中 文 信 息 学 报
第7卷 第4期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.7 No.4

# 汉字认知心理研究对机器自动识别汉字的启示

韩布新　　陈一凡

(中国科学院心理研究所)　(北京信息工程学院)

【摘要】几项认知心理学实验研究从不同角度一致证实，方块汉字的四个等分象限所含的字形特征信息量不同。在人类识别汉字时作用也不一样。其中以左上象限最重要，右下象限的作用则要弱得多。本文结合部件的象限位置频率，讨论了这些结果对汉字机器识别的一些启示。

心理学的实验研究证实，整字的空间部位、字形、结构类型、笔画数等字形因素，以及整字频率、部件频率、部件组合频率、部件位置频率等均对汉字识别有重要的影响[1][2]。现代认知心理学的研究成果已经成功地应用于建立汉字键盘输入的认知模型[3]，因此本文结合几个心理学实验，谈谈汉字空间部位在汉字识别中的作用，及其对于计算机自动识别汉字领域研究可能存在的启发。

## 一、关于汉字识别中空间部位作用的几个心理学实验

汉字的不同空间部位在汉字识别中的作用在汉字的心理学研究之初便受到重视。周先庚（1929）在美国斯坦福大学进行的一项研究中，将汉字切成上下或左右两半，结果发现阅读只露出左半部或上半部所费的时间短于阅读只露出右半部或下半部的时间，前者，比较容易重构出整个汉字。他认为这是由于汉字字根大部分都在左边或上边，并且书写的笔画顺序也绝大多数是先写左边或上边的缘故[4]。因此字的空间部位是汉字阅读方面重要的影响因素之一。

六十年代认知心理学兴起后，对阅读尤其是对字词识别的研究掀起了新的高潮。一些心理学家结合汉字机器识别的需要开展了研究，试图从人类再认汉字的心理特点得到有益的启发。曾性初等（1965）用笔画省略法研究表明，在省掉同等比例的笔画情况下，被试比较容易恢复出"略后式"（按笔顺将该字最后一些笔画省掉）字的完整结构来；而"省前式"（按笔顺将该字前面的一些笔画省掉）字则相对较难[6]。表1节录了其实验的部分

①本文1993年3月29日收到

结果，从中可以看出，在省略比例为10%、20%和30%的情况下，两种方式的差别不明显．但若省略比例超过40%，则差别逐渐增大．从图1我们可以更直观地看出这一变化规律．

表1　不同省略条件下正确恢复字百分率中数及全距

| 省略百分比 | 略后式 | | 省前式 | |
|---|---|---|---|---|
| | 百分率中数 | 全距 | 百分率中数 | 全距 |
| 10 | 100 | 78–100 | 100 | 79–100 |
| 20 | 100 | 79–100 | 100 | 69–100 |
| 30 | 100 | 77–100 | 100 | 43–100 |
| 40 | 100 | 50–100 | 94 | 0–100 |
| 50 | 82 | 0–100 | 61 | 0–100 |
| 60 | 39 | 0–100 | 31 | 0–100 |

* 据文献[5]表1,数据改编

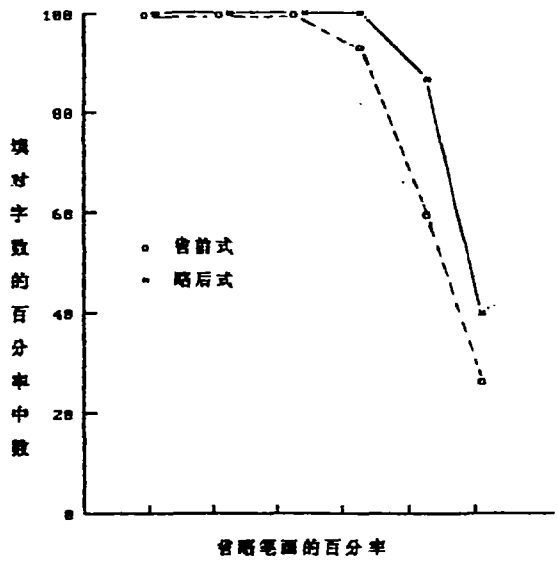


图1　不同省略百分率下填对字数示意图

彭瑞祥、喻柏林（1980）将汉字分为左上角、右上角、左下角和右下角等四个面积相等的象限．分别掩盖每个象限一次（参见表2示例），要求被试再认被掩盖后的不完整汉字，并根据再认的难易程度作出主观评价[8]．实验结果如表3所示，从中可以看出独体字无论掩盖哪个象限，对再认都没有明显影响，说明其特征信息不是集中分布于某一固定象限内．而左右型或上下型字的左上角被掩盖后，再认百分比明显下降，特别是左右字．说明在此类结构的字中，左上角所含的特征信息比较重要．掩盖右下角后，字仍很容易被

再认，因此这部分信息相对不太重要。就左右或上下字而言，左边的特征或信息比右边的重要一些，上边的比下边的重要一些。图 2 直观地表示了这一结果。

表2 掩盖不同象限后汉字示例

| 整字 | 先 | 去 | 走 | 段 | 胆 | 放 | 态 | 备 | 青 |
|---|---|---|---|---|---|---|---|---|---|
| 左上角 | | | | | | | | | |
| 右上角 | | | | | | | | | |
| 左下角 | | | | | | | | | |
| 右下角 | | | | | | | | | |

表 3 再认不完整汉字的百分比率

| 被掩盖的象限 | 结构类型 | | |
|---|---|---|---|
| | 左右型 | 上下型 | 独体字 |
| 左上角 | 61.58 | 72.28 | 84.74 |
| 右上角 | 71.58 | 82.11 | 80.35 |
| 左下角 | 77.02 | 84.56 | 87.02 |
| 右下角 | 86.32 | 95.26 | 92.46 |

* 引自文献〔6〕表 2

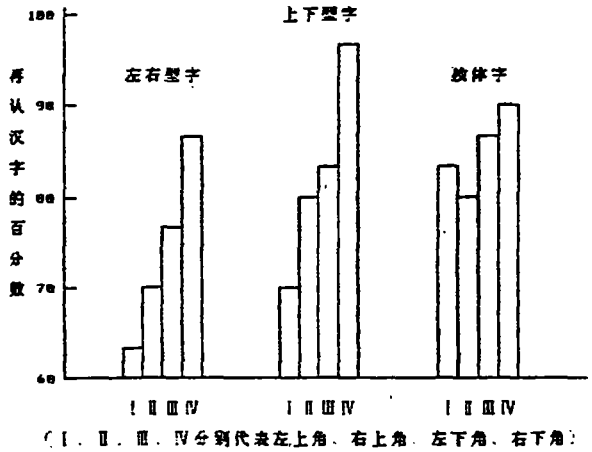


图 2 掩盖不同象限对再认的影响（* 引自文献[6]图 1）

刘英茂(1983)将 12 篇（每篇约 120 字）故事印在纸上，消去各字的某一象限，令被

试阅读整篇故事．结果如表 4 [7]．破坏左上角或右上角使错误率和阅读时间明显增加．

所以，一个字的左上角和右上角线索比左下角和右下角的线索重要．而字的左上角或左下角所提供的线索和其右上角或右下角所提供的线索没有很大的差异．

表 4　破坏字的不同部分对阅读的影响

| 破坏部分 | 左上角 | 右上角 | 左下角 | 右下角 |
|---|---|---|---|---|
| 增加阅读时间(秒) | 7.13 | 8.13 | 6.68 | 5.83 |
| 增加错误字数(均值) | 2.90 | 2.36 | 1.22 | 1.49 |

* 据文献[7]表 2 改编

于国丰（1986）发现认读画面显示字符时的认读路线大多数(67.25%)先从左侧开始，经过上、中、下位置顺序，最后止于右侧．记忆也如此 [8]．那么，识别汉字时如果可以排除眼动因素的话，由于我们日常书写笔划顺序习惯的潜移默化，是否可以说对每个汉字都有一个类似的心理扫描路线，或者说对汉字中的每个部件都按其空间部位系列地加工呢？或许空间部位的作用，正是通过这种系列的加工才得以体现出来．消除了字的左上象限，这种心理扫描过程自左向右、自上而下的自然顺序被中断，便难以开始或者需要其它部分的反馈才能得以进行下去，因而使不完整字的识别受到较大的影响．而消除字的右下象限，对这种系列扫描的前一部分没有什么影响，这前面的部分是却是识别整字的关键．所以，此时整字识别所受的影响不大．

## 二、汉字部件的位置频率在汉字识别中的作用

国外学者发现，横排形式的英文单词，其不同位置在人的识别过程中之作用是不同的．开头部分提供最佳线索，末端其次，中间部分的线索最差 [9] [10]．以此类推，则汉字不同部位提供的信息应以左上角最佳，右下角其次，左下、右上角最差．然而汉字识别的心理学实验结果证实了左上角确实最重要，但右下角却是最不重要的．这种不一致，可能因为汉字是二维平面结构，每一笔画在空间位置上联系非常紧密，不适用于所谓系列位置(serial position)的概念．

彭瑞祥(1982)统计分析了小学课本中的三千个汉字左上角象限、右下角象限的笔画构成形状，并将形状相同或类似的归为一类，称为子模式 [11]．表 5 所示结果表明，左上角模式的组字能力比右下角模式强，但在机器识别时的初分类宜以右下角为依据，因为左上角的子模式结构复杂，常常带有部首以外的笔画，而右下角的子模式比较稳定．此结果也可以比较合理地解释彭瑞祥、喻柏林（1980）的实验结果．因为左上角特征信息丰富，所以掩盖后难以重构整字；而右下角子模式相对比较单纯，特征信息不多．因此掩盖后容易通过前面几部分的提示，从记忆中得到填补，所以对于重构整字的影响不大．说明被试对汉字再认成绩很多程度上取决于汉字的结构特点．这个实验的结果说明依据不完整的汉字

信息也可以识别出汉字来，因为人脑中贮存着关于某字特定部位是特定的部件或笔画的知识．反过来，只根据特定部位上出现的某一个或某几个部件等不完整的信息也可以判断很可能是某字．

表 5　左上角和右下角子模式的出现频率统计

| | 频率水平 | 0.5–0.99% | 1–1.99% | 2.0–4.99% | 5%以上 |
|---|---|---|---|---|---|
| 左上角 | 子模式数目 | 25 | 13 | 7 | 2 |
| | 字数 | 551 | 584 | 722 | 357 |
| | 占总字数百分比 | 18.37 | 16.78 | 24.07 | 11.9 |
| 右下角 | 子模式数目 | 36 | 19 | 3 | |
| | 字数 | 748 | 756 | 262 | |
| | 占总字数百分比 | 24.93 | 25.19 | 8.73 | |

* 据文献[11]之表 2,表 5 改编

韩布新(1993)探讨了速示条件下两部件合体字中两个部件的部件频率对整字识别的影响[12]，发现第二部件（下半或右半字）比第一部件（上半或左半字）对整字识别反应时有更显著的影响；而且第二部件频率越高，则识别整字反应时越长．表明第一部件的加工可能是自动的，而第二部件的加工则是一个系列的过程．

## 三、可能的启示

汉字的机器自动识别一般必须经过输入、分割预处理、初分类、单字识别和后处理等过程．以上实验结果，对汉字自动识别的初分类和特征匹配算法设计提供了一定的依据．首先，目前单字识别过程中采用的距离或相似度等指标的算法，将汉字二维图形中的各部位特征等权对待．但认知心理学的实验结果表明合体字中左上角包含的特征信息比较重要，也就是说汉字二维平面结构中信息量的分布是不均匀的．因此，在设计多阶段处理的识别方案时，可对汉字不同部位的特征予以加权处理．同时由于左右字在汉字集中占40.1%，在初分类时若能确定结构类型，则可使匹配范围缩小近一半．由于独体字的左上角相对并无特别的重要性，在初分类阶段还可以将合体字和独体字区别开来．混淆字大多出现在占汉字集 25.9%的独体字中，因此在减少识别错误率时应优先考虑这类问题．

其次，掩盖汉字的某一象限，大部分的字仍可能被认出，说明根据不完整的局部信息，可以（至少部分地可以）唯一地确认某个汉字．因此在识别时可以对某一部分的局部特征（比如左上角象限），加强匹配密度或权值，以提高正确识别率和识别速度．另外，速示条件下仍能清楚地辨别汉字，说明人类在阅读汉字文本时只提取汉字的部分图形信息

便能识别或重构汉字，是人大脑中的知识库在起作用。机器识别系统并且无此知识库，其识别与重构能力远远不如人类。在某种意义上讲，由于印刷质量水平不一、不同文本的字体、大小等特征的差异，自动识别时实际上也是对不完整汉字的重构。

第三，在汉字机器识别系统的字典信息中，除了字的特征矢量外，还可以加入部件频率、部件组合频率、部件位置频率等动态参数信息。字的特征矢量可以分成两大类，一类为汉字部件的特征矢量集合，可用于预处理后的初分类。另一部分则可以是合体字的字形特征矢量集合。这样前面所说的动态频率参数可以在根据部件的特征矢量进行初分类之后产生作用，由于限定了匹配目标范围，大量的排除了无关的匹配操作。字形特征矢量匹配的运算量大，在初分类中利用部件的频率信息大量地减少了待匹配的字形特征量，使识别效率大大提高。

第四，心理学的实验研究表明，在字词识别中存在自上而下(top-down)和自下而上(bottom-up)两种知觉加工过程[13]。前者指的是对所识别字词的前后联系、性质等已经具备一定知识，这些存在大脑中的知识，指导着加工过程；后者是指由提取字词材料的笔画数、结构关系、清晰度对比、大小等物理刺激特征开始逐渐上升到抽象理解的加工过程。这两种过程密不可分，尤其是自上而下加工在字词识别的准确性和速度等方面有重要的作用。目前的机器识别汉字方案多要经过单字识别这一过程，主要利用单字的各种不同字形特征量值与标准模板逐个进行匹配（如距离、相似度等）。如果摹拟人类识别汉字和阅读时的自上而下加工过程，在字典信息中依据汉字的前后文词义联系、笔画组合、部件组合等规律而建立起相应的网络结构，虽然这种智能化匹配使算法复杂性增加，但由于可以在初分类的基础上进一步区分特匹配的标准特征量模板集合，因此识别的准确性和速度都将大大提高。

比如可以每个部件作为节点，将有关部件组合的信息（如某个部件可能和哪些部件组合成字）形成网络，则可以利用有关频率的知识，仅靠第一部分部件集的字形特征矢量便可完成匹配。

最后，认知心理学对于人类如何利用部件或部件组合等局部信息来识别整字的认知策略的研究结果，或可为汉字机器识别提供一些可能的启发或依据。

## 参考文献


[1] 韩布新，汉字识别中部件及其组合的频率效应，中科院心理所博士学位论文，1993.

[2] 谭力海，现代心理学关于单词识别的研究—影响单词再认的因素，心理学探新，1988,2,32.

[3] 张　侃、陈一凡，汉字键盘输入的认知模型，中文信息学报，1991,5(4),13-19

[4] Chou,Sicgen K.(周先庚),Reading and Legibility of Chinese Characters Ⅱ. Reading half-characters, Journal of Experincntal Psychology, 1938,Vol X.Ⅲ,No.4,332-351.

[5] 曾性初等，汉字的讯息分析Ⅱ、文句中汉字笔画的省略与恢复，心理学报，1965,4.281-298

[6] 彭瑞祥、喻柏林，不同结构的汉字再认的研究，中国心理学会普通心理学与实验心理学论文集，1983年9月，甘肃人民出版社，兰州，第一版

[7] 刘英茂，字的各部分在阅读时所提供的线索，中华心理学刊，1983,25,85-98

[8] 于国丰，画而字母显示的认读及认读后记忆效果的探讨，心理科学通讯，1986,1.

[9] Horowitz, L.M.,et al., Word fragments as aids to recall:the orgnization of a word. Journal of Experimental Psycholosy, 1976, 76,219–226

[10] Liu,i., Cusing function of frgnents of verbal itens. Journal of fxperimental Psychology, 1969.82,187–114.

[11] 彭瑞祥，汉字结构的统计，心理学报，1982,4,1–8

[12] 韩布新，GB2312–88 中部件频率的统计分析，中国心理学会第七届学术会议论文，1993

[13] 荆其诚主编，简明心理学百科全书，湖南教育出版社，1991.4,长沙,第 1 版

## The Effects of the Spacial Position of Chinese Character and its implication to the Computer recognition of Chinese Character

With introduction of the results of several experiments about the effects of the spacial position of Chinese character in the visual recognition of the character, which indicated that the left upper quadrant in the square of chinese charater in most important while the right lower quadrant, it was discussed that the effects of the constituents positional frequency and their implication to the computer recognition of Chinese characters.