漢字グリフ自動生成による多漢字処理環境の提案

上地 宏一
kamichi@fonts.jp
慶應義塾大学

2003 年 10 月 31 日

概要

漢字部品を組み合わせてさまざまな漢字フォント（グリフ）を動的に生成する KAGE システムについて紹介する。 KAGE システムは、主に既存の符号化文字集合（文字コード）には含まれない文字を利用することを目的としているが、既存の文字コードのグリフセットをライセンスフリーで提供するための準備も進めている。また CHISE プロジェクト [1] が掲げている、特定の符号化文字集合に依存しない文字処理技術の文字表示部分に関して共同研究を行っている。ここでは現状と今後の展望について説明する。

1　はじめに

現在の文字処理は、すべて符号化文字集合の枠組みで行われている。文字表現についても、多くは符号化文字集合をもとにしたフォントセットが利用されている。日本語のフォント環境を欧文のそれと比較すると、商用ベースのものに関しては、複数の会社からさまざまな書体バリエーションを持つフォントが市販されているが、ライセンスフリーのフォントに関しては、まだ貧弱であり、とくにアウトラインフォントに関してはほとんど存在しないのが現状である。

ライセンスフリーのフォントというと、フォントベンダーがライセンスフリーで配布する、個人がボランティアで作成する、公的な機関が作成、または権利を購入し配布する、などの形態が考えられるが、日本語の場合、標準では JIS X 0208 の 6,355 漢字や、さらに JIS X 0212 を含めた 12,156 漢字と、欧米の文字集合と比較して文字種数が多く、コストが高いので、個人が作成することは難しく、またフォントベンダーからの完全なライセンスフリーという形態での配布は望めない。一方で、公的機関がフリーライセンスフォントを配布することに対していろいろな意見があるが、ここでは触れないことにする。またライセンスフリーのフォントの必要性に関しても意見が分かれるが、従来のフォントでは、フォントの埋め込みとコンテンツの公開が制限される、ライセンスによって使用できる OS や利用できるクライアントの数が制限される、など、利用規約によってさまざまな制限が存在する。ここで、ライセンスフリーのフォントが存在することは、日本語環境の発展、ひいてはフリーソフトウェアの発展に役立つのではないかと考えられる。

一方で符号化文字集合に採録されていない文字や、字形の異なるグリフを利用したい場合、従来はユーザー定義文字（外字）として自由に定義することができたが、情報交換の障害となるため、現在では適切な手段とはいえない。この場合、ある程度流通している外字フォントといわれる製品などを利用することになるが、グリフの提供者となるフォントベンダーによって利用できる字形が制限されることになり、ユーザは受動的立場に置かれ

ている。

これらの問題に対して、真の多漢字字形処理環境を目指すためには、ユーザが任意の漢字グリフを自由に作成 (利用) でき、さらに情報交換が保証される処理系の開発が必要である。そしてテキストデータを具視化するフォントは自由に利用できることが重要である。このためには、自分の手元で漢字グリフを生成できる仕組みが必要であるが、アウトラインフォント（グリフ）をデザインすることは大変である。そこで、複数の漢字部品や筆画などを合成してグリフを低コストで生成する方法が現実的である。

このような観点に立ち、著者は 1998 年から KAGE システムの開発に着手し、 2001 年にはエンジン部分が完成した [4]。その後、生成文字品質を向上させるために研究を継続しているが、現在は、生成されたグリフのデザインをユーザが自由に調整できるシステムを開発中である。さらに、 2002 年から CHISE プロジェクトに参加し、多文字処理技術に KAGE システムを応用するための研究を行っている。

2 KAGE システム

KAGE(Kanji-glyph Automatic Generating Engine) システムは、漢字部品を合成して漢字グリフを生成する KAGE エンジンを基本としている。複数の漢字部品を組み合わせて漢字グリフを生成する手法は、 19 世紀の中国における活字字母製作でも提案されていた技術であり、近年では「和田研フォント」と称される（当時）東京大学和田研究室で開発されたフォント生成技術 [5, 6] が記憶に新しい。

2.1 新たな生成エンジンの構築

「和田研フォント」は、当時高価だったアウトラインフォント（ベクトルフォント）を低コストでデザインする目的で研究されたが、その後のアウトラインフォントの普及や生成される文字品質の面から、技術そのものは普及しなかった。著者は、人名地名や中国の古典文献などでしか使われない頻度の低い漢字を処理するためにフォント自動生成技術が応用できると考えたが、すでに和田研フォントの研究は終了しており、またデータも公開されていなかったため（現在は公開されている）、論文などを参考に、新たにグリフ自動生成エンジンを作り直した。和田研フォントの研究は基本的に JIS X 0208 符号化文字集合などの閉じた漢字集合のグリフを生成するものであったが、 KAGE システムでは漢字集合からもれた漢字を生成することを主目的とするため、結果として異なる特徴を持つシステムとなった。

2.2 拡張 IDS による漢字記述

KAGE エンジンは、ユーザが必要とする漢字字形を動的に生成する必要がある。ここで生成すべき漢字字形を表現するために IDS(Ideographic Description Sequence) [2, 3] の拡張表現を利用している。本来の IDS と異なる点としては、 IDC(Ideographic Description Character) のうち、 U+2FFB(ideographic description character overlaid) は字形が特定できないために廃止した。また漢字部品として任意の漢字が利用できるが、元来 UCS のコードポイントは具体的な字形を特定せず、例えば国や地域によっても想定される字形が異なってくることから、字形を固定するために、 4 桁の字形詳細番号を付加して部品の字形を特定することができる。

KAGE エンジンでは以上で述べた IDS 拡張表現によってさまざまな漢字を記述することが可能である。漢字部品そのものは有限の集合であるため、全く未知の漢字に対して必ずしも IDS で表現できるわけではないが、その場合は漢字部品の追加によって対処することになる。

2.3 字形の骨格表現と肉付けによる書体表現

IDS で記述された漢字の構造に沿って、 KAGE エンジンが漢字部品を任意の大きさに拡大・縮小して配置

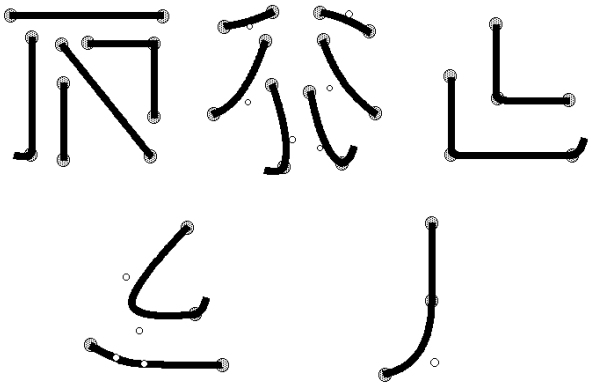

図 1: 応用ストローク 5 種（左上から直線、曲線、かぎ、複雑曲線、縦はらい

することになるが、ここで部品データを輪郭形式で保持している場合、拡大・縮小処理によって筆画の太さがバラバラになってしまう。そこで KAGE エンジンでは和田研フォントと同様に、部品をスケルトン形式で保有する。和田研フォントでは 16 種類のストロークデータがあるが、 KAGE システムでは、さらに共通化して直線と曲線の 2 種類の基本ストロークを元に、 5 種類の応用ストロークを用意している (図 1)。各ストロークは始点、終点および 1,2 つの制御点からなる 2 〜 4 点の座標で表現する。ストロークの頭部と尾部の形状を指定することで「撥ね」や「払い」、「ウロコ」を表現する。

また、箱型の縦棒のトメ（下に突き出す部分）は、自動的に付加されるため、例えば「口」という文字は 4 つの座標からなる 4 線分の組み合わせで表現することで、左下と右下のトメの突き出しは自動的に再現される (図 2)。

曲線は、 TrueType フォントと親和性を高めるために 2 次 B スプラインを用いている。

いわゆる書体を表現するための肉付け方式は、ゴシック体と、明朝体の 2 種類を用意している。ゴシック体は、単純に等幅の線を付加しているだけである。明朝体は、簡略化した装飾を施している (図 3)。一般の商

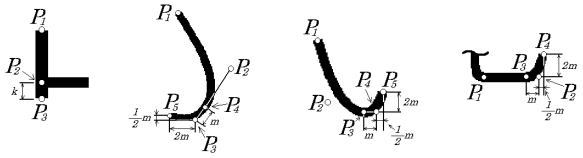

図 2: 筆末の自動処理（縦棒のトメ、左はね、右はね、かぎはね）

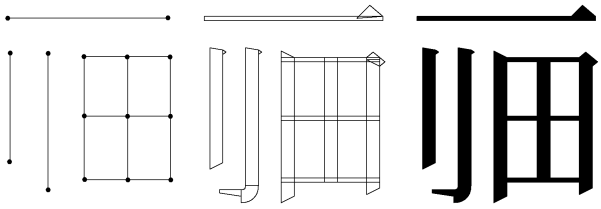

図 3: 明朝体風の肉付け

業印刷物で用いるようなフォントのデザインでは、ゴシック体、明朝体ともに複雑なデザインが必要であり、さらに、錯視などを考慮した細かい調整が必要とされているが、 KAGE システムで想定しているのは、 300 〜 600dpi の 11 ポイント本文における印刷用途であるため、解像度としては 100 ドット程度となり、それほど複雑な肉付けを施しても結果的には意味をなさないと考えている。そのためサーバから出力されるグリフ画像は 200 × 200 ドットの解像度で表現されている。ただし、エンジンの設定を変えることによって 1000 × 1000 ドットなどの高解像度の出力にも対応が可能である。

2.4 グリフ生成のルール

漢字部品の配置や大きさの決定は、エンジンが自動的に行っている。ユーザの要求を動的に処理し、グリフを配信する必要があるため、ルールは非常に単純なものとしている。具体的には、部品を X 軸（左右結合の場合）、 Y 軸（上下結合の場合）それぞれの方向にスキャンし、線分をまたぐ回数を数える。ここで部品ごと

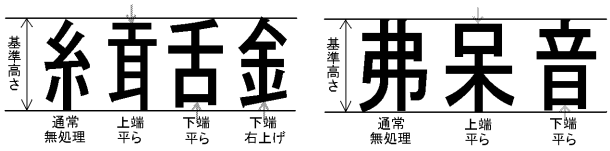

図 4: 部品調整処理（左部品、右部品）

の最大本数を部品複雑係数とし、係数比を基準に部品ごとの大きさを計算する。例えば左右 2 部品結合では以下の計算式を用いている。

$$ratio = \frac{(x_l \times 0.7^{b_{lL}+b_{lR}})^{0.6}}{(x_r \times 0.7^{b_{rL}+b_{rR}})^{0.6} \times 1.05}$$

(l: 左部品, r: 右部品, L: 左端, R: 右端, x: 部品複雑係数（X 軸）, b: 平ら (箱型) ならば 1, もしくは 0)

このルールだけでは実用的でないので、このほかに数種類の例外ルールを用いている。例えば、部品の形状や位置に応じて、位置や大きさを調整している。

さらに常用漢字の集合をサンプルに、生成されたグリフの中で違和感の大きい字に関して、含まれる部品の部品複雑係数に対して手動による調整を施している (図 4)。

2.5 グリフのネットワーク配信

KAGE システムの当初の目的は、符号化文字集合に沿ったフォントセットを生成するのではなく、集合に含まれない文字を利用することであり、さらに用いた文字の情報交換が保証される必要があった。そこでシステム実現のための第一段階として、ネットワーク上で文字を共有するためにグリフ配信サーバを開発した。サーバは拡張 IDS による文字の記述、書体、データ形式を入力すると、動的にグリフを生成し、クライアントに配信する。

配信するグリフのデータ形式は、 png または svg 形式が用意されており、主に html 文書内の`<img>` タグを利用して、文字を画像として呼び出すことを想定している。具体的には、`<img>` での URL に、 IDS などの情報を盛り込むことで、グリフを一つのファイルとして要求する。

図 5: ブラウザでの呼び出し

2.6 デザインの調整と再利用

2001 年に KAGE エンジンの実装が完了したが、生成した漢字グリフの文字品質は、ただちに利用できるような実用的レベルまでは達しなかった。部品が小さくなったときに生じるひずみの調整や、部品の大きさ・配置を決定するルールに改善の余地が見られる (図 6)。

漢字グリフ生成エンジンを実用的なものとするためには、印刷・出版向けの高解像度フォントや、市販されている一般的なフォントの品質までは必要ないが、少なくとも一般のユーザが見たときに問題とならない文字品質水準に達する必要がある。そこで、 KAGE システムでは、合成ルールの見直しだけでなく、生成グリフのデザインをユーザが任意に調整できる手段を用意することにした。

具体的に調整できる項目としては、部品の大きさ、位置だけでなく、部品同士が接触する場合の末端形状の変更や部品内の筆画の調整も想定している。調整したデータは、サーバに登録することによってその後に他のユーザが呼び出したときの生成にも反映されることを考えている。現在、デザイン調整ツールを実装している段階で

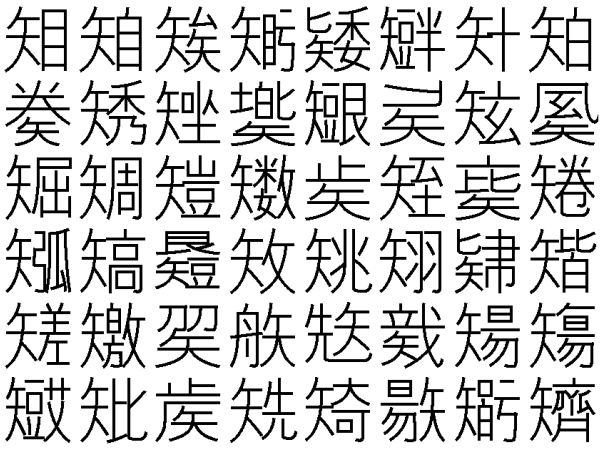

図 6: 生成されたグリフの例

ある。

さらに、構想の段階であるが、ユーザが施したデザインの調整を分析し、グリフ生成ルールそのものを再構築する機構を考えている。第一段階としては、部品複雑係数に対する各部品の適応度を計算し、誤差の大きい部品に対して補正が施されるようにする。

デザイン調整ツールは、そのまま漢字部品のデザインツールとして流用することが可能で、先に述べた IDS で表現しきれない漢字が生じたときに、新規に部品をデザインすることも可能となる。

3 CHISE プロジェクトとの連携

現在の KAGE システムでは、グリフをネットワークで配信する形態を考えているが、クライアントの OS やアプリケーションそのものを拡張するわけではないため、 HTML 文書内での利用のみを想定していて、あまり実用的ではない。

2002 年の秋から、 CHISE プロジェクトの文字合成サブプロジェクトへ参加することになった。ここでは、KAGE エンジンが CHISE モデルの任意の文字オブジェクトに付加された IDS 情報をもとにグリフを生成して具視化する機能を提供することを目指している。

現在 CHISE プロジェクトで用いている IDS 表記では指定できる漢字部品として ISO 10646 に含まれる文字だけでなく CDP データベースで利用されている文字や、 GT コードの文字も含まれている。さらに重なりの構造をあらわす IDC の U+2FFB の利用も認められている。一方で、 KAGE システムの IDS 表記では、拡張された字形詳細番号が用いられている。このため両者の表記法の差異を吸収する機構を考えている。具体的には、変換テーブルを用意することで U+2FFB を含む IDS を丸ごと 1 つの部品として用意することで回避する。また CDP 文字や GT 文字の部品を KAGE エンジン側に相当する部品漢字＋字形詳細番号に変換することで当面は運用可能であると考えている。

また、グリフを画像として渡すだけでなく、 SVG 形式などの輪郭情報に変換して渡すことで色や大きさをアプリケーション側で制御するための出力機能も追加した。

3.1 フォントのパッケージング

CHISE プロジェクトの XEmacs CHISE 実装では、JIS X 0208 や ISO 10646 の符号化文字集合だけでなく、 JIS X 0213、 Big5 や中国 GB 規格、 GT コードなどのさまざまな符号化を用いたデータを利用できるが、具視化するためのフォントが入手できないために、不満を感じることがある。一方、 CHISE プロジェクトでは漢字構造情報データベースと称して諸橋大漢和辞典や、 ISO 10646 の Ext.B 集合などの IDS 情報を持っている。そこで、これらのデータをもとに KAGE エンジンでグリフを生成し、 TrueType 形式のフォントにパッケージングして、ライセンスフリーで配布できないかと考えている。

また、特定の符号化文字集合の漢字フォントに関しては、不特定多数のメンバーによってデザイン調整を行うことのできる仕組みを考えている。当然、メンバーが増えるにつれて、字形の細部やデザインバランスに不統一

感が生じてくる。そこで、熟練したメンバーをメンテナとして昇格し、調整されたグリフの選別を行うことである程度のデザイン性の統一を図ることができないかと考えている。この新しい手法で漢字に代表される大規模文字集合のフォントをデザインすることが可能かどうかの検証を行いたい。

4 今後のスケジュール

今まで実装してきたKAGEエンジンと、現在実装中のデザイン調整ツールを2003年度中に公開する予定である。さらにCHISEの漢字構造データベースを処理できるための漢字部品データの大幅な拡充を行う。

5 おわりに

KAGEシステムによる多漢字環境の構築に向けた現状と展望について紹介した。

将来的には、著者の本来の目的であった漢字グリフ自動生成の文字品質の向上に向けて、生成ルールのアイデアを簡単に検証できるためのツールキット的な機構を用意することで、漢字の字形に対する知識を蓄積し、共有財産として活用できる形にしたいと考えている。生成ルールに関して、諸氏のアイデアを拝聴できれば幸いである。

参考文献

[1] CHISEプロジェクト.
http://www.kanji.zinbun.kyoto-u.ac.jp/projects/chise/

[2] International Organization for Standardization (ISO). *Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane (BMP)* , March 2000. ISO/IEC 10646-1:2000.

[3] The Unicode Consortium. *The Unicode Standard, Version 3.0* , February 2000.

[4] 上地宏一.「漢字フォント自動生成サーバ“影KAGE”の構築 －文字コードの枠組みを越える次世代漢字処理の提案－」, （漢字文献情報処理研究第3号, pp.4-13, 漢字文献情報処理研究会, 好文出版）, 2002.

[5] 田中哲朗、石井裕一郎、竹内幹雄、和田英一.「プログラム肉付けによる複数漢字書体間のスケルトンデータの共有」（情報処理学会論文誌 Vol.36 No.1 pp.177-186）, 1995.

[6] 田中哲朗、岩崎英哉、長橋賢児、和田英一.「部品合成による漢字スケルトンフォントの作成」（情報処理学会論文誌 Vol.36 No.9 pp.2122-2131）, 1995.