单位代码	10475
学　　号	104753160693
分 类 号	TP391

河南大學
硕士学位论文

大规模甲骨文数据集构建及算法研究

学科、专业：计算机应用技术
研究方向：文字识别
申请学位类别：理学硕士
申　请　人：王慧慧
指导教师：张重生 教授

二〇一九 年 六 月

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Large-scale Oracle Bone Inscriptions Dataset Construction and Algorithm Research

A Dissertation Submitted to
the Graduate School of Henan University
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science

By

Wang Huihui

Supervisor: Prof.Zhang Chongsheng

June,2019

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

关于学位论文独创声明和学术诚信承诺

本人向河南大学提出硕士学位申请。本人郑重声明：所呈交的学位论文是本人在导师的指导下独立完成的，对所研究的课题有新的见解。据我所知，除文中特别加以说明、标注和致谢的地方外，论文中不包括其他人已经发表或撰写过的研究成果，也不包括其他人为获得任何教育、科研机构的学位或证书而使用过的材料。与我一同工作的同事对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

在此本人郑重承诺：所呈交的学位论文不存在舞弊作伪行为，文责自负。

学位申请人（学位论文作者）签名：王慧慧

2019年6月24日

关于学位论文著作权使用授权书

本人经河南大学审核批准授予硕士学位。作为学位论文的作者，本人完全了解并同意河南大学有关保留、使用学位论文的要求，即河南大学有权向国家图书馆、科研信息机构、数据收集机构和本校图书馆等提供学位论文（纸质文本和电子文本）以供公众检索、查阅。本人授权河南大学出于宣扬、展览学校学术发展和进行学术交流等目的，可以采取影印、缩印、扫描和拷贝等复制手段保存、汇编学位论文（纸质文本和电子文本）。

（涉及保密内容的学位论文在解密后适用本授权书）

学位获得者（学位论文作者）签名：王慧慧

2019年6月24日

学位论文指导教师签名：张重生

2019年6月24日

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

摘 要

甲骨文是源自中国的古老文字，与古埃及的象形文字、古印度的哈拉巴铭文、古巴比伦的楔形文字并列为四大文明古国的文字文明标志。甲骨文的研究从 1899 年首次发现从未停止，然而甲骨文研究相当依赖古文字专业人员。当前甲骨文领域待解决的问题仍有很多，如甲骨缀合、字迹鉴定、未释字破译，语义分析等。目前已发掘出土甲骨共十五万余片，因年代久远甲骨比较脆弱，颠沛流离过程中易破损等等，极少有人可以直接接触甲骨，拓本图像具有及其重要的作用。绝大多数甲骨文研究者都是基于拓片图像进行相关研究。随着计算机的发展，甲骨文研究有了字库、输入法及大量的摹本资料等等。本文旨在构造一个大规模的甲骨文基准数据集并开展相关的甲骨文识别算法对比实验，验证已有算法的性能与优缺点。

已有的甲骨文识别算法所使用的数据来源于甲骨文数据库提取部分文字对应的少量图片样本或者人工生成数据集，文字图片背景较为单一，类别较少，噪声较少，无法将其研究成果直接应用于实际场景。而实际甲骨文拓片中的文字类别较多，仅可释字就有 1500-2000 字左右，还有 3000 左右的未释字。数据分布极度不均衡，部分文字出现频率极低。另外龟甲本身纹路、占卜过程中烧灼产生的裂纹、保存过程中破损、拓印技术的好坏等等多种原因造成拓片中噪声比较严重，且部分文字残损。

基于此现状，本文研究内容分为以下三个方面：

一、基于拓本图像，构建了一个按字标注的甲骨文基准数据集。该数据集可进行甲骨文检测与识别的相关研究。实验证明使用深度学习算法在六千余张训练样本的情况下可以取得较高的检测准确率，但识别准确率较低。

二、在构造甲骨文数据集和基准实验的基础上，进行了两个较小的改进工作：第一：改进基追踪降噪算法。在简单基准实验的基础上，确定了稀疏表示算法是解决甲骨文识别任务的有效方法。在此基础上改进基追踪降噪算法的支持集更新方式，使其在小样本数据集中准确率得到提升，适用于强噪声多类不均衡样本的文字识别，在较少的特征数目下取得较高的准确率。实验结果表明，改进后的基追踪降噪算法取得比深度学习算法及基追踪降噪算法更加优异的结果。并在准确率、运行时间、资源使用率三方面达到均

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

衡。第二：将批处理策略应用于非负二次规划活动集算法。为了解决基于稀疏表示的非负二次规划活动集算法在样本量过大时占用内存过多而造成内存溢出的问题，本文提出将批处理策略应用于非负二次规划活动集算法并设计多个分类器，取最优分类器实验结果作为最终预测准确率。实验证明，批处理非负二次规划活动集方法在解决样本量过大内存溢出问题的同时，在部分数据集上性能比原始算法更加优越。

三、在此基础上，进行了基于非稀疏表示、深度学习和稀疏表示三类算法共 33 种算法的实验对比分析。根据算法准确率和运行时间等条件筛选出 13 种算法进行 15 个数据集的大规模实验并分析其准确率、运行时间和资源利用情况等以探寻最适合甲骨文数据集的算法。

关键词：甲骨文拓片文本识别，甲骨文数据集，稀疏表示，有效集，批处理

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ABSTRACT

Oracle bone inscriptions is an ancient script originated from China, along with the hieroglyphics of ancient Egypt, the Harabar inscriptions of ancient India, and the cuneiforms of Babylon, are considered as the symbols of the four civilizations. The study of oracle bone inscriptions started since its first discovery in 1899. The study of oracle bone inscriptions relies heavily on professionals of ancient Chinese characters. At present, there are still many problems to be solved in the field of oracle bone inscriptions, such as oracle bone conjugation, handwriting identification, unscrambled character interpretation, semantic analysis and so on. At present, more than 150,000 pieces of oracle bones have been unearthed. Because of the fragility of oracle bones, they were easy to be damaged during the process of displacement. Few people can directly investigate the original Oracle bones, but utilize rubbing image is specific and important. With the development of computer technology, the research of oracle bone inscriptions has a font library, input method and a large number of copy materials. The aim of this paper is to construct a large-scale oracle bone inscription benchmark data set and carry out comprehensive experiments on this dataset to verify the performance of existing algorithms.

Existing oracle bone recognition algorithms use data from oracle bone inscriptions database to extract a small number of sample crops corresponding to part of the scripts on the oracle bones. These character crops present simple background, few category, less noise, so the research algorithms developed on such dataset can’t be directly applied to the real rubbings. In fact, there are many types of characters in the actual Oracle rubbings, only 1500-2000 words can be deciphered, with the other 3000 characters undeciphered. The distribution of data is extremely uneven, and the frequency of some words is very low. In addition, on the oracle bones, cracks caused by burning in divination process, or damaged in preservation process, and because of the quality of rubbing techniques and so on, there exists serious noise in rubbings.

Based on the above, the research content of this paper can be divided into three aspects:

First, based on the rubbing images, a character-level-labeled oracle bone inscription data set is constructed. The data set can be used for text detection and recognition. Baseline experiments show that the

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

deep learning algorithm can achieve higher detection accuracy with 6,000 training samples, but the recognition accuracy is relatively low.

Second, based on oracle bone inscription benchmark data set and the baseline experiments, two minor improvents works have been proposed: First, we improve the basis pursuit de-noising algorithm. We find the sparse representation algorithm to be an effective method to solve the oracle script recognition task. On this basis, the updating method of the support set of the base tracking denoising algorithm is adjusted to improve the accuracy of the small sample data set. It is suitable for the text recognition of the strong noise and multi-class unbalanced samples, and achieves a higher accuracy under a small number of features. The experimental results show that the adapted base-tracking denoising algorithm achieves better results than the deep learning algorithm and the original base-tracking denoising algorithm. And it achieves a balance in accuracy, running time and resource utilization. Next, the batch strategy is applied to the activity set algorithm using non-negative quadratic programming. In order to solve the problem of memory overflow caused by excessive memory occupied by the sparse representation-based method when the sample size is too large, this paper proposes to apply batch processing strategy to the non-negative quadratic programming activity set algorithm, we design multiple classifiers, and take the experimental results of the optimal classifier as the final prediction accuracy. Experiments show that the batch non-negative quadratic programming activity set method has better performance than the original algorithm on a few data sets more importedly, it mitigates solving the problem of memory overflow with too large sample size.

Third, we carry out benchmark experiments on 33 algorithms which include non-sparse representation, deep learning and sparse representation ones. Then, 13 algorithms are selected to conduct large-scale experiments on 15 data sets. We report and compare their accuracy, running time and resource utilization, to find the most suitable algorithm for Oracle data sets.

Keywords: Oracle rubbing recognition; Text recognition; Sparse representation; Effective set; Batch processing

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目 录

摘 要 I
ABSTRACT III
1 绪论 1
1.1 研究背景及意义 1
1.2 国内外研究现状 3
1.3 本文研究内容 5
1.4 文本组织结构 5
2 相关工作 9
2.1 相关基础符号说明 9
2.2 具有不同范数正则的稀疏表示问题 11
2.2.1 具有 l_0 范数最小化的稀疏表示 11
2.2.2 具有 l_1 范数最小化的稀疏表示 13
2.2.3 具有 l_p 范数（$0 < p < 1$）最小化的稀疏表示 13
2.2.4.具有 $l_{2,1}$ 范数和最小化的稀疏表示 14
2.2.5.具有 l_2 范数最小化的稀疏表示 14
2.3 基于同伦思想的稀疏表示 14
2.3.1 最小绝对收缩选择算子同伦算法 14
2.3.2 基追踪降噪同伦算法 17
2.4 非负二次规划活动集算法 18
2.5 本章小结 21
3 甲骨文数据集构建 23
3.1 甲骨文数据集的提出背景 23
3.1.1 甲骨文研究数据现状 24
3.1.2 甲骨文拓片本身的特征及其造成的原因 26
3.2 数据集描述 29
3.2.1 数据集标注方式及工具 29
3.2.2 文本检测数据集 31
3.2.3 文本识别数据集 31
3.3 基线实验 35
3.4 本章小结 35
4.基于甲骨文数据集的大规模实验对比分析 37
4.1 改进的基于同伦思想的基追踪降噪算法简述 37
4.2 非负二次规划活动集算法批处理策略与分类器设计 39
4.3 实验设计 40
4.3.1 评价指标 40
4.3.2 软硬件配置 41
4.3.3 数据集 41
4.3.4 实验参数 41

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.4 实验结果分析 42
4.4.1 基准实验对比分析 42
4.4.2 大规模实验对比分析 44
4.5 本章小结 49
5 总结与展望 51
5.1 总结 51
5.2 展望 51
参考文献 53
附 录 59
附录 A 基准算法基本信息表 59
附录 B 大规模实验基本信息表 61
致 谢 63
攻读学位期间发表的学术论文 65

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1 绪论

大数据时代，机器学习技术得到进一步研究发展，并与其他学科快速交叉融合。甲骨文是源自中国的古老文字，与古埃及的象形文字、古印度的哈拉巴铭文、古巴比伦的楔形文字并列为四大文明古国的文字文明标志。其相关研究从 1899 年首次发现殷商龟甲从未停止，然而甲骨文的研究依赖专业人员，一个专业的甲骨文研究人员的培养需要数十年的时间。随着计算机的发展，甲骨文的研究也越来越现代化，有了专属的字库和输入法，可以在计算机终端输入呈现，但在甲骨文自动识别方面研究进展缓慢。基于此现状，本文开展了相关研究。

1.1 研究背景及意义

文字识别的目的是让计算机自动化识别图片中的文字。按照其不同场景，可以分为：打印版文件的扫描件、广告宣传页、手写字体、自然场景等。每种类型的图片都有其相应的算法。当前深度学习算法研究较为热门，很多研究领域尝试使用基于深度学习的方法。图像文字的检测与识别问题也是深度学习的一个重要研究领域。与这类图像相比，甲骨文拓片中的甲骨文识别面临一些独有的问题：一、甲骨文本身是一种雕刻文字，本质是一种手写体文字，同一个字的字形多样，形态不一；二、因为甲骨宝贵，相关研究更多是基于拓片的研究，拓印技术也会影响字的清晰程度；三、甲骨本身纹路以及甲骨保存过程中造成的裂损对于识别有一定的影响；四、数据量的问题，甲骨文的文字字频基本符合齐普夫定律，即文字字频的分布不均衡，且相差极大，所以甲骨文文字识别是多类不均衡的文字识别问题；五、因为甲骨文研究属于专业性很强的研究，其标注比较困难，样本数据量比较少。综上所述，甲骨文拓片中的文字识别面临同类样本之间差异性大，样本量少，样本种类多等一些不同于以往文字识别的特点。本文整理并标注了一个甲骨文拓片数据集并尝试了一些算法并进行相关基准实验研究，在 2019 年甲骨文发现 120 周年纪念的时间点上，为中国古文字的传承与发展贡献绵薄之力。

趁着深度学习的兴起和不断发展,本文也尝试使用了深度学习算法进行甲骨文识别。但因为甲骨文数据集自身的一些特性，其准确率并不理想。故针对甲骨文拓片数据集，在大量对比分析了多类基准算法之后选用了准确率比较高的稀疏表示方法进行研究并

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

取得比其他一些基准算法更加优越的结果。

近年来稀疏表示备受关注，根据多个研究领域很多例子可以发现，稀疏表示[1]是可行的且效果良好。其中一个典型示例是图像分类[2]，图像分类的基本目标是将给定的测试图像赋予已经预先定义多种类别标签的一种。研究成果显示，从视觉神经元的特性的角度可以稀疏地表示自然图像。稀疏表示分类（Sparse Representation Classifier ，SRC）方法[3]假设测试样本可以由训练样本充分表示，SRC 利用训练样本的线性组合来表示测试样本并计算线性表示的稀疏表示系数，然后使用稀疏表示系数和训练样本计算每个类的重建残差，测试样本将被归类为最小重建残差的类。Wright[3]也证明了 SRC 法在处理残损或遮挡图像上的图像分类问题时具有很大的优势。在这种情况下，可以稀疏地表示每个自然图像，并且可以利用稀疏表示理论来完成图像分类任务。基于稀疏表示理论处理残损或遮挡图像上的图像分类问题时的优势，将其应用于到甲骨文拓片的文字识别，有望解决包含甲骨文一些损坏或者残损的文字的识别问题，获得比传统方法更高的识别准确率。

稀疏表示理论可以从不同的角度进行分类。由于不同的方法出自于各自的动机、想法和关注点，因此从分类学的角度将现有的稀疏表示方法分成不同的类别有各种各样的策略。比如：第一：从“原子”的角度来看，稀疏表示方法可以分为两大类：简单样本的稀疏表示和字典学习的稀疏表示。第二：从是否使用样本标签的角度来看，稀疏表示方法可以分为三类：监督学习，半监督学习和无监督学习。第三：从不同的稀疏约束方法，可以分为两类：结构约束的稀疏表示和稀疏编码的稀疏表示。第四：在图像分类领域，基于表示的分类方法在开发“原子”的方式上包括两类：整体表示和局部表示[4]，详细来说就是基于整体表示的方法利用所有类的训练样本来表示测试样本，而基于局部表示的方法仅使用每个类或几个类的训练样本（或原子）来表示测试样本。大多数稀疏表示方法是基于整体表示的方法。典型的局部稀疏表示方法是两段测试样本稀疏表示（two-phase test sample sparse representation，TPTSR）方法[5]。考虑到不同的方法，稀疏表示方法可以分为两类：纯稀疏表示和混合稀疏表示，它借助于其他方法改进了预先存在的稀疏表示方法。文献[6]认为稀疏表示算法大致分为三类：凸松弛，贪婪算法和组合方法。在文献[7],[8]中，从稀疏问题建模和问题求解的角度来看，稀疏分解算法一般分为两个部分：贪心算法和凸松弛算法。另一方面，如果考虑优化的观点，稀疏表示的问题可以分为四个优化问题：平滑凸问题，非光滑非凸问题，平滑非凸问题和非光滑凸问题。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

此外，施密特等人[9]回顾了一些求解 l_1 范数正则化问题的优化技术，并将这些方法粗略地划分为三种优化策略：子梯度方法，无约束近似方法和约束优化方法。本文在尝试了贪婪策略近似，约束优化策略，基于邻近算法的优化策略和基于同伦算法的稀疏表示四种稀疏表示比较有的代表性的算法之后选择了时间效率最高的基于同伦策略的基追踪降噪算法并加以优化。同伦算法的一般框架是通过连续调整同伦参数来迭代地追踪从初始点到最佳点的最终期望解[7]。在基于同伦算法的稀疏表示中，同伦算法用于求解具有 k 稀疏性质的 l_1 范数最小化问题。在此基础上将 l_1 范数最小化的非负二次规划活动集方法应用于甲骨文拓片识别，实验表明，稀疏表示方法是解决甲骨文拓片识别的有效算法。

1.2 国内外研究现状

随着数学研究和计算机性能的提升与更新，线性表示方法[10],[11]（Linear Representation Methods，LRBM）发展极快，很多研究人员已经对该方向有所研究并提出多种改进方法。LRBM 算法的代表方法之一的稀疏表示方法，是应用于多个领域多个具体问题的强大解决方案：信号处理、图像处理、机器学习、计算机视觉等多个领域；图像去噪、去模糊、图像修复、图像恢复、超分辨率、视觉跟踪、图像分类和图像分割[12]-[19]等具体问题，稀疏表示在处理这些问题方面显示出巨大的潜力。

稀疏表示理论从其兴起与发源来看，与压缩感知（Compressed Sensing，CS）直接相关[20]-[22]，Donoho[20]提出压缩感知的概念，CS 理论认为，如果信号稀疏或可压缩，则可以通过一些测量值重建原始信号。Candès 等[22]从数学的角度证明 CS 理论的基本原理，即通过利用傅立叶变换系数的一小部分，可以精确地重建原始信号。Baraniuk[21]对压缩感知进行具体分析，且分析了不同信号重构算法。所有这些文献[20]-[26]为 CS 理论奠定了基础，为今后的研究提供了理论依据。当前已提出很多解决不同问题的基于 CS 理论的算法。CS 理论包括三个基本成分：稀疏表示、编码测量、重构。作为 CS 理论的先决条件，稀疏表示理论[13],[16]-[19],[26]是用于克服许多领域出现的困难的最优秀技术之一。例如，作为一种用于稀疏或可压缩信号的新型信号采样方法，稀疏表示的方法已成功应用于信号处理[13]-[15]。

对于信号处理，一个重要的任务是根据不同的要求从大量杂波信号或复杂信号组中提取关键分量。在稀疏表示出现之前，信号稳定技术（Signal Sustain Technology，SST）和 Nyquist 采样法（NSL）是传统的信号采集方法，一般程序包括采样，编码压缩，传

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

输和解码。在 SST 和 NSL 的框架下，信号处理的最大困难在于从海量数据中进行尽量节约存储器的有效采样。在这种情况下，稀疏表示理论可以同时打破传统采样规则即 SST 和 NSL 的瓶颈，因此具有非常广阔的应用前景。稀疏表示理论提出了信号采样和编码压缩过程的集成。尤其是稀疏表示理论采用更高效的采样率，通过放弃 SST 和 NSL 的原始测量来测量原始样本，然后采用最优重建算法进行样本重构。在压缩感知的背景下，首先假设所有信号都是稀疏的或近似稀疏的[13],[15],[16]。与主信号空间相比，在稀疏性的约束下，可以大大减小可能信号集的大小。因此，已经提出了基于稀疏表示理论的大规模算法来有效地解决诸如信号重建和恢复之类的信号处理问题。为此，稀疏表示技术可以节省大量的采样时间和样本存储空间，该方法可行并效果显著。

文字识别是当前计算机视觉研究的一个热点，其中包括印刷体文字识别，手写体文字识别[27]，车牌识别，验证码识别，古籍识别，视频文字识别，自然场景文本识别等多种不同场景的文本识别。传统的文字识别，最初是从印刷体文字识别开始的，为后来其他的文字识别场景打下了坚定的基础，其基本流程大致为几部分：图像预处理（灰度化，二值化，倾斜检测与矫正，行字切分，平滑，规范化等）；版面处理（版面分析、版面理解、版面重构）；图像切分（行（列）切分和字切分）；特征提取（HOG，SIFT，直方图等）、匹配及模型训练、匹配；识别后处理等。当下数据爆炸的时代，需要海量样本的神经网络发展迅速，文本识别任务中使用的基础网络有：卷积神经网络（CNN,Convolutional Neural Network）[28],全卷积网络（FCN,fully convolutional network）[29]，空间变换网络（STN，Spatial Transformer Networks）[30]等；文本识别模型有：CRNN 模型，RARE 模型等；端到端模型有：FOTS Rotation-Sensitive Regression，STN-OCR 模型等。这些深度学习算法都是基于海量样本。然而甲骨文拓片的研究，因其没有大规模的标注数据，使用深度学习算法的准确率并不太理想。当前基于深度学习的甲骨文识别算法，使用的数据样本类别比较少，且使用了一定的数据增益手段[31]。而非深度学习的甲骨文算法研究，也是由印刷体的识别[32],[33]、纯净背景的拓片识别[34],[35]、一直发展到复杂背景的拓片识别[36]。相关算法也是多种多样：图论[32]、拓扑特征[33]、曲率直方图的傅里叶描述子[34]、网格特征提取[35]、Hopfield 网络[36]、 模板匹配[37]、电路模拟[38]、拓扑编码[39][39]、使用生成数剧集的草图识别[40][40]、支持向量机[41][41]、霍夫变换聚类[42][42]等。然而这些算法都是相对基本纯净的背景的甲骨文拓片的文字识别，而实际甲骨文拓片上的文字，很多因为龟甲本身纹路或者保存不善以及雕刻技术，拓印技术等造成的噪

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

声及其严重，当前算法的实用性不太强[31]-[42]。基于此现状，在导师的指导下，提出并整理了一个存在强噪音、高残损、多类不均衡的甲骨文检测与识别数据集，基于该数据集运行了多种算法进行基准实验，改进基追踪降噪算法增加其鲁棒性以提升识别准确率，且将批处理策略应用于非负二次规划活动集算法以解决其在数据量较大时内存溢出问题。实验结果表明，基于稀疏表示的基追踪降噪算法和非负二次规划活动集算法能够在一定程度上提升甲骨文拓片文字识别的效果。

1.3 本文研究内容

本文首先阐述了文字识别及甲骨文识别的研究现状，重点构建了一个大规模甲骨文基准数据集并使用大量已有方法对其进行了基准实验，揭示了基于稀疏表示的算法准确率较高。在此基础上改进了有效抗噪的基追踪降噪算法的活动集更新方式以增加算法的鲁棒性，提高了识别准确率。将非负二次规划活动集算法应用于甲骨文识别，并为了解决内存溢出问题将批处理策略加入算法且设计多个分类器，解决了内存溢出的问题，在部分数据集上准确性也有一定提高。而且，本文在运行多种基准实验的基础上筛选算法进行大规模实验以分析不同算法在甲骨文数据集上的准确率、运行时间等性能指标。

本文研究内容主要分为三方面：第一，标注并整理出一个基于真实甲骨文拓片的图像文本检测与识别数据集并运行部分算法作为基准实验。该部分主要介绍数据集和特征以及基准算法简单介绍。第二，本文对两个相关算法进行了细微的改进。具体而言，首先在分析研究抗噪性较强的基追踪降噪算法的基础上，改进算法支持集更新方式，增加算法的鲁棒性，提升了算法准确率。使用整理出来不同数据量的 15 个甲骨文拓片文字识别数据集，对改进前后的算法进行了实验，结果表明，改进后的基追踪降噪算法在甲骨文数据集上表现出比原始的基追踪降噪算法更加优越的性能。另外将原本用来高维生物信息分类的核稀疏表示分类经典算法之一非负二次规划活动集算法应用于甲骨文数据集，并添加批处理策略减少算法对于内存的依赖以解决内存溢出问题，设计多个分类器以获得更高准确率。第三，对 33 种算法在甲骨文基准数据集上实验对比分析，并从中筛选 13 种算法在 15 个子数据集上进行了大规模实验对比分析，以探寻适用于拓片甲骨文识别的算法和技术。

1.4 文本组织结构

由于甲骨文数据集与现有已公开的数据集具备的数据特征不同，造成使用现有算法

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

直接识别甲骨文拓片文字的准确率较低。为了提升甲骨文数据集的识别准确率，在运行多种基准算法的基础上，最终选择了抗噪性能强的稀疏表示方法典型代表之一：基于压缩感知理论的基追踪降噪算法作为研究的算法并加以改进，最终改变算法支持集的更新方式提升了算法的准确率。将非负二次规划活动集算法应用于甲骨文文字识别并添加批处理策略重新设计分类器以取得较为理想的实验结果。

第一章为绪论部分。首先分析了文字识别和甲骨文拓片文字识别研究现状，其次简单介绍稀疏表示的发展历程和相关分类，最后概述本文研究内容。

第二章为相关工作。首先规定了文中使用的一些公式符号基本定义，其次介绍一些不同范数正则的稀疏表示问题，再者概述基于同伦思想算法的稀疏表示，其中重点介绍了最小绝对收缩选择算子同伦算法、基追踪降噪同伦算法和非负二次规划活动集算法的公式推导和算法基本流程。

第三章为甲骨文数据集详细介绍。首先是数据集的提出背景：甲骨文研究使用的一些数据的现状、甲骨文拓片本身存在的特征及其造成的相关原因。其次是数据集的一些详细描述：先介绍了数据集标注的方式及其使用的工具，又详细说明文本检测数据集的一些特点及其文本识别数据集的一些数据特征和数据分布等情况。最后简单介绍基于该数据集运行的基准算法实验结果，以证明甲骨文数据集的挑战性。

第四章为算法改进与多种算法的实验设计及相关结果分析。首先是改进的基于同伦思想的基追踪降噪算法简述。其次非负二次规划活动集算法批处理策略与基于该策略下的分类器设计。再者，介绍了实验设计方面包含算法性能度量指标、软硬件环境、数据集的简单介绍、相关参数等。最后介绍了实验结果分析，实验分为两类，一类为多种算法基准实验，另外一类为筛选出部分算法大规模数据集实验，其结果分析包含算法的准确率、运行时间等。

第五章为总结和展望部分。首先对本文的工作进行总结：第一、整理并提出了一个比较通用的真实甲骨文拓片数据集。第二、改进基追踪降噪算法支持集更新方式以提升在小样本不均衡强噪声的情况下准确率。第三、将非负二次规划活动集算法应用于甲骨文识别并添加批处理策略以解决部分数据集样本量较多的情况下内存溢出问题并设计多个分类器以谋求最佳准确率。第四、进行了多种算法的基准实验和基于部分算法的大规模数据集的实验及其实验结果分析。最后提出本文对未来的工作展望，一方面甲骨文数据集目前的识别准确率较低，怎样获得更高的识别率是未来工作的一大重点。另一方

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

面稀疏表示算法在数据量较大的时候容易发生内存溢出问题，如何在保持准确率不下降的同时减少算法对于计算机资源的占用也是未来工作研究的一个重点，除此之外，稀疏表示算法运行时间较长，如何缩短运行时间也是未来的工作之一。除此之外，其它一些方法可能也能在一定程度上甲骨文识别的准确率。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 相关工作

基于甲骨文拓片的一些基本特征，本文将抗噪声能力比较强的稀疏表示算法应用于甲骨文拓片的识别。本章首先规定文中使用的符号说明，其次简单概述不同范数正则的稀疏表示问题，再者简单介绍基于同伦思想的最小绝对收缩选择算子算法和基于同伦思想的基追踪降噪算法，最后给出非负二次规划活动集算法简单伪代码。本论文中的主要算法的名称及其英文简称，以列表的形式附于附录 A 和 B 中。

2.1 相关基础符号说明

首先需要声明的是，本文中用到的变量都是实数。在此基础上，向量用小写字母表示：向量 a，表示为小写字母 a，向量 a 中第 i 个元素表示为 a_i。矩阵用大写字母表示：矩阵 A，表示大写字母 A；矩阵 A 中第 i 个元素表示为 A_i。

稀疏向量表示向量的一些元素为零且非零元素较少。使用一个线性组合的基础矩阵 $A \in R^{N\times N}$ 来表示一个样本 $y \in R^{N\times 1}$，其中 $y=As$，$s \in R^{N\times 1}$ 是加权系数的列向量。如果只有 k（$k<<N$）个元素非零，称样本或者信号 y 是 k-稀疏的。

两个 n 维向量 q 和 p 的标准内积被定义为：

$$\langle q, p\rangle = q^T p = q_1 p_1 + q_2 p_2 + \cdots + q_n y_n \tag{2-1}$$

来自实数矩阵 $m\times n$ 的的两个 n 维向量 Q（$Q \in R^{m\times n}$）和 P（$P \in R^{m\times n}$）的标准内积被定义为：

$$\langle Q, P\rangle = tr(Q^T P) = \sum_{i=1}^{m}\sum_{j=1}^{n} Q_{ij} P_{ij} \tag{2-2}$$

其中，$tr(A)$在线性代数被称为矩阵的迹，表示矩阵 A 左上到右下对角线（主对角线）上所有元素总和。

假设 $v=[v_1,v_2,\cdots,v_n]$是欧几里得空间中的 n 维向量，向量 v 的 p 范数或者说 l_p（$1\le p\le\infty$）范数表示如下：

$$\|v\|_p = \left(\sum_{i=1}^{n}|v_i|^p\right)^{1/p} = \sqrt[p]{(|v_1|^p + |v_2|^p + \cdots + |v_n|^p)} \tag{2-3}$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

当 p=2 时，公式（2-3）被称为 l_2 范数或者欧几里德范数，表示向量中每个元素平方之和最后再开平方的值；

当 p=1 时，公式（2-3）被称为 l_1 范数，表示向量中每个元素绝对值之和；

当 $p\to 0$,（$0<p<1$）时，公式（2-3）被称为 l_p 范数，表示向量中每个元素绝对值的 p 次方的和；

当 p=0 时，公式（2-3）被称为 l_0 范数，表示向量中非零元素的个数；

根据公式（2-3）,此时假设 $f(x)$是参数向量 x 上的 l_p（$p>0$）范数的函数，公式（2-3）可以写为：

$$\mathrm{f}(x)=\|x\|_p^p=(\sum_{i=1}^{n}|x_i|^{p}) \tag{2-4}$$

根据数学相关公式特性可知：

l_2 范数是凸、平滑、全局可微函数；

l_1 范数是凸、非平滑、全局不可微函数；

l_p 范数是非凸、非光滑、全局不可微函数；

l_0 范数是非凸、非光滑、不连续、全局不可微函数。

假设矩阵

$$\mathrm{V}=\begin{bmatrix} V_{11} & \mathrm{V}_{12} & \cdots & \mathrm{V}_{1\mathrm{n}} \\ \mathrm{V}_{21} & \mathrm{V}_{22} & \cdots & V_{2n} \\ \cdots & \cdots & & \cdots \\ V_{\mathrm{m}1} & V_{m2} & \cdots & V_{mn} \end{bmatrix}$$

其对应的 *forbenius* 范数，l_1 范数，l_2 范数，$l_{2,1}$ 范数或者说 R 范数分别被定义为：

$$\|\mathrm{V}\|_{\mathrm{F}}=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}V_{j,i}{}^{2}} \tag{2-5}$$

$$\|\mathrm{V}\|_{l_1}=\sum_{i=1}^{n}\sum_{j=1}^{m}|V_{j,i}| \tag{2-6}$$

$$\|\mathrm{V}\|_2=\delta_{\max}(V)=\sqrt{\lambda_{\max}(V^TV)} \tag{2-7}$$

其中 δ 是奇异值算子，V 的 l_2 范数是其最大奇异值。

$$\|\mathrm{V}\|_{2,1}=\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m}V_{j,i}{}^{2}} \tag{2-8}$$

如上所示，范数可以被视为向量 v 的长度的度量。两个向量 x 和 y 或矩阵 X 和 Y 之间的距离，可以通过它们的长度来测量差异，即

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\mathrm{dist}(x,y)=\|x-y\|_2^2, dist(X,Y)=\|X,Y\|_F \quad (2\text{-}9)$$

如上所示，两个向量 x，y 之间的长度可以使用 l_2 范数来表示，两个矩阵 X，Y 之间的长度可以使用 *forbenius* 范数来表示。

假设 $X\in R^{m\times n}$，并且 X 的秩表示为 $rang(X)=r$。X 的 *SVD* 被计算为

$$\mathrm{X}=U\Lambda \mathrm{V}^T \quad (2\text{-}10)$$

其中 $U\in R^{m\times n}$ 满足 $U^TU=I$，$V\in R^{n\times r}$ 满足 $V^TV=I$。U 和 V 的列分别称为 X 的左右奇异向量。另外 Λ 是对角矩阵，其元素由 X 的奇异值组成，即 $\Lambda=diag(\lambda_1,\lambda_2,\cdots,\lambda_r)$满足 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r>0$。此时，奇异值分解可以重写为：

$$\mathrm{X}=\sum_{i=1}^{r}\lambda_i u_i v_i \quad (2\text{-}11)$$

其中，λ_i，u_i，v_i 分别表示第 i 个奇异值，U 的第 i 列和 V 的第 i 列。

2.2 具有不同范数正则的稀疏表示问题

在本节中，简单描述了稀疏表示，并根据使用的范数正则化将其分组为不同的类别。稀疏表示的一般框架是利用一些样本的线性组合或“原子”来表示探针样本，以计算表示解，即这些样本或“原子”的表示系数，然后使用表示解决方案来重建所需的结果。表示导致稀疏表示，但是，可以通过强加于表示解[43],[44]的正则化器（或优化器）来主导。因此，根据优化器中使用的不同规范，稀疏表示方法可以大致分为五大类：

1.具有 l_0 范数最小化的稀疏表示[45],[46]；

2.具有 l_p 范数（$0<p<1$）最小化的稀疏表示[47]-[49]；

3.具有 l_1 范数最小化的稀疏表示[50]-[53]；

4.具有 $l_{2,1}$ 范数最小化的稀疏表示[54]-[58]；

5.具有 l_2 范数最小化的稀疏表示[错误!未找到引用源。] [5],[59],[60]。

2.2.1 具有 l_0 范数最小化的稀疏表示

假设 $x_1,x_2,\cdots,x_r\in R^d$ 是所有已知样本，矩阵 $X\in R^{d\times n}$（$d<n$）是由已知样本构成的矩阵或者说基础字典，且 X 应该为一个过完备字典。X 的每一列是一个样本，探针样本是 $y\in R^d$，是列向量。因此如果用所有已知样本用于近似代表探针样本，则应该表示为：

$$\mathrm{y}=x_1\alpha_1+x_2\alpha_2+\cdots+x_n\alpha_n \quad (2\text{-}12)$$

其中,$\alpha_i=(i=1,2,\cdots,n)$是 x_i 的系数。此时，公式（2-12）可以重写为：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\mathrm{y} = X\alpha \tag{2-13}$$

其中，矩阵 $X=[x_1,x_2,\cdots,x_n]$, $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_n]^T$。

然而，公式（2-13）是一个不确定的线性方程组，主要问题是如何解决它。从线性代数的角度来看，如果没有任何先验知识或任何约束强加于表示解，公式（2-13）是一个不确定的问题，永远不会有一个唯一的解决方案。也就是说，不可能利用方程式（2-13）来使用测量矩阵 X 唯一地表示探针样本 y。为了减轻这种困难，可以在表示方案上施加适当的正则化约束或正则化函数。稀疏表示方法要求获得的表示解应该是稀疏的。此后，“稀疏”或“稀疏性”的含义是指当利用测量矩阵的线性组合来表示探针样本时，许多系数应该为零或非常接近于零且条目很少的条件。在表示解决方案中差异很大。

通过利用 l_0 范数最小化约束[61]求解线性表示系统（2-13），可以获得最稀疏表示解。因此（2-13）可以转换为优化问题（2-14）：

$$\hat{\alpha} = \arg\min \|\alpha\|_0 \ s.t. y = X\alpha \tag{2-14}$$

其中$||\cdot||_0$是向量中非零元素的数量，也被称做稀疏度的度量。此外，如果仅使用来自测量矩阵 X 的 k（$k<n$）个原子来表示探针样本，则问题（2-14）将等同于以下优化问题：

$$\mathrm{y} = X\alpha \ s.t. \|\alpha\|_0 \ \le \ k \tag{2-15}$$

问题（2-15）被称为 k-稀疏近似问题。因为实际数据总是包含噪声，所以在大多数情况下表示噪声是不可避免的。因此，可以通过表示将原始模型（2-13）修改为关于可能的小噪声的修改模型

$$\mathrm{y} = X\alpha + s \tag{2-16}$$

其中 $s \in R^d$ 指的是表示噪声，其界限为$||s||_2 \le \varepsilon$。在存在噪声的情况下，公式（2-14）和（2-15）的稀疏解可以通过解决以下优化问题来近似获得：

$$\hat{\alpha} = \arg\min \|\alpha\|_0 \ s.t. \|y - X\alpha\|_2^2 \le \varepsilon \tag{2-17}$$

或者说

$$\hat{\alpha} = \arg\min \|\mathrm{y} - X\alpha\|_2^2 \ s.t. \|\alpha\|_0 \le \varepsilon \tag{2-18}$$

此外，根据拉格朗日乘数定理，存在适当的常数 λ，使得公式（2-17）和（2-18）等效于具有适当 λ 值的下述无约束最小化问题。

$$\hat{\alpha} = \mathrm{L}(\alpha,\lambda) = \arg\min \|\mathrm{y} - X\alpha\|_2^2 + \lambda\|\alpha\|_0 \tag{2-19}$$

其中 λ 指与$||\alpha||_0$相关的拉格朗日乘数。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2.2 具有 l_1 范数最小化的稀疏表示

l_1 范数源于 LASSO 问题[50],[51]并且它已被广泛用于解决模式识别，机器学习和统计学中的问题[62],[63]。尽管稀疏表示方法具有 l_0 范数最小化可以在矩阵 X 上获得的基本稀疏解，问题仍然是一个 NP-hard 问题，并且难以达到全局最优[64]。相关文献[3],[65],[66]已经证明，当使用 l_1 范数最小化约束得到的表示解也满足于稀疏条件时，使用具有足够稀疏性的 l_1 范数最小化的解可以是相当于通过 l_0 范数最小化和完全概率获得的解。此外，l_1 范数优化问题具有解析解，并且可以在多项式时间内求解。因此，已经提出了具有 l_1 范数最小化的广泛稀疏表示方法来丰富稀疏表示理论。具有 l_1 范数最小化的稀疏表示的应用非常广泛且非常普遍。相应地，具有 l_1 范数最小化的稀疏表示的主要流行结构，类似于具有 l_0 范数最小化的稀疏表示，通常用于解决以下问题：

$$\hat{\alpha} = \arg\min \|\alpha\|_1 \quad s.t. y = X\alpha \tag{2-20}$$

$$\hat{\alpha} = \arg\min \|y - X\alpha\|_1 \quad s.t. \|y - X\alpha\|_2^2 \le \varepsilon \tag{2-21}$$

或者说

$$\hat{\alpha} = \arg\min \|y - X\alpha\|_2^2 \quad s.t. \|\alpha\|_1 \le \tau \tag{2-22}$$

$$\hat{\alpha} = L(\alpha, \lambda) = \arg\min \frac{1}{2}\|y - X\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{2-23}$$

其中，λ 和 τ 都是较小的常数。

2.2.3 具有 l_p 范数（$0 < p < 1$）最小化的稀疏表示

一般的稀疏表示方法是求解具有 l_p 范数最小化问题的线性表示系统。除了 l_0 范数最小化和 l_1 范数最小化之外，一些研究人员正在尝试用 l_p 范数（$0<p<1$）最小化来解决稀疏表示问题，尤其是 p =0.1,1/2,1/3 或 0.9[67],[68]。也就是说，l_p 范数（$0<p<1$）最小化的稀疏表示问题是解决问题（2-24）：

$$\hat{\alpha} = \arg\min \|y - X\alpha\|_p^p \quad s.t. \|y - X\alpha\|_2^2 \le \varepsilon \tag{2-24}$$

或者说

$$\hat{\alpha} = \arg\min \|y - X\alpha\|_p^p \quad s.t. \|y - X\alpha\|_2^2 \le \varepsilon \tag{2-25}$$

尽管具有 l_p 范数（$0<p<1$）最小化的稀疏表示方法不是获得稀疏表示解的主流方法，但它极大地影响了稀疏表示理论的改进。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2.4.具有 $l_{2,1}$ 范数和最小化的稀疏表示

$l_{2,1}$ 范数也称为旋转不变 l_1 范数，其被提出用于克服对异常值的鲁棒性的困难[69]。具有 $l_{2,1}$ 范数最小化的稀疏表示问题的目标函数是解决问题（2-26）：

$$\arg\min\left\|Y-XA\right\|_{2,1}+\mu\left\|A\right\|_{2,1} \tag{2-26}$$

其中 $Y=[y_1,y_2,\cdots,y_N]$是指由样本组成的矩阵，$A=[\alpha_1,\alpha_2,\cdots,\alpha_n]$是 X 的相应系数矩阵，μ 是一个较小的常数。$l_{2,1}$ 范数最小化的稀疏表示可以通过文献[54]-[56]中提出的算法来实现。

2.2.5.具有 l_2 范数最小化的稀疏表示

通过 l_2 范数最小化获得的表示解不是严格稀疏的。它只能获得一种“有限稀疏”的表示解，即解决方案具有判别性和可区分性，但实际上不够稀疏[70]。具有 l_2 范数最小化的稀疏表示方法的目标函数是解决以下问题：

$$\hat{\alpha}=\arg\min\left\|\alpha\right\|_2^2 \quad s.t.\left\|y-X\alpha\right\|_2^2\le\varepsilon \tag{2-27}$$

或者说

$$\hat{\alpha}=\mathrm{L}(\alpha,\lambda)=\arg\min\left\|\mathrm{y}-X\alpha\right\|_2^2+\lambda\left\|\alpha\right\|_2^2 \tag{2-28}$$

2.3 基于同伦思想的稀疏表示

同伦的概念源于拓扑，同伦技术主要用于解决非线性方程组问题。同伦方法最初被提出用于解决 l_1 惩罚的最小二乘问题[71]。同伦的主要思想是通过跟踪解决方案的连续参数化路径以及变化的参数来解决原始优化问题。与常规稀疏表示方法（如最小角度回归（LAR）[51]，正交匹配追踪（OMP）[72]）具有高度密切的关系，同伦算法已成功用于解决 l_1 范数最小化问题。与 LAR 和 OMP 相比，同伦方法更有利于通过添加或移除活动集中的元素来顺序更新稀疏解。利用基于同伦的策略来解决具有 l_1 范数正则化的稀疏表示问题的代表方法显示如下。

2.3.1 最小绝对收缩选择算子同伦算法

由于参数在 l_1 范数最小化中的重要性，提出了一种众所周知的最小绝对收缩选择算子（Least absolute shrinkage and selection operator，LASSO）同伦算法，通过在参数 λ 的递减值范围内跟踪整个同伦解路径来解决（2-20）中的 LASSO 问题。证明具有适当参数值的问题（2-23）等同于问题（2-20）[73]。此外，很明显，当将 λ 从非常大的值改为零时，问题（2-23）的解决方案正在收敛到问题（2-20）的解决方案[73]。变量值的集合

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

构成解决方案路径，解决方案路径上的任何点都是问题（2-23）的最优性条件。更具体地说，LASSO 同伦算法从参数 λ 较大的初始值开始，并且沿着同伦解路径终止于接近零的 λ 点，使得最优解收敛到问题（2-20）的解。同伦算法的基本原理是同伦解路径是具有离散数量运算的分段线性路径，而同伦参数的值改变，并且每段的方向和绝对步长由符号序列和分别对相应段的解决方案的支持集决定[74]。

基于凸优化问题的基本思想，零向量应该是问题（2-23）的目标函数的次梯度的解的必要条件。因此，对于任何给定的 λ 值，可以得到关于 α 的目标函数的次级梯度，即

$$\frac{\partial \mathrm{L}}{\partial \alpha}=-X^{T}\left(y-\mathrm{X}\alpha\right)+\lambda\partial\left\|\alpha\right\|_{1} \tag{2-29}$$

其中第一项 $r=x^{T}(y-X\alpha)$称为残差相关向量，$\partial||\alpha||_1$ 是由下式获得的子梯度：

$$\partial\left\|\alpha\right\|_{1}=\left\{\theta\in \mathrm{R}^{N}\left|\begin{array}{l}\theta_{\mathrm{i}}=\operatorname{sgn}(\alpha_{i}),\alpha_{i}\neq 0\\ \theta_{\mathrm{i}}=[-1,1],\ \alpha_{i}=0\end{array}\right.\right\} \tag{2-30}$$

令 Λ 和 u 分别表示 α 的支持集和在该支持集 Λ 上 α 的符号序列。X_{Λ} 表示 X_{Λ} 中所有样本的索引都包含在支持集 Λ 中。如果分析问题（2-23）的 KKT 最优性条件，可以得到以下两个问题（2-29）的等效条件，即

$$\mathrm{X}_{\Lambda}(y-X\alpha)=\lambda u;\ \ \left\|\mathrm{X}_{\Lambda^{c}}^{T}(y-X\alpha)\right\|_{\infty}\leq\lambda \tag{2-31}$$

其中 Λ^{c} 表示 Λ 的补集。因此问题（2-31）中的最优性条件可以被划分为 N 个约束，并且同伦算法在任何 $\lambda\geq 0$ 的情况下保持沿着最优同伦解路径的两个条件。当将 λ 的值减小到 $\lambda-\tau$ 时，对于小的 τ 值，应满足以下条件

$$\begin{array}{ll}\mathrm{X}_{\Lambda}(y-X\alpha)+\tau\mathrm{X}_{\Lambda}^{T}\mathrm{X}\delta=(\lambda-\tau)u & \text{（a）}\\ \left\|\mathrm{p}+\tau q\right\|_{\infty}\leq\lambda-\tau & \text{（b）}\end{array} \tag{2-32}$$

其中 $p=X^{T}(y-X\alpha)$，$q=X^{T}X\delta$，δ 是更新方向。

通常，同伦算法被迭代地实现并且它通过将参数 λ 从大值减小到期望值来更新支持集来遵循同伦解决方案路径。解决方案的支持集将仅在 λ 的临界点更新和更改，其中现有的非零元素缩小为零，或者将新的非零元素添加到支持集中。两个最重要的参数是步长 τ 和更新方向 δ。在第 1 阶段（如果存在$(X_{\Lambda}^{T}X_{\Lambda})^{-1}$），同伦算法首先计算更新方向，这可以通过求解得到

$$\mathrm{X}_{\Lambda}^{\mathrm{T}}\mathrm{X}_{\Lambda}\delta_{\mathrm{i}}=u \tag{2-33}$$

因此，问题（2-33）的解决方案可以写成

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\delta_1 = \begin{cases} (\mathrm{X}_\Lambda^{\mathrm{T}}\mathrm{X}_\Lambda)^{-1}u, & \mathrm{on}\Lambda \\ 0, & other \end{cases} \tag{2-34}$$

随后，同伦算法通过跟踪同伦解路径来计算到下一个临界点的步长 τ。即同伦算法沿更新方向移动，直到不满足（2-32）中的约束之一。在该临界点，新的非零元素必须进入支持集 Λ，或者 α 中的非零元素之一将收缩到零，即必须从支持集 Λ 中移除该元素。两个典型案例可能会导致一个新的临界点，其中违反了（2-32）的任何一个条件。通过计算 $\tau_1^*=\min(\tau_1^+,\tau_1^-)$可以很容易地获得导致临界点的最小步长，并且 τ_1^+和 τ_1^-由以下公式计算所得：

$$\tau_1^+ = \min_{i\in\Lambda^c}\left(\frac{\lambda - p_i}{1 - x_i^T X_\Lambda \delta_l}, \frac{\lambda + p_i}{1 + x_i^T X_\Lambda \delta_l}\right)_+ \tag{2-35}$$

$$\tau_1^- = \min_{i\in\Lambda}\left(\frac{-\alpha_t^i}{\delta_t^i}\right)_+ \tag{2-36}$$

其中 $p_i=x_i^T(y-x_i\alpha_l^i)$和 $\min(\cdot)_+$表示最小值仅在正参数上运算。τ_1^+是将索引 i^+中的非活动元素转换为活动元素的最小步长，即索引 i^+应该被添加到支持集中。τ_1^-是在索引 i^-处将非零有效元素的值收缩为零的最小步长，并且索引 i^-应该从支持集中移除。解决方案即稀疏系数由公式（2-37）更新，其支持和符号序列相应更新。

$$\alpha_{t+1} = \alpha_t + \tau_t^* \delta \tag{2-37}$$

同伦算法迭代地计算步长和更新方向，并更新同伦解及其相应的支持和符号序列，直到满足条件$||p||_\infty=0$，从而达到问题（2-20）的解。

LASSO 同伦算法的伪代码总结如下：

算法目标：解决 LASSO 问题（公式（2-22））

输入：探针样本 y,度量矩阵 X

初始化：l=1,初始化稀疏系数 α_1 及其支持集 Λ_1

重复：

第一步：公式（2-34）计算更新方向

第二步：公式（2-35）、公式（2-36）计算 τ_1^+和 τ_1^-

第三步：计算更新步长：$\tau_1^*=\min(\tau_1^+,\tau_1^-)$

第四步：公式（2-37）更新稀疏系数

第五步：更新支持集：

如果 $\tau_1^+==\tau_1^-$，从 Λ 中移除 i^-即 $\Lambda_{t+1}=\Lambda_t\backslash i^-$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

如果 $\tau_1^+ != \tau_1^-$，将 i^+添加入支持集 Λ 中即 $\Lambda_{t+1}=\Lambda_t \cup i^+$

第六步：$l=l+1$

直到：$||X^T(y-X\alpha)||_\infty=0$

输出：α_{t+1}

2.3.2 基追踪降噪同伦算法

问题（2-22）在信号处理中称为基追踪去噪（Basis Pursuit De-Noising，BPDN），是 LASSO 问题（2-20）的无约束拉格朗日函数，它是一个无约束问题。BPDN 同伦算法与 LASSO 同伦算法非常相似。如果考虑问题（2-23）的 KKT 最优性条件，则对于解 α 应满足以下条件

$$\left\|\mathrm{X}^T(y-X\alpha)\right\|_\infty \le \lambda \tag{2-38}$$

对于任何给定值 λ 和支持集 Λ，还需要满足以下两个条件：

$$\begin{aligned} &\mathrm{X}_\Lambda^{\mathrm{T}}(y-X\alpha)=\lambda u; && \text{（a）}\\ &\left\|\mathrm{X}_{\Lambda^c}^T(y-X\alpha)\right\|_\infty \le \lambda && \text{（b）} \end{aligned} \tag{2-39}$$

基追踪降噪同伦算法直接通过如下公式计算同伦解：

$$\alpha = \begin{Bmatrix} (\mathrm{X}_\Lambda^T X_\Lambda)^{-1}(\mathrm{X}_\Lambda^T y - \lambda u), & \mathrm{on}\Lambda \\ 0, & other \end{Bmatrix} \tag{2-40}$$

同伦参数 λ 的值用较大的值初始化，满足 $\lambda_0 > ||X^T y||_\infty$。随着同伦参数 λ 的值减小，基追踪降噪同伦算法在$(X_\Lambda^T X_\Lambda)^{-1}u$ 的方向上追踪解，直到获得临界点。当非活动元素被转移到活动元素时，达到每个临界点，即，应将其相应的索引添加到支持集中，或者 α 中的非零活动元素值收缩到零，即应从支持集中删除其相应的索引。因此，在每个临界点处，仅更新一个元素，即，一个元素从活动集中移除或添加到活动集中，并且每个操作在计算上非常有效。当同伦参数的值低于其期望值时，终止算法。基追踪降噪同伦算法的伪代码总结如下：

算法目标：解决 LASSO 问题（公式（2-23））

输入：探针样本 y,度量矩阵 X

初始化：$l=0$、稀疏系数 α_0 及其支持集 Λ_0、λ_0 值、初始步长 τ、停止参数 ε

重复：

第一步：公式（2-34）计算更新方向

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第二步：公式（2-40）更新稀疏系数

第三步：公式（2-35）、公式（2-36）计算 τ_1^+和 τ_1^-

第四步：计算更新步长：$\tau_1^*=\min(\tau_1^+,\tau_1^-)$

第五步：更新支持集：

如果 $\tau_1^+==\tau_1^-$，从 Λ 中移除 i^-即 $\Lambda_{t+1}=\Lambda_t\backslash i^-$

如果 $\tau_1^+!=\tau_1^-$，将 i^+添加入支持集 Λ 中即 $\Lambda_{t+1}=\Lambda_t\cup i^+$

第六步：公式（2-33）更新符号序列集 u

第七步：更新 $\lambda_{t+1}=\lambda_{t-\tau}$，l=l+1

直到：$\lambda\leq\varepsilon$

输出：α_{t+1}

2.4 非负二次规划活动集算法

非负二次规划活动集算法的核心思想是保证数据非负性的前提下，将存在不等式约束的二次规划问题，在每一次迭代中，以当前已知可行点为起点，把在该点起作用的约束作为等式约束，将不起作用的约束去掉，在此等式约束下极小化目标函数，解得新的较好可行点之后重复上述做法。

非负二次规划活动集算法（Non-negative quadratic programming activity set algorithm，NNQPAS）是解决 l_1 范数最小化的有效方法，其目标是解决公式（2-23）的最优化问题，但公式（2-23）中 α 的值可以为 0，正或负。其中满足稀疏系数 α 大于等于 0。在此处可以表述为

$$\hat{\alpha}=\mathrm{L}(\alpha,\lambda)=\arg\min\left\|y-X\alpha\right\|_2^2+\lambda\left\|\alpha\right\|_1 \text{ s.t.}\alpha\geq 0 \qquad (2\text{-}41)$$

公式（2-41）在非负二次规划活动集问题中可以重写为（2-42）：

$$\hat{\alpha}=\mathrm{L}(\alpha,\lambda)=\arg\min\frac{1}{2}\alpha^{\mathrm{T}}H\alpha+g^{\mathrm{T}}\alpha \ s.t.\alpha\geq 0 \qquad (2\text{-}42)$$

其中，$H=X^TX$，$g=-X^Ty+\lambda$。此时 H 是还需要满足 $f=min(y-X\alpha)$，其等式约束表述为 $\Lambda=\{i|x_i=0\}$;不等式约束表述为 $\Lambda^c=\{i|x_i>0\}$。

假设 α^*是二次规划问题的局部极小点，且在 α^*处有效集集合为 $\Lambda\cup\Lambda^c$，则 α^*也是问题（2-43）的极小点。

$$\hat{\alpha}=\min\frac{1}{2}\alpha^TH\alpha+g^T\alpha \ s.t.x_i^T\alpha=y_i,i\in\Lambda\cup\Lambda^c \qquad (2\text{-}43)$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

如果 α^*是问题（2-42）的可行点，同时是问题（2-43）的 K-T 点，相应拉格朗日乘子应 λ^*满足

$$\lambda^* \geq 0, i \in \Lambda(\alpha^*) \tag{2-44}$$

有效集的主要难点有效集的获取，所以需要构造一个集合序列去逼近有效集，即从初始点出发，计算有效集，解决其对应的等式约束子问题，依次迭代，直至 $\alpha_0 \rightarrow \alpha^*$，以获得原始问题的最优解。

有效集方法每个迭代点要求是可行点，每次迭代求解一个等式约束的二次规划。如果等式约束二次规划的解是原约束问题的可行点，则判断对应的 λ^*是否满足（2-44），其中 λ^*满足公式（2-45）如果满足，则停止计算，否则去掉最小的约束重新求解。当等式约束二次规划不是原问题的可行点，则需要增加约束重新求解。

$$\mathrm{H}\alpha^* + g = \sum_{i \in \Lambda^c \cup \Lambda(\alpha^*)} x_i \lambda_i^* \tag{2-45}$$

在第 k 次迭代，有可行点 α_k 及其工作集 w_k，w_k 由所有等式约束和有效不等式约束组成。设 p_k 是问题（2-46）的 K-T 点，$\lambda_i^k (i \in w_k)$是对应的拉格朗日乘子。

$$\hat{\alpha} = \operatorname{argmin} \frac{1}{2}(\alpha_k + p)^T H(\alpha_k + p) + g^T(\alpha_k + p) \; s.t. \mathrm{x}_i^T p = 0, i \in w_k \tag{2-46}$$

如果 $\mathrm{p_k}$=0，此时 α_k 是问题（2-47）的 K-T 点。

$$\hat{\alpha} = \operatorname{argmin} \frac{1}{2}\alpha^T H\alpha + g^T\alpha \; s.t. \mathrm{x}_i^T\alpha = 0, i \in w_k \tag{2-47}$$

如果此时对于一切 $i \in w_k \cap \Lambda$，$\lambda_i^k \geq 0$ 都成立，则 α_k 也是（2-42）的最优解。否则令 $i \in w_k \cap \Lambda$，使得

$$\lambda_{ik}^{(k)} = \min\left\{\lambda_i^k \,|\lambda_i^k < 0, i \in w_k \cap \Lambda\right\} \tag{2-48}$$

且令 $w_{k+1}=w_k \backslash \{i_k\}$,然后重新计算问题（2-46）。

设（2-46）的解 $p_k \neq 0$，此时 $\alpha_k + p_k$ 可能不是（2-42）的可行点，此时在 α_k 和 $\alpha_k + p_k$ 之间的线段上取靠近 $\alpha_k + p_k$ 最近的可行点作为下次的迭代点 α_{k+1}。

$$\alpha_{\mathrm{k+1}} = \alpha_{\mathrm{k}} + \mathrm{x_k} p_k \tag{2-49}$$

其中，

$$\mathrm{x}_k = \min\left\{1, \min \frac{y_i - x_i^T \alpha_k}{x_i^T p_k}\right\} \tag{2-50}$$

（2-46）的简化计算，更新 $\alpha=\alpha_k+p$，则

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$q(\alpha)=\operatorname{argmin}\frac{1}{2}(\alpha_k+p)^T H(\alpha_k+p)+g^T(\alpha_k+p)$$

$$=\frac{1}{2}p^T Hp+\nabla q(\alpha_k)^T p+\frac{1}{2}\alpha_k{}^{\mathrm{T}} H\alpha_k+g^T\alpha_k$$

其中 $1/2\alpha_k{}^T H\alpha_k+g^T\alpha_k$ 为常数项，不影响最优解，所以求解方程可以简化为公式（2-51）

$$\hat{\alpha}=\min\frac{1}{2}p^T Hp+g_k{}^T p\ \ s.t.x_i^T p=0,i\in w_k \qquad (2\text{-}51)$$

非负二次规划活动集算法的伪代码简单总结如下：

算法目标：解决非负二次规划活动集算法公式（2-42）

输入：探针样本 y,度量矩阵 X

初始化：较小的正实数 λ; $H=X^TX$; g=$-X^Ty+\lambda$; 稀疏系数 $\alpha=[(H)^{-1}(-g)]_+$; 有效集 $\Lambda=\{i|x_i=0\}$；非有效集 $\Lambda^c=\{i|x_i>0\}$；拉格朗日乘数 $\mu=Hx+g$；停止参数 ε

重复：

第一步：获得负面影响最小的乘数：$j=\operatorname{argmin}_{i\in\Lambda}(\mu_i)$;

如果 j<-ε:

第二步：更新支持集：将 j 加入支持集 Λ 中即 $\Lambda=\Lambda\cup j$;

将 j 移出非支持集 Λ^c 中即 $\Lambda^c=\Lambda^c\backslash j$;

第三步：重新计算关联稀疏系数：$t_\Lambda=(H_\Lambda)^{-1}(-g_\Lambda)$ ，$t_\Lambda{}^c=0$;

如果 $argmin\ t_\Lambda\leq 0$：

第四步：$\rho=min_{i\in\Lambda,ti\leq 0}(\alpha_i/(\alpha_i-t_i))$;

第五步：可能存在一个或几个索引对应于 ρ：$\kappa=argmin_{i\in\Lambda,ti\leq 0}(\alpha_i/(\alpha_i-t_i))$;

第六步：更新稀疏表示系数：$\alpha=\alpha+\rho(t-\alpha)$;

更新支持集：将 κ 移出支持集 Λ 中即 $\Lambda=\Lambda\backslash\kappa$;

将 κ 加入非支持集 Λ^c 中即 $\Lambda^c=\Lambda^c\cup\kappa$;

第七步：重新计算关联稀疏系数：$t_\Lambda=(H_\Lambda)^{-1}(-g_\Lambda)$ ，$t_\Lambda{}^c=0$;

第八步：更新系数表示系数 $\alpha=t$；更新拉格朗日乘数 $\mu=Hx+g$;

直到：$\Lambda\neq\emptyset$，且 $j=\operatorname{argmin}_{i\in\Lambda}(\mu_i)$;

输出：α

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.5 本章小结

本章主要介绍了稀疏表示的相关基本概念和稀疏表示的一些基本公式介绍以及最小绝对收缩选择算子同伦算法、基追踪降噪同伦算法和非负二次规划活动集算法。首先是本文的一些数学符号的定义与说明；其次介绍了不同范数最小化的稀疏表示的相关公式；最后介绍了最小绝对收缩选择算子同伦算法、基追踪降噪同伦算法和非负二次规划活动集算法的数学公式的推导以及伪代码描述。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3 甲骨文数据集构建

很长时间以来，甲骨文研究作为一个相对比较“冷门”的学科，主要依赖专家多年积累的经验，计算机与之结合的相关研究一直比较少，且相对集中在一些背景较为单纯的甲骨文拓片文字识别或者电子版扫描件识别或者摹本识别，而实际甲骨文拓片图像中，因为殷商时期雕刻技术的发展与雕刻技师自身雕刻水平差异、保存不当等造成龟甲断裂或者损坏、龟甲或者牛骨自身纹路或者裂纹、现代甲骨文研究人员拓印技术的水平差异等等多种原因造成甲骨文拓本有自己独特的噪声。该噪声不同于已有的一些图像数据集的特征，也不同于一些常见的文本识别数据集特征。最突出的特征是强噪声、部分字体残缺比较严重、黑白图片特征相比起彩图比较少，数据分布不均衡且字体形态较多等。基于此现状，本文标注并整理了一个基于真实甲骨文拓片的甲骨文基准数据集。本章主要介绍甲骨文数据集的提出背景、标注方式、数据特点、样本示例、数据分布等基本信息并给出在该数据集十五个子数据集中样本量最小的子数据集 big50-99 运行的多种算法作为数据集的基准实验。

3.1 甲骨文数据集的提出背景

1899 年甲骨文首次被发现开始，据学者胡厚宣统计，共计出土甲骨 154600 多片，其中中国大陆 97600 多片，宝岛台湾 30200 多片，香港 89 片，总计中国共 127900 多片，此外，日本、加拿大、英、美等国家共 26700 多片。目前为止这些甲骨上刻有的单字约 4500 个，可解释的字约有 2000 个。中国作为甲骨文发源地拥有最多的甲骨资源、最优秀的研究团队。然而，甲骨文研究特别依赖专业人员，传统的甲骨文研究都是专业人员凭借经验以及良好的记忆力从浩瀚的资料中查找相关信息，相关人工智能研究较为薄弱。目前甲骨文研究面临几大问题，其中最基本的问题就是甲骨文的智能化检测与识别。基于该背景，本人在导师带领下开始着手做基于拓片的甲骨文识别相关研究，期望在 2019 年甲骨文发现 120 周年之际，致敬甲骨文的研究。研究的最终目标是实现甲骨文拓片文字的自动化识别，本部分进行基础性的数据集构建工作。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

下面介绍甲骨文数据集的提出背景。包括：1.现有甲骨文研究使用的数据现状，主要介绍现有甲骨文研究中使用的一些数据样本来源及其局限性。2.甲骨文拓片本身的特征及其造成的原因，主要是为了说明当前甲骨文识别研究使用的数据是局限不全面的，缺少一个相对全面、基于真实拓片的甲骨文基准数据集。

3.1.1 甲骨文研究数据现状

在导师的带领与指导下，查阅了大量的资料与信息，发现基于甲骨文拓片的文本识别研究一直在进行，但没有一个合适的有代表性的公开数据集。研究甲骨文识别相关文献发现，大多数甲骨文识别算法都是基于科研人员从甲骨文合集或者相关数据库中筛选几个比较有代表性的字作为研究，比如人、卜、貞等一些出现次数比较频繁的字。且筛选的大都是比较完整且较清晰的图片。

第一个比较有代表性的例子是 2018 年日本情报处理学会第 80 回全国大会上的演讲论文使用深度学习来识别甲骨文[31]。其中用到的甲骨文图片均为手动筛选的 114 种不同字形的甲骨文文字的共 1358 张图片，其中 114 中不同字形如图 3-1 所示。



图3-1 日本情报处理学会使用的114种字形样本示例[31]

每张图片包含一定的噪点，甚至面积较大的噪声，但所有样本均为完整的样本，即每张图片中的文字是完整的，无残缺的，如图 3-2 所示。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

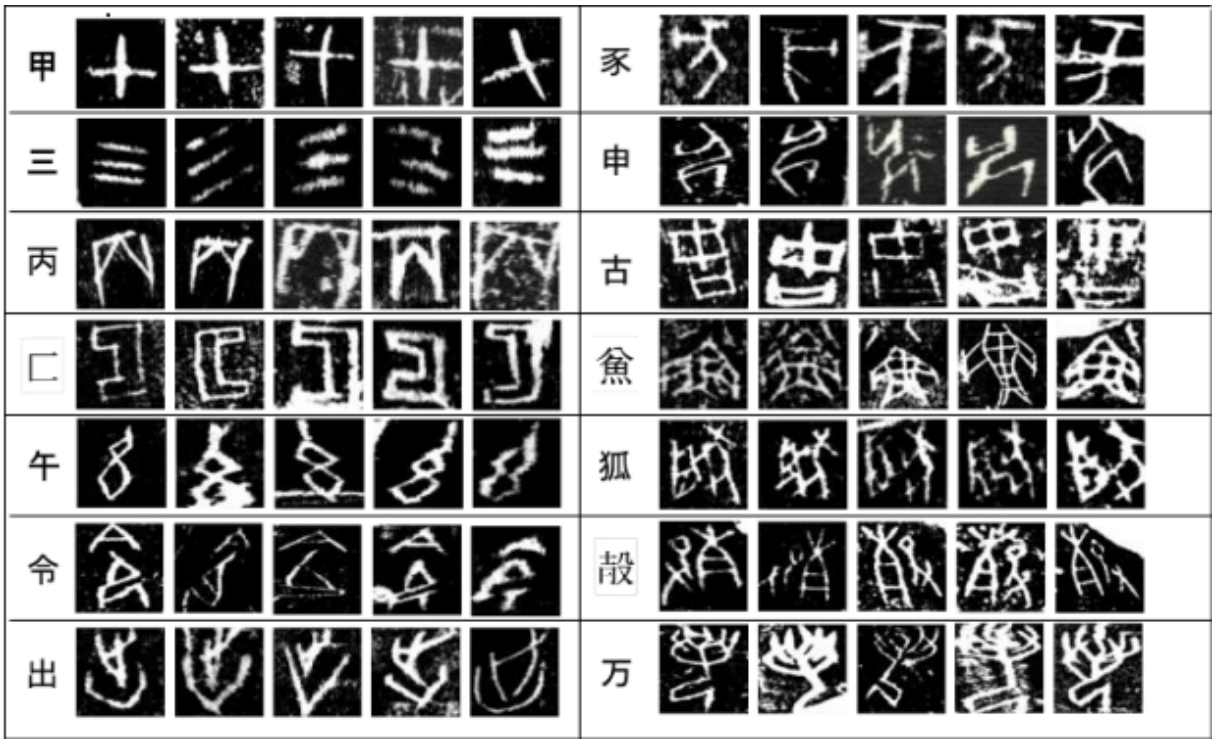


图3-2 日本情报处理学会使用的部分字形多个样本示例[31]

然后通过裁剪，反转，旋转，不同灰度对比度等多种方式做了数据增强，如图 3-3 所示。然后再使用卷积神经网络进行甲骨文深度学习的训练与识别。

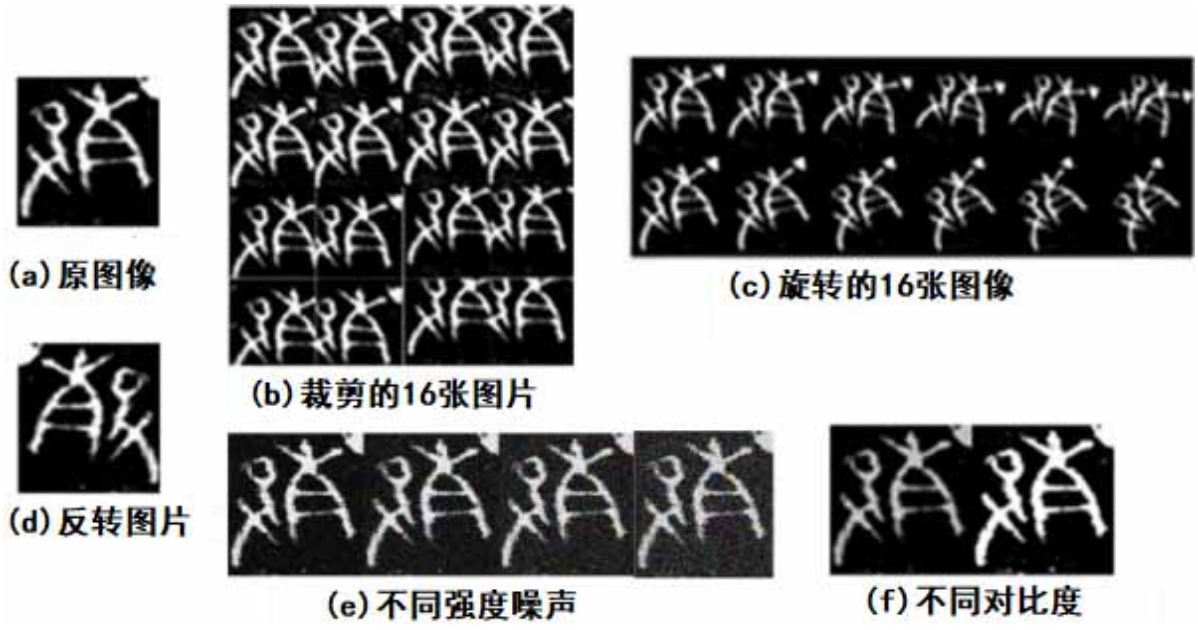


图3-3 日本情报处理学会使用的图像增强样本示例[31]

第二个比较典型的例子是 ICPRAM 2017 中的文献[42][42]中使用的甲骨文数据，其图片数据显示近乎黑白图片，极少有噪点和一些干扰的纹路之类的。其典型样本显示如图 3-4 所示：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

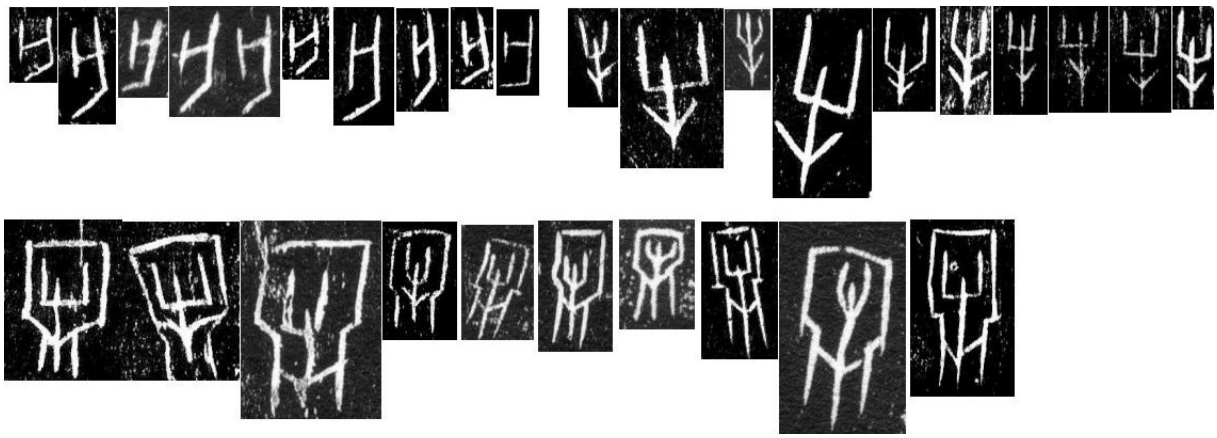
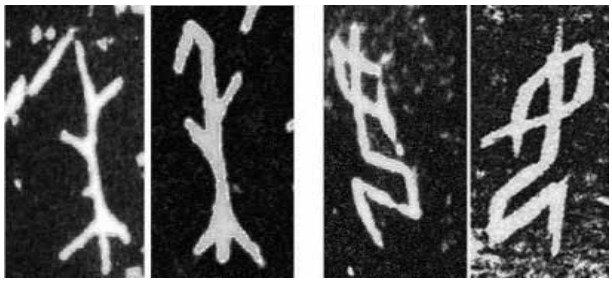


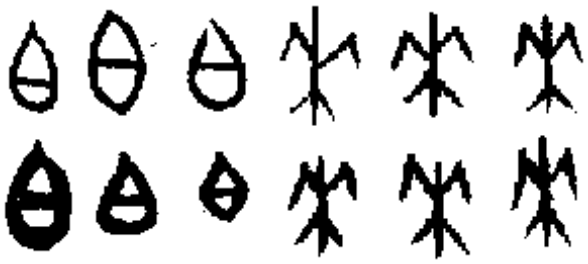
图3-4 纯净背景的甲骨文识别图像样本示例[42]

第三个比较典型的例子是 2010 年北京大学几位学者的文献[34]中使用的数据，其包含一定的噪点，但数据样本量极少，共选取了 8 个字，每个字共选取 30 张图片作为样本，其中三分之二作为训练，三分之一作为测试。样本量极少、噪点较少、无大面积遮挡或残损。其典型样本显示如图 3.2-5 所示，其中图 3-5（a）表示图中所示两个样本为字“禾”的同一字形，图 3-5（b）表示图中所示两个样本为字“女”的两种写法。



（a）同字形示例“禾”（b）不同字形示例“女”

图3-5 小样本多字形甲骨文识别图像样本示例



（a）无噪声示例“白”（b）无噪声示例“禾”

图3-6 无噪声简单甲骨文识别样本示例

第四个经典例子是 2015 年郭俊[40]使用的甲骨文数据集，其数据背景纯净，采用了二值化，图片几乎没有噪声，字形及其完整。此时的甲骨文识别及其类似于手写体文字识别，且不用关心其文字间的粘连。相关示例如上图 3-6 所示。

而现实情况中，甲骨文拓片识别研究面临的数据情况并非那么理想更多的是以上多种情况的综合，甚至包含更为复杂的情况。

3.1.2 甲骨文拓片本身的特征及其造成的原因

因甲骨文是远古先民手工雕刻的一种文字，其本质上属于手写体。但甲骨文与现有一些手写体识别具有不同的特征。其造成的原因分析如下：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（1）它所书写的载体是龟甲牛骨以及少数的虎骨、猪骨、鹿骨等，骨骼本身具有的天然纹路，尤其以龟甲最甚，其背甲和腹甲本身就具有特定纹路，如图 3-7 所示。

（2）甲骨文的书写随着时代的发展越来越成熟，其雕工技术的发展最直观的影响就是龟甲上面雕刻字数、字的清晰度、字的大小以及雕刻纹路深浅等差异，如图 3-8 所示，其中左侧龟甲字数较多，字体较小，而右侧龟甲则字数较少，字体较大。



图3-7 原始龟甲图片样本示例[75]

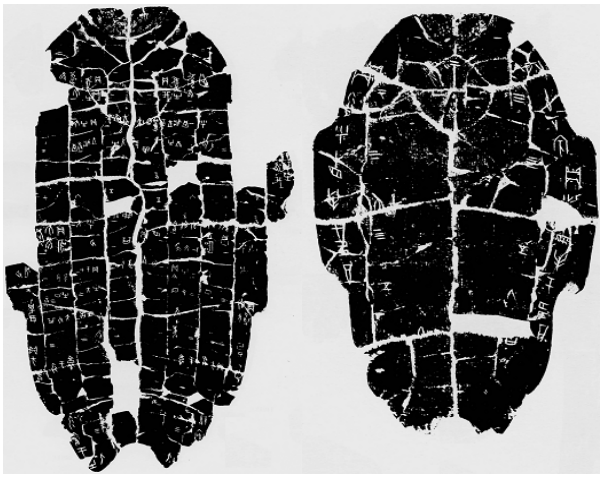


图3-8 不同字数与字体大小的龟甲拓片样本示例

（3）甲骨文最早出现并不是为了作为记录历史使用，而是为了占卜，首先在骨头上雕刻文字，然后灼烧受热之后骨头出现裂纹（有时并非灼烧而是在龟甲背面钻坑，随着钻的坑增大，龟甲正面慢慢产生裂纹），纹路与卜辞作为占卜结果。这就造成甲骨文字的断裂与损害，如图 3-9 所示，左侧显示为带有龟裂的龟甲，右侧显示的是同一张龟甲背面钻的坑。

（4）殷商时代距今时代久远，甲骨出土发现之后又历经战争，颠沛流离之中甲骨破碎损坏比较严重，造成甲骨上的文字被破坏的比较严重，如图 3-10 所示。

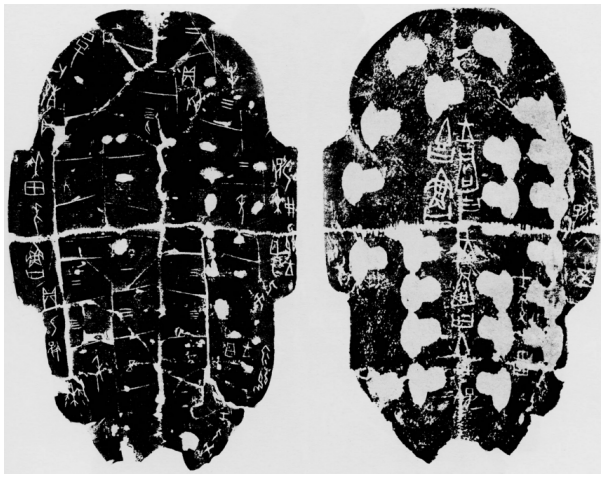


图3-9 龟甲正面纹路及背面钻坑印记样本示例

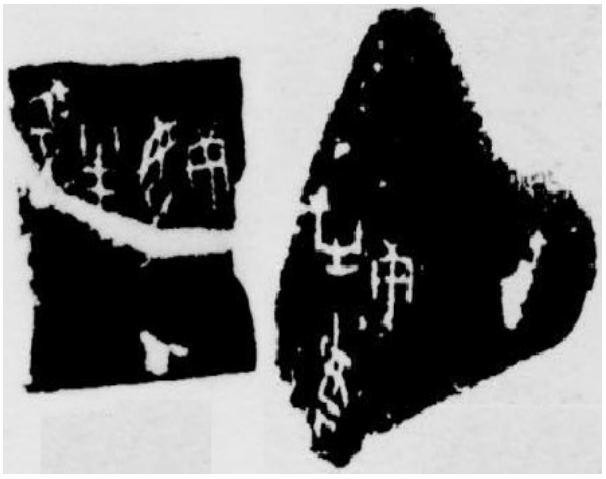


图3-10 破碎的甲骨拓片样本示例

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（5）甲骨文发现之后人为的在甲骨上设置了标签或者标号，这些标记本身也会造成纹路的不平整或者破坏，拓印之后也会有一些纹路形成干扰，如图 3-11 所示。

（6）因为龟甲的重要性与珍稀程度，大部分的甲骨文研究都是基于甲骨文拓片的研究。拓印人员的拓印技术好坏对于甲骨文拓片的质量及其重要。拓印质量较好的拓片，一些比较细小的纹路也会被完成的保留，而拓印质量较差的拓片甚至都不太清晰，如图 3-12 所示。

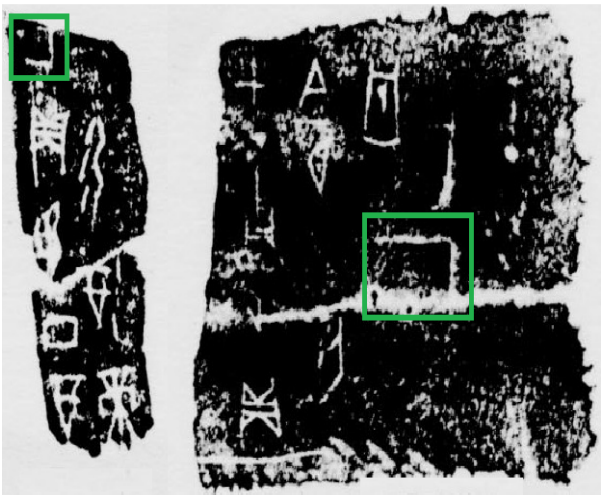


图3-11 甲骨出土标记拓印噪声样本示例

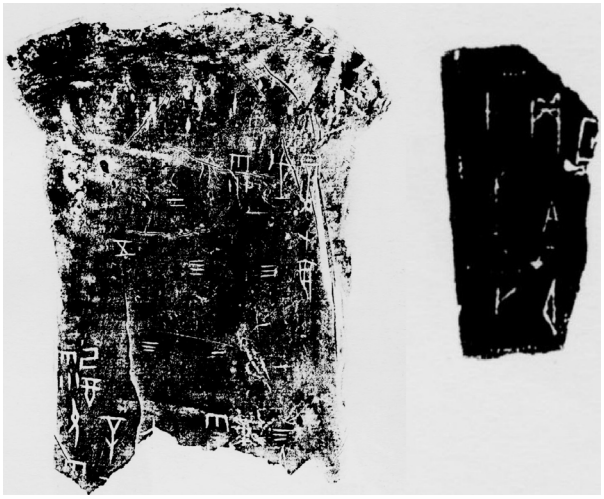


图3-12 拓印技术对于拓片质量的影响样本示例

（7）甲骨文的一句话并不像是传统的横平竖直的一句话，而是横竖都有甚至为了贴合龟甲的边缘而呈现弧形。一句话可能在同一片区域，但可能被拆为几列。即语句前后联系比较弱，如图 3-13 所示。

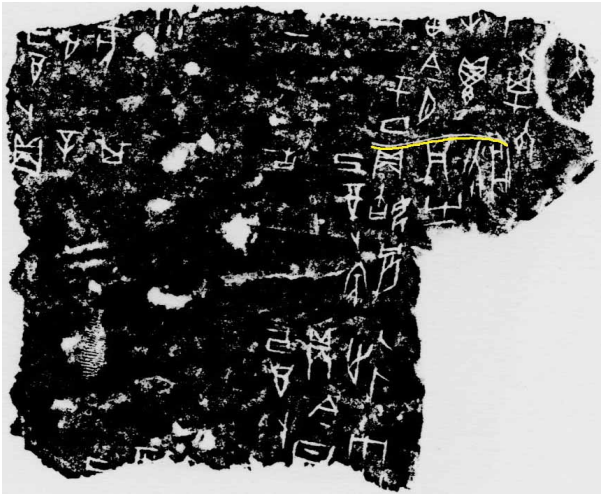


图3-13 同一列文本包含多句话不同词汇样本示例

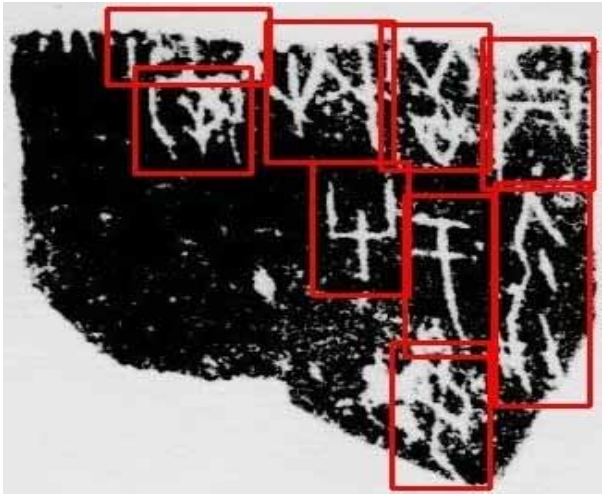


图3-14 数据集标注结果样本示例

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.2 数据集描述

本部分描述数据集标注方式以及相关子数据集划分标准。

3.2.1 数据集标注方式及工具

数据集图像来源于甲骨文合集，相关图像和文字说明可以从国学大师网站上查找。在43041张图片中筛选了7883张图片，最终共完成标注7824张图片，其中本人标注1558张。数据集整理团队由三位老师和 18 名学生组成。图像标注采取的是单字标注，即每个字作为一个标注框，结果即对应的甲骨文文字，如图 3-14 所示。本数据集采取的是四点标注法，标注结果记录了左上右下两点坐标及其对应的文字，其格式为：左上 x,左上 y,右上 x,右上 y,"字"。每张图片对应有一个与图片同名的 txt 文本文件用来记录所有标注信息。标注完成之后又经过古文字专业人员的校对，最后又将标注结果经过一遍非技术校对，筛选去除掉其中的一些因为标注过程中复制粘贴甲骨文而造成的乱码情况等，又为了扩大甲骨文数据集的可应用性，将标注结果做了一份编码（对应的给每个甲骨文字一个数字编码）以使得标注结果可以脱离使用的指定字体在其他一些平台或者没有安装指定字体的机器上使用，标注质量较高。

使用的标注软件为本人自主开发的基于 C#的 winform 窗体应用程序，其中包含功能如下：标注：在图片上画矩形框并将其指定甲骨文字板上的字点击输入作为标注结果存入文件；矫正最新矩形框：点击之后可以修改之前标注时候位置不太理想的标注结果；清除标注：清除当前图片对应的所有标注结果，包括文字和记录的坐标；修改标注：将某个矩形框对应的文字覆盖掉换成新的文字等多种基本功能。因为甲骨文输入困难，所以还设计了一个甲骨文字典面板去将 word 文档中的甲骨文转换为对应的按钮以协助文字的输入加快标注速度同时避免一些非甲骨文的乱码数据的输入。其基本界面如图 3-15 介绍如下：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

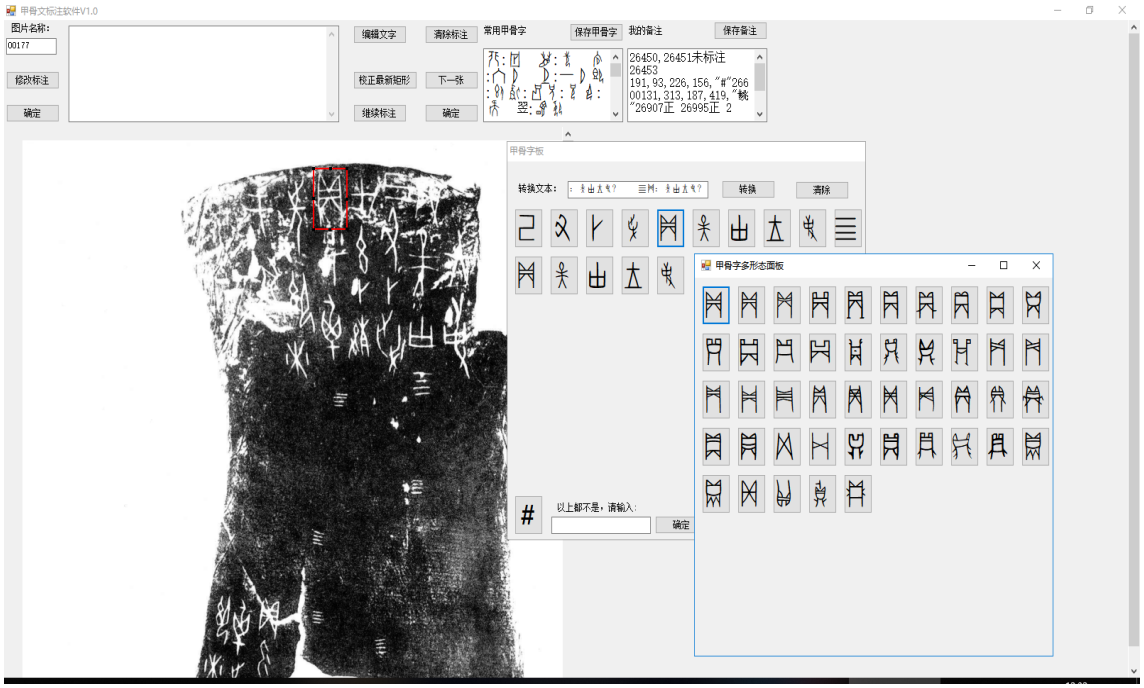


图3-15 甲骨文标注软件主界面示例

使用的校对软件为自主开发的基于 C#的 winform 窗体应用程序，该软件是在甲骨文标注软件的基础上重新设计改进的。其中包含功能如下：下一条标注：点击该按钮可以逐条显示每个矩形框及其标注文字；新增标注：将之前遗漏忘记标注的文字记录其坐标和文字写入标注结果文件；显示全部：一次性显示所有标注框和文字；重新开始：将对当前图片所做的所有操作都取消掉，重新从文件中读取标注结果然后重新校对等多种基本功能。因为甲骨文输入困难，所以还设计了一个甲骨字面板保存一些常用的甲骨文字以便快速查找相关文字。其基本界面如图 3-16 介绍如下：

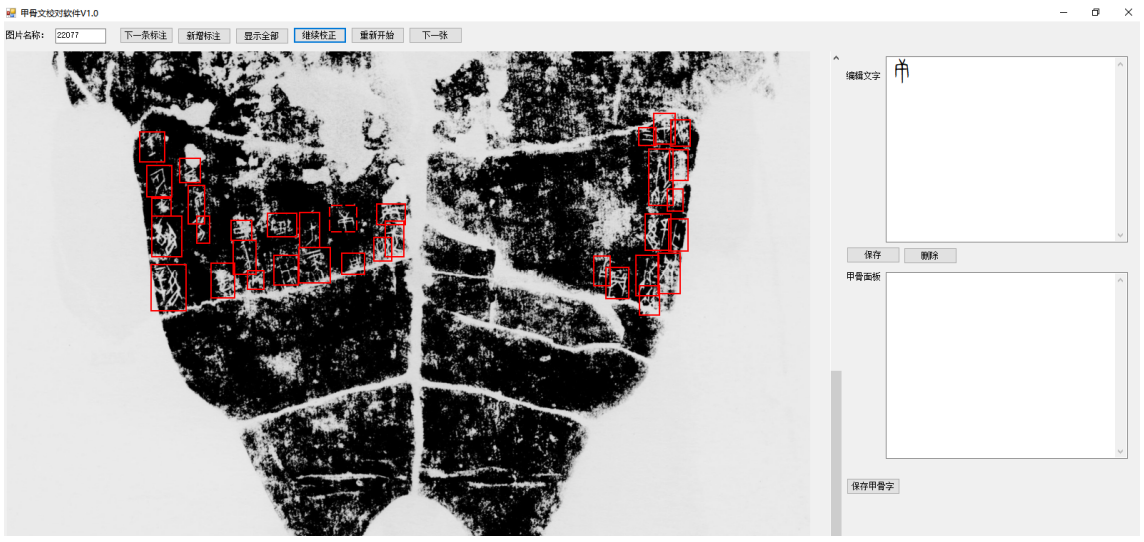


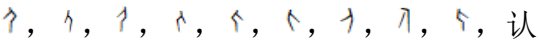
图3-16 甲骨文校对软件主界面示例

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.2.2 文本检测数据集

因甲骨文拓片图像本质上是一张图片，将其每个文字的坐标位置信息提取即可整理成为一个文本检测数据集。该数据集为全部为灰度图，其中每张图片具有多个检测框，且部分检测框可能存在部分重叠现象。检测数据集将标注框添加在图片上之后显示如图3-14 所示。在一些基准实验上证明该数据集用来做文本检测挑战性相对于甲骨文识别而言较小，因此将主要注意力集中在文本识别方面。

3.2.3 文本识别数据集

根据标注结果，将每张图片中对应的甲骨文裁剪出来制作成了甲骨文识别数据集。甲骨文拓片识别数据集被分为三大类。分别为字数据集、字形数据集和均衡字形数据集。字数据集是指同一个甲骨字作为一类的数据集：因甲骨文中存在一字多形的情况，字表示同一个字有多种不同形态，如图 3-5（b）中“女”字，或者甲骨文“人”字就有多个形态：，认为这些不同形态的字都属于同一个字；字形数据集指甲骨文某一个字的某一种字形作为一类的数据集：如甲骨文中人字的某个形态；均衡字形数据集是指在字形数据集的基础上，将不同类别的之间的数据差别缩小，保证每个类样本量相近或者说比较均衡，比如原本甲骨文字形样本量在 10-49 期间的共有 587 类，样本数目最少的是 10，最多的是 49，均衡之后的数据保证样本数目最多的不超过最少的二倍。

数据集主要分为三类，每类共五种。

三大类分别分为：

甲骨文字数据集：big;

甲骨文字形数据集:small；

甲骨文均衡字形数据集:balance。

每类分别按照样本出现次数（字频）不同划分为不同的数据集：

样本量在 10-49 之间；

样本量在 50-99 之间；

样本量在 100-499 之间；

样本量大于等于 500；

样本量大于等于 1000。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

根据以上组合，将数据集依次命名为表 3-1 所示：

表 3-1 数据集名称命名规则及名称

类别/样本数	10-49	50-99	100-499	500	1000
字	big10-49	big50-99	big100-499	big500	big1000
字形	small10-49	small50-99	small100-499	small500	small1000
均衡	balance10-49	balance50-99	balance100-499	balance500	balance1000

共十五个数据集详细信息展示如下表 3-2 所示，其中类别数代表该数据集共有多少类、样本总数代表该数据集中共有多少样本、样本最小类数代表该数据集中样本数目最少的一类样本数目、样本最大类数代表该数据集中样本数目最多的一类样本数目、类间平均数代表该数据集中所有类的平均数目。

表 3-2 数据集基本信息

数据集/指标	类别数	样本总数	样本最小类数	样本最大类数	类间平均数
big10-49	254	5496	49	10	21.64
big50-99	49	3425	97	50	69.9
big100-499	77	17517	476	102	227.49
big500	58	94695	9337	526	1632.67
big1000	31	73671	9333	1107	2376.48
small10-49	587	12844	49	10	21.88
small50-99	129	9102	99	50	70.56
small100-499	151	30961	499	100	205.04
small500	52	62451	5549	505	1200.98
small1000	18	40104	5549	1090	2228
balance10-49	587	7044	17	10	12
balance50-99	129	6930	59	50	53.72
balance100-499	151	18213	179	100	120.62
balance500	52	33474	1513	505	643.73
balance1000	18	19620	1090	1090	1090

其每个数据集中对应的数据数目显示如下图 3-17（a）、图 3-17（b）、图 3-17（c）、图 3-17（d）、图 3-17（e）所示,其中每张图代表一种字频的三类数据的分布。根据数据样本数和类别的分布可以看出十五个数据集中，除了均衡数据集样本相对比较均衡之外，字与字形数据集都属于分布较为不均衡的数据。所以甲骨文拓片数据集本质上是一种多类不均衡的数据集。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

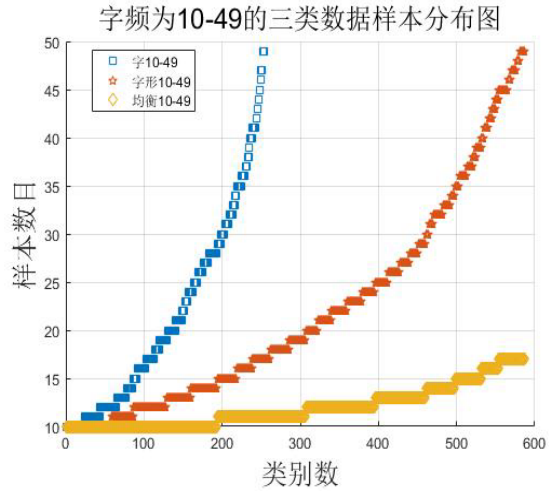


图3-17（a） 10-49的数据样本分布图

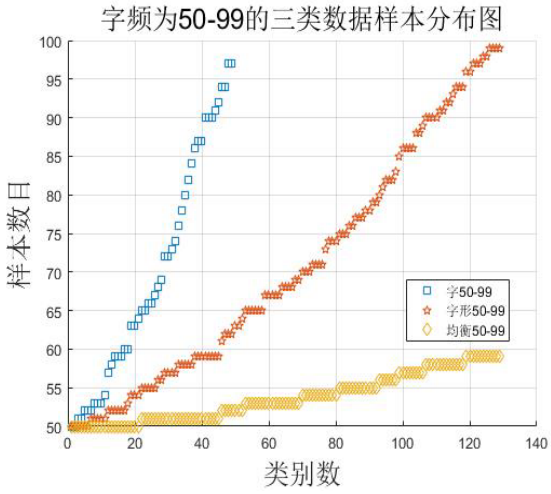


图3-17（b） 50-99的数据样本分布图

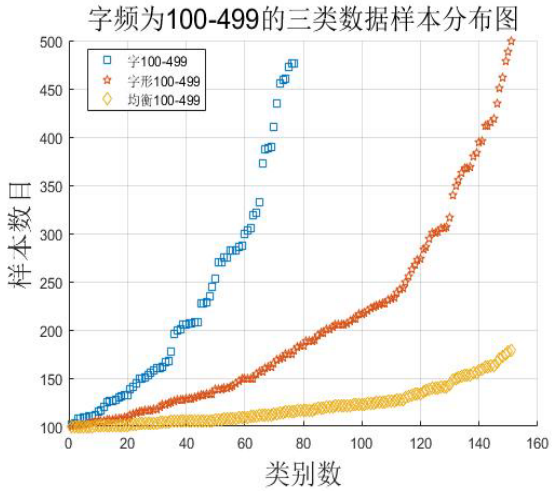


图3-17（c） 100-499的数据样本分布图

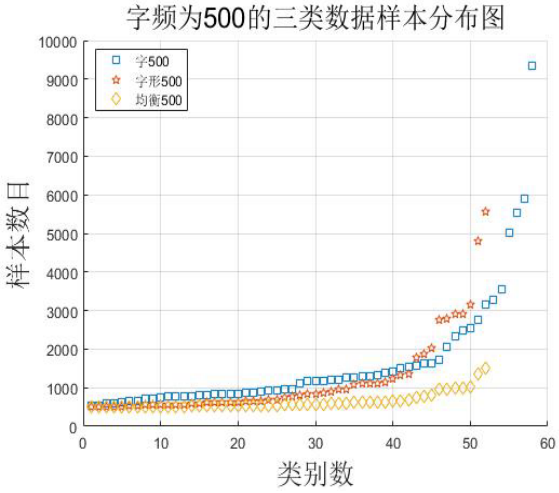


图3-17（d） 500的数据样本分布图

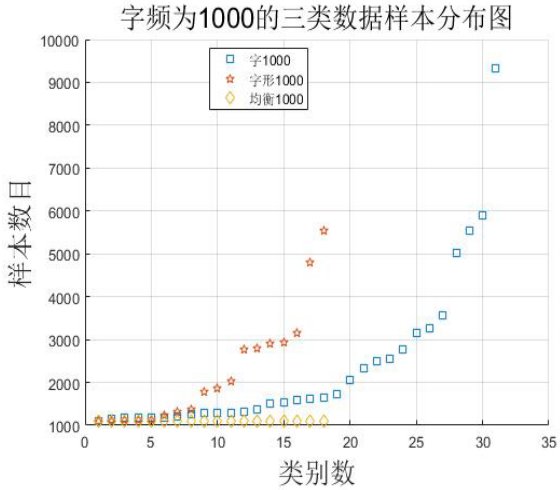


图3-17（e） 1000的数据样本分布图

数据集中抽取样本示例如图 3-18 所示：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net



图3-18 字、字形、均衡字形数据集样本示例

其中，0，1，2 代表从不同类别的数据集中随机各抽取的三类样本。

如上图 3-18 所示：样本存在多种噪音，其造成的原因已在章节 3.1.2 中阐述。相关特征简单总结如下：

（1）形态不一：因甲骨文本质是一种手写体，所以同一个字，即便是同一个字形，在不同的人手中写出来的形态也并非完全一致。

（2）字体残缺：如图 3-18 第四行中的图片 1、2、5、8 所示，字体残缺比较严重。甚至其残字的相对位置也不一定正确，如第一行中的图片 1 所示，残缺的部首本应该在该字的图片右下角，但其当前处于图片的中间位置。

（3）字体清晰度不一：如图 3-18 中第六行中的图片 9、10 所示，因拓印技术或者其他原因造成的字体清晰度不一致，部分字很模糊甚至线条断裂等。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（4）形似字的影响：部分甲骨文文字过于形似，造成识别困难，比如“人”字的其中一个字形为：，与“匕”字的其中一个字形，其相似性甚至超过了同一个字的相似性（“人”字的各个字形：、、、、、、、、；“匕”字的各个字形：、、、、、、、、、、、、）。再加上雕刻技术的影响等等。

（5）各种其他噪声：因拓印或者龟甲残缺等各种原因造成字的笔画之间粘连或者缺失等。

3.3 基线实验

在确定使用稀疏表示相关算法作为基准算法加以研究之前，本文在 big50-99 数据集上进行简单的基准实验。相关算法及其实验结果显示如下，证明甲骨文拓片数据集用来做识别任务具有挑战性，其结果如表 3-3 所示。以下算法中，除深度学习算法（VGG 和 Inception）外，均采用相同的数据预处理策略，即特征脸方法(Eigenface)。算法全称请参考附录部分表 A-1。

表 3-3 基本基准实验方法及准确率

方法	准确率（%）
VGG	35.49
Inception	12.03
OMP	31.13
EB	4.96
LRC	17.59
KNN	29.32
RF	21.05
NB	20.75
DA	27.22
BPDN	37.74

3.4 本章小结

本章首先介绍了甲骨文拓片数据集的提出背景，即为什么整理该数据集的问题：其中包括甲骨文研究现状，以及甲骨文拓片的特征及其造成的原因分析。其次介绍了整理数据集的过程中使用的工具以及最后整理出来的检测与识别数据集样本数据示例。最后介绍了在甲骨文识别数据集中运行的一些基准实验以证明甲骨文数据集的挑战性。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.基于甲骨文数据集的大规模实验对比分析

在第三章所构建的甲骨文数据集的基础上，本章进行了相关算法研究。其中包括对于算法的改进及 33 种算法的基准实验和筛选出 13 种算法在 15 个数据集上的大规模实验。实验结果证明，本文提出的 IBPDN 算法和 BRNAK 算法是解决甲骨文拓片文字识别的有效算法，同时证明了稀疏表示在甲骨文识别是可行的。本部分的主要贡献在于对 13 个文字识别算法进行大规模实验对比分析，该工作对于学术界甲骨文识别相关的研究有一定的借鉴作用和参考价值。需要说明的是，本部分中涉及的主要算法的名称及英文简称，已以列表的形式附于附录 A 和 B 中。

4.1 改进的基于同伦思想的基追踪降噪算法简述

根据表 3-3 中基准实验结果显示，稀疏表示算法在解决甲骨文识别问题中具有更高的准确率，分析其可能原因为稀疏表示算法对于噪声的敏感度弱于其他几种算法。但在实验过程中发现因为数据集中残字的影响，部分测试样本的预测的最小重建误差对应的预测结果并不准确，但次小重建误差对应的预测结果是测试样本真实标签的比重较大。尤其在样本量较少的时候，一些较大的纹裂和残字大面积模糊的部分对于算法干扰性较强。因此本文改进了基追踪降噪算法的数据集更新方式，提升算法在小样本数据集中的识别准确率。

算法改进过程中基本上是遵循了原始追踪降噪算法的核心思想。在 BPDN 算法的基础上，为了增加算法的鲁棒性，改变了支持集的更新方式，将基追踪去噪算法中关键的一个步骤做了变动：关键步骤为：（1）计算更新方向（2）更新稀疏系数（3）计算更新步长（4）更新支持集和符号序列集。本章改进了更新支持集的方式，提升了算法的识别准确率。

改进算法主要用来解决问题（2-23）的最优化问题。在凸优化问题中，零向量是问题目标函数的次梯度解的必要条件。所以，给定 λ 值，可以得到关于稀疏系数 α 的目标函数的次级梯度，即公式（2-29）。其中 $r=x^T(y-X\alpha)$是残差相关向量，由公式（2-30）获得子梯度$\partial||\alpha||_1$。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Λ 和 u 分别表示稀疏系数 α 的支持集和在该支持集 Λ 上稀疏系数 α 的符号序列。X_Λ 表示 X_Λ 中样本索引包含在支持集 Λ 中。分析满足 KKT 最优性条件的问题（2-23），得到问题（2-29）的等效条件公式（2-31）。Λ^c 表示 Λ 的补集。问题（2-31）中的最优性条件可以划分成 N 个约束，同时同伦算法在任意 $\lambda \geq 0$ 的情况下保持沿着最优同伦解路径的两个条件。当将 λ 的值减小到 $\lambda-\tau$ 时，对于小的 τ 值，应满足公式（2-32）的两个条件。同伦算法最重要的参数步长 τ 和更新方向 δ。在第 l 阶段（如果存在$(X_\Lambda^T X_\Lambda)^{-1}$，同伦算法首先计算更新方向 δ，这通过求解问题（2-33）得到。因此，问题（2-33）的解决方案写成公式（2-34）。随后，同伦算法跟踪同伦解路径来计算到下一个临界点的步长 τ。即算法沿更新方向移动，直到不满足问题（2-32）中的任一约束。在该临界点，新的非零元素必须进入支持集 Λ。两个典型案例可能会导致一个新的临界点，其中违反了（2-32）的任何一个条件。通过计算 $\tau_l^*=\min(\tau_l^+,\tau_l^-)$可以很容易地获得导致临界点的最小步长，并且 τ_l^+和 τ_l^-由公式（2-35）公式（2-36）计算所得。其中 $p_i=x_i^T(y-x_i\alpha_i^l)$和 $\min(\cdot)_+$表示最小值仅在正参数上运算。τ_l^+是将索引 i^+中的非活动元素转换为活动元素的最小步长，即索引 i^+应该被添加到支持集中。τ_l^-是在索引 i^-处将非零有效元素的值收缩为零的最小步长。解决方案由公式（2-40）更新。同伦算法迭代地计算步长和更新方向，并更新同伦解及其相应的支持和符号序列，直到满足条件（2-39b），或者说满足对于 λ 的临界值 ε 从而达到问题（2-23）的解。

改进 BPDN 同伦算法的伪代码总结如下

输入：探针样本 y,度量矩阵 X

初始化：l=0、稀疏系数 α_0 及其支持集 Λ_0、λ_0 值、初始步长 τ、停止参数 ε

重复：

第一步：公式（2-34）计算更新方向

第二步：公式（2-40）更新稀疏系数

第三步：公式（2-35）、公式（2-36）计算 τ_l^+和 τ_l^-

第四步：计算更新步长：$\tau_l^*=\min(\tau_l^+,\tau_l^-)$

第五步：更新支持集：

如果 $\tau_l^+ != \tau_l^-$，将 i^+添加入支持集 Λ 中即 $\Lambda_{t+1}=\Lambda_t \cup i^+$

第六步：公式（2-33）更新符号序列集 u

第七步：更新 $\lambda_{t+1}=\lambda_{t-\tau}$，l=l+1

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

直到：$\lambda \leq \varepsilon$

输出：α_{t+1}

4.2 非负二次规划活动集算法批处理策略与分类器设计

在实验 l_1 范数最小化的同伦基追踪降噪算法、确定稀疏表示算法可行的基础上，采用解决生物信息高维数据降维的稀疏表示框架[76]并筛选出时间效率和准确率都比较好的非负二次规划活动集算法来运行大规模实验。实验证明该方法是解决甲骨文文字识别的有效方法。但该算法的输入是半正定方阵，在样本量过多的时候及其占据内存，在运行数据集部分样本量较大的数据集时直接内存溢出无法运行。为了解决在样本量较多内存溢出问题，在该算法中添加批处理策略以解决数据集过大占据内存过多问题。因添加了批处理策略，在子训练样本作为稀疏字典的时候，实验准确率有所下降，为了尽可能高的保持实验准确率，本文设计多个分类器，尽可能的在减少样本量的同时保持算法的识别准确率。实验结果表明，在部分数据集中，本文设计的分类器可以取得比原始算法更优越的结果。

为了解决非负二次规划活动集算法占据内存过多容易造成内存溢出的问题，在算法中添加了批处理的策略并设计使用多个分类器，最终选择效果最好的分类器的结果作为识别准确率。其中六个分类器简单概述如下：

Classifier1	1.本批次对每个类所有样本的稀疏值求平均值 2.多个批次对每个类的稀疏值相加 3.求稀疏值最大的对应的类的下标预测类
Classifier2	1.本批次对每个类所有样本的稀疏值求平均值 2.多个批次对每个类的稀疏值求最大值 3.求稀疏值最大的对应的类的下标预测类
Classifier3	1.本批次对每个类所有样本稀疏值求平均值 2.求稀疏值最大的对应的类的下标预测类
Classifier4	1.本批次所有训练样本稀疏值 KNN 聚类预测
Classifier5	1.本批次所有训练样本稀疏值最大值预测
Classifier6	1.本次迭代最后一次迭代 acc1 的值 2.多次迭代求平均值

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

批处理算法的伪代码简单总结如下：

算法目标：解决非负二次规划活动集算法内存不够的问题

输入：迭代次数，批处理 batch 百分比

重复：

第一步：加载数据集

重复：

第二步：抽取对应批次的训练样本及全量的测试样本

第三步：执行非负二次规划活动集算法返回该批次训练样本对每个测试样本的稀疏系数 spar

第四步：将每个类对应的训练样本稀疏系数叠加求平均值作为该类样本对于测试样本的稀疏系数 classspar

第五步：求解 classifier3=max(classspar), classifier4=max(spar), classifier5=KNN(spar)

第六步：classifier1=classifier1+classspar；classifier2=max(classifier2,classspar)

如果是最后一次迭代：

classifier6=classifier6+classifier1；

直到：所有批次结束

直到：所有迭代结束

classifier6=mean(classifier6)

4.3 实验设计

本节展示算法的实验设计相关细节。在 4.3.1 章节中介绍度量算法好坏的评判标准；4.3.2 中表明算法运行的软硬件环境配置；4.3.3 中简单介绍算法使用的数据集。

4.3.1 评价指标

度量一个算法的好坏，最直观的感受就是准确率。本文的评价标准也是如此。准确率计算为测试样本预测标签与实际标签一致的数目除以测试样本的总数目然后再乘以 100 作为最后的准确率。用公式表示如下：

$Acc=100*(sum(predict_label)==test_label)/N$

其中，Acc 表示准确率，N 表示测试样本的总数目，predict_label 表示算法预测标签，test_label 表示测试样本实际标签。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

除此之外，在准确率相当的情况下，运行时间变成了重要的评判依据，此处运行时间单位为秒。

4.3.2 软硬件配置

本文所有实验均在三种配置下运行，其软件环境保持一致，硬件配置有所差异，其相关配置显示如下：

硬件配置 1：

系统：Release Linux Mint 19.1 Tessa 64-bit

Kernel Linux 4.15.0-39-generic x86_64

MATE 1.20.1

处理器：Inter(R) Xeon(R) CPU E5-2667 v2 @ 3.30GHz *16

内存：62.8GB

硬件配置 2：

系统：Windows Server 2008 R2 Enterprise

处理器：Inter(R) Xeon(R) CPU E5-2660 v2 @ 2.20GHz *4

内存：32.0GB

硬件配置 3：

系统：Windows10 Professional

处理器：AMD Phenom(tm) X4 II B97 Processor 3.20GHz*4

内存：8.0GB

软件配置：

Matlab 版本：Matlab R2016b

4.3.3 数据集

实验使用的甲骨文数据集的具体详细信息，请参考第三章，此处不再赘述。

4.3.4 实验参数

改进的基追踪降噪同伦算法中有多个参数，大部分参数都可在初始化之后随着算法自动调整，唯有停止参数 ε，需要设置，相关实验中设置停止参数 ε=0.03。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

与改进基追踪降噪算法类似，批处理非负二次规划活动集算法中有多个参数，但大部分参数均可在算法自动调整，唯有停止参数 ε 需要手动指定。需要输入每次实验迭代次数与每次迭代过程中需要从剩余的训练样本中抽取数据的百分比。比如：本次迭代将数据分为两份，则 batch=[0.5,1]。

4.4 实验结果分析

为了更快的找到合适的算法应用于甲骨文数据集，本文在运行多种基准算法之后选取了比较合适的稀疏表示算法进行相关研究。实验结果包含两部分，基准实验和大规模实验，其中基准实验运行共 33 种算法，大规模实验运行共 13 种算法。实验结果分析包含准确率和运行时间。其中基准实验基于硬件配置 3 运行，大规模实验对比基于硬件配置 2 运行，大规模实验对比中部分实验基于硬件配置 1 运行。

4.4.1 基准实验对比分析

为了证明稀疏表示算法在甲骨文数据集上的优越性，本文做了大量的实验做比对，包括基于深度学习神经网络算法、经典贪婪算法 OMP 等。实验准确率是多次实验的平均值，多种基准算法的准确率和算法运行时间显示如下表 4-1 至表 4-3 所示。由表 4-1 可知，传统经典机器学习方法在基准数据集上运行时间短，准确率低，五种经典算法准确率均低于 30%。综合表现最优秀的算法为 KNN 算法（k=1），在运行时间较短的情况下可以保持较高准确率，达到 29.32%。EB 算法准确率和运行时间表现均不理想。

表4-1 传统非稀疏表示算法基准实验准确率和运行时间

算法名称	准确率（%）	时间（s）
EB	4.96	164.55
RF	20.45	75.50
NB	20.75	0.89
DA	27.22	4.28
KNN	29.32	1.33

由表 4-2 可知，深度学习算法准确率差距较大，其中 Inception 算法准确率低于 15%，而 VGG 算法准确率突破了 35%，但其准确率仍不理想。因其运行软硬件环境于其他算法不一致，此处不分析其运行效率。

表4-2 深度学习算法基准实验准确率和运行时间

算法名称	准确率（%）	时间（s）
Inception	12.03	--
VGG	35.49	--

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

传统非稀疏表示机器学习方法和深度学习分类算法，在甲骨文数据集上的准确率都较低，为了进一步提升甲骨文数据集识别的准确率，本文尝试了抗噪能力较强的基于稀疏表示的算法。由表 4-3 可知，稀疏表示算法准确率差距也极大，但其表现最优的算法较多。除 RNIK、RNIM、LNIK、LNIM、ADL-KSVD、LRC、WSRC 准确率较低之外，其余稀疏表示算法准确率均超过 30%。其中多个算法的准确率超过非稀疏表示算法，甚至多个算法准确率高于 40%。但部分算法在数据集样本量较多的时候运行时间过长。典型例子为 PALM 算法，运行时间为 793104.22 秒，折合为 9.18 天，显然该算法并不适合较大的数据集。综合运行时间和准确率，准确率在 35%以上，运行时间在 1000 以内的算法有 NNLS、BPDN、IBPDN、LNDK、RNDK、RNAK 多种，其中 RNDK、RNAK 算法在准确率和运行时间均表现优异。证明稀疏表示算法是解决甲骨文识别的有效算法。

表4-3 稀疏表示算法基准实验准确率和运行时间

算法名称	准确率（%）	时间（s）
RNIK	1.50	516.24
RNIM	1.50	508.79
LNIK	1.50	508.49
LNIM	1.50	508.29
ADL-KSVD	17.59	1191.07
LRC	17.59	3.62
WSRC	27.07	6078.25
OMP	31.88	10.86
LNDM	32.63	75.41
LNAM	32.63	73.02
NMNNLS	33.08	52197.36
SMNNLS	33.98	503.04
MSRC	33.98	503.04
RNDM	35.34	50.81
RNAM	35.34	86.07
PALM	35.64	793104.22
NNLS	37.14	218.00
BPDN	37.74	245.84
IBPDN	39.55	675.09
L1LS	40.00	6067.34
LNDK	40.75	75.45
FISTA	42.56	38506.55
DALM	42.56	7837.91
RNDK	45.26	46.01
RNAK	45.26	85.84
BRNAK	45.41	146.74

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

如表4-1、表4-2、表4-3中多种基准算法实验所示，深度学习或者传统方法在big50-99中效果都不算特别理想，证明数据集是存在一定难度的。与此同时，我们发现基于稀疏表示的研究方法在该数据集上保持更高的准确率。分析其原因是稀疏表示算法本身具有一定的抗噪能力，在噪声比较强的时候仍能保持较高的准确率。基准实验所使用算法详细信息来源于附录 A 中表 A-1,所有基准实验准确率和运行时间汇总结果来源于表 A-2。所有基准算法准确率比较如下图 4-1 所示，图中 1-33 代表不同算法的对应序号，序号与算法对应情况请参考表 A-2。

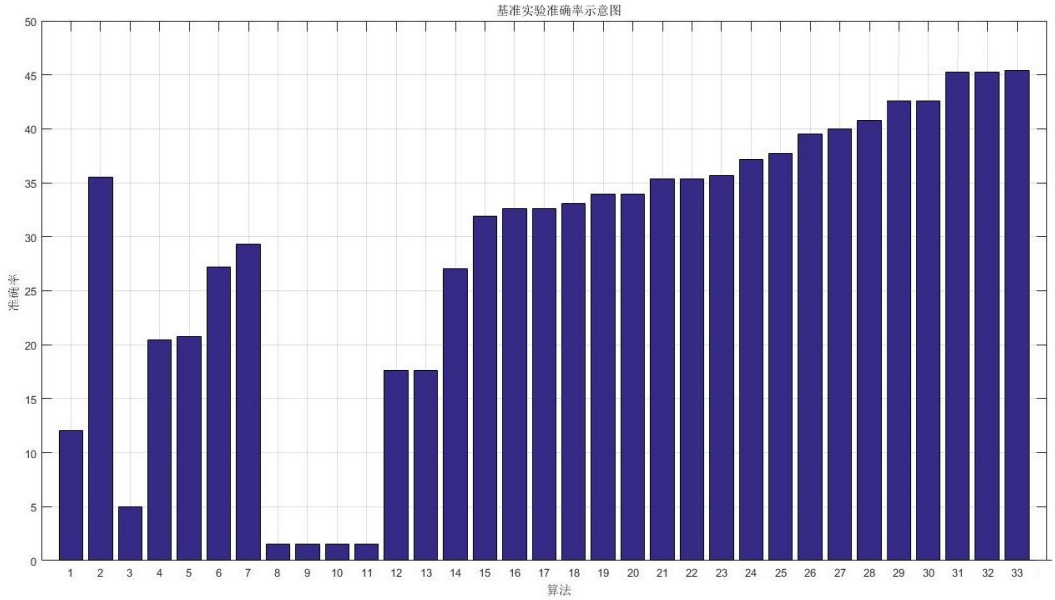


图 4-1 基准实验准确率示意图

4.4.2 大规模实验对比分析

基于算法准确率、运行时间、内存开销等方面的考虑，本文仅将部分算法进行大规模实验。该实验基于硬件配置其中深度学习的两种算法、传统方法中表现次优的 NB 和最优的 KNN 算法、稀疏表示算法中运行速度最快的 LRC、占用内存资源较少的 BPDN、IBPDN 以及稀疏表示算法中比较有代表性的线性内核的 NNLS 算法和高斯核函数的 RNAK 算法以及添加批处理 RNAK 的 BRNAK 算法。为了更清楚的看到随着数据集样本量对算法的影响，本节中所有实验结果将数据集按照数据量排序。有关数据集的基本信息，请参考表 3-2。

由表 4-4 可知，基于传统非稀疏表示的算法中我们将表现较为优秀的 NB 和 KNN 算法（k=1）进行了十五个数据集的大规模实验，其中实验结果包含算法运行时间和准

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

确率。由表 4-4(a)可知，NB 算法在准确率上低于 KNN。由表 4-4(b)但随着数据量的增加，运行时间比 KNN 快。

由表 4-5 可知，深度学习算法中 VGG 比 Inception 更加适合甲骨文识别任务。VGG 在十五个数据集中最高可达 64.20%的准确率，而同比 Inception 准确率仅有 42.79%。分析其数据集特征可知，两种算法在类别少，样本量较多的时候能够达到较高的准确率。因算法均在 GPU 加速的服务器运行，故未统计其相关运行时间信息。

表4-4 传统非稀疏表示算法大规模实验(a)准确率和(b)运行时间

数据集/准确率（%）	NB	KNN
big50-99	20.75	29.32
big10-49	10.09	14.44
balance50-99	14.86	25.82
balance10-49	4.34	8.48
small50-99	18.64	29.21
small10-49	9.07	14.40
big100-499	21.77	32.40
balance100-499	15.57	25.19
balance1000	36.57	66.87
small100-499	18.47	30.85
balance500	25.89	48.94
small1000	35.63	72.31
small500	28.65	56.75
big1000	27.92	63.81
big500	24.21	55.82

数据集/运行时间（s）	NB	KNN
big50-99	22	25
big10-49	60	58
balance50-99	100	94
balance10-49	68	66
small50-99	214	217
small10-49	82	86
big100-499	535	555
balance100-499	1373	1400
balance1000	2757	2847
small100-499	10757	10692
balance500	581	688
small1000	640	747
small500	749	1010
big1000	799	1162
big500	905	1504

表4-5 深度学习算法大规模实验准确率

数据集/准确率（%）	Inception	VGG
big50-99	12.03	35.49
big10-49	0.89	6.03
balance50-99	5.88	6.00
balance10-49	0.60	7.20
small50-99	6.95	26.95
small10-49	0.30	3.28
big100-499	12.51	38.38
balance100-499	3.51	22.53
balance1000	34.83	60.03
small100-499	6.77	27.41
balance500	16.57	40.66
small1000	42.79	64.20
small500	27.96	47.97
big1000	33.76	56.38
big500	22.38	44.25

所有非稀疏表示算法在 15 个数据集上准确率比较与示意显示如图 4-2 所示。由图中可知 KNN 和 VGG 算法在四种算法中占据优势。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

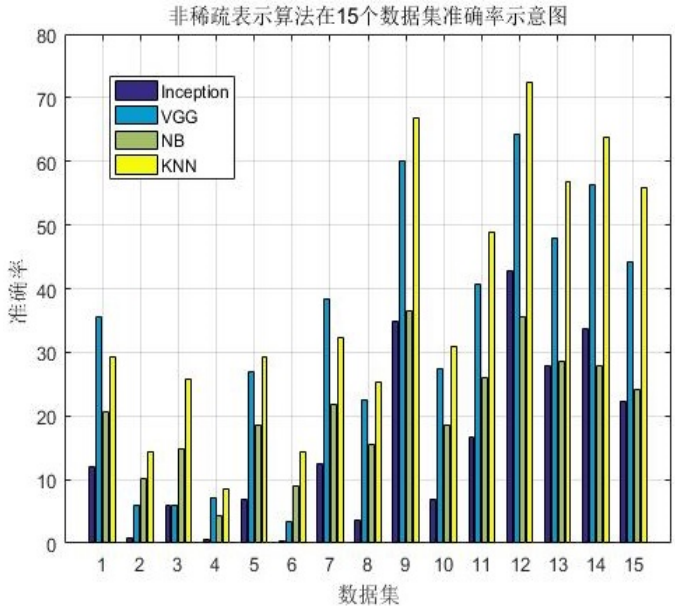


图 4-2 非稀疏表示算法在 15 个数据集准确率示意图

由表 4-6 可知，并非所有稀疏表示算法都适合甲骨文识别任务。但通过不同数据集准确率与运行时间的比较，发现一些规律。

LRC 算法准确率极低，但该算法在运行时间上占据绝对优势。 OMP 算法在算法准确率上优于 LRC 算法，其运行时间也远远超过 LRC，且随着数据量增加其运行时间成正比增加。NNLS 算法准确率高于 LRC 算法和 OMP 算法,但部分数据集准确率略低于 BPDN 和 IBPDN 算法，且算法运行时间较长，占据内存过多造成内存溢出。RNDK 算法准确率略微低于 RNAK 算法、高于 NNLS，且算法运行时间与 NNLS 相比也占据优势。RNAK 算法在多种算法中准确率占据优势，因其在多个配置环境下共同完成其实验，故没有统计其运行时间。BPDN 和 IBPDN 两种算法以及 BRNAK 是仅有的可在硬件配置 2 上完全运行十五个数据集的算法。其余算法均因为内存溢出无法完全运行其样本量较多的实验。BPDN 和 IBPDN 两个算法中，IBPDN 在部分数据集中保持较高准确率，部分数据集保持与 BPDN 算法准确率一致，但部分数据集上运行时间比 BPDN 算法要长。

BPDN 算法和 IBPDN 算法在准确率和运行时间以及资源占用方面取得了一个均衡，二者中 IBPDN 算法总体优于 BPDN。证明我们 4.4.1 节中的改进是有效的。

RNAK 算法保持最高准确率，但在硬件配置 2 中因占用资源过多而无法完整运行所有实验。BRNAK 算法解决了 RNAK 算法占用内存资源过多而无法运行数据量较大的算

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

法的问题。并在部分数据集中表现与 RNAK 算法相当甚至略优于 RNAK，并在数据量较大的 big1000 和 big500 可运行所有实验。与此同时，在这两个数据集中保持较高的准确率。

表4-6 稀疏表示算法大规模实验(a)准确率

数据集/准确率（%）	LRC	OMP	BPDN	IBPDN	NNLS	LNDK	RNDK	RNAK	BRNAK
big50-99	17.14	31.88	37.74	39.55	37.14	40.75	45.26	45.26	45.41
big10-49	17.71	15.23	17.71	19.98	18.79	21.17	22.55	23.44	18.40
balance50-99	34.01	24.97	25.82	25.82	30.45	35.08	36.84	37.57	37.29
balance10-49	13.20	8.18	8.86	10.95	9.20	12.01	13.12	13.37	13.80
small50-99	3.90	28.42	29.21	29.21	31.81	38.81	41.64	42.94	41.19
small10-49	9.28	13.88	16.10	18.23	15.97	19.59	21.76	22.44	17.80
big100-499	1.73	34.30	43.10	41.78	36.46	45.49	48.53	49.37	46.69
balance100-499	11.71	24.98	29.51	33.13	26.94	35.04	38.56	39.13	39.57
balance1000	6.42	66.82	66.87	66.87	67.15	77.40	79.08	79.77	80.30
small100-499	1.37	30.98	30.85	30.85	32.14	42.02	44.84	45.82	42.43
balance500	2.90	47.91	48.94	48.94	48.93	59.37	62.19	62.88	62.58
small1000	6.26	71.02	72.31	72.31	66.28	79.66	81.94	82.92	79.01
small500	3.06	55.82	56.75	56.75	--	65.78	69.38	69.76	61.14
big1000	7.95	62.65	63.81	63.81	--	--	--	--	71.07
big500	--	--	55.81	55.81	--	--	--	--	63.65

表4-6 稀疏表示算法大规模实验 (b)运行时间

数据集/运行时间（s）	LRC	OMP	BPDN	IBPDN	NNLS	RNDK	BRNAK
big50-99	2	31	449	1243	94	32	147
big10-49	5	152	1044	2642	484	83	503
balance50-99	9	374	314	318	983	161	1399
balance10-49	12	399	2332	6361	1611	197	2333
small50-99	13	603	499	490	1960	241	2959
small10-49	22	1127	4792	12524	5325	486	7962
big100-499	54	2311	9998	26112	10890	1014	15756
balance100-499	94	3483	14391	39966	24737	1484	34785
balance1000	155	3052	3757	3753	13704	1312	11141
small100-499	422	8453	12760	12798	58215	3604	49724
balance500	157	14493	4551	4526	172793	5271	101483
small1000	361	15050	3964	3923	144074	5686	58425
small500	940	35510	8032	8005	--	24358	233537
big1000	2241	56427	10306	10261	--	--	414266
big500	--	--	15845	15833	--	--	--

所有稀疏表示算法在 15 个数据集上准确率比较与示意显示如图 4-3 所示。由图中可知 BPDN 算法、IBPDN 算法、BRNAK 算法三个算法可以完整运行所有实验，其余算法均无法在硬件配置 2 下运行成功。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

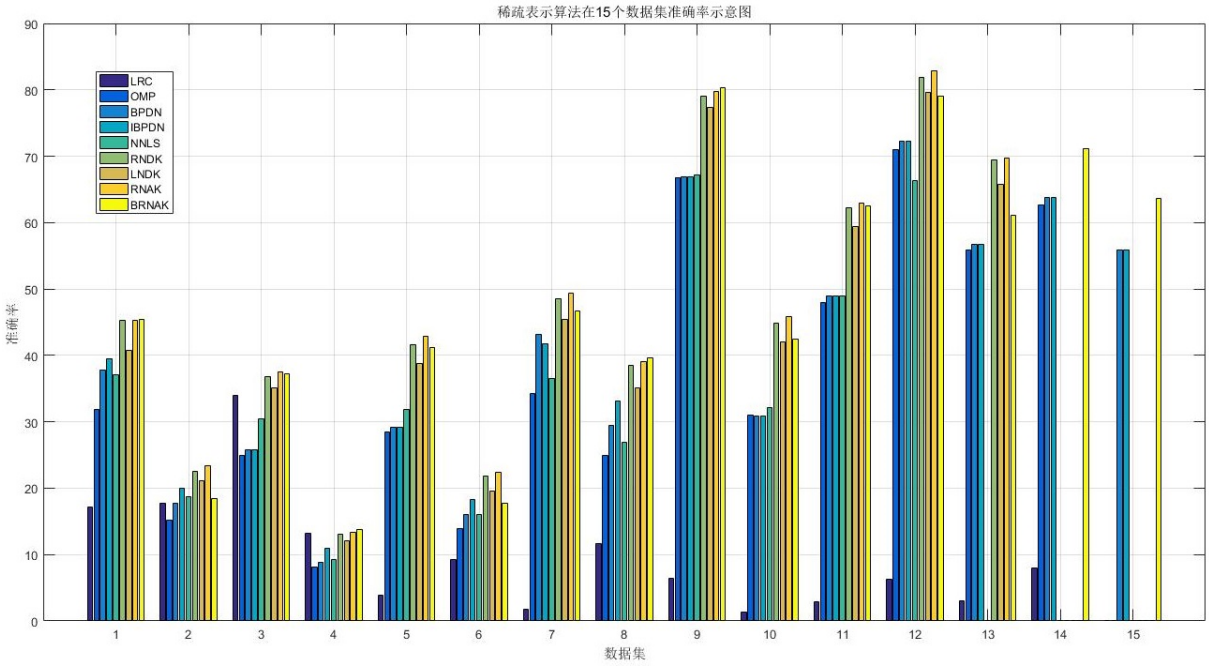


图4-3 稀疏表示算法在15个数据集准确率示意图

分析附表 B-1 可知，在 15 个数据集上准确率排行前五的算法依次排列如下表 4-7 所示。由表 4-7 可知，在 15 个数据集上，RNAK 算法和 BRNAK 算法准确率最高，RNDK 算法和 LNDK 算法准确率略逊于前两者，IBPDN 和 BPDN 算法准确率不算最高，但也基本都在前五之内。从中可以得出结论，基于稀疏表示的 BRNAK 算法在均衡数据集上较优，而 RNAK 算法在不均衡数据集上表现更为优异。

表4-7 大规模实验算法准确率分析表

数据集/准确率排名	1	2	3	4	5
big50-99	BRNAK	RNDK	RNAK	LNDK	IBPDN
big10-49	RNAK	RNDK	LNDK	IBPDN	NNLS
balance50-99	RNAK	BRNAK	RNDK	LNDK	LRC
balance10-49	BRNAK	RNAK	LRC	RNDK	LNDK
small50-99	RNAK	RNDK	BRNAK	LNDK	NNLS
small10-49	RNAK	RNDK	LNDK	IBPDN	BRNAK
big100-499	RNAK	RNDK	BRNAK	LNDK	BPDN
balance100-499	BRNAK	RNAK	RNDK	LNDK	IBPDN
balance1000	BRNAK	RNAK	RNDK	LNDK	NNLS
small100-499	RNAK	RNDK	BRNAK	LNDK	NNLS
balance500	RNAK	BRNAK	RNDK	LNDK	IBPDN、BPDN、KNN
small1000	RNAK	RNDK	LNDK	BRNAK	IBPDN、BPDN、KNN
small500	RNAK	RNDK	LNDK	BRNAK	IBPDN、BPDN、KNN
big1000	BRNAK	IBPDN	BPDN	KNN	OMP
big500	BRNAK	KNN	IBPDN	BPDN	VGG

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.5 本章小结

本章主要介绍了基于甲骨文数据集的大规模实验。首先介绍了改进后的基追踪降噪算法和非负二次规划活动集算法批处理策略与分类器设计，其次对比分析了在基准数据集 big50-99 上运行的共 33 种算法的准确率与运行时间，最后在 33 种算法中选出部分比较有代表性的算法进行共 15 个数据集的大规模实验，分析其准确率与运行时间。通过实验结果的对比分析可知，稀疏表示算法准确率高于其他算法。在稀疏表示算法上，15 个数据集准确率最高的三个算法为 RNAK、BRNAK、RNDK；运行时间最短的三个算法为 LRC、BPDN、IBPDN；占用资源最小的三个算法为 BPDN、IBPDN、BRNAK。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5 总结与展望

5.1 总结

稀疏表示的相关研究，起于压缩感知，但应用广泛，本文在此将稀疏表示算法应用于强噪声高残缺的甲骨文拓片文字识别。本文主要研究内容如下：

（1）在导师的指导下，与其他人合作标注整理了一个甲骨文检测与识别数据集。该数据集相关特征及其数据分布在第三章有所体现。

（2）基于甲骨文数据集的识别任务，本文对两种算法进行了微小的改进：其一是将基追踪降噪算法作为基准算法改进以提升准确率。根据第四章实验结果所示，改进基追踪降噪算法在小样本的数据集上具有较大优势，且改进前后的基追踪降噪算法在运行时间、准确率、资源利用率达到均衡。其二是在改进基追踪降噪算法的基础上，继续研究稀疏表示相关算法，将非负二次规划算法应用于甲骨文拓片识别任务，采用高斯核函数提取特征并使用 KNN 做为分类器，并为了解决算法占用内存问题，添加了批处理策略并尝试多种分类器。实验表明，该算法在多个数据集上表现最优。

（3）在两种算法的优化后分析了多种算法的性能。其中分为基准实验和大规模实验。基准实验部分主要分析其算法准确率和运行时间；大规模实验部分分析了算法在不同数据集上的准确率、运行时间，并分析其资源占用情况等，初步探索了准确率较高的 5 种算法。

5.2 展望

本文虽然在甲骨文拓片数据集上使用稀疏表示取得了一定的研究成果，但甲骨文拓片的识别仍有很多的问题需要解决，基本问题显示如下：

（1）稀疏表示算法在强噪声高残缺的甲骨文拓片识别中准确率并非十分理想，仍需要相关研究以提升甲骨文拓片数据集的准确率直至其可以应用于实际应用。

（2）稀疏表示算法在甲骨文数据集上尤其在样本量较多的数据集上，其对于内存和时间的要求比较高，如何在保持准确率不降低的前提下，降低算法对于内存的需求，以及快速收敛以节约时间，仍值得探讨。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（3）稀疏表示算法运行时间比较长，如何降低算法的运行时间也是未来的研究工作之一。

（4）当前甲骨文识别是基于单字的识别，如果甲骨文识别能够实现基于整张拓片的识别，加入版面分析及其语义信息在一定程度上提高识别的准确率。

（5）在样本量超过一千的甲骨文数据集上，深度学习算法也勉强可用，若是能扩大甲骨文数据集的样本量，再使用一些数据增强手段，深度学习方法在甲骨文拓片的识别上，同样有用武之地。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

[1] Lu X , Wu H , Yuan Y , et al. Manifold Regularized Sparse NMF for Hyperspectral Unmixing[J]. IEEE Transactions on Geoscience & Remote Sensing, 2013, 51(5):2815-2826.

[2] Yuan Y , Li X , Pang Y , et al. Binary Sparse Nonnegative Matrix Factorization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(5):772-777.

[3] Wright J, Ganesh A, Zhou Z, et al. Demo: Robust face recognition via sparse representation[C]// IEEE International Conference on Automatic Face & Gesture Recognition. 2009.

[4] Z.Zhang,Z.Li,B.Xie,L.Wang,and Y.Chen.Integrating globality and locality for robust representation based classifiation.Math.Problems Eng.,vol.2014,Mar.2014,Art.ID 415856.

[5] Xu Y , Zhang D , Yang J , et al. A Two-Phase Test Sample Sparse Representation Method for Use With Face Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(9):1255-1262.

[6] Cheng H , Liu Z , Yang L , et al. Sparse representation and learning in visual recognition: Theory and applications[J]. Signal Processing, 2013, 93(6):1408–1425.

[7] Tropp J A , Gilbert A C , Strauss M J . Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit[J]. Signal Processing, 2006, 86(3):572-588.

[8] Tropp J A . Algorithms for simultaneous sparse approximation. Part II: Convex relaxation[J]. Signal Processing, 2006, 86(3):589-602.

[9] M.Schmidt,G.Fung,and R.Rosaless.Optimization methods for 1-regularization. Univ. British Columbia,Vancouver,BC,Canada,Tech.Rep.TR-2009-19,2009.

[10] Natarajan B K . Sparse Approximate Solutions to Linear Systems[J]. SIAM Journal on Computing, 1995, 24(2):227-234.

[11] Brain extraction based on locally linear representation-based classification[J]. NeuroImage, 2014, 92:322-339.

[12] Lu X , Li X . Group sparse reconstruction for image segmentation[J]. Neurocomputing, 2014, 136:41-48.

[13] Elad M , Member S , IEEE, et al. On the Role of Sparse and Redundant Representations in Image Processing[J]. Proceedings of the IEEE, 2010, 98(6):972-982.

[14] Mallat S. A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way[M]. 2008.

[15] Wakin, Michael. Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity (Starck, J.-L. et al; 2010) [Book Reviews][J]. IEEE Signal Processing Magazine, 2011, 28(5):144-146.

[16] Elad M . Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing[J]. 2010.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[17] Bruckstein A M , Donoho D L , Elad M . From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images[J]. SIAM Review, 2009, 51(1):34-81.

[18] Xu Y , Zhang D , Yang J , et al. A Two-Phase Test Sample Sparse Representation Method for Use With Face Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(9):1255-1262.

[19] Wright J , Ma Y , Mairal J , et al. Sparse Representation for Computer Vision and Pattern Recognition[J]. Proceedings of the IEEE, 2010, 98(6):1031-1044.

[20] Donoho D L . Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4):1289-1306.

[21] Baraniuk R G . Compressive Sensing [Lecture Notes][J]. IEEE Signal Processing Magazine, 2007, 24(4):118-121.

[22] Candes E J , Romberg J , Tao T . Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information[M]. IEEE Press, 2006.

[23] E.J.Candès and M.B.Wakin.An introduction to compressive sampling.IEEE Signal Process. Mag. ,vol.25,no.2,pp.21-30,Mar.2008.

[24] Tsaig Y , Donoho D L . Extensions of compressed sensing.[J]. Signal Processing, 2006, 86(3):549-571.

[25] E.J.Candès.Compressive sampling.in Proc.Int.Congr.Math.Madrid,Spain,Aug.2006, pp. 1433 - 1452.

[26] Candès, Emmanuel, Romberg J . Sparsity and incoherence in compressive sampling[J]. Inverse Problems, 2007, 23(3):969-985.

[27] Michael J, Labahn R, Grüning T, et al. Evaluating Sequence-to-Sequence Models for Handwritten Text Recognition[J]. 2019.

[28] Liu W , Chen C , Wong K Y K . SAFE: Scale Aware Feature Encoder for Scene Text Recognition[J]. 2019

[29] Yan W , Chu H , Xinlong L , et al. A Hierarchical Fully Convolutional Network Integrated with Sparse and Low-Rank Subspace Representations for PolSAR Imagery Classification[J]. Remote Sensing, 2018, 10(2):342-.

[30] Verma A, Sharma M, Hebbalaguppe R, et al. Automatic Container Code Recognition via Spatial Transformer Networks and Connected Component Region Proposals[C]// IEEE International Conference on Machine Learning & Applications. 2017.

[31] http://res.imtt.qq.com/m_download_qb/release/20190109/qbload_new_1.html

[32] 李锋, 周新伦. 甲骨文自动识别的图论方法[J]. 电子与信息学报, 1996(s1):41-47.

[33] 周新伦, 李锋, 华星城, et al. 甲骨文计算机识别方法研究[J]. 复旦学报(自然科学版), 1996(5):481-486.

[34] 吕肖庆, 李沫楠, 蔡凯伟, et al. 一种基于图形识别的甲骨文分类方法[J]. 北京信息科技大学学报(自然科学版), 2010(S2):92-96.

[35] Guo L. Characters Feature Extraction Based on Neat Oracle Bone Rubbings [J].Telkomnika Indonesian Journal of Electrical Engineering,2013,11(9).

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[36] 高峰, 吴琴霞, 刘永革, et al. 基于语义构件的甲骨文模糊字形的识别方法[J]. 科学技术与工程, 2014, 14(30):67-70.

[37] Lin Meng. Recognition of Oracular Bone Inscriptions Using Template Matching. International Journal of Computer Theory and Engineering ,Vol.8,No.1,February 2016

[38] 商慧亮, 施若画, 鲍申杰, et al. 基于电路模拟法的甲骨文识别[J]. 2016.

[39] 顾绍通. 基于拓扑配准的甲骨文字形识别方法[J]. 计算机与数字工程, 2016, 44(10).

[40] Guo J , Wang C , Roman-Rangel E , et al. Building Hierarchical Representations for Oracle Character and Sketch Recognition[J]. IEEE Transactions on Image Processing, 2015, 25(1):104-118.

[41] 刘永革, 刘国英. 基于 SVM 的甲骨文字识别[J]. 安阳师范学院学报, 2017(2):54-56.

[42] Lin Meng.Recognition of Oracle Bone Inscriptions by Extracting Line Features on Image Processing.In Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)

[43] Chen J , Huo X . Theoretical Results on Sparse Representations of Multiple-Measurement Vectors[J]. IEEE Transactions on Signal Processing, 2007, 54(12):4634-4643.

[44] Yang J , Zhang L , Xu Y , et al. Beyond sparsity: The role of L1-optimizer in pattern classification[J]. Pattern Recognition, 2012, 45(3):1104-1118.

[45] Tropp J A , Gilbert A C . Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit[J]. IEEE Transactions on Information Theory, 2007, 53(12):4655-4666.

[46] Needell D , Vershynin R . Uniform Uncertainty Principle and Signal Recovery via Regularized Orthogonal Matching Pursuit[J]. Foundations of Computational Mathematics, 2007, 9(3).

[47] Saab R , Chartrand R , Yilmaz, Özgür. Stable sparse approximations via nonconvex optimization[C]// IEEE International Conference on Acoustics. IEEE, 2008.

[48] Chartrand R . Exact Reconstruction of Sparse Signals via Nonconvex Minimization[J]. IEEE Signal Processing Letters, 2007, 14(10):707-710.

[49] Xu Z . Data Modeling : Visual Psychology Approach and L 1 / 2 Regularization Theory[J]. Proceedings of the International Congress of Mathematicians Icm, 2010:3151-3184.

[50] Tibshirani R J . Regression Shrinkage and Selection via the LASSO[J]. Journal of the Royal Statistical Society. Series B: Methodological, 1996, 73(1):273-282.

[51] Efron B , Hastie T , Tibshirani J R . Least Angle Regression[J]. The Annals of Statistics, 2004, 32(2):407-451.

[52] Yang J, Zhang Y. Alternating Direction Algorithms for ℓ_1-Problems in Compressive Sensing[J]. Siam Journal on Scientific Computing, 2009, 33(1):250-278.

[53] Schmidt M, Fung G, Rosales R. Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches[C]// European Conference on Machine Learning. 2007.

[54] Nie F , Huang H , Cai X , et al. Efficient and Robust Feature Selection via Joint ℓ2, 1-Norms Minimization[C]// Advances in Neural Information Processing Systems 23: 24th Annual Conference

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada. Curran Associates Inc. 2010.

[55] Yang Y . l2,1-norm regularized discriminative feature selection for unsupervised learning[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2011.

[56] Shi X , Yang Y , Guo Z , et al. Face recognition by sparse discriminant analysis via joint L2,1-norm minimization[J]. Pattern Recognition, 2014, 47(7):2447-2453.

[57] Liu J, Ji S, Ye J. Multi-Task Feature Learning Via Efficient l2,1-Norm Minimization[J]. 2012.

[58] Hou C , Nie F , Li X , et al. Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection[J]. IEEE Trans Cybern, 2013, 44(6):793-804.

[59] Zhang L , Yang M , Feng X . Sparse Representation or Collaborative Representation: Which Helps Face Recognition?[C]// IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011. IEEE, 2011.

[60] Naseem I , Togneri R , Bennamoun M . Linear Regression for Face Recognition[J]. IEEE Transactions on Software Engineering, 2010, 32(11):2106-2112.

[61] [58] Donoho D L , Elad M . Optimally sparse representation in general (nonorthogonal) dictionaries via ???[J]. Proceedings of the National Academy of Sciences, 2003, 100(5):2197-2202.

[62] Liu L , Shao L , Zheng F , et al. Realistic action recognition via sparsely-constructed Gaussian processes[J]. Pattern Recognition, 2014, 47(12):3819-3827.

[63] Yuan Y , Lu X , Li X . Learning Hash Functions Using Sparse Reconstruction[J]. 2014.

[64] Amaldi E , Kann V . On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems[J]. Theoretical Computer Science, 1998, 209(1-2):237-260.

[65] Donoho D , Kakavand H , Mammen J . The Simplest Solution to an Underdetermined System of Linear Equations[J]. Communications on Pure & Applied Mathematics, 2007, 59(6):797-829.

[66] Candes E J , Tao T . Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?[J]. IEEE Transactions on Information Theory, 2006, 52(12):5406-5425.

[67] Lyu Q , Lin Z , She Y , et al. A comparison of typical ℓ p, minimization algorithms[J]. Neurocomputing, 2013, 119(16):413-424.

[68] S.Guo,Z.Wang,and Q.Ruan.Enhancing sparsity via `p (0<p<1)minimization for robust face recognition.Neurocomputing,vol.99,pp.592-602,Jan.2013.

[69] Ding C, Ding Z, He X, et al. R1-PCA:rotational invariant L1-norm principal component analysis for robust subspace factorization[C]// International Conference on Machine Learning. 2006.

[70] Zhang Z , Wang L , Zhu Q , et al. Noise modeling and representation based classification methods for face recognition[J]. Neurocomputing, 2015, 148:420-429.

[71] Osborne, M. A new approach to variable selection in least squares problems[J]. IMA Journal of Numerical Analysis, 2000, 20(3):389-403.

[72] Y.C.Pati,R.Rezaiifar,and P.S.Krishnaprasad.Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition.in Proc.27th Asilomar Conf.Signals,Syst.,Comput.,Nov.1993,pp.40-44.

[73] D. L. Donoho and Y. Tsaig,'Fast solution of 1 -norm minimization problems when the solution may be sparse,'IEEE Trans. Inf. Theory,vol. 54, no. 11, pp. 4789–4812, Nov.2008.

[74] Asif M S , Romberg J . Fast and Accurate Algorithms for Re-Weighted L1-Norm Minimization[J]. IEEE Transactions on Signal Processing, 2012, 61(23):5905-5916.

[75] http://www.baicaolu.com/article/219520.html

[76] Li Y , Ngom A . Sparse representation approaches for the classification of high-dimensional biological

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

data[J]. BMC Systems Biology, 2013, 7 Suppl 4(Suppl 4):S6.
[77] Szegedy C , Vanhoucke V , Ioffe S , et al. [IEEE 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Las Vegas, NV, USA (2016.6.27-2016.6.30)] 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Rethinking the Inception Architecture for Computer Vision[J]. 2016:2818-2826.
[78] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.
[79] Wang Q , Guo Y , Guo J , et al. Synthesis K-SVD based analysis dictionary learning for pattern classification[J]. Multimedia Tools and Applications, 2018, 77(13):17023-17041.
[80] Lu C Y , Min H , Gui J , et al. Face recognition via Weighted Sparse Representation[J]. Journal of Visual Communication and Image Representation, 2013, 24(2):111-116.
[81] Yang, AllenGanesh, ArvindSastry, et al. Fast L1-Minimization Algorithms and an Application in Robust Face Recognition: A Review[J]. 2010.
[82] Kim S J, Koh K, Lustig M, et al. An Interior-Point Method for Large-Scale l1-Regularized Least Squares[J]. IEEE Journal of Selected Topics in Signal Processing, 2008, 1(4):606-617.
[83] Beck A , Teboulle M . A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems[J]. SIAM Journal on Imaging Sciences, 2009, 2(1):183-202.
[84] Yang A Y , Zhou Z , Balasubramanian A G , et al. Fast L1-Minimization Algorithms For Robust Face Recognition[J]. IEEE Transactions on Image Processing, 2013, 22(8).
.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

附 录

附录 A 基准算法基本信息表

该部分介绍了运行基准实验的多种算法的基本介绍，算法中文名称，英文名称以及本文中为了描述方便而使用的简称，以及不同算法在 4.3.2 中硬件配置 3 上运行的算法实验时间与准确率。其中表 A-1 记录算法中英文对照及算法简称。表 A-2 记录基准实验中使用的算法分类、算法的英文简称、准确率和在硬件配置 3 上的运行时间。

表A-1 算法名中英文对照及算法简称

算法中文名称	算法英文名称	简称
InceptionNet V4[77]	Inception NetV4	Inception
视觉几何组[78]	Visual Geometry Group	VGG
集成学习	Ensembles for Boosting	EB
随机森林	Random Forest	RF
朴素贝叶斯	NaiveBayes	NB
鉴别分析	Discriminant Analysis	DA
K 近邻	k-Nearest Neighbor	KNN
高斯核非负二次规划内点 KNN[76]	Rbf NNQP Interior-point+KNN	RNIK
高斯核非负二次规划内点最大值[76]	Rbf NNQP Interior-point+MAX	RNIM
线性核非负二次规划内点 KNN[76]	Linear NNQP interior-point+KNN	LNIK
线性核非负二次规划内点最大值[76]	Linear NNQP interior-point+MAX	LNIM
基于合成 K-SVD 的分析词典学习[79]	K-SVD+Analysis Dictionary Learning	ADL-KSVD
线性回归分类[76]	Linear regression classifier	LRC
加权稀疏表示[80]	Weighted Sparse Representation Classification	WSRC
正交匹配追踪[43]	Orthogonal Matching Pursuit	OMP
线性核非负二次规划分解最大值[76]	Linear NNQP Decomposition+MAX	LNDM
线性核非负二次规划活动集最大值[76]	Linear NNQP active-set+MAX	LNAM
基于 NMF 元样本的线性分类器[76]	NMF meta-sample based NNLS	NMNNLS
基于 SVD 元样本的线性分类器[76]	SVD meta-sample based NNLS	SMNNLS
基于 SVD 元样本的稀疏分类器[76]	SVD meta-sample based SRC2	MSRC
高斯核非负二次规划分解最大值[76]	Rbf NNQP Decomposition+MAX	RNDM
高斯核非负二次规划活动集最大值[76]	Rbf NNQP Active-set+MAX	RNAM
原始增广拉格朗日方法[81]	primal augmented Lagrangian method	PALM
线性核非负二次规划活动集 KNN[76]	Linear NNQP active-set+KNN	NNLS
基追踪降噪[45]	Basis Pursuit De-Noising	BPDN
本文的改进基追踪降噪	Improve Basis Pursuit De-Noising	IBPDN
基于 L1 正则的最小二乘[82]	ℓ1-Regularized Least Squares	L1LS
线性核非负二次规划分解 KNN[76]	Linear NNQP Decomposition+KNN	LNDK
快速迭代收缩阈值[83]	Fast Iterative Shrinkage Thresholding	FISTA

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

	Algorithm	
双增广拉格朗日[84]	Dual Augmented Lagrangian Method	DALM
高斯核非负二次规划分解 KNN[76]	Rbf NNQP Decomposition+KNN	RNDK
高斯核非负二次规划活动集 KNN[76]	Rbf NNQP Active-set+KNN	RNAK
批处理高斯核非负二次规划活动集 KNN	Batch Rbf NNQP Active-set+KNN	BRNAK

表 A-2 基准实验算法准确率与时间

算法分类	序号	算法名称	准确率（%）	时间（s）
深度学习	1	Inception	12.03	--
	2	VGG	35.49	--
传统非稀疏表示	3	EB	4.96	164.55
	4	RF	20.45	75.50
	5	NB	20.75	0.89
	6	DA	27.22	4.28
	7	KNN	29.32	1.33
稀疏表示	8	RNIK	1.50	516.24
	9	RNIM	1.50	508.79
	10	LNIK	1.50	508.49
	11	LNIM	1.50	508.29
	12	ADL-KSVD	17.59	1191.07
	13	LRC	17.59	3.62
	14	WSRC	27.07	6078.25
	15	OMP	31.88	10.86
	16	LNDM	32.63	75.41
	17	LNAM	32.63	73.02
	18	NMNNLS	33.08	52197.36
	19	SMNNLS	33.98	503.04
	20	MSRC	33.98	503.04
	21	RNDM	35.34	50.81
	22	RNAM	35.34	86.07
	23	PALM	35.64	793104.22
	24	NNLS	37.14	218.00
	25	BPDN	37.74	245.84
	26	IBPDN	39.55	675.09
	27	L1LS	40.00	6067.34
	28	LNDK	40.75	75.45
	29	FISTA	42.56	38506.55
	30	DALM	42.56	7837.91
	31	RNDK	45.26	46.01
	32	RNAK	45.26	85.84
	33	BRNAK	45.41	146.74

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

附录 B 大规模实验基本信息表

其中表 B-1 表示在十五个数据集上运行的所有算法的准确率。表 B-2 表示在十五个数据集上运行的所有算法的运行时间。其中大部分实验结果来源于硬件配置 2，部分实验结果来源于硬件配置 1，硬件配置 1 运行的算法，其运行时间在表 B-2 中显示为‘--’。

表B-1 大规模实验算法准确率总表

数据集/准确率（%）	Inception	VGG	NB	KNN	LRC	OMP	BPDN	IBPDN	NNLS	RNDK	LNDK	RNAK	BRNAK
big50-99	12.03	35.49	20.75	29.32	17.14	31.88	37.74	39.55	37.14	45.26	40.75	45.26	45.41
big10-49	0.89	6.03	10.09	14.44	17.71	15.23	17.71	19.98	18.79	22.55	21.17	23.44	18.40
balance50-99	5.88	6.00	14.86	25.82	34.01	24.97	25.82	25.82	30.45	36.84	35.08	37.57	37.29
balance10-49	0.60	7.20	4.34	8.48	13.20	8.18	8.86	10.95	9.20	13.12	12.01	13.37	13.80
small50-99	6.95	26.95	18.64	29.21	3.90	28.42	29.21	29.21	31.81	41.64	38.81	42.94	41.19
small10-49	0.30	3.28	9.07	14.40	9.28	13.88	16.10	18.23	15.97	21.76	19.59	22.44	17.80
big100-499	12.51	38.38	21.77	32.40	1.73	34.30	43.10	41.78	36.46	48.53	45.49	49.37	46.69
balance100-499	3.51	22.53	15.57	25.19	11.71	24.98	29.51	33.13	26.94	38.56	35.04	39.13	39.57
balance1000	34.83	60.03	36.57	66.87	6.42	66.82	66.87	66.87	67.15	79.08	77.40	79.77	80.30
small100-499	6.77	27.41	18.47	30.85	1.37	30.98	30.85	30.85	32.14	44.84	42.02	45.82	42.43
balance500	16.57	40.66	25.89	48.94	2.90	47.91	48.94	48.94	48.93	62.19	59.37	62.88	62.58
small1000	42.79	64.20	35.63	72.31	6.26	71.02	72.31	72.31	66.28	81.94	79.66	82.92	79.01
small500	27.96	47.97	28.65	56.75	3.06	55.82	56.75	56.75	--	69.38	65.78	69.76	61.14
big1000	33.76	56.38	27.92	63.81	7.95	62.65	63.81	63.81	--	--	--	--	71.07
big500	22.38	44.25	24.21	55.82	--	--	55.81	55.81	--	--	--	--	63.65

表B-2 大规模实验算法运行时间总表

数据集/运行时间（s）	NB	KNN	LRC	OMP	BPDN	IBPDN	NNLS	RNDK	LNDK	BRNAK
big50-99	22	25	2	31	449	1243	94	32	49	147
big10-49	60	58	5	152	1044	2642	484	83	134	503
balance50-99	100	94	9	374	314	318	983	161	285	1399
balance10-49	68	66	12	399	2332	6361	1611	197	340	2333
small50-99	214	217	13	603	499	490	1960	241	446	2959
small10-49	82	86	22	1127	4792	12524	5325	486	884	7962
big100-499	535	555	54	2311	9998	26112	10890	1014	1938	15756
balance100-499	1373	1400	94	3483	14391	39966	24737	1484	2897	34785
balance1000	2757	2847	155	3052	3757	3753	13704	1312	2861	11141
small100-499	10757	10692	422	8453	12760	12798	58215	3604	7391	49724
balance500	581	688	157	14493	4551	4526	172793	5271	11820	101483
small1000	640	747	361	15050	3964	3923	144074	5686	13868	58425
small500	749	1010	940	35510	8032	8005	--	24358	43628	233537
big1000	799	1162	2241	56427	10306	10261	--	--	--	414266
big500	905	1504	--	--	15845	15833	--	--	--	--

表B-3表示批处理二次规划活动集算法在六个分类器中算法的准确率，其中大规模算法比对中，采用的是分类器6的最终结果作为算法实验结果。其中batch表示分批实验

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中第几批次。cla表示分类器，time表示当前批次算法运行时间，sumtime表示该算法多批次运行时间相加的值，即算法总体的运行时间。

表B-3 BRNAK算法准确率总表

数据集/准确率(%)	batch	cla1	cla2	cla3	cla4	cla5	time	cla6	sumtime
big50-99	1	35.94	35.94	35.94	35.79	30.08	73.65	45.41	146.74
	2	45.41	41.80	36.54	36.09	28.72	73.08		
big10-49	1	11.47	11.47	11.47	17.11	14.24	246.75	18.40	502.53
	2	18.40	15.83	13.55	18.40	14.84	255.78		
balance50-99	1	29.15	29.15	29.15	29.15	25.03	701.73	37.29	1399.47
	2	37.29	35.65	30.79	30.79	25.48	697.74		
balance10-49	1	8.86	8.86	8.86	8.86	7.54	1187.65	13.80	2332.82
	2	13.80	12.61	9.97	9.97	8.56	1145.17		
small50-99	1	34.12	34.12	34.12	35.25	28.53	1426.09	41.19	2959.05
	2	41.19	39.55	33.33	34.69	27.63	1532.96		
small10-49	1	12.69	12.69	12.69	17.42	13.12	3990.55	17.80	7961.97
	2	17.80	14.40	11.41	16.14	13.33	3971.42		
big100-499	1	39.59	39.59	39.59	42.58	33.21	7573.90	46.69	15756.01
	2	46.69	42.29	40.71	43.90	34.39	8182.11		
balance100-499	1	32.69	32.69	32.69	32.69	24.44	17394.03	39.57	34785.49
	2	39.57	37.34	32.41	32.41	25.21	17391.47		
balance1000	1	77.78	77.78	77.78	77.78	65.37	5584.67	80.30	11141.02
	2	80.30	79.61	77.55	77.55	65.42	5556.35		
small100-499	1	35.35	35.35	35.35	40.91	31.09	24792.49	42.43	49723.78
	2	42.43	37.86	34.75	40.00	30.03	24931.29		
balance500	1	55.74	55.74	55.74	55.74	44.44	51297.74	62.58	101482.91
	2	62.58	61.52	57.22	57.21	46.70	50185.17		
small1000	1	74.75	74.75	74.75	78.32	68.86	29337.59	79.01	58425.44
	2	79.01	76.29	76.28	78.72	69.42	29087.85		
small500	1	54.36	54.36	54.36	64.05	53.25	117620.83	61.14	233536.57
	2	61.14	55.14	55.75	64.63	54.20	115915.73		
big1000	1	64.70	64.70	64.70	68.99	58.59	198132.35	71.07	414265.56
	2	71.07	68.08	66.29	70.38	60.12	216133.21		
big500	1	59.46	59.46	59.46	59.46	46.65	73794.45	63.65	147553.26
	2	63.65	62.74	58.44	58.44	45.82	73758.81		

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

致 谢

时间转瞬即逝，回顾研究生生涯之初的新奇与喜悦，三年来积淀成为学术上上下求索的动力与支撑。三年生活中，有过期望与失望，有过犯错与迷茫，但最终收获成功的喜悦，回过头来看，曾经的那些好的坏的，都变成了成功路上点缀的花，拂过的风。正是这些完美与不完美的交集，才是人生，三年来我的能力得到了综合的提升，现在回顾以往，感谢三年的时光，将我雕塑成了一个更加强大的人。感谢那些陪伴过一段光阴的人，让我的生活和学术都得到了帮助与提升。

首先感谢我的导师，张重生教授。在企业项目和科研项目中，张老师井井有条的安排与严谨的思维逻辑给我造成很深的影响，使我做事考虑更加周全，锻炼了我的团队协作能力和独立作业能力。在科研上，张老师是一个非常有想法与创造力的导师，具有超凡的精力与活跃的思维。在张老师的指导下，我的代码实现能力与理论分析能力得到极大提高。

其次要感谢试验室的师兄弟姐妹们：毕璟君、王朋友、王弯弯、赵冬冬、于珂珂、张愿、肖怀远、裴宸平、樊力文、彭国文、郭志强、闫红松、陶月锋、付飞飞、闫璟哲、王陆敏、田园、曹爽、黄雅冰、纵瑞兴、牛钦、徐仕鑫。在我初入试验室，是师兄师姐带我融入环境，引导我如何学习、查资料，在科研的道路上有一个良好的开端。是我的同届小伙伴与我一同成长，在你们的陪伴下，我度过了一段很开心快乐的时光。是我可爱的师弟师妹们给实验室注入了新的活力，替我分担压力，为实验室营造了良好的学习氛围，无论是在科研还是项目实践，都给予我很多帮助。另外感谢我的室友们，在生活上精神上给予我的帮助；感谢宿管阿姨，在我偶尔忘记带钥匙的时候默默地将备用钥匙借给我用；感谢学院张蕾老师，帮我们修空调修网线抓耗子等。最后还要感谢我的家人，他们是我心灵的净化者也是我背后强大的支撑者。

面对即将到来的告别和分离，莫名有些感伤，临表涕零，不知所言，祝大家在以后的工作和生活中，心想事成，万事如意！

作者：王慧慧

日 期：2019 年 6 月 10 日

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

明德　新民
止於　至善

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net