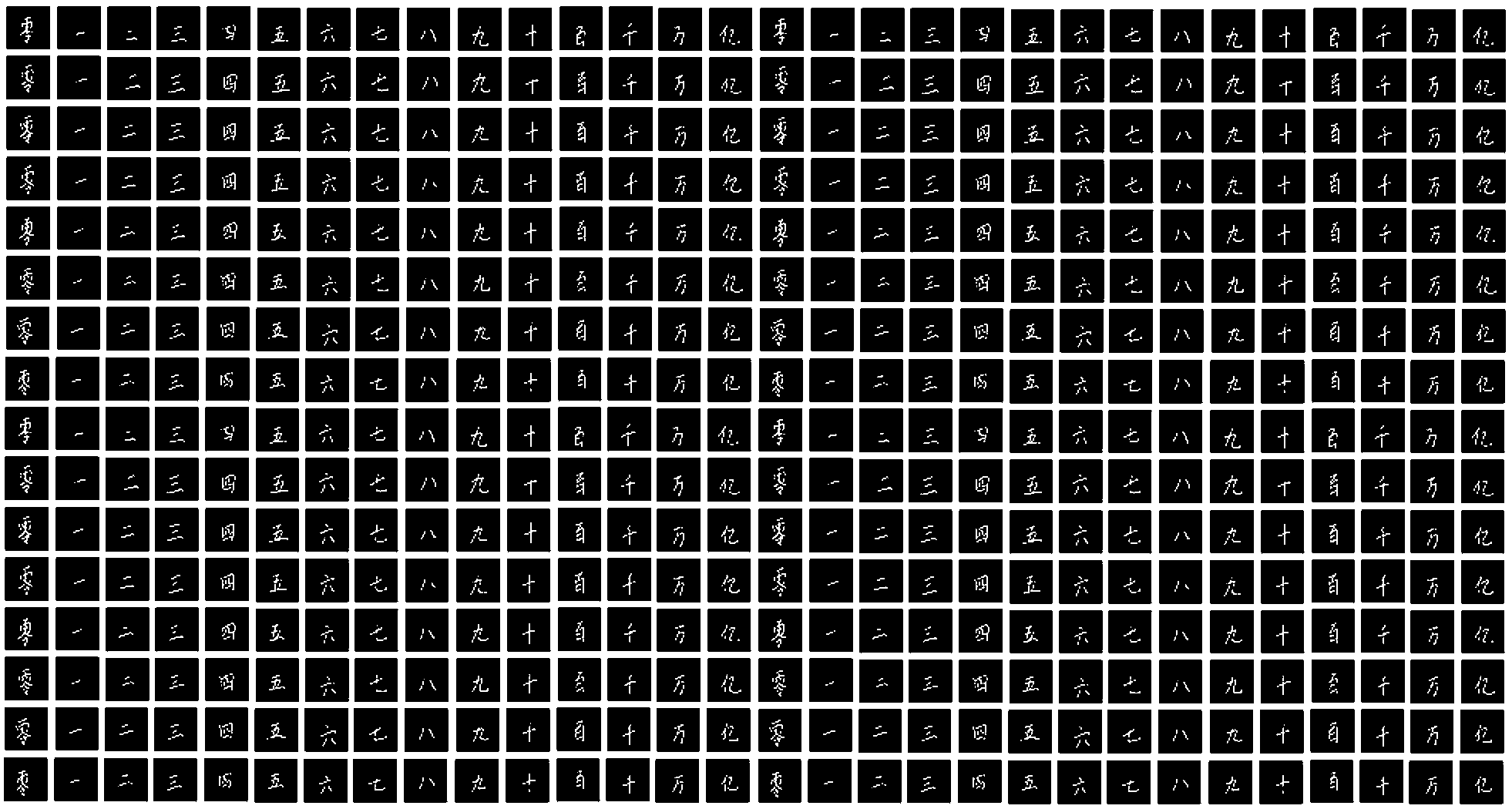
中文数字手写字符图像项目

**1 数据集**



数据集网址：<https://www.kaggle.com/gpreda/chinese-mnist>

数据集使用[在纽卡斯尔大学](https://data.ncl.ac.uk/articles/Handwritten_Chinese_Numbers/10280831/1" \t "_blank)的[项目](https://data.ncl.ac.uk/articles/Handwritten_Chinese_Numbers/10280831/1)框架中收集的数据。

一百名中国人参加了数据收集。每个参与者用标准的黑色墨水笔在一张白色A4纸上绘制的表格中用15个指定区域书写所有15个数字。每个参与者重复此过程10次。每张纸都以300x300像素的分辨率进行扫描。  
 结果是一个包含15000张图像的数据集，每个图像代表一组15个字符中的一个字符（每个志愿者有10个样本，每组100个志愿者）。

从[原始项目页面](https://data.ncl.ac.uk/articles/Handwritten_Chinese_Numbers/10280831/1)下载了原始图像。基于图像名称，为每个图像创建了一个索引，如下所示：

original name (example): Locate{1,3,4}.jpg

index extracted: suite\_id: 1, sample\_id: 3, code: 4

resulted file name: input\_1\_3\_4.jpg

* 索引文件 chinese\_mnist.csv
* 包含15,000张jpg图像的文件夹，尺寸为64 x64。有关详细信息，请参见images文件夹说明。

感谢：纽卡斯尔大学的K Nazarpour博士和M Chen博士，他们收集了数据。



映射

**2 实现方法**

* 本实验项目旨在评估同学们对于知识课上讲过的机器学习模型的掌握与应用的程度。项目实现方法鼓励采用课程内的机器学习模型，包括但不限于逻辑回归，支持向量机，聚类，神经网络等。
* 对于数据集的导入和数据预处理，可以参考网页<https://www.kaggle.com/gpreda/chinese-mnist>的discuss。
* 项目实现编程语言包括但不限于python，matlab，c++等。

**3 项目要求**

* **不允许简单复制网络代码，不允许复制同学代码，避免学术不端行为。**
* **不允许调用机器学习相关库函数，课程相关的关键步骤（模型的建立，模型的训练等）简单使用已有库函数或者API将大幅降低项目成绩。**
* 项目采用分组形式：每组1~3人，项目里应包含代码，测试结果，各成员列表以及分工，发表所用的ppt等内容。
* 项目材料最后按组各自打包提交。

**4 评估**

* 项目完成之后需要作演示，演示的内容包括测试集的识别准确度(这个需要和实验报告一致)，模型的具体思路讲解，额外加分项部分等。
* 项目中**准确率（以及其他机器学习评估指标）**，**算法执行过程或模型可视化**，以及**多种算法的比较**将作为主要的评估标准。
* 代码需要成员自主完成，对自己团队代码的熟悉程度将一定程度影响最后评分。
* 鼓励适当减少训练集数量，用较少的图片实现较高的准确率。【额外加分项，鼓励有能力的同学尝试】
* 任何你觉得能让项目更好的功能，都可以作为加分项。