

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成24年 5月14日現在

機関番号：15401

研究種目：若手研究（B）

研究期間：2009～2011

課題番号：21700113

研究課題名（和文）和文かな書体の自動的な特長量抽出と定量的評価の研究

研究課題名（英文）A Study on Automatic Detection of Typographic Properties
for Japanese Typeface

研究代表者

鈴木 俊哉 (SUZUKI SHUNYA)

広島大学・情報メディア教育研究センター・助教

研究者番号：70311545

研究成果の概要（和文）：

本研究は抽象的なフォント参照を行うための書体分類について、従来欧文タイプフェイスしか想定されていなかった Panose 分類などをもとに、既存の実装と相互運用可能な形で和文タイプフェイスを分類するための指標について、既存のタイプフェイスの調査を行い、また、既存実装との互換性について検討した。その成果は最終年度に情報処理学会試行標準 IPSJ-TS 0013 として発表された。

また、調査の過程で、具体的なフォント参照として設計された電子文書中の埋め込みフォントについて、表示・印刷処理の過程で書体に変形する可能性が発見され、これを特定するための照合手法を開発した。この照合手法を利用し、JTC1/SC2/WG2/IRG へ重複符号化の可能性のある漢字群の指摘を行った。

研究成果の概要（英文）：

The font management frameworks in modern operating system are strongly dependent with the font reference by their “family names”. Some frameworks have the features to refer the fonts with specific typographic characteristics; “Panose” specification is one of the most widely referred standards, because it calculates the property values systematically and the classification is objective. In this study, the existing Japanese typefaces, classifications, terminologies and the implementations were investigated and the font resource reference scheme was drafted. The result is published the Trial Standard of Information Processing Society Japan, IPSJ-TS 0013.

During the investigation, it was found that the rendering result of the embedded fonts in PDF can be changed and loss its identity. The identification method of the embedded fonts in PDF is developed in this study. By using the developed method, some duplicated possible duplicated encodings of CJK Unified Ideographs are reported to ISO/IEC JTC 1/SC 2/WG 2/IRG.

交付決定額

(金額単位：円)

	直接経費	間接経費	合計
2009年度	1,000,000	300,000	1,300,000
2010年度	1,200,000	360,000	1,560,000
2011年度	1,100,000	330,000	1,430,000
総計	3,300,000	990,000	4,290,000

研究分野：総合領域

科研費の分科・細目：情報学・メディア情報学・データベース

キーワード：フォント、かな書体、特徴量抽出

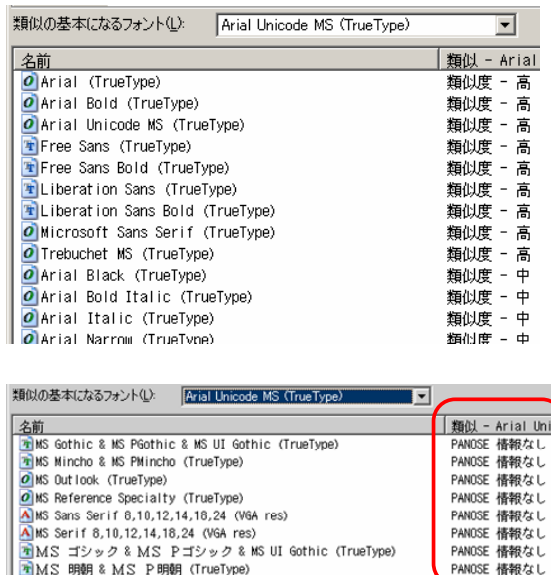
1. 研究開始当初の背景

本申請の着想は ISO/IEC JTC1/SC34/WG2 小委員会において、フォント情報交換の既存標準である ISO/IEC 9541 の書体分類に対し、昨年国際標準に採択された ISO/IEC 14496-22(いわゆる OpenType 規格)の書体分類や書体置換メカニズムが適合せず、その適合性を議論する際に和文書体での書体特性の議論の材料の不足の指摘に基づく。

CSS3 や XSL-FO ではフォントの抽象的な指定のための属性を多数定義しているが、和文フォントを抽象的に指定するためにどのように利用できるのかははっきりせず、抽象的なフォント指定は明朝・ゴシック、固定幅・プロポーショナルといった粒度でしか信頼性がない。

欧文フォントについては、そのデザインを幾何学的なパラメータで数値的に分類し、書体の類似性の評価や、書体の置換の自動化手法が標準化されているが、和文フォントについてはそのような分類方法や自動置換の方法が存在しない。現在のフォントに関する国際標準では欧文書体以外は書体をまったく分類しないか、あるいは主観的な分類にとどまっている。

たとえば、TrueType フォント中には書体分類のための Panose 値が書き込まれているにもかかわらず、フォントフレームワークはそれを無視している実装は少なくない。一例を下に示す。



和文フォントの Panose 値が無視されるフォントフレームワークの例

この結果、フォントファミリ名にノイズが加わった場合、ゴシック・明朝の対比も含めてまったく書体分類が機能しない場合が多い。同一プラットフォームでの例、異種プラットフォームでの例を下に示す。

元のWord XMLファイル	Word XMLのフォント名を破壊したファイル
This is Century	This is Century
This is Times New Roman	This is Times New Roman
This is Liberation Serif	This is Liberation Serif
This is Garamond	This is Garamond
This is Book Antiqua	This is Book Antiqua
This is Bookman Old Style	This is Bookman Old Style
This is Tahoma	This is Tahoma
This is Arial	This is Arial
This is Liberation Sans	This is Liberation Sans
This is Courier	This is Courier
This is Courier New	This is Courier New
This is Lucida Console	This is Lucida Console
This is Liberation Mono	This is Liberation Mono
This is Lucida Sans	This is Lucida Sans

Windows上の表示	Office: mac v. X上の表示
This is Century	This is Century
This is Times New Roman	This is Times New Roman
This is Liberation Serif	This is Liberation Serif
This is Garamond	This is Garamond
This is Book Antiqua	This is Book Antiqua
This is Bookman Old Style	This is Bookman Old Style
This is Tahoma	This is Tahoma
This is Arial	This is Arial
This is Liberation Sans	This is Liberation Sans
This is courier	This is Courier
This is Courier New	This is Courier New
This is Lucida console	This is Lucida Console
This is Liberation Mono	This is Liberation Mono
This is Lucida Sans	This is Lucida Sans

元のWord XMLファイル	Word XMLのフォント名を破壊したファイル
この行はリュウミンL-KLです。	この行はリュウミンL-KLです。
この行は中ゴシック BBB です。	この行は中ゴシック BBB です。
この行はじゅん 101 です。	この行はじゅん 101 です。
この行は平成明朝 W5 です。	この行は平成明朝 W5 です。
この行は平成角ゴシック W3 です。	この行は平成角ゴシック W3 です。
この行は HGS 平成明朝 W3 です。	この行は HGS 平成明朝 W3 です。
この行は HGS 平成角ゴシック W3 です。	この行は HGS 平成角ゴシック W3 です。
この行は HGS 平成丸ゴシック W4 です。	この行は HGS 平成丸ゴシック W4 です。
この行は HGS 行書体です。	この行は HGS 行書体です。
この行は新正楷書 CBSK1 です。	この行は新正楷書 CBSK1 です。

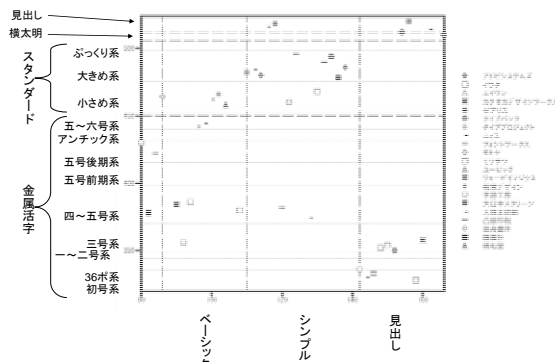
Windows上の表示	Office: mac v. X上の表示
この行はリュウミンL-KLです。	この行はリュウミンL-KLです。
この行は中ゴシック BBB です。	この行は中ゴシック BBB です。
この行はじゅん 101 です。	この行はじゅん 101 です。
この行は平成明朝 W5 です。	この行は平成明朝 W5 です。
この行は平成角ゴシック W3 です。	この行は平成角ゴシック W3 です。
この行は HGS 平成明朝 W3 です。	この行は HGS 平成明朝 W3 です。
この行は HGS 平成角ゴシック W3 です。	この行は HGS 平成角ゴシック W3 です。
この行は HGS 平成丸ゴシック W4 です。	この行は HGS 平成丸ゴシック W4 です。
この行は HGS 行書体です。	この行は HGS 行書体です。
この行は新正楷書 CBSK1 です。	この行は新正楷書 CBSK1 です。

欧文フォント名を破壊した場合の修復例と和文での修復例

2. 研究の目的

欧文と比較して和文の金属活字やフォントは文字数が非常に多く、特定の書体の特徴を把握するためのパラメータ数が見積もることが難しい問題がある。しかし、実際にいくつかの分類を調査すると、分類者や用途が異なる分類の間でも相関性が見られ、注目して

いる図形的特長は共通していることが予想される。一例として祖父江分類と小宮山分類の関係を下に示す。



祖父江慎のかな書体分類値(縦軸)と漢字書体分類値(横軸)の比較。かなのスタンダード書体と漢字のシンプル書体に強い相関が見られる

和文タイポグラフィの研究は、欧文のローマン体に基づく研究に比較すると、図形の特徴量把握が困難なために、書体デザインや可読性・視認性に対する影響に関して非常に主観的なレベルに留まっている。本研究では、現在の和文の大半を占めるひらがなについて、書体デザインの特徴を把握するためのパラメータの定義と抽出方法の確立を目指す。

3. 研究の方法

フォントからのアウトラインおよび特徴点抽出により、キャラクタレベルで共通の方法で計測できる特徴量を検討し、また、既存の書体分類で注目している特長量との関連性を調査した。既存の和文タイプデザインで用いられる用語から書体印象にかかわる属性を検討し、数値化できると判断されたものは計測文字と計測方法を検討し、さらに ISO/IEC 9541 や Panose によって交換できるものとできないものに分類した。数値化できないと判断されたものについては、少数のグリフによって例示できるかどうかを検討し、例示できるものについては指標文字を選定した。

さらに、それらと欧文フォントでの特徴量との類似性を比較し、既存の欧文フォント用の特徴量のみが伝達されるフォントフレームワークによって伝達するための対応関係を整理した。

また、電子文書中に埋め込まれたフォントについて、サブセット化に影響を受けない部分を検出し、そのハッシュ値によってレンダリングの際のアウトライン変形を検出する手法を開発した。

4. 研究成果

(1) アウトライン抽出およびグリフ同定手法の開発

オープンソースのフォントラスタライザである FreeType2 をもとに、TrueType フォントおよび Post-Script フォントのアウトライ

ンから制御点を抽出する方法を開発した。PostScript や PDF などのページ記述言語に TrueType フォントを埋め込んだ場合にアウトライン記述方式が変更される場合があるが、その際に制御点が移動される場合や、描画ベクトル方向そのものも変形される場合があり、フォントの特定に課題があることを発見した。

電子文書中に埋め込まれたフォントの特定手法として、TrueType フォントのサブセット化に際して維持される描画サブルーチンを検出し、このハッシュ値によってフォントを特定できることを示した。また、ページ記述言語はフォントをページ単位で埋め込むため、アウトライン描画ルーチン抽出やハッシュ計算よりもフォントの抽出処理が処理時間に大きく影響することを見出し、フォントキャッシュを用いることで電子文書をラスタした結果を比較するよりも高速にフォントを同定することが可能であることを示した。

(2) 既存研究におけるタイプデザイン用語の整理と対応する特徴量の計測方法の検討

欧文フォントのタイプデザインは全てが数値化できるわけではないが、Panose などの分類においては指標文字とどの部分に注目して評価するかが明らかである量が用いられている。典型的なものでは xHeight があるが、本来は主観的な用語であったウエイトやカウンタも Panose の中では計測方法が定められている。

和文フォントのタイプデザイン用語では注目する部分や計測方法が明らかになっているものはほとんど無い。例外的に矢作勝美や佐藤敬之助による金属活字・写真植字字形の計測を行っているが、前者は築地活版活字・秀英舎活字のデザイン傾向の差、後者はデザインの時代変遷の大きな傾向を明らかにするためのものであり、以後の書体分類ではそれらの計測方法によって書体分類を行おうという活用はほとんど為されていない。

本研究では広く使われるにもかかわらず定義が明らかでないタイプデザイン用語として「ふところ」「重心」をとり、指標文字の選定と計測方法を検討した。1945 年以前の金属活字、および現在市販のデジタルフォントについて、出版での出現頻度の高い常用漢字から「口」「冂」を含む漢字の矩形、および、「十」「元」「未」などの中央の横画を持つ字形について調査し、単一のフォントでも全てが同一の傾向を持ってデザインされているわけではなく、複数のグループがあることを明らかにした。

この調査結果を踏まえ、計測文字として写真植字の文字デザインの慣習において初期段階に作字される「東」「国」を選定した。

(3) 既存実装を通じた和文フォントの特徴量
伝達方法の検討

Panose 規格それ自体は欧文フォント以外には全く異なる分類パラメータを持つことを念頭に置いて設計されている。たとえば、セリフ体・サンセリフ体のパラメータと、筆記体やオーナメント(装飾書体)のパラメータは全く異なる。しかし、既存実装を調査したところ、現在国際標準となっているフォントフォーマット(ISO/IEC 14496-22)はこれらの分類以前に設計されたものであり、欧文セリフ体・サンセリフ体のパラメータしか格納できないことがわかった。

そのため、和文フォントのパラメータも欧文セリフ体・サンセリフ体のパラメータに対応づけなければ既存実装に伝達できない。この対応づけについて、TrueType Consortium Japan の規格に追加する形での検討を行い、試行標準に反映させた。

また、Panose 規格と国際標準フォントフォーマット間の非互換問題については上記国際標準をメンテナンスしている ISO/IEC JTC 1/SC 29/WG 11 に指摘を行い、改善方法を議論している。

(4) ISO/IEC 10646 に提案された統合漢字拡張 E への改善提案

漢字タイプフェイスの変遷調査の過程で、古漢字を強引に明朝体化した場合、本来はデザイン差であったものが字体差として扱われることが多いことが判明した。ISO/IEC 10646 の統合漢字に提案されている漢字の中にこの問題を持つものが 100 字以上見つかっており、ISO/IEC JTC 1/SC 2/WG 2/IRG に指摘を行い、提案改善を議論している。

5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文] (計 9 件)

1. 鈴木俊哉 「電子文書中の TrueType グリフ照合とその高速化手法の検討」、情報処理学会第 74 回全国大会講演論文集、1 (2012)、p.533-534、査読無
2. 鈴木俊哉、小町祐史、長村玄 「情報処理学会試行標準フォントリソース参照方式原案のファイナライズ」、画像電子学会 VMA 研究会 第 31 回講演概要集、31-4 (2011)、p.1-8、査読無
3. 鈴木俊哉 「電子文書中に埋め込まれたフォントの描画プログラムによる特定手法の考察」、情報処理学会研究報告、DD-83(2) (2011)、p.1-5、査読無
4. 鈴木俊哉 「電子文書中のフォント特定とヒント制御」、画像電子学会 第 258 回研究会講演予稿、11-2-10 (2011)

p.51-57、査読無

5. 鈴木俊哉 「Type1 埋め込みされた TrueType フォントの同定手法の考察」、情報処理学会研究報告、DD-79(8) (2011)、p.1-7、査読無
6. 鈴木俊哉 「TrueType の UVS 拡張を用いた Adobe CID 字形集合の代替処理の高速化」、情報処理学会研究報告、DD-76(5) (2010)、p.1-6、査読無
7. 鈴木俊哉、小町祐史、長村玄 「フォントリソース参照方式の試行標準(TS)原案」、画像電子学会 VMA 研究会第 25 回概要集、25-5 (2009)、p.1-12、査読無
8. 鈴木俊哉 「Investigation of Opaque Glyphs Synthesized from Old Hanzi」 Internationalization & Unicode Conference 33, S05-3 (2009)、p.1-37、査読有
9. 鈴木俊哉 「統合漢字に申請された殷周金文集成引得図形文字の調査」 電子情報通信学会 技術研究報告、109(160) (2009)、p.35-42、査読無

[学会発表] (計 5 件)

1. 鈴木俊哉 「西夏・女真・契丹文字の国際標準コード化動向について」 東京外国語大学アジア・アフリカ言語文化研究所共同研究課題「契丹語・契丹文字研究の新展開」、2011 年 7 月 9 日、東京外国語大学アジア・アフリカ言語文化研究所
2. 鈴木俊哉 「国際標準における契丹文字符号化の動向」 契丹文字研究会、2011 年 1 月 29 日、京都大学
3. 鈴木俊哉 「Comment on IRG N1519 part 1, 2, 3 and 4, Chinese evidences for CJK Unified Ideographs Extension E」 ISO/IEC JTC 1/SC 2/WG 2 IRG#35、2010 年 11 月 9 日、マカオサイエンスセンター(中国、マカオ特別行政区)
4. 鈴木俊哉 「Proposal to improve source information of ZJW characters」 ISO/IEC JTC 1/SC 2/WG 2 IRG#34、2010 年 6 月 21 日、長岡技術科学大学
5. 鈴木俊哉 「Comments on Proposed CJK Unified Ideographs from the Source “Index to the collection of Inscriptions in Yin-Zhou Period”」 ISO/IEC JTC 1/SC 2/WG 2 IRG#33、2009 年 11 月 23 日、ダナンソフトウェアパーク(ベトナム、ダナン)

6. 研究組織

(1) 研究代表者

鈴木 俊哉 (SUZUKI SHUNYA)

広島大学・情報メディア教育研究センター・助教

研究者番号：70311545