手書き文字認識用辞書のネットワーク分散データ収集システム

Network data collection system
for the Japanese hand-written character recognition system

福居 宏和
Hirokazu FUKUI

株式会社 アックス (〒604-0857 京都府京都市中京区烏丸通二条上ル蒔絵屋町 280 カーニープレイス京都烏丸 8F　E-mail: fuku@axe-inc.co.jp)

ABSTRACT. We developed a system that collects pattern data of Japanese hand-written character, and generates a dictionary of Japanese hand-written character recognition system from collected pattern data in this system, and distributes generated dictionary. This system is aimed at collecting Japanese hand-written character pattern data that anyone is able to apply and redistribute. Already we have collected 34,842 character pattern data under our system.

１．背景

近年は携帯マシンで Linux が動作することが当たり前になっている。いわゆる PDA 型携帯マシンには、キーボードがなく、タッチパネルによってすべての入力を行わなければならない。

我々は携帯 Linux に必要なソフトウェアをオープンソース(GPL)で開発し提供するプロジェクト「式神(しきがみ)[1]」を行っている。手書き文字認識システム「布目(ぬのめ)」も式神システムの一部として開発中である。

現在、布目を含む式神システムは、第 2 版をリリースし、広く無償で配布を行っている。現在の布目の課題の一つは、手書きデータを広く収集することである。

２．目的

本プロジェクトでは、手書き文字データを、インターネットを介して、広く収集するシステムを開発する。また、そのシステムを実際に運用し、大規模にデータを収集する。収集したデータは手書き文字認識用の辞書とするために機械処理する。さらに、辞書用として加工したデータを広く誰もが取得し、利用できるように、本システムにより配布を行う。

３．成果

（1） 手書き文字データ収集システム

インターネットを介して手書き文字データの収集を行うシステム開発した。システムの概要を図 1 に示す。サーバ・クライアント間の通信は TCP のポート 80 番を通して行う。

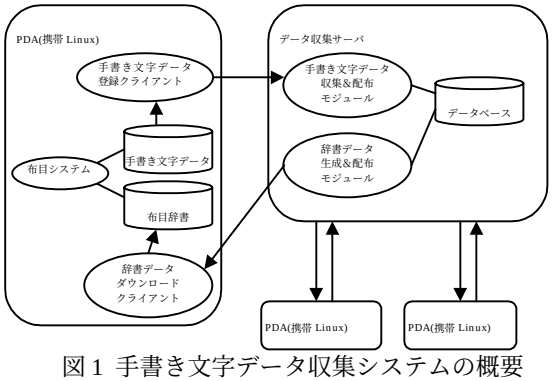

図 1 手書き文字データ収集システムの概要

（2） Web による投票システム

サーバに登録された手書き文字データの妥当性を投票するシステムを開発した。投票は Web ブラウザを介して行う。誤って登録された手書き文字データを除外するためにこのシステムを作成した。

Web による投票システムは、
http://nnmnd.hwchar.org/hwcvote/ にて運用されている。

Web によるシステムの投票を行う画面を図 2 に示す。

図 2　Web による投票システム

図 3 布目用 GUI フロント・エンド(動作中の画面)

（3） データ収集システムのサーバの運用

本データ収集システムの運用は 2002 年 10 月 31 日から開始した。2003 年 4 月 20 日までに延べ 34842 文字の手書き文字データを収集した。収集したデータには、6 セット分の JIS 第一水準（ただし、一部の記号やギリシャ文字、キリル文字を除く）の文字が含まれている。
サーバに収集された手書き文字データは、Web による投票システムを使用することで、ビットマップ形式で閲覧することができる。手書き文字データ・ダウンロード・クライアントを用いることで、誰でもサーバに収集された手書き文字データをダウンロードし、利用することができる。

（4） 布目用 GUI フロント・エンドの開発

手書き文字データ収集システムに対する布目用 GUI フロント・エンドを開発した。布目用 GUI フロント・エンドは、以下の機能を持つ。

- サーバ上の布目用辞書の一覧を表示
- サーバ上にユーザ辞書を作成
- 布目でこれまで入力した手書き文字データをサーバに登録
- サーバ上の布目ユーザ辞書を更新
- サーバ上の布目ユーザ辞書をダウンロード

ユーザ辞書の更新を行うと、新規に登録された手書き文字データを反映させた辞書がサーバ内で再生成される。
このプログラムにより布目ユーザは、手書き文字の登録、サーバ上の辞書の更新、辞書のダウンロード、布目用ユーザ辞書の置き換えのサイクルを簡便に行うことができる。
Hewlett-Packard 社製の PDA である iPAQ 上で布目用 GUI フロント・エンドを起動した画面を図 3、図 4 に示す。

図 4 布目用 GUI フロント・エンド
(スクリーン・ショット)

（5） 手書き文字データ入力用のツールの改良

式神プロジェクトから配布されているプログラム nnmsample の改良を行った。このプログラムは、JIS X 0208 の文字を順番にユーザに入力させ、手書き文字データを収集するツールである。
本プロジェクトで行った主な改良は以下の 2 点である。

- 大きなフォントを用いて入力すべき文字を表示するように変更した
- 入力の進捗状況を示すバーを追加した

プログラム nnmsample の画面を図 5 に示す。本プロジェクトでサーバに登録した手書き文字データの大部分はこのプログラムによって入力した。

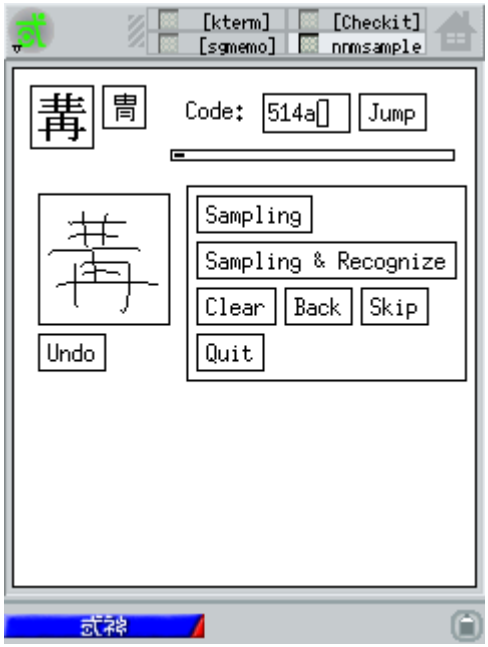

図 5　nnmsample

（6） フリーソフトウェア、フリーなデータとして成果物を配布

本プロジェクトで作成したプログラムの全てのソースコードは、フリーソフトウェア(GPL)として http://www.sikigami.com/~fuku/nnmnd/ で配布している。

また、全ての収集した手書き文字データも同様に誰もが自由に利用および再配布可能なデータとして配布している。

（7） 仮名漢字変換入力システム用辞書データ収集および辞書の生成および配布モジュール

仮名漢字変換入力システム用辞書データ収集および辞書の生成および配布を行う本システムのための追加モジュールを開発した。

このモジュールのサポートする仮名漢字変換入力システム用辞書は以下の通り。

- FreeWnn[2]
- Canna[3]
- Anthy[4]

現在の実装では、辞書の生成時に異なる仮名漢字変換システムへ辞書データの変換を行わない。 例えば、FreeWnn の辞書データとして収集されたデータは、Anthy 向けに配布する辞書に含めることはできない。

４．今後の課題

（1） 本システムの普及

本プロジェクトの作業として、手書き文字データの入力を行い延べ 6 人分のフリーな手書き文字データを収集することができた。手書き文字認識システムの開発を開始するには十分な量であるが、手書き文字認識システムの認識率向上させるには、辞書作成のためやより実際に近い認識率の測定により多くの手書き文字データが必要である。

これまでに収集した手書き文字データの大部分は、本プロジェクトで入力したものである。一般のユーザから提供されたデータはほとんど含まれていない。

一般のユーザの入力がない理由として、本プロジェクトのターゲットとした PDA の iPAQ を実際に使用しているユーザ数が少ないことが原因ではないかと考えられる。本プロジェクトを開始時には Linux が動作し、手書き文字認識システムの実用的に動作している PDA は iPAQ しか選択肢が存在しなかった。

より多くの手書き文字データを集めるために、よりユーザ数の多い PDA への移植を行うことが今後の課題として残っている。2002 年 12 月にシャープ株式会社より発売された Linux の動作する SL シリーズザウルスが活動しているユーザ数も多く、移植先の有力候補である。

（2） 仮名漢字変換システム用辞書サポートの改善

本データ収集システムの仮名漢字変換システム用辞書サポートの実装は、辞書ファイルの登録およびダウンロードが確認できたにとどまる。

５．参加企業及び機関

株式会社アックス

６．参考文献

[1] 式神. http://www.sikigami.com/
[2] FreeWnn. http://www.freewnn.org/
[3] Canna. http://www.nec.co.jp/canna/
[4] Anthy. http://anthy.sourceforge.jp/