，我们演示了如何将此文本匹配与流行的检索功能（如 vlad）结合使用简单的学习设置，实现与 vlad 或单独文本相比的检索准确性显著提高。少

2014 年 11 月 19 日提交;最初宣布 2014 年 11 月。

78. 利用 k-近邻域的光学字符识别

作者:王伟

文摘: 光学字符识别问题 ocr 在文献中得到了广泛的讨论。该程序有一个手写的文本，旨在识别文本。尽管有几种方法可以解决这个问题，但它仍然是一个悬而未决的问题。在本文中，我们提出了一种使用 k 最近的邻域算法，其精度超过 90% 的方法。训练和运行时间也很短。少

2014 年 11 月 5 日提交;**最初宣布** 2014 年 11 月。

79. 利用高斯模糊滤波器提高 captcha 的安全性

作者:阿里扬·扎雷伊

摘要: 为 web 服务器提供安全性，防止不需要的自动注册已成为一个大问题。为了防止这类虚假注册，许多网站使用 captcha。在各种基于 **appcha 或**视觉的 captcha 中，这种情况非常普遍。

实际上，视觉验证码是包含一系列字符的图像。到目前为止，大多数可视化 captcha 为了抵御 **ocr** 程序，使用一些常见的实现，如包装字符、随机放置和字符旋转等。本文将图像变换高斯模糊滤波器应用于视觉验证码，通过 **ocr** 程序降低其可读性。我们的结论是，这种技术使 captcha 程序的 captcha 几乎无法读取，但人类用户的可读性仍然很高. 少

2014 年 10 月 16 日提交;最初宣布 2014 年 10 月。

80. 三个运行时间的故事

作者:nicolas vasilache, muthu baskaran, tomhenretty, benoit meister, m. harper langston, sanket tavarageri, richard lethin

摘要: 此贡献讨论了从顺序 c 规范自动生成面向任务的执行模型的事件驱动、基于元组空间的程序。我们开发了一个分层映射解决方案，使用自动并行化编译器技术来定位三个不同的运行时，依赖于事件驱动的任务 (edt)。我们的解决方案受益于循环类型编码 edt 之间简短、可传递关系的重要观察，这些关系在运行时进行了紧凑且高效的评估。在这种情况下，可变循环特别重要，因为它们会立即转化为距离 1 的保守点对点同步。我们的解决方案将调用生成到一个运行时无关的 c++ 层，我们已将其重定向到英特尔的并发集合 (cnc)、eti 的 swarm 和开放社区运行时 (**ocr**)。其他运行时系统的经验促使我们在 cnc 中引入了对分层

异步完成的支持. 提供了实验数据，以显示自动生成的基于 edt 的运行时代码以及跨运行时比较的好处。少

2014 年 9 月 5 日提交;**最初宣布** 2014 年 9 月。

81. 通过机器绕过验证码的证明

作者:ahmad b. a. h 哈桑 at

摘要: 在过去的十年中, captcha 被网站广泛使用，以防止其数据被计算机自动更新。通过应该只允许人类这样做, captcha 利用了反向图灵测试（tt)，知道人类比机器更聪明。一般来说，captcha 已经打败了机器，但随着技术的进步，情况正在迅速变化。因此，光学字符识别（**ocr**）的高级研究正在超越加强 captcha 以抵御机器攻击的尝试。本文研究了基于 tt 故障的 captcha 的免疫研究。我们表明，一些 captcha 很容易被破坏使用一个简单的 **ocr** 机器，为本研究的目的。通过回顾其他技术，我们表明，使用先进的 **ocr** 机器可以破坏更困难的 captcha。**目前 ocr** 技术的进步应该使机器能够通过图像识别领域的 tt，而这正是机器寻求克服 captcha 的地方。我们不仅采用字符，而且还采用自然语言和同一 captcha 中的多个对象来增强传统的 captcha。拟议的 captcha 或许能够抵御机器，至少在完全通过 tt 的机器出现之前是如此。少

2014 年 9 月 2 日提交;**最初宣布** 2014 年 9 月。

82. 从运行长度压缩打印文本文档中直接提取线条文字符段

作者:mohammed javed, p. nabhhushan, b. b. chaudhuri

摘要: 将文本文档分割成行、字和字符，这被认为是光学字符识别（ocr）中的关键预处理阶段，传统上是在未压缩的文档上进行的，尽管大多数由于传输和存储效率等原因，在现实生活中的文档以压缩形式提供。但是，这意味着应解压缩压缩图像，这将缩进其他计算资源。这一局限性促使我们利用压缩文档进行文档图像分析研究。本文认为，在运行长度压缩打印文本文档中，我们以一种新的方法对行、词和字符进行分割。我们从压缩文件中提取水平投影轮廓曲线，并使用局部极小值点执行线分割。但是，跟踪垂直信息，从而在运行长度的压缩文件中跟踪单词字符，并不是很直接。因此，我们提出了一种新的方法，通过弹出一个智能序列从每一行的列运行进行单词和字符的同步分割。从 35 个噪音和 35 个免费压缩文件的孟加拉语、卡纳达语和英语脚本中，使用 1101 条文本线、1409 个单词和 7582 字符对算法进行了验证。少

2014 年 3 月 30 日提交;最初宣布 2014 年 3 月。

83. 一种识别独立手写阿拉伯字符的新方法

作者:ahmed salol, cheng suen

摘要: 手写的阿拉伯识别系统面临着许多困难，如人类笔迹的无限变化、不同字符形状的相似性、相邻字符的相互联系及其在单词中的位置。典型的光学字符识别（ocr）系统主要基于预处理、特征提取和识别三个阶段。本文提出了一种基于新的预处理操作的手写阿拉伯字符识别新方法，包括不同种类的噪声去除，以及不同类型的特征，如结构、统计和形态特征从主体的字符，也从次要组分。对所选要素的准确性进行了评估。利用 cenprmi 数据集的反向传播神经网络对该系统进行了训练和测试。该算法能够准确识别我们 88% 的测试集，取得了很有希望的结果。在与其他相关作品的比较中，我们发现我们的结果在其他已出版作品中是最高的。少

2014 年 2 月 26 日提交;最初宣布 2014 年 2 月。

84. 从运行长度压缩文本文档直接自动检测字体大小

作者:mohammed javed, p. nabhhushan, b. b. chaudhuri

摘要: 字体大小的自动检测在智能 ocreing 和文档图像分析领域有许多应用，传统上在未压缩的文档上进行，尽管在现实生活中，文档以压缩形式存在高效的存储和传输。如果能够直接从这些文档的压缩数据中执行字体大小检测任务，而不进行解压缩，将节省大量的处理时间和空间，这将是新颖和智能的。因此，本文提出了一种新的思路，即利用简单的线高特征，在线路级别直接从运

行长度压缩文本文档中学习和检测字体大小，为智能 oring 和文档分析铺平了道路。直接从压缩文档。在所提出的模型中，将给定的不同字体大小的混合大小文本文档分割成压缩的文本行，并使用提取的线高和提升高度等特征以回归线的形式捕获字体大小的图案，使用它在识别阶段进行字体大小的自动检测。该方法由 50 个压缩文档组成的数据集进行实验，其中包括 780 条单字符大小的文本行和 375 条混合字体大小的文本行，总精度为 99.67。少

2014 年 2 月 18 日提交;最初宣布 2014 年 2 月。

85. 视频序列中的孟加拉语识别：一个新的焦点

作者:souvik Bhowmick, purnendu banerjee

摘要: 由于背景复杂、分辨率低等特点，从视频帧图像中提取和识别孟加拉语文本具有挑战性。本文提出了一种复杂背景下的孟加拉语文本的提取和识别算法。在这方面，提出了两步走的办法。首先，利用基于线条轮廓的信息将文本行分割成单词。文本块的一阶梯度值用于查找单词间隙。接下来，在每个单词上应用局部二值化技术，并使用这些单词重建文本行。其次，将此二值化文本块发送到 ocr 进行识别。少

2014 年 1 月 6 日提交;最初宣布 2014 年 1 月。

86. 对古代文献进行分类

作者:nizar zaghden, remi mullot, mohamed adel alimi

摘要: 鉴于可以提取的信息的重要性以及各机构对保护其遗产的重视，对历史文件的分析仍然是一个热点问题。为了在试图清理图像后对古代文献图像的内容进行描述，主要思想是从同一图像中分割块文本，并试图在同一图像或整个图像数据库中找到类似的块。大多数离线手写识别方法都是通过将单词分割成较小的部分（通常是字符）来进行的，这些片段是单独识别的。然后，识别一个单词需要识别组成它的所有字符（ocr）。我们的工作主要集中在旧文档图像中的类的表征上。我们使用 som 工具箱在文档中查找类。我们还应用分形维数和兴趣点对古代文献进行分类和匹配。少

2013 年 8 月 28 日提交;最初宣布 2013 年 8 月。

87. k 算法一种改进的手写文件噪声去除技术

作者:kanika bansal, rajiv kumar

文摘: 自过去几十年以来，ocr 一直是一个活跃的研究领域。ocr 执行对扫描文档图像中文本的识别，并将其转换为可编辑的形式。ocr 过程可以有几个阶段，如预处理、分段、识别和后处理。预处理阶段是 ocr 成功的关键阶段，主要处理噪声去除问题。本文提出了一种改进的噪声去除技术--k 算法，该技术分为滤波和二值

化两个阶段。与中值滤波技术相比，该技术显示了即兴的结果。少

2013 年 6 月 6 日提交;最初宣布 2013 年 6 月。

88. 阅读古代钱币传奇：对象识别与 ocr

作者:albert kavelar, sebastian zambanini, martin kambel

摘要: 标准 **ocr** 是计算机视觉的一个研究性的课题，可以被认为是机器印刷文本的解决方法。但是，当应用于不受约束的图像时，识别率会大幅下降。因此，基于对象识别的技术的应用已经成为场景文本识别应用的最先进的领域。本文提出了一种适合古代钱币传说的场景文本识别方法，并将字符和单词识别实验的结果与标准 **ocr** 发动机进行了比较。实验表明，该方法在一组 180 枚铸币图例词上优于标准 **ocr** 引擎。少

2013 年 4 月 26 日提交;最初宣布 2013 年 4 月。

89. 识别手写卡纳达数字的分类器融合方法

作者:h. r. mamatha, s. karthik, murthy k. srikanta

文摘: 光学字符识别（**ocr**）是图像处理和模式识别领域的重要领域之一。手写字符识别一直是一项具有挑战性的任务。只有一点点的工作可以追溯到承认手写的字符南印度语言。卡纳达就是这

样一种南印度语言，也是印度的官方语言之一。由于卡纳达字符之间的高度相似性，对卡纳达字符的准确识别是一项具有挑战性的任务。因此，需要提取出高质量的特征，并需要更好的分类器来提高 kannada 字符 ocr 的准确性。本文探讨了运行长度计数（rlc）和定向链码（dcc）等特征提取方法在手写 kannada 数字识别中的有效性。本文采用分类器融合方法，提高了识别率。对于分类器融合，我们考虑了 k 最近邻居（knn）和线性分类器（lc）。该方法的新颖性是利用分类器融合方法，在几乎没有特征的情况下获得更好的精度。提出的方法平均识别率为 96%。少

2013 年 1 月 1 日提交;最初宣布 2013 年 1 月。

90. nf-savo：阿拉伯语视频 ocr 的神经模糊系统

作者:mohamed ben halima, hichem kj 弄,adel。m. alimi, ana fernández vila

文摘: 本文提出了一种鲁棒的视频剪辑文本提取和识别方法，即阿拉伯语视频 ocr 的神经模糊系统。在阿拉伯语视频文本识别中，一些噪声组件提供了相对复杂的文本与背景分离。此外，字符可以移动或以不同颜色、大小和字体的形式显示，而这些颜色、大小和字体并不统一。除此之外，背景通常在移动，这使得文本提取成为一个更加复杂的过程。视频包括两种文本，场景文本和人工文本。场景文本通常是在拍摄场景时记录的成为场景本身一部分的文本。但人工文本是单独制作的，远离现场，在稍后阶段或处理

后的时间内放置在场景上。因此，人为文本的出现受到了警惕的指导。这种类型的文本带有重要的信息，有助于视频引用、索引和检索。少

2012 年 11 月 9 日提交;**最初宣布** 2012 年 11 月。

91. 数字化旧报文章提取的逻辑分割

作 者 :thomas palfray, david hébert,stphane nicolas, pierrick tranouez,thierry paquet

摘要: 报纸是由新闻和信息文章制成的文件。它们并不是反复变红的: 读者可以按照他喜欢的任何顺序挑选他的物品。忽略了这一结构属性，大多数数字化的报纸档案只提供访问的问题，或最多按页面访问其内容。我们已经建立了一个数字化工作流程，自动从图像中提取报纸文章，从而允许在文章级别对信息进行索引和检索。我们的后端系统提取页面的逻辑结构，以生成信息性单元: 文章。每个图像都在像素级标记，通过基于机器学习的方法，然后通过检测结构实体（如水平和垂直分隔符、标题和文本行）来构造页面逻辑结构。此逻辑结构存储在与系统生成的 alto 文件（包括 **ocred** 文本）关联的 mets 包装中。我们的前端系统提供了一个网络高清可视化的图像，文本索引和检索设施，搜索和阅读的文章水平。文章转录可以合作更正，因此可以更好地编制索引。我们目前正在法国当地最大报纸之一的《鲁昂杂志》的档案中测

试我们的系统。这 250 年的出版量达 30 万页，图像质量和布局复杂度极易变。测试年 1808 可在 plair.univ-rouen.fr 咨询。少

2012 年 10 月 3 日提交;最初宣布 2012 年 10 月。

92. 对文字图像数据集的基准识别结果进行标记

作者:deepak kumar, m n anil prasad, a g ramakrishnan

摘要: 我们使用手动分割和当前可用的商业 **ocr**，在各种文字图像数据集上对可获得的最大识别精度进行了基准测试。我们开发了一个带有图形用户界面的 matlab 程序，用于单词图像的半自动像素级分割。我们讨论了像素级注释的优点。我们已经覆盖了五个数据库，这些数据库的图像加起来超过 3600 字。这些文字图像是从相机拍摄的场景、原始数字和街景图像中裁剪出来的。我们使用 nuance omnipage **ocr** 的试用版识别分段字图像。我们还讨论了在获取过程中引入的降级或在创建单词图像过程中引入的不准确如何影响图像中存在的单词的识别。还讨论了不同类型的退化和纠正词的倾斜和曲线性质的单词图像。在 icdar 2003、符号评估、街景、出生数字和 icdar 2011 数据集上获得的单词识别率分别为 83.9、8.9.3%、17.6%、八 8.5% 和 18.6.7%。少

2012 年 8 月 30 日提交;最初宣布 2012 年 8 月。

93. 利用主观人的注释化对历史报纸文章的聚类分析

作　者：haimonti dutta, william chan, deepak shankargouda, manoj pooleery, axinia radeva, kyle rego, boyi xie, rebecca passonneau, austinlee, 芭芭拉·塔兰托

摘要：纽约公共图书馆正在参与按时间排列的美国倡议，开发一个具有历史意义的报纸文章的在线可搜索数据库。对报纸的缩微胶卷副本进行扫描，并在报纸上运行高分辨率光学字符识别（ocr）软件。ocr 的文本为研究人员和历史学家提供了丰富的数据和意见。但是，ocr 引擎提供的文章的分类尚不成熟，大量文章被标记为编辑，而无需进一步分组。考虑到语料库的大小，手动将文章分类为细粒度类别即使不是不可能，也是非常耗时的。本文研究了报纸文章自动分类的技术，以加强对档案的检索和检索。我们探索无监督（例如 k 男男女女）和半监督（例如约束聚类）学习算法，以开发针对最终用户需求的文章分类方案。设计了一项试点研究，以了解顾客之间是否就如何对物品进行分类达成了一致意见。研究发现，这项任务非常主观，因此使用了能够处理主观标签的自动化算法。虽然小规模的试点研究对机器学习算法的设计非常有帮助，但需要开发一个更大的系统来收集档案用户的注释。目前正在开发的 "bodhi" 系统是朝着这一方向迈出的一步，允许用户纠正错误扫描的 ocr ，并为经常使用的报纸文章提供关键字和标签。在成功实施该系统的测试版时，我们希望它能够与正在为编年史美国项目开发的现有软件集成。少

2012 年 8 月 17 日提交;**最初宣布** 2012 年 8 月。

94. ocr 系统对其他印度语言（卡纳达语和印地语）的歧视

作者:ankit kumar, tushar patnaik , vivek kr verma

摘要: 印度是一个多语言多脚本国家。在印度的每个州都有两种语言，一种是州当地语言，另一种是英语。例如，在印度的安得拉邦，该文件可能包含英文和泰卢固语脚本中的文本词。对于这种双语文档的光学字符识别（ocr），有必要先识别脚本，然后再将文本单词输入到各个脚本的 ocr 。本文介绍了一种简单有效的印刷文档卡纳达、英语和印地语文本词的脚本识别技术。所提出的方法是基于水平和垂直投影剖面的歧视三个脚本。特征提取是根据每个文本单词的水平投影轮廓进行的。为了提取歧视特征，发展知识库，我们分析了卡纳达语、英语词和印地语的 700 个不同单词。我们使用每个文本单词的水平投影配置文件，并根据水平投影配置文件提取适当的特征。该系统在 100 种不同的文档图像上进行了测试，每个脚本包含 1000 多个文本单词，卡纳达语、英语和印地语的分类率分别为 98.25%、99.25 和 98.87%。少

2012 年 5 月 10 日提交;最初宣布 2012 年 5 月。

95. 预测直方图在马拉雅拉姆提高近距离字符识别中的频谱分析

作者:sajilal divakaran

摘要: 用于打印的 malalyam 文档的光学字符识别（**ocr**）系统的成功率令人印象深刻，各种文档的最先进的精度水平在 85–95 之间。然而，对于实际应用，需要进一步提高这一精度水平。其中一个瓶子颈在进一步提高精度被确定为紧密匹配的字符。本文描述了马拉雅拉姆语中的近配字符，并报告了这些紧密匹配字符的专用分类器的开发。获取 **ocr** 的最新技术的输出，并将放入紧密匹配字符集中的字符进一步输入到此专用分类器中，以提高准确性。该分类器基于支持向量机算法，利用近距离匹配特征投影直方图信号的光谱系数导出的特征向量。少

2012 年 5 月 8 日提交;**最初宣布** 2012 年 5 月。

96. 孟加拉 ocr 开发的完整工作流程

作者:farjana yeasmin ome, shiam shabbir himel, md. abu naser bikas

摘要: 开发孟加拉 **ocr** 需要大量的算法和方法。为发展孟加拉 **ocr** 作出了许多努力。但他们都没有提供一个没有错误的孟加拉 ocr . 他们每个人都有一些缺乏。我们讨论了现有孟加拉 **ocr** 的问题范围。在本文中，我们介绍了开发孟加拉 **ocr** 所需的基本步骤和开发孟加拉 **ocr** 的完整工作流程，并提到了所需的所有可能算法。少

2012 年 4 月 5 日提交;**最初宣布** 2012 年 4 月。

97. 基于谷歌在线拼写建议的 ocr 后处理纠错算法

作者:youssef bassil, mohammad alwani

摘要: 随着数字光学扫描仪的出现，许多纸质书籍、教科书、杂志、文章和文档正在被转换为电子版本，可由计算机操作。为此，开发了 ocr，光学字符识别简称，将扫描的图形文本转换为可编辑的计算机文本。不幸的是，ocr 仍然是不完善的，因为它偶尔识别字母和错误识别扫描的文本，导致 ocr 输出文本中的拼写错误和语言学错误。提出了一种基于上下文的后处理纠错算法，用于检测和纠正 ocr 非字和真实字错误。该算法是基于谷歌的在线拼写建议，利用内部数据库，其中包含从网络上收集到的大量术语和单词序列集合，方便地建议可能的替换的话，有在 ocr 过程中拼写错误。实验表明，ocr 纠错率有显著提高。今后的研究可以对该算法进行大量的改进，使其能够在多处理平台上实现并行化和执行。少

2012 年 4 月 1 日提交;最初宣布 2012 年 4 月。

98. 基于谷歌网站 1t 5 克数据集的 ocr 上下文敏感误差修正

作者:youssef bassil, mohammad alwani

摘要: 自计算时代开始以来，信息一直以数字方式表示，以便由电子计算机处理。当时大量出版了大量的纸质书籍和文件;因此，

有必要将它们转换为数字格式。**ocr** 是光学字符识别的简称，旨在将纸质书籍翻译成数字电子书。令人遗憾**的是, ocr** 系统仍然是错误和不准确的，因为它们在公认的文本中产生拼写错误，特别是在原始文件印刷质量较低的情况下。提出了一种后**处理 ocr** 上下文相关纠错方法，用于检测和纠正非字和真字 **ocr 错误**。这种建议方法的基石是使用谷歌 web 1t5 克数据集作为字典的单词拼写检查 **ocr** 文本。谷歌数据集包含了非常大的词汇和完全从互联网上获得的词汇和词汇统计，使其成为执行基于字典的错误校正的可靠来源。该解决方案的核心是三种算法的组合：错误检测、候选拼写生成器和错误校正算法，它们都利用从 google web 1t 5-gram 数据集中提取的信息。对不同语言书写的扫描图像进行的实验表明，**ocr** 错误校正率有了显著提高。随着未来的发展，该算法将进行并行化，以支持并行和分布式计算体系结构。少

2012 年 4 月 1 日提交;**最初宣布** 2012 年 4 月。

99. 使用基于 mlp 的分类器手写孟加拉语字母识别

作 者 :subhadip basu, nibaran das, ram sarkar, mahantapas kundu, mita nasipuri, dipak kumar basu

文摘: 这里介绍的工作涉及设计一个基于多层感知器（mlp）的分类器，用于识别手写的孟加拉字母使用 76 元素功能集孟加拉是第二最流行的脚本和语言在印度次大陆和第五世界上最流行的语言。为表示孟加拉字母的手写字符而开发的功能集包括 24 个阴影特

征、16 个质心特征和 36 个运行时间最长的特征。在训练样本和测试装置上，实验观察到设计用于该功能集的 mlp 的识别性能分别为 8646% 和 75. 05%。这项工作在开发手写孟加拉文的完整 ocr 系统方面具有重要的应用价值。少

2012 年 3 月 5 日提交;**最初宣布** 2012 年 3 月。

100. 一种基于 mlp 的手写 "孟加拉" 数字识别方法

作者:subhadip basu, nibaran das, ram sarkar, mahantapas kundu, mita nasipuri, dipak kumar basu

文摘: 本文的工作涉及设计一种基于多层感知器（mlp）的模式分类器，用于使用 76 元素特征向量识别手写的孟加拉语数字。孟加拉语是印度次大陆第二流行的文字和语言，也是世界第五大最流行的语言。为在这里表示手写的孟加拉语数字而开发的功能集包括 24 个阴影特征、16 个质心特征和 36 个运行时间最长的特征。在对 6000 个样本数据库进行实验时，该技术在对结果进行了 3 倍交叉验证后，平均识别率为 967.5%。它适用于手写的孟加拉语数字的 ocr 相关的应用，也可以扩展到包括手写的孟加拉字母字符的 **ocr** 。少

2012 年 3 月 5 日提交;**最初宣布** 2012 年 3 月。