ヒューマンインタフェース 73－1
（1997. 7. 11）

電子テキスト及びディジタルインクに対する統一的検索機能

張 スン　　中川 正樹
東京農工大学工学部電子情報工学科

本稿では，ディジタルインクと電子テキストに対する統一化された文字検索機能について述べる。指定された文字列について，検索対象が文字コードからなる電子テキストならば，Boyer-Moore 法により検索し，対象がディジタルインクファイルの場合は，文字列の標準パタンとインクファイルを照合してその類似度から検索する。後者の検索機能をオンライン手書き文字パタンデータベース（TUAT Nakagawa Lab. HANDS-kuchibue_d-96-02）に適用してみて，その性能を評価する。

Unified Search Function for Electronic Text and Digital Ink

Shen Zhang and Masaki Nakagawa
Department of Computer Science,
Tokyo University of Agriculture & Technology

This paper proposes a unified search function for both of electronic text and digital ink. It employs the Boyer-Moore algorithm to search for occurrences of a word within an electronic text as well as applies pattern matching for digital ink. The performance of the latter method is evaluated using the on-line handwritten character pattern database (TUAT Nakagawa. Lab.HANDS-kuchibue_d-96-02).

1．はじめに

ペン入力の PDA や電子手帳，タブレットなどが普及するにつれ，手書き入力されたパタン（ディジタルインク）を送受信し，それらをファイルに保存する利用が増えることが考えられる。そうなると，それらを検索する機能が不可欠になる。このとき，従来の文字コード列の電子テキストに対する検索も利用でき，ユーザには検索対象がどちらであるかを意識する必要がないことが大切であろう。また，自分の手書きパタンだけでなく，他人から電子メールなどで受け取った手書きパタンも検索できるように，個人へに依存性を緩和した検索機能である必要があろう。

電子テキストの検索は，検索対象となるデータが文字コード列であって，アルゴリズムはさまざま報告されている，本稿では，電子テキストの検索アルゴリズムには，Boyer-Moore 法を採用した[7]。

一方，ディジタルインクの検索における検索アルゴリズムは，電子テキストと同じような単純なコード比較処理では済まない。検索対象となる各々の文字パタンの類似性を評価する方法が必要とする。文字パタンの類似性を評価するためには，本研究室で開発されたオンライン文字パタン認識エンジンが利用できる[1,2,4,6]。

なお，本研究では，ディジタルインクとして，文字の区切りが挿入された文字パタン列を対象にする。これは，最近の PDA などで手書きパタンをその場では認識させないものの後から認識させ易いように，一文字一文字を筆記枠の中に書かせる入力方式[5]が普及してきていることから現実性を欠くものではない。一方，枠なしの状況で筆記された文字パタン列の検索も，枠なし認識を適用することで可能であるが，その検索への利用はこの報告では対象外とする。

2．システム概要

本システムの設計方針は，次の３点である。

- 検索機能の設計・実現では，ディジタルインクの検索を重視する。
- システムの拡張を容易にするため，システムの構成上，各部分の機能を明確にし，違う種類のものは，同じブロックにしない。
- ユーザインタフェースを重視し，操作はなるべく統一する。

システムは，インタフェース部，文字認識部，そして検索部３部分から成る。

インタフェース部は，ファイルのオープン，表示，検索文字列の入力，検索結果を確認するための入出力部分を担当する。文字認識部は，ペンで入力された文字パタンをテキストの文字列に認識するときに利用される。また，検索部は，電子テキストまたはディジタルインクファイルに対し，それぞれの適応する検索エンジンを利用して検索処理を行う。

システム全体の構成を図１に示す。今回文字認識部には，本研究室で開発したオンライン文字認識エンジン[1,2]を採用した。なお，この部分は交換可能である。

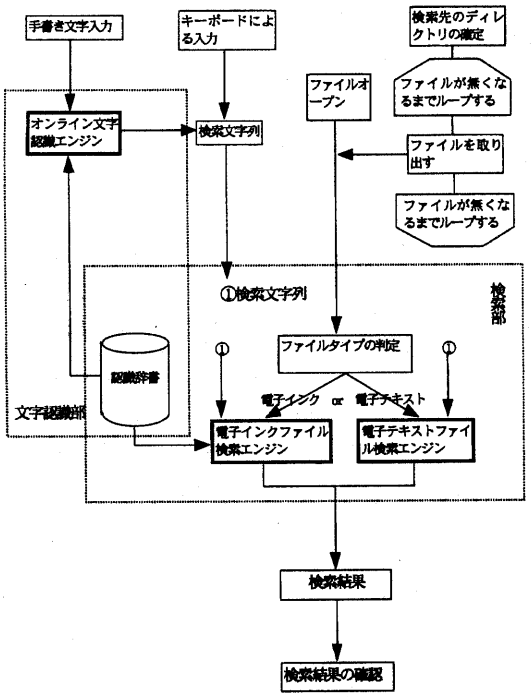

図１システム全体の構成

Fig. 1 Flow diagram of the system.

３．手書きパタンの検索アルゴリズム

文字認識処理の流れを，次に示す[1,2]。

① 正規化・特徴点抽出（入力パタンの大きさを辞書パタンの大きさに合わせて，特徴的な筆点だけを抽出する）

② 認識候補の大分類（辞書カテゴリのうち，入力パタンに該当しそうなものとそうでないものを分ける）

③ 特徴点対応付け（入力パタンの特徴点と辞書パタンの特徴点を対応づける）

④ 類似度算出（入力パタンと辞書パタンの間の類似度を算出する）

⑤ 前後の認識結果からの文脈判断

文字認識は，手書きパタンを文字コードに変換する処理である。一文字の手書きパタンに対して，認識候補の大分類処理を行なった後，残った候補に入っている標準文字パタンとマッチングを行い，類似度を算出する。

手書きパタンファイルから，文字列を検索する場合，検索先のディジタルインクファイルを文字コードのテキストファイルに認識する必要はない。検索したい文字列の標準手書き文字パタンと検索先のディジタルインクファイルの各々の文字パタンとの間の類似度を算出し，ある程度の基準値を満たしていれば，検索できたと判定できる。この方法で検索を行う場合，処理時間を大幅に短縮することができる。

ところが，実際に検索実験を行ったところ，画数によって，類似度の基準値が違ってくることがわかった。したがって，類似度がある基準値を超えるかどうかで検索を行う場合，すべての文字に対して同じ値の基準値を設定すると，検索結果が不安定になる。そこで，類似度基準の調整が必要となる。しかし，文字ごとに基準値を設定するのは，大変な手間を要する。そこで，とりあえず比較的画数が多い漢字，文字の特徴が近い，平仮名，片仮名，ローマ字などで分類してみた。

本稿で提案する，手書きパタンに対する検索処理の流れは，次のようになる。

① 検索文字列の各々の文字に対応する認識辞書の標準パタンのカテゴリ番号を確定し，配列Ｃに入れる（図２）。

② 検索文字列の総合類似度基準値Ｓを求める。本稿で提案する検索アルゴリズムでは，基準値設定を字種別に分類するため，検索文字列の各文字の字種に対応する基

準値の合計が総合類似度基準値Sとなる。

③ ディジタルインクファイルの手書き文字パタンを指す番号 i と①で求めた配列 C の要素番号 j を初期化する。

④ 検索先のディジタルインクファイルの1文字の手書きパタン P_i と認識辞書のカテゴリ番号 C_j の標準パタン $S_{C[j]}$ との類似度 R を算出する（図3）。

⑤ もし，④で求めた類似度 R が一定の値に満たさないなら， i を次の式によって，設定する。

$$i = i - j + 1。$$

j を初期値に戻す，累積類似度を 0 にする。④の処理へ戻る。

⑥ 類似度 R を累積類似度 A に足し， j に 1 を足し， i に 1 を足す。

⑦ j は検索文字列の文字数と等しかより大きければ ⑨ の処理へ。

⑧ 検索先のディジタルインクファイルから，文字の手書きパタン P_i を次の類似度を計算する対象として指定し，④の処理に戻る。

⑨ 累積類似度が②で求めた総合類似度基準値を超えれば（A > S），検索文字列が検出したと判定する。そうでなければ，i を③で設定した初期値より次の文字の手書きパタン番号に設定し， j を初期値に戻し，④の処理へ戻る。

以上

検索の流れを次の図4に示す。

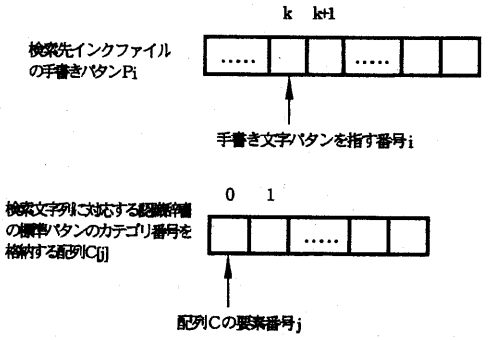

図3 文字列の標準パタンとインクファイルの照合例

Fig. 3 Example of compared standard handwritten pattern with digital ink.

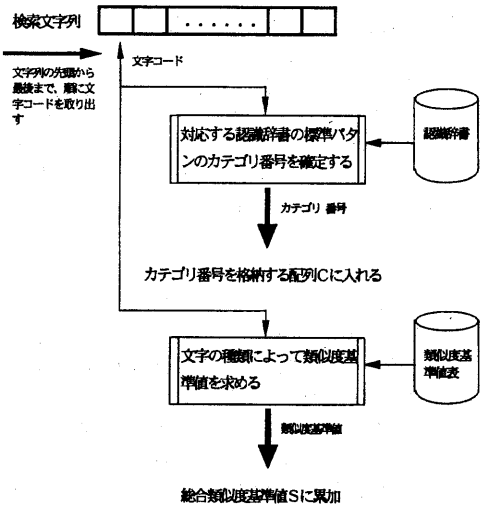

図2 検索処理の準備処理

Fig. 2 Preparation for search process.

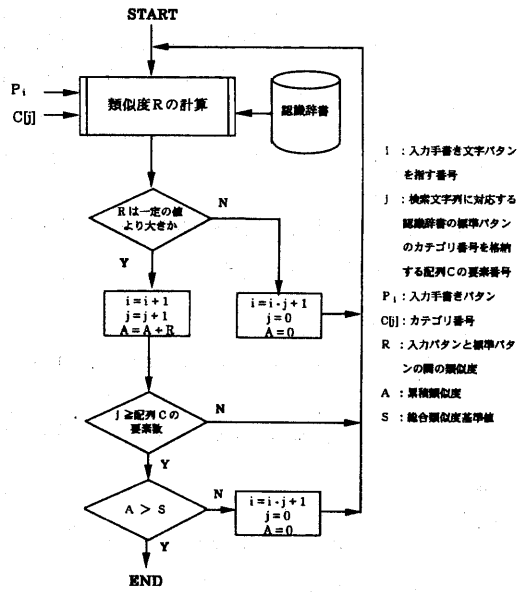

図4 検索処理流れ図

Fig. 4 Flow diagram of search process.

4 検索アプリケーションの実現

本アプリケーションの実現は，MS-WINDOW 95 日本語版の上で行った。開発言語は，Borland C++ ver 5.0 である。

アプリケーションの設計方針を次に示す。

- 検索文字列の入力は，キーボードから入力する他，ペンによる手書き入力もできるようにする。つまり，本検索アプリケーションの検索機能において，ペンだけで操作ができるようにする。
- オープンしたファイルに対して検索をかけるローカル検索と，あるディレクトリ(Windows 95 ではフォルダ)以下のファイルに対して検索文字列を検索するグローバル検索を両方サポートする
- 検索エンジンは，内部構造上電子テキストとディジタルインク検索エンジンを一緒にまとめずに，それぞれ別にするが，使うユーザには，検索ファイルが電子テキストファイルであろうと，ディジタルインクファイルであろうと意識せずに使えるようにする。
- 類似度基準の設定はユーザが自由にできるようにする。もちろん，初期状態では，平均的に検索率が高い類似度基準値を標準値として設定する。

実際に作成したアプリケーションの対話画面を図５に示す。

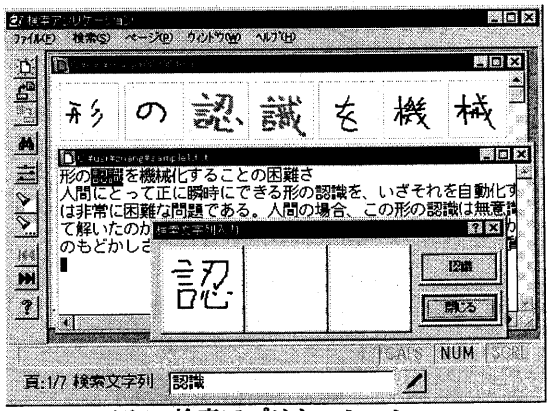

図５ 検索アプリケーション
Fig 5 Interaction Window.

5．ディジタルインクにおける検索の評価

本研究では，実験対象として，オンライン手書き文字パタンデータベース TUAT Nakagawa Lab. HANDS-kuchibue_d-96-02[3]を利用した。このデータベースには，各人約 10,000 文字パタンが文章列で筆記されている．このデータベース 10 名分を実験に使用した（言葉を統一するため，以降オンライン手書き文字パタンデータファイルをディジタルインクファイルと言葉を変える）。ファイル名は Nk10w.ipd ～ Nk19w.ipd である。ディジタルインクファイルと同じ文字列の電子テキストファイルは，Asahi93.txt である。

検索では，検索したい文字を見付けることが重要である。検索文字と検索された文字パタンが合っている場合，その間の類似度はどれぐらいあるかを測定することによって，類似度の基準を設定するのに参考となる。

電子テキストファイル Asahi93.txt と実験対象のディジタルインクファイルには，同じ内容の文章が文字コード列および手書きパタンで格納されている，したがって電子テキストファイルの最初の文字は，ディジタルインクファイルの最初の手書きパタンと同じ文字であるというふうに，一対一である。順番に電子テキストファイルの文字を一個ずつ認識辞書から標準パタンを確定して，それをディジタルインクファイルの文字パタンと類似度を算出することで，文字が合っているとき，類似度がどれぐらいあるかはわかる。

この測定方法で，漢字，平仮名，片仮名，ローマ字，その他，各種類の文字について，類似度の平均値を表１にまとめた。最後の行は，10 セットファイルのすべて文字パタンについて，字類別の類似度平均値と標準偏差値である。

図６は表１をグラフにしたもので，字類別の類似度の分布図であり，図７は字類別の類似度平均値と標準偏差値をグラフにしたものである。

実験結果でわかるように，文字の種類によって，類似度が異なり，偏差値の差が大きい。人（実験対象のディジタルインクファイル）によって，バラ付きがあるが，比較的，漢字，平仮名，片仮名の類似度が高く，ローマ字は平均的に少し低く，その他の文字（主に符号や，特殊記号や，ギリシャ文字など）は最も低いことが

表１実験結果（各種類の文字類似度の平均値）

Table 1 Results of experiments.

ファイル名	漢字個数(5643)	平仮名個数(4492)	片仮名個数(614)	ローマ字個数(166)	その他個数(1047)
nk10w	837±140	858±66	859±136	837±123	810±233
nk11w	786±181	759±148	755±173	759±156	663±304
nk12w	836±180	861±71	839±138	820±138	630±346
nk13w	802±191	824±135	845±105	834±99	765±261
nk14w	722±273	827±95	815±136	781±203	770±235
nk15w	834±156	810±144	813±134	812±175	758±253
nk16w	814±180	829±72	809±122	804±166	667±307
nk17w	869±134	874±57	880±93	874±95	813±222
nk18w	793±208	830±178	830±193	809±200	774±262
nk19w	861±146	873±57	868±112	791±205	831±239
平　均	815±176	835±115	831±141	812±164	748±277

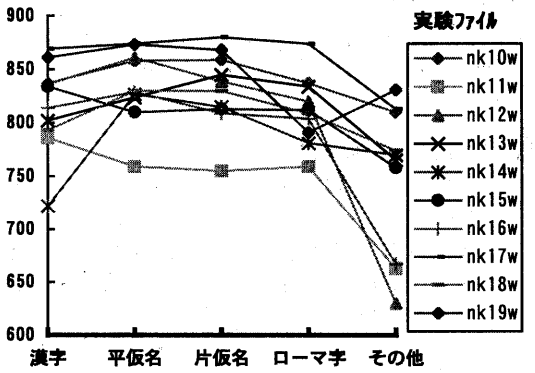

図６ 字種別の類似度の分布

Fig.6 Similarity distribution.

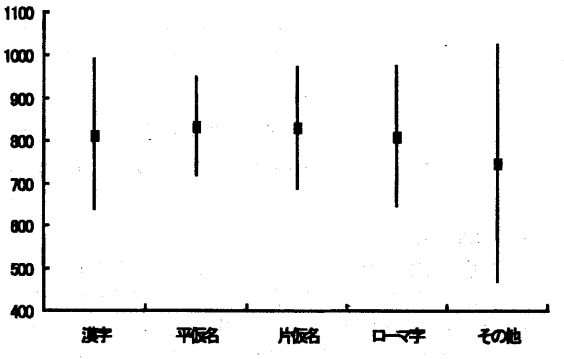

図７ 字種別の類似度平均値

Fig.7 Average of similarity.

わかる。これは，文字の画数の違いによって，差が出たと考えられる。画数の多い文字では，文字パタンの情報量が多い，逆に画数の少ない文字は，文字パタンの情報量が少ない。また，句読点などは，人の書く癖がある。たとえば，セミコロンを書くとき，上の丸部分はペンをタブレット上で軽くタッチする人もいれば，太く塗り潰したような点を書く人もいる。結果的に，その他に分類された文字の類似度は低くなる。

次に，実験用の１０セットディジタルインクファイルについて，比較的出現した頻度が高い１０種類の２文字熟語および同じく１０種類の４文字熟語を検索文字列として，手書き文字検索エンジンにかけてみた。検索の基準値を変えることによって，検索すべきものが検索できない誤り（第一種の誤り）と誤って検索された不正解（第二種の誤り）の割合を測定した。

結果のグラフ（図８）で分かるように，基準値を上げるに連れて，第二種誤りが減少していき，一方，第一種誤りが増加し，両方とも基準値が７００から８００の範囲で最適になった。したがって，本稿で提案した手書きパタンに対する検索アルゴリズムにおいて，基準値を適正に設定することにより，第一種および第二種の誤りを抑えて，ディジタルインクの検索が可能となる。

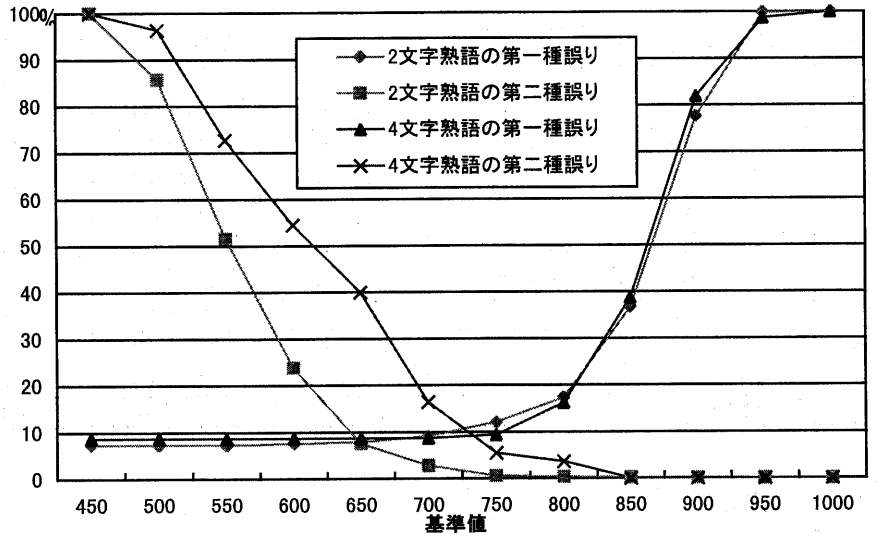

図８ 基準値に対する検索の第一種及び第二種の誤り
Fig 8 First-order & second-order errors in results of searching words by changing threshold value.

6．おわりに

本稿では，ディジタルインクファイルにおける検索アルゴリズムを提案し，これを利用して，検索先を意識することなく検索できるアプリケーションを試作した。発展課題として，枠なしのディジタルインクファイルへの対応などが考えられる。

本研究は，情報処理振興事業協会による創造的ソフトウェア育成事業の一部補助による。

参考文献

[1] 秋山 勝彦，中川 正樹：ストロークのつながりに寛容なオンライン手書き文字認識，画像の認識・理解シンポジウム（MIRU'94）Ⅰ，pp.67-74，(1994.7).

[2] M. Nakagawa, K. Akiyama, L. V. Tu, A. Homma and T. Higashiyama: "Robust andhighly customizable recognition of on-line handwritten Japanese characters," Proc. 13th ICPR, Vol. Ⅲ pp.269-273 (1996.8).

[3] 中川正樹，東山孝生，山中由紀子，澤田伸一，レー・バン・トゥー，秋山勝彦："文章形式字体制限なしオンライン手書き文字パターンの収集と利用，" 電子情報通信学会信学技法，95, 278, 43-48 (1995.9).

[4] 東川 レ-バン：ストローク数非依存の高速オンライン手書き文字認識手法，情報処理学会第 50 回(平成 7 年前期)全国大会，4D-4, 2-61(1995.3).

[5] 曽谷俊男，福島英洋，高橋延匡，中川正樹："遅延認識方式を用いた手書きユーザインタフェースの基本設計，" 情報処理学会論文誌，34, 1, 158-166 (1993.1).

[6] M. Nakagawa, T. Oguni and A. Homma, A Coarse Classification of On-line Handwritten Characters, Proc 5th IWFHR, pp.417-420(1996.9).

[7] 河西 朝雄：C言語による初めてのアルゴリズム入門，技術評論社，pp.148-149(1994).

[8] 加藤 直樹，田中 宏，中川 正樹：手書き電子メール環境の試作，計測自動制御学会第１２回ＨＩシンポジウム論文集，pp.189-194(1996.10).