东巴象形文字特征曲线简化算法研究

杨玉婷[1]， 康厚良[2]， 廖国富[3]

(1. 云南开放大学文化旅游学院，云南 昆明 650000；
2. 苏州市职业大学体育部，江苏 苏州 215000；
3. 昆明理工大学津桥学院电气与信息工程学院，云南 昆明 650000)

摘 要：东巴文作为一种原始的图画象形文字，在检索和识别方面的研究较多，且从不同角度应用各类算法进行了实现，但是在文字特征提取和简化方面的研究却很少。由于字符特征提取的精练性和完全性将直接影响识别算法的精度和复杂度，因此结合计算机视觉中形状简化的相关研究成果，给出了适用于东巴象形文字特征曲线简化的改进算法。该算法以离散曲线演化算法为基础，进一步给出了区域最大面积差的临界点选取法和二次简化算法，有效去除了东巴字符特征曲线中的冗余点和潜在异常点。通过通用性和鲁棒性实验表明，该算法在保留原有字符特征的基础上可以去除曲线中 87%以上的冗余点，实现了特征曲线的最简化，从而为东巴文字的相似性度量奠定基础。

关 键 词：东巴文字特征提取；特征曲线简化；离散曲线演化算法；二次简化
中图分类号：TP 391 **DOI：**10.11996/JG.j.2095-302X.2019040697
文献标识码：A **文 章 编 号：**2095-302X(2019)04-0697-07

Research on Simplification Algorithm of Dongba Hieroglyphic Feature Curve

YANG Yu-ting[1], KANG Hou-liang[2], LIAO Guo-fu[3]

(1. Culture and Tourism College, Yunnan Open University, Kunming Yunnan 650000, China;
2. Sports Department, Suzhou Vocational University, Suzhou Jiangsu 215000, China;
3. Faculty of Electrical and Information Engineering, Oxbridge College, Kunming University of Science and Technology, Kunming Yunnan 650000, China)

Abstract: Dongba hieroglyph is a primitive ideographic script. Many researchers have done a lot of research on the retrieval and recognition of Dongba hieroglyph, and applied various algorithms from different angles, but few of them are on the feature extraction and simplification. In view of the fact that the succinctness and completeness of character feature extraction will directly affect the accuracy and complexity of the recognition algorithm, we combine the related research of shape simplification in computer vision and present an improved algorithm which is helpful to the simplification of Dongba hieroglyph feature curve. Based on the discrete curve evolution algorithm, our algorithm further gives the critical point selection method based on the maximum area difference and the second simplification algorithm, which effectively remove the redundant points and potential anomalies in the character feature curve. The universality and robustness experiments show that our algorithm can remove more than 87% redundant points in the curve while retaining the original character features, and achieving the most simplified feature curve. It lays the foundation for retrieval and recognition of Dongba hieroglyph.

收稿日期：2018-10-16；**定稿日期：**2018-12-08
基金项目：云南省科学研究基金项目(2018JS748，2019J1152)；国家社会科学基金项目(15BTY038)
第一作者：杨玉婷(1983–)，女，云南昆明人，副教授，硕士。主要研究方向为图形图像处理、计算机视觉等。E-mail：tudou-yeah@163.com
通信作者：康厚良(1979–)，男，四川泸州人，教授，硕士。主要研究方向为民族体育与民族文化。E-mail：kangfu1979110@163.com

Keywords: extracting feature of Dongba hieroglyphs; feature curve simplification; discrete curve evolution algorithm; secondary simplification algorithm

纳西东巴文字是通行于纳西族西部方言区的一种古老的图画象形文字[1]，其保留了人类早期文字演变的珍贵信息[2]，是当今世界上唯一存活的象形文字[3]。2003 年，使用东巴文撰写的东巴古籍被联合国教科文组织列入世界记忆遗产名录[4]。

东巴文字具有较为浓厚的原始图画意味，从文字的结构要素分析[5]，字素可细分为轮廓型和结构型 2 类。轮廓型字素通过临摹物体的外在形状来表达实际含义，具有轮廓特征明显且闭合性好的特点；而结构型字素一般使用简单的字符笔划通过描绘事物的结构或骨架来表达含义，其中人形字最具代表性[6]，见表 1。

表 1 东巴字素的分类

类型	东巴字	注释	东巴字	注释	东巴字	注释	东巴字	注释
轮廓型字素		麻雀		号角		男神		五福
结构型字素		人		脖子		碗		坡脚

当前，东巴文字在检索和识别方面的研究较多[7-11]，且应用各类算法进行了实现，但是在文字特征提取和简化方面的研究却相对较少，即使在相关文献中包括与文字特征提取有关的内容，也仅仅是使用常见的灰度化、二值化、直方图等通用方法[12-15]。显然，字符特征提取的不精练、不完全，不仅影响识别算法的精度，同时也会增加识别算法的复杂度。

基于链码的连通域优先级标记(connected domain priority marking, CDPM)算法[6]为东巴文字的特征提取提供了一种新的思路，针对东巴文字的结构特征，准确提取 2 类不同结构东巴文字的特征曲线。但是，曲线中过多的顶点序列仍会对文字识别的效率和准确性造成影响。因此，本文在 CDPM 算法的基础上，结合东巴文字的书写习惯及计算机视觉中形状简化的相关研究成果[16]，给出了适用于东巴象形文字特征曲线的简化算法。该算法能有效去除字符特征曲线中的大量冗余点及潜在异常点，实现使用最少的特征点序列表示最多的字符特征。另外，该算法也可与其他特征曲线提取算法相结合用于曲线的简化和特征提取。

1 东巴象形文字特征曲线简化算法

东巴象形文字特征曲线简化算法是基于 CDPM 算法在东巴文字特征提取方面的进一步优化，其思想是：首先采用离散曲线演化算法(discrete curve evolution, DCE)和区域最大面积差的临界点选取法去除特征曲线中的大量冗余点，统称为一次简化；然后，使用二次简化算法进一步去除特征曲线中的剩余冗余点及潜在异常点。

1.1 离散曲线演化算法

DCE 在保持曲线特征的同时，可快速去除曲线中的大量冗余点[17]，即在演化的每一个阶段，使用一条新的线段替换原有的一对相邻的线段 s_1 和 s_2，该线段是通过连接 $s_1 \cup s_2$ 的 2 个端点而得到的。演化过程中，线段的合并顺序由度量 K 决定，即

$$K(s_1,s_2)=\frac{\beta(s_1,s_2)l(s_1)l(s_2)}{l(s_1)+l(s_2)}$$

其中，$\beta(s_1,s_2)$为线段 s_1 和 s_2 的顶点转向角；l 为归一化之后的线段长度[18-19]。

由于东巴字符的特征曲线中除了存在大量冗余点之外，还有一些潜在异常点，若直接使用 DCE 算法的结束条件可能会产生曲线中的关键点被删除，而异常点仍存在的问题，容易导致曲线的过度简化或简化结果异常。因此，通过分析东巴字符特征曲线在实际演化过程中所反映的外在变化，采用基于区域最大面积差的临界点作为演化的结束条件，达到使用最少特征点表示最多字符特征的目的。

1.2 基于区域最大面积差的临界点选取法

特征曲线在每一阶段的演化都会导致曲线形态的变化，并进一步引起曲线所围面积的变化。因此，当两次演化中曲线所围面积的差值最大时，说明此时丢失的字符细节特征最多，从而得出基于区域最大面积差的临界点选取法的核心思想，即：

设字符的特征曲线包含 n 个特征点，在演化的每一个阶段，若每次去除总量 1%的特征点，则第

$i\ (i=\{50,51,52,\cdots,99\})$ 次演化将去除 $\left\lfloor \frac{i}{100}\times n \right\rfloor$ 个特征点，剩余特征点所围成的面积为 $Area_i$。计算第 i 次和 i+1 次演化时特征曲线所围面积的差值，若差值最大，说明第 i+1 次演化丢失的字符细节特征最多，则第 i 次演化后得到的特征曲线为最简，计算公式为

$$f(critica_{\text{point}})=f\left(\max\left(\left|Area_i-Area_{i+1}\right|\right)\right), i\in[50,99]$$

显然，当曲线中的特征点较多时，每次删除部分特征点不会对曲线产生太大影响。但是，当特征点较少时，每次曲线演化都可能造成曲线中字符特征的大量丢失。因此，在曲线演化过程中，当简化曲线的特征点数量小于原有总量的 50% (即，完成第 50 次演化)时，则在后续的每次演化中加入曲线所围面积的计算，以减少算法的总体计算量；另外，当完成第 99 次曲线演化时，曲线中仅剩原有总量 1%的特征点，若继续进行演化，则特征点将被全部删除。因此，演化次数最大值为 99。以字符  为例，图 1 显示了在曲线演化的每个阶段，字符特征曲线所围面积的值。其中，当完成第 89 次演化(即，去除了总量 89%的特征点)时，曲线所包含的字符特征丢失最多。因此，第 88 次演化(去除总量 88%的特征点)得到的字符特征曲线为最简。

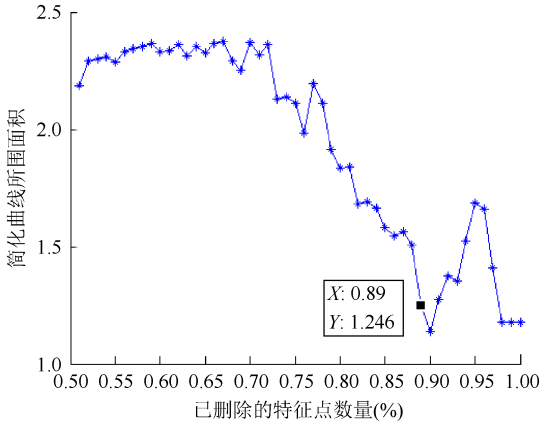
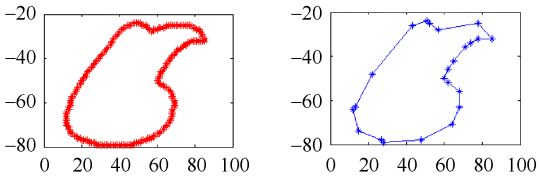


图 1　在曲线演化的各个阶段特征曲线所围成的面积差值

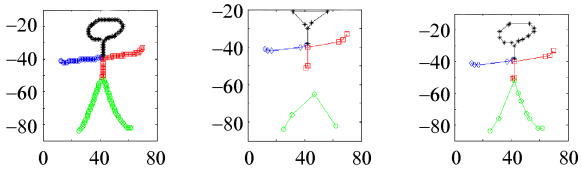
图 2(a)为字符的原始特征曲线，图 2(b)为采用基于区域最大面积差的临界点选取法去除特征曲线中 88%的特征点后的简化结果。可以看出，在去除曲线中的大量冗余点后，字符的特征曲线并未丢失过多的细节，说明该临界点选取法对于轮廓型字素是可行的。

在结构型字素中，由于字符的特征曲线已被划分为多条局部曲线，不同曲线间差异较大，如图 3(a)所示。因此，以字符的各局部曲线为单位，首先使用基于区域最大面积差的临界点选取法计算各条局部曲线的临界点完成曲线的简化，然后再进行曲线的拼接。



(a) 字符的原始特征曲线　(b) 基于临界值的 DCE 演化(去除总量 88%的特征点后的简化曲线)

图 2　字符的特征曲线及基于临界值的 DCE 演化效果



(a) 原轮廓(初始特征曲线段)　(b) 区域最大面积差的临界点(进行 DEC 演化后的效果)　(c) 改进的临界点(简化效果)

图 3　线演化中临界值选取规则的效果对比

由于结构型字素各局部曲线包含的特征点总体较少，加之曲线中潜在异常点的存在，当去除的特征点超过总量的 90%时，部分局部曲线中的关键点出现了误删，使曲线特征发生了丢失，如图 3(b)所示，且该问题在结构型字素的演化过程中经常发生。因此，为了避免过度简化，定义了曲线演化临界值选取规则，即：在东巴字素的曲线演化中，若特征曲线中去除的特征点大于或等于总量的 90%时，选择次大面积差所对应的值作为新的临界值。若新的临界值仍然大于等于 90%，则重复上述过程直到所选择的临界值小于 90%为止。

通过限制，图 3(b)得到了优化，效果如图 3(c)所示。此时，字符的头部和脚部的过度简化得到了改善，但其他部分并未受到影响。

1.3　二次简化

分析 DCE 的演化过程可知，简化时无法删除的无效特征点大多为凹点，这是因为特征点的权值与转向角的大小成正比(即，$K_i\propto\beta_i$)，使得曲线中的凹点具有较大的权值，即便是一个无效点也无法被及时删除，如图 3(b)所示。结合东巴字的书写习惯可知：东巴字的书写一般使用线描法，即“靠线条状物、靠线条造型，以线条描绘的方法作为其基本

特征”。由于东巴字的书写过程具有笔画线条流畅、连贯，曲线幅度变化均匀等特点。因此，可结合其书写习惯进一步去除字符特征曲线中的异常点。

通过对基础图库中东巴字符的反复实验发现：在字符的特征曲线中，若特征点 P_i 所对应的转向角极小($\beta_i \in [0°,15°]$)，则为异常点；另外，与 P_i 相邻的顺时针方向特征点 P_{i+1} 所组成线段 s 的归一化长度 $l(s) \leqslant \frac{2}{Len}$ (Len 为曲线的总长度)时，说明 P_i 和 P_{i+1} 非常接近点，则 P_i 也被视为异常点。

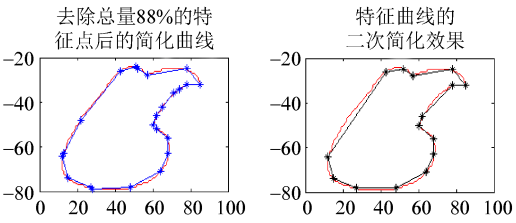
因此，对于包含 m 个特征点的字符简化曲线 Cur，设曲线的总长度为 Len，若曲线中的特征点 P_i 满足以下条件，则为异常点，即

$$P_i = \begin{cases} noise, & l(s) \leqslant \dfrac{2}{Len} \text{ or } \beta_i \in [0°,15°] \\ P_i, & \text{other} \end{cases}$$

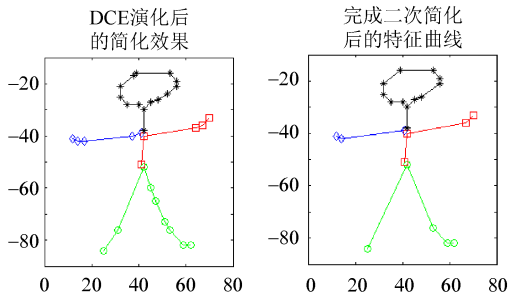
其中，由 P_i 和 P_{i+1} 所组成的线段 s 的归一化长度 $l(s)$ 有如下约束

$$l(s) = \begin{cases} \dfrac{|\overline{P_iP_{i+1}}|}{Len}, & 1 \leqslant i \leqslant m-1 \\ \dfrac{|\overline{P_mP_1}|}{Len}, & i = m \end{cases}$$

对图 2(b)和图 3(c)中的字符使用二次简化算法后的效果如图 4 所示。此时，轮廓型字素和结构型字素中的冗余点均得到了进一步去除，简化效果明显。但需要注意的是，二次简化算法仅能用于包含少量冗余点的曲线优化，并不能直接用于原始字符特征曲线的简化。



(a) 轮廓型字素的二次简化对比



(b) 结构型字素的二次简化对比

图 4 特征曲线二次简化前后对比

1.4 复杂度分析

东巴象形文字特征曲线简化算法包括 4 个步骤(设字符的特征曲线包含 n 个顶点)：

(1) 顺序计算曲线中每个顶点的权值及曲线所围成的面积，由于仅遍历曲线顶点序列 1 次，其时间复杂度 $O(n_1)=O(n)$。

(2) 查找曲线中权值最小的顶点 P_i 并删除，连接 P_i 的相邻顶点 P_{i-1} 和 P_{i+1}，重新计算顶点 P_{i-1} 和 P_{i+1} 的权值及简化后曲线所围成的面积，并记录简化前后曲线面积的差值。若假设仅剩 1 个顶点时，简化停止，则时间复杂度为 $O(n_1)=n+(n-1)+(n-2)+\cdots+1=n+\frac{n-1}{2}\approx O(n)$。

(3) 遍历曲线面积差值序列确定 DCE 算法的结束条件，得到曲线的一次简化结果。由于仅需遍历 1 次，时间复杂度 $O(n_3)=O(n)$。

(4) 依次遍历简化曲线中的特征点，判断并删除简化曲线中的潜在冗余点和异常点，得到二次简化结果。假设一次简化后，特征曲线中剩余的顶点数为 m($m<<n$)，但在二次简化中，删除噪音点的同时仍需重新计算相邻点的权值，因此其时间复杂度与步骤(2)相似，则 $O(n_4)\approx O(m)< O(n)$。

上述 4 个计算步骤相互独立且在计算中没有交叉，因此，东巴象形文字特征曲线简化算法的整体时间复杂度为：$O(n_1)+O(n_2)+O(n_3)+O(n_4)\approx 4\times O(n)\approx O(n)$。由此可知，整个算法的时间复杂度是线性的。

另外，由于曲线的二次简化是在一次简化的基础上完成的，而一次简化的处理结果又取决于临界点的选取，错误的临界点将会影响最终的简化效果，如图 3(b)所示。因此，临界点选取是东巴文字特征曲线简化算法的核心。

2 实验与分析

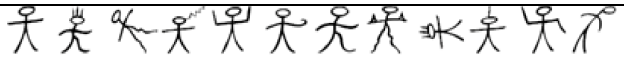
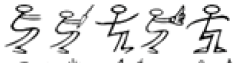
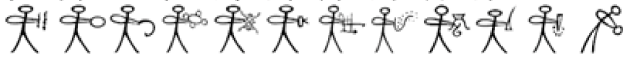
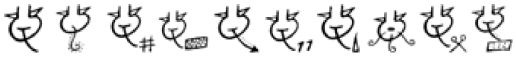
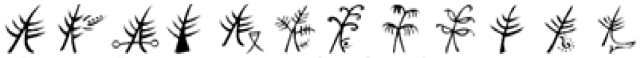
2.1 通用性测试

为测试算法的通用性，从 1 340 个东巴文字中为结构型字素和轮廓型字素分别选取 10 类字符，每类字符包括 5~12 个数量不等的东巴字。其中，结构型字素包括人、蹲、单手持物、双手持物、右侧偏移、头戴冠、心、植物、行走和坐等，轮廓型字素包括鱼虫、鸟、花、手、山、房屋、水、东巴祭祀、牲畜和山坡等，见表 2。

首先，从 2 类东巴字的 20 个子类中，每类随机提取 1 个东巴字；然后，使用 CDPM 算法提取东巴字中字素的特征曲线；接着，结合基于区域最大

面积差的临界点选取法和 DCE 演化算法去除特征曲线中的大量冗余特征点；最后，采用二次简化算法进一步去除曲线中的冗余点和潜在异常点，具体效果如图 5 和图 6 所示。

表 2　轮廓型和结构型字素的 10 种类型

类别	结构型——东巴字符
人	
蹲	
单手持物	
双手持物	
右侧偏移	
头戴冠	
心	
植物	
行走	
坐	

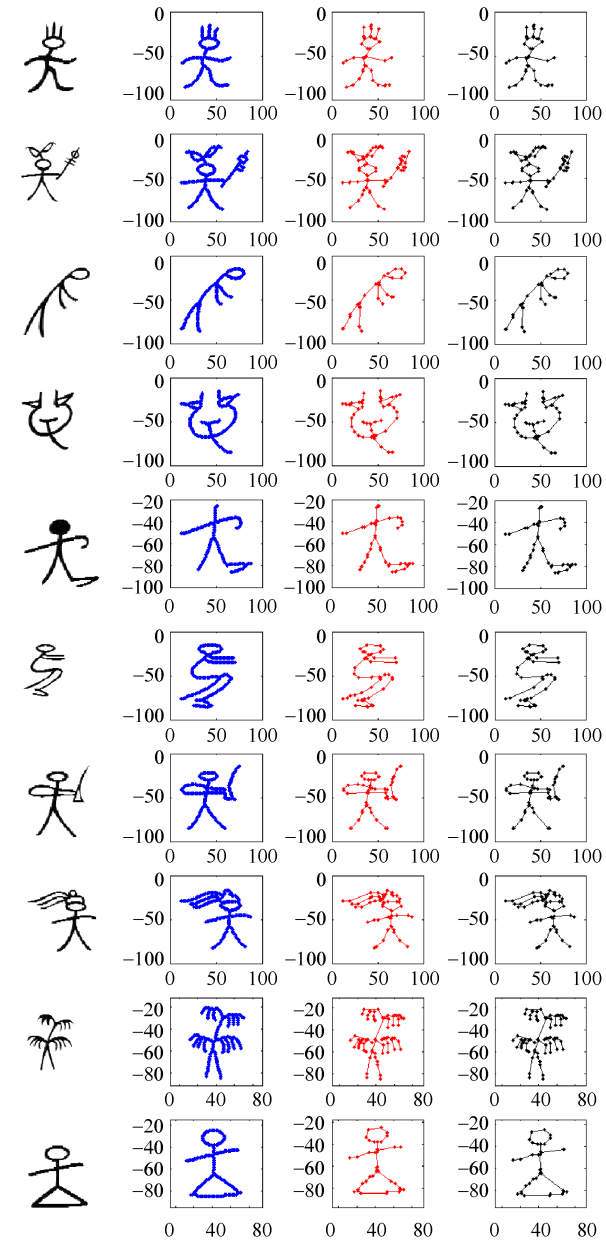
类别	轮廓型——东巴字符
鱼虫	
鸟	
花	
手	
山	
房屋(封闭)	
水	
祭祀	
牲畜	
房屋(不封闭)	

与原始曲线相比，经过两次简化处理，特征曲线中的大量冗余特征点和部分异常点被去除。同时，在简化曲线中，幅度变化较大的部分保留的特征点较多，而较为平滑的部分保留的特征点较少，保证仅用较少的特征点就能完整表示字符的原有特征。另外，通过测试说明，东巴象形文字特征曲线简化算法能够用于不同类型、不同结构和具有不同特征的东巴字符的特征曲线简化。

2.2　鲁棒性测试

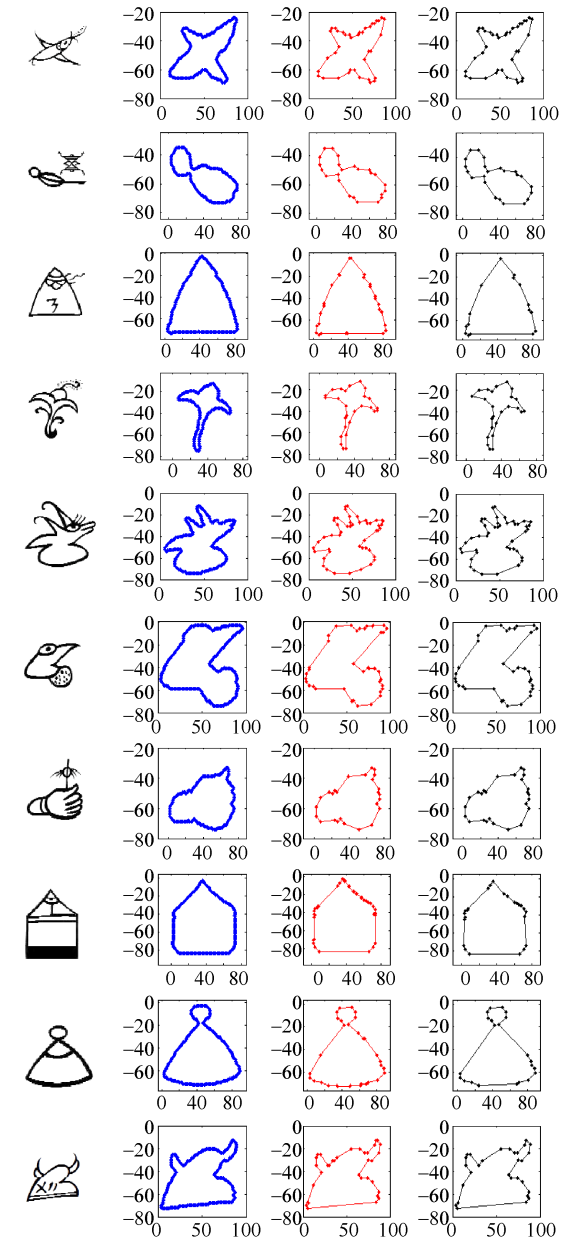
为进一步测试特征曲线简化算法的鲁棒性，从 1 591 个东巴字符图片中随机选取 100 个东巴字符的特征曲线作为测试对象，比较字符特征曲线简化前后的差异。首先，使用 DCE 算法结合区域最大面积差的临界点选取法实现字符特征曲线中大量冗余点的去除。在 100 个随机选取的字符中，最多去除了原曲线总量 87.83%的特征点，最少去除了 81.03%的特征点；平均去除量为 85.33%。

其次，使用二次简化算法实现字符特征曲线中部分冗余点和异常点的去除。100 个随机选取的字符中，最多去除了原曲线总量 94.55%的特征点，最少去除了 82.16%的特征点；平均去除量为 87.75%。与第一阶段的简化相比，使用二次简化算法，在 DCE 简化的基础上平均又减少了原有总量 2.42%的特征点，进一步剔除特征曲线中的冗余点和潜在异常点，使字符特性更加显著，如图 7 所示。



(a) 东巴字原图 (b) 特征曲线 (c) 基于 DCE 的简化曲线 (d) 二次简化曲线

图 5 结构型东巴字的演化过程



(a) 东巴字原图 (b) 特征曲线 (c) 基于 DCE 的简化曲线 (d) 二次简化曲线

图 6 轮廓型东巴字的演化过程

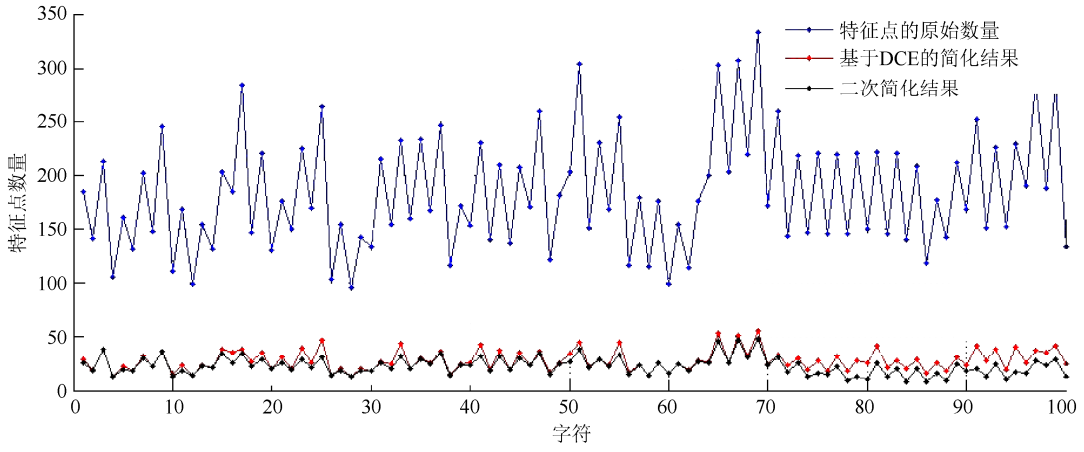


图 7 东巴文字特征曲线简化前后的比较

3 结 束 语

东巴象形文字特征曲线简化算法充分利用了现有算法的优势，同时又结合了东巴文字的书写习惯和文字特征，更好的适用于东巴字的特征曲线简化。但是，由于不同文字自身结构的差异性，使得其简化精炼度和完全性各不相同，这将直接影响文字准确识别的精度和复杂度。因此，为了将字符特征曲线的简化效果限定在一定范围内，可通过计算曲线在简化过程中所围面积的累计差值(即，特征损失)在原始曲线面积中所占的比例来判断曲线简化的精炼度，但这一判断方法的合理性还需要在后续工作中进一步进行验证。

参 考 文 献

[1] 方国瑜. 纳西象形文字谱[M]. 和志武, 参订. 2 版. 昆明: 云南人民出版社, 2005: 56-87.

[2] YANG Y T, KANG H L. The digital measures for protection and heritage of dongba culture [M]// Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2018: 1203-1208.

[3] 和力民. 试论东巴文化的传承[J]. 云南社会科学, 2004(1): 83-87.

[4] 王元鹿. 汉古文字与纳西东巴文字比较研究[M]. 上海: 华东师范大学出版社, 1988: 20-35.

[5] 和志武. 试论纳西象形文字的特点: 兼论原始图画字、象形文字和表意文字的区别[J]. 云南社会科学, 1981(3): 67-78.

[6] YANG Y T, KANG H L. A novel algorithm of contour tracking and partition for dongba hieroglyph [M]// Image and Graphics Technologies and Applications. Singapore: Springer Singapore, 2018: 157-167.

[7] 郑飞洲. 关于建设东巴文字字素检索数据库的构想[J]. 中国文字研究, 2002(3): 76-81.

[8] GUO H, ZHAO J Y. Research on feature extraction for character recognition of NaXi pictograph [J]. Journal of Computers, 2011, 6(5): 947-954.

[9] GUO H, YIN J H, ZHAO J Y. Feature dimension reduction of NaXi pictograph recognition based on LDA [J]. International Journal of Computer Science, 2012, 9(1): 90-96.

[10] 郑飞洲. 关于纳西族东巴文字信息处理的设想[J]. 学术探索, 2003(2): 83-86.

[11] LI X, GUO H, SUOG J, et al. The design and realization of NAXI pictograph character recognition preprocessing system [M]//Computer Science for Environmental Engineering and EcoInformatics. Heidelberg: Springer, 2011: 54-59.

[12] SONG W J, WANG K Q, XU R P, et al. The analysis and application of associative element combining configuration of Dongba Characters [C]//2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design (CAIDCD). New York: IEEE Press, 2010, 1(11): 753-756.

[13] 杨萌, 徐小力, 吴国新, 等. 东巴象形文字识别方法[J]. 北京信息科技大学学报: 自然科学版, 2014, 29(3): 72-76.

[14] 王海燕, 王红军, 徐小力. 基于支持向量机的纳西东巴象形文字符识别[J]. 云南大学学报: 自然科学版, 2016, 38(5): 730-736.

[15] DA M J, ZHAO J Y, SUO G J, et al. Online handwritten Naxi pictograph digits recognition system using coarse grid [M]//ComputerSciencefor Environmental Engineering and EcoInformatics. Heidelberg: Springer, 2011: 390-396.

[16] 周瑜, 刘俊涛, 白翔. 形状匹配方法研究与展望[J]. 自动化学报, 2012, 38(6): 889-910.

[17] LATECKI L J, LAKÄMPER R. Polygon evolution by vertex deletion [M]//Scale-Space Theories in Computer Vision. Heidelberg: Springer, 1999: 398-409.

[18] LATECKI L J. Shape similarity measure based on correspondence of visual parts [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(10): 1185-1190.

[19] LATECKI L J, LAKÄMPER R. Convexity rule for shape decomposition based on discrete contour evolution [J]. Computer Vision and Image Understanding, 1999, 73(3): 441-454.