[论文名称] 文文件图像大数据分析：研究现状与趋势

[作者]：刘成林　中国科学院自动化研究所 模式识别国家重点实验室
北京中关村东路 95 号，北京 100190

[摘要] 文文件图像普遍存在于人们的日常生活，包括个人笔记、公文、档案、票据、图书数据、古籍、网络文文件、自然场景文本等，具有数据量大、形式和风格多样、结构复杂等特点，是一个典型的大数据问题。半个多世纪以来，文文件图像分析和文字识别研究提出了大量有效的方法，相关技术取得了一些成功的应用。近年来深度学习的广泛使用带来了文字识别性能的大幅提升，但是相对文字识别的巨大应用需求，现有技术还有很多不足，从图像处理、分割、模式识别和机器学习的角度还有很多问题值得深入研究。本报告先介绍文文件图像大数据的特点、应用背景、相关的研究问题（图像处理、分割、图文内容和风格识别、语义分析、检索等），然后对代表性的方法和当前性能作简要介绍，最后对将来的研究和应用趋势进行讨论。