基于关键偏旁部首分析的汉字识别系统

一.成果完成人及所在单位： 罗国锋，计算机与信息工程学院 指导教师：尹华一

二.成果简介：

汉字识别是文献数字化的重要环节之一，其精度影响数字文献的质量。文献数字化过程面临识别未见字的难点，这极大影响汉字的识别精度，尤其是在古籍文献数字化中，存在许多难以收集的生僻字。因此，提高汉字识别系统识别未见类汉字的能力有利于促进文献数字化进程。本研究基于零样本汉字识别系统，以汉字的偏旁部首为研究对象，将信息论和汉字识别方法相结合，探讨了偏旁部首对于识别汉字的重要性，提出了基于关键偏旁部首分析的汉字识别系统。这项研究使汉字识别系统具有识别未见字的能力，对解决现实生活中汉字类别不断增多但是汉字识别系统类别固定的矛盾具有重要意义。

三. 主要研究内容及创新点：

针对汉字类别不断变化的问题，本研究从汉字组成角度，将汉字识别问题转换成汉字中的偏旁部首及其结构信息的识别问题（图 1 为汉字的树形结构图及一些典型汉字的偏旁部首分解序列）。为了提升偏旁部首预测的准确性，本系统为每个偏旁部首建立可学习的原型表示，并且通过距离度量方式来识别偏旁部首。在偏旁部首序列到汉字的翻译过程中，本研究受人类识别汉字过程的启发，对汉字识别过程进行建模并且提出偏旁部首关键程度度量方法和汉字不确定性消除算法。汉字不确定性消除算法将预测的偏旁部首序列作为系统获取到的信息，利用偏旁部首的关键程度迭代选择出信息量最大的偏旁部首来消除汉字的不确定性，直到识别出这个汉字。基于关键偏旁部首分析的汉字识别系统（见图 2）的识别过程分三步。首先利用卷积神经网络提取输入图片的高维特征，然后使用带有注意力机制的循环神经网络预测其中包含的偏旁部首，最后使用汉字不确定性消除算法确定汉字类别。研究结果指出，原型学习和汉字不确定性消除算法有助于提高汉字识别系统的识别能力。

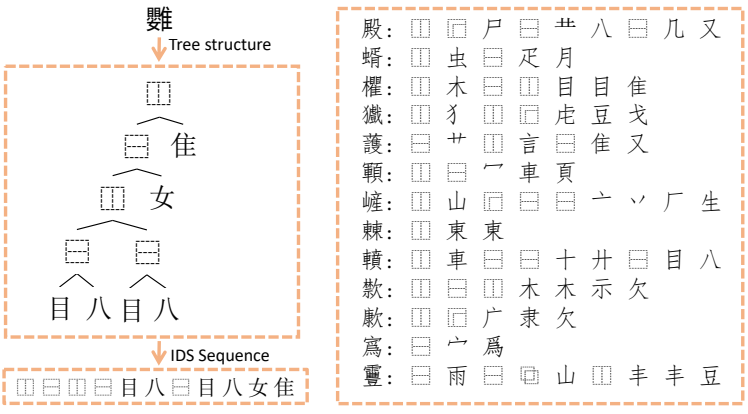


图 1 汉字结构图

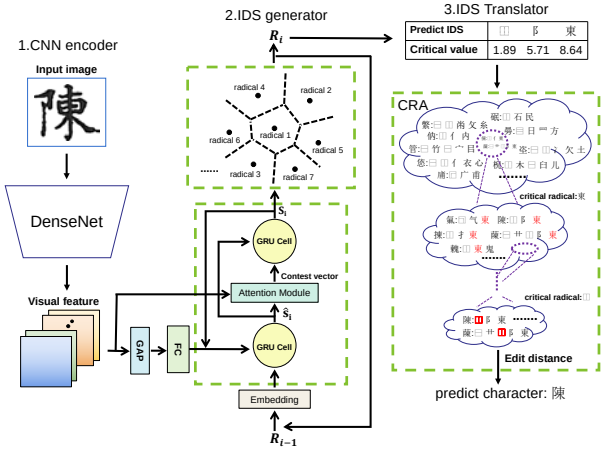


图 2 汉字识别系统图

四. 主要研究成果及应用情况

在 ICPR2022 发表论文 1 篇。

联系人：罗国锋， 电 话：18030190219 Email: 1120947176@qq.com