CCF中文信息技术开放基金课题
结题报告

项目名称：汉字笔画的自动提取

梁燕 谢山山 徐宗懿 吕娜 陈修远

答辩日期：2014.12.09

汇报内容

- 课题目标及目标完成情况
- 课题考核指标及指标完成情况
- 主要创新点及取得的成果
- 经费使用情况说明
- 问题和未来的工作

一、课题目标及目标完成情况

● 研究目标

➢ 研究汉字笔画结构，确定汉字拆分和交叉点区域提取算法。

➢ 研究笔画部件已提取的笔画信息，挖掘出汉字具有交叉区域的笔画段的组合规律并制定相应组合规则。

➢ 研究汉字笔画部件的特征表达和匹配，建立基于汉字拆分匹配的模型并实现系统。

一、课题目标及目标完成情况

- 研究内容

  - 对汉字结构和拆分深入分析，研究汉字拆分和交叉区域提取。

  - 对笔画部件中分割的笔画段的组合进行研究，制定相应的笔画段组合规则。

  - 对汉字拆分的笔画部件特征表达和如何匹配进行研究，并对基于汉字拆分匹配的笔画提取方式进行建模。

一、课题目标及目标完成情况

- 预期研究成果

  - 确立基于标准字库的笔画提取模型和算法理论。

  - 提出基于汉字结构是否连接的汉字拆分算法及笔画段组合算法。

  - 开发一个实际可操作的基于汉字拆分匹配的笔画提取系统。

一、课题目标及目标完成情况

●目标完成情况

- 建立基于拆分匹配的汉字笔画提取模型。
- 提出汉字拆分算法，交叉区域提取及笔画段组合算法。
- 利用VS2010和OpenCV开发出一个实际可操作的基于汉字拆分匹配的笔画提取系统。

一、任务考核指标及完成情况

●目标完成情况

➢ 建立基于拆分匹配的汉字笔画提取模型

汉字笔画的自动提取模型包括汉字拆分，形状匹配，交叉点区域提取算法和笔画段组合模块。

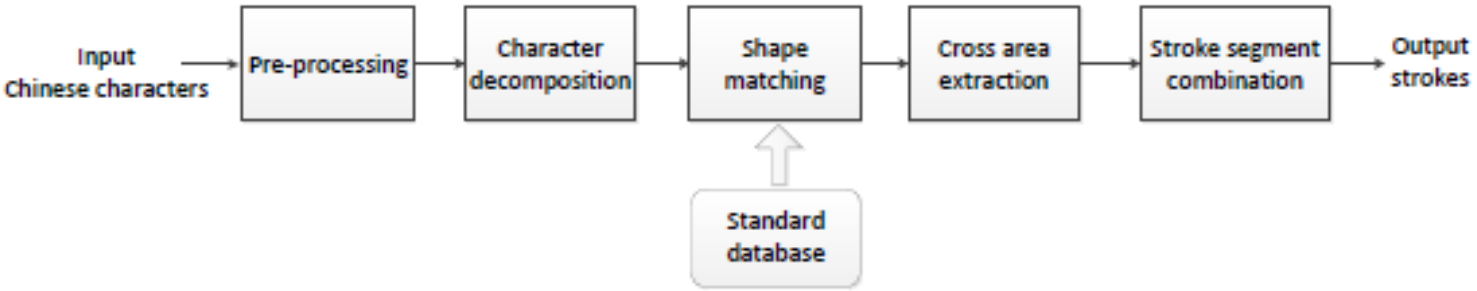

图1. 基于拆分匹配的汉字笔画自动提取系统框架
（标准库中存有已正确提取笔画的笔画部件的交叉点和笔画段的组合方式）

一、任务考核指标及完成情况

●目标完成情况

➢汉字拆分和交叉点提取

✓ 根据笔画是否连接将汉字拆分为多个笔画部件。

✓ 计算每个笔画部件的交叉点。

每个像素的相交数：

$$N_c(\mathrm{p}) = \frac{1}{2}\sum_{i=0}^{7}|p_{i+1} - p_i|, \quad P = \{p \mid N_c(\mathrm{p}) > 2\}.$$

将属于同一个交叉区域的多个交叉点进行合并:

$$(\mathrm{x}_k, \mathrm{y}_k) = (\frac{1}{n_k}\sum_{i=1}^{n_k} x_i, \frac{1}{n_k}\sum_{i=1}^{n_k} y_i).$$

✓ 若笔画部件交叉点数为0，则该笔画部件为单个笔画，可以直接输出；否则继续笔画提取。

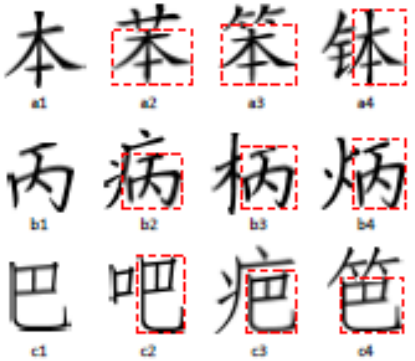

图2. 笔画部件重复例子

p_2	p_1	p_0
p_3	p	p_7
p_4	p_5	p_6

图3. 像素邻域

一、任务考核指标及完成情况

●目标完成情况

➢ 汉字交叉点区域提取

交叉区域随着汉字变化和字体变化有不同的大小和形状。

✓ 计算交叉点的点到边界的方向距离(PBOD)直方图。

$$h_{i,j}^{PBOD} = dist(j), j = 1, 2, \cdots R.$$

整个区域被划分为R的小方向，$dist(j)$表示第j个方向内交叉点到轮廓的距离 。

✓ 计算该PBOD直方图的波谷找出交叉区域的分割点。

一、任务考核指标及完成情况

●目标完成情况

➢对笔画部件进行特征提取和匹配

✓ 计算笔画部件的shape-context特征，与标准库中笔画部件的shape-context特征进行匹配。

✓ 匹配成功则利用标准字库存储的交叉点和笔画组合信息提取笔画。

✓ 匹配不成功则根据制定的笔画段组合规则进行提取，通过用户反馈将正确的提取笔画的笔画部件添加到标准字库。

一、任务考核指标及完成情况

●目标完成情况

➢汉字笔画段组合

在对大量的汉字进行分析的基础上，我们总结出一种大部分笔画段都满足的组合规律，即两个笔画段属于同一笔画的概率与笔画段间的斜率有关。

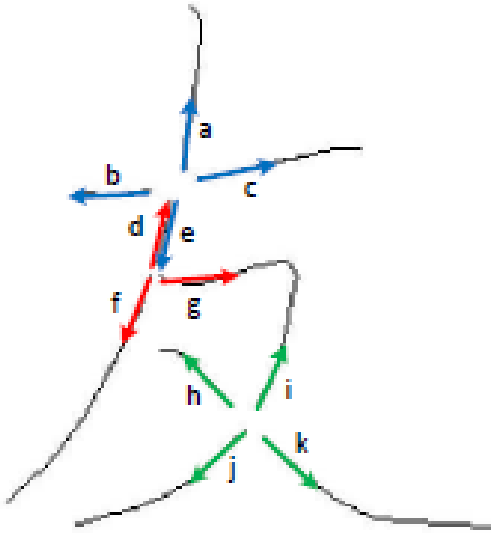

图4. 分割的笔画段向量图

✓计算笔画段向量之间夹角的cos绝对值。

$$\mathrm{a}_{\theta} = \left| \frac{\boldsymbol{a} \bullet \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} \right|.$$

✓若两个笔画段之间的a_{θ}值大于设定的阈值，则将这两个笔画段进行组合。

二、课题考核指标及指标完成情况

- 考核指标

  - 技术指标：算法的准确性与实用性。

  - 成果指标：根据资助机构提供的笔画已经提取好的500个参考楷体汉字（在KaiTi-Ref-500文件夹中）信息，自动提取出对应的仿宋（在FS-Target-500文件夹中）和硬笔楷书（在YK-Target-500文件夹中）汉字图片的所有笔画，并以所提供的参考数据的格式进行存储。通过资助机构的测试验收，提交技术文档与源代码，投稿到NLP&CC 2014论文一篇。

二、课题考核指标及指标完成情况

- 指标完成情况

➢ 课题组开发了一个实际可操作的基于汉字拆分匹配的笔画提取系统，根据资助机构提供的笔画已经提取好的500个参考楷体（KT）汉字信息，自动提取出对应的仿宋（FS）和硬笔楷书（YK）汉字图片的所有笔画，并以所提供的参考数据的格式进行存储。

➢在会议NLPCC 2014上投稿论文Decomposition and Matching: towards Efficient Chinese Character Stroke Automatic Extraction以阐述提出理论的意义和应用价值。

三、创新点及取得的成果

- 创新点
  - 提出了一种基于拆分匹配的汉字笔画自动提取机制
    - 与参考文献[8]的笔画提取算法的效率对比非常明显，平均提取时间减少了50%左右。
    - 对结构相对简单的汉字笔画提取效果非常理想。
    - 对于不同字体的汉字均能达到较好的提取效果。
    - 具有很好的可扩展性。

[8] Zhenghua, L., Qiguang, H.: Algorithm and Implementation in Chinese Characters Order of Strokes Recognition. Computer Applications and Software, vol. 21, no. 7, pp. 96-97 (2004) (In Chinese)

三、创新点及取得的成果

- 提出了一种的自适应交叉区域提取算法
  - 克服了交叉区域形状不同大小不相同造成的困难。
  - 减少了交叉区域提取的计算开销，仅花了传统基于PBOD的交叉区域提取算法5%左右的时间。

图5. 交叉区域提取结果
（左边是的固定大小形状交叉区域提取方法的结果，右边为提出的基于自适应的交叉区域提取方法的结果）

三、创新点及取得的成果

- 提出了一种基于夹角的笔画段组合方法

  - 通过笔画段间夹角的绝对值判断笔画段组合的概率。

  - 计算简单，避免了随机的两两组合方式带来的大量计算开销，减少了大约75%的时间开销。

三、创新点及取得的成果

●课题成果

➢开发出一个实际可操作的基于汉字拆分匹配的笔画提取系统。

➢采用500楷体常用字构建标准库，采用500个仿宋（FS字体）和硬笔楷书（YK字体），进行测试，验证了系统的有效性和高效性。

➢预将研究成果投稿在SCI期刊Multimedia Tools and Applications 。

Le Dong, Yan Liang, Ling He, Zongyi Xu, Shanshan Xie, Na Lv, and Ning Feng, “Decomposition and Matching: towards Efficient Chinese Character Stroke Automatic Extraction,” Multimedia Tools and Applications, ready to submit.

三、创新点及取得的成果

- 结果展示

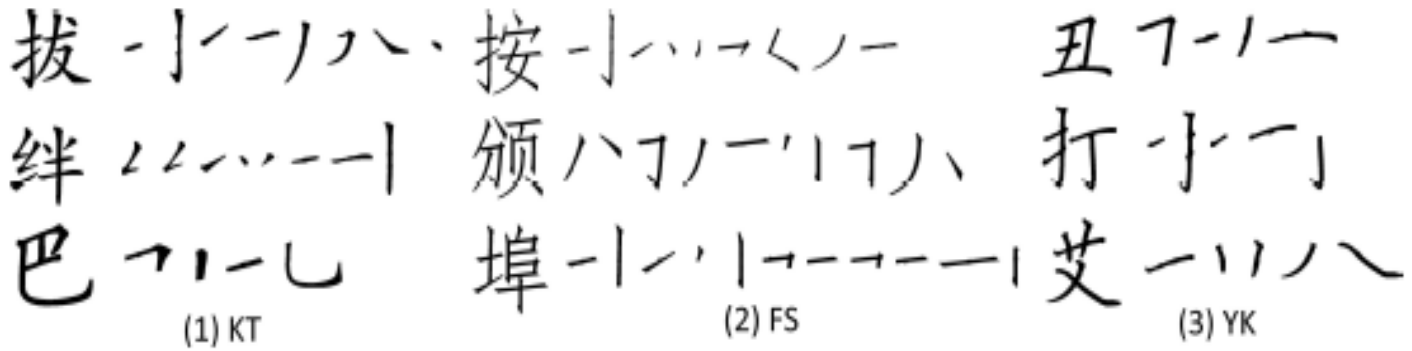

图6. 不同字体的笔画提取结果

Model \ Characters / Time(ms)		KT scripts	FS script	YK script
Method in [8]	Cross area extraction	652	613	245
	Stroke segment combination	244	258	245
	Total1	1335	1303	1327
This paper	Cross area extraction	32	27	30
	Stroke segment combination	56	52	54
	Total2	674	635	680
Ratio(Total2/Total1)		50.49%	48.73%	51.24%

图7. 汉字笔画自动提取时间开销比较

四、经费使用情况说明

- 情况说明

实验材料费0.6w，会议费/差旅费1.4w，出版/文献/信息传播/知识产权事务费0.25w，劳务费0.45w，管理费0.3w，共计3w经费使用完毕，每项均未超出预算。

五、问题和未来的工作

- 由于现有算法并不能保证匹配正确率为100%，因而这块还有很大的提升空间。我们未来的工作，会通过采用更有效的描述特征进行笔画部件匹配，提高匹配的准确率。

- 我们的方法目前在算法效率以及应用场景等方面仍然具有非常大的发展空间。未来的工作将通过结合机器学习以及模式识别等领域方法进行优化，得到更高的笔画提取精度以及准确率。

谢谢！