論　文

切出しと認識を同時に行う活字デーヴァナーガリ文献の認識法

正員　鈴木　昭浩[†]　正員　金井　浩[††]　正員　川添　良幸[††]
正員　牧野　正三[†]　正員　城戸　健一[†]

Printed Devanagari Text Recognition Method by Simultaneous Extraction and Recognition Procedures

Akihiro SUZUKI[†], Hiroshi KANAI[††], Yoshiyuki KAWAZOE[††], Shozo MAKINO[†] *and* Ken'iti KIDO[†], *Members*

あらまし　デーヴァナーガリは，サンスクリット語やヒンディー語などの記述文字である．この文字の文字幅は字種によって大きく異なり，かつ文節は文字上部の横線によって接続されているため文字列からの文字の切出しを機械的に行うのは困難である．本論文では，文字の切出しと認識を同時に行う手法を用いたデーヴァナーガリ文字認識法を提案する．また，本手法を高頻度に出現するデーヴァナーガリ文字89字種9431文字に適用した結果，99.4％の認識率を得，本手法の有効性を確認した．

1. まえがき

古代の仏典の写本等はデーヴァナーガリ文字で記述されている．これらの資料は専門家以外には判読が困難であるため，ローマ字転写が仏教研究者から強く要望されている．しかし，人手による作業には限界があり，膨大な量の写本すべてをローマ字転写することは非常に困難であるため，転写作業の自動化が望まれている．また，デーヴァナーガリは，現在でもインドではヒンディー語等の記述のために広く用いられており，文字の自動認識が実現すれば広い用途がある．

デーヴァナーガリ文字の自動認識については，既にいくつかの報告[1],[2]がなされているが，いずれも認識対象とする字種数が少ない，文字の切出しに関する検討が不十分，といった点において実用的な文字認識システムまでには至っていない．そこで，本論文では，実用的なデーヴァナーガリ文字自動認識システム作成への第1段階として，現在まで考察されたことのなかった文字の自動切出しに着目して，その検討を行い，一つの実現方法として文字切出しと認識を同時に行う新しい認識法を提案する．

文字候補の抽出には，文字列を1画素単位にずらしながらパターンマッチングをとる手法を導入することにより，個々のストロークが分離できない文字列からの文字候補の抽出を実現する．また，文字候補抽出結果について，グラフとDPを適用して最適文字列を決定するが，この最適文字列の評価の基準には，テキスト未知文字列と，辞書の候補文字列を連結した文字列との，画素単位のユークリッド距離に関する厳密な整合を考えた評価関数を導入し，それに伴い，新たに，「画素上の空白」および「画素上の候補文字の重なり」という概念を導入する．

2. 認識対象

今回認識対象としたのは，1912年に出版されたケルン-南條本[3]と呼ばれるデーヴァナーガリによる法華経の活字校訂本である．以下，デーヴァナーガリ文字の概要と，認識実験のために作成したデータについて述べる．

デーヴァナーガリは，サンスクリット語やヒンディー語の表記に用いられる文字である[4]．この文字は，母音文字と子音文字に大きく分けられ，母音文字には独立体と半体がある．母音独立体文字は語頭に用いられ，

† 東北大学応用情報学研究センター，仙台市
Research Center for Applied Information Sciences, Tohoku University, Sendai-shi, 980 Japan
†† 東北大学情報処理教育センター，仙台市
Education Center for Information Processing, Tohoku University, Sendai-shi, 980 Japan

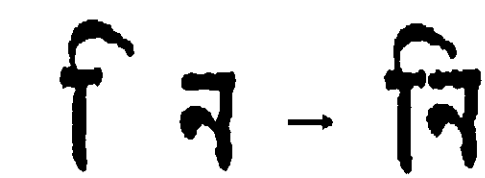


図1 デーヴァナーガリ文字例
Fig. 1 Example of Devanagari character.

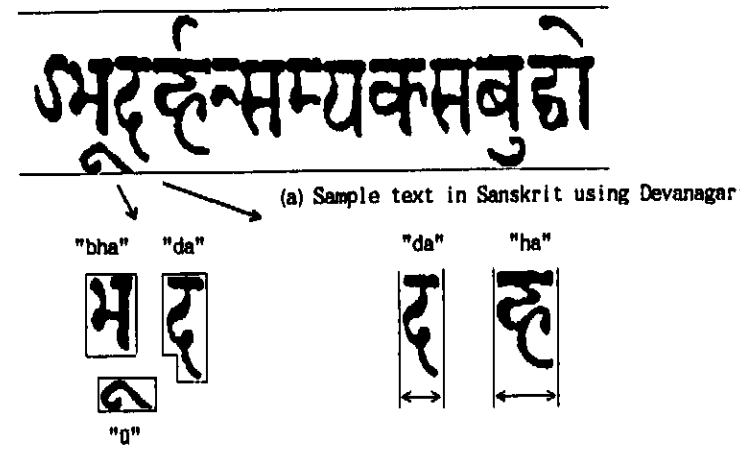


図2 デーヴァナーガリ文字列の例および文字の特徴
Fig. 2 An example of Devanagari text and feature of the characters.

半体は子音文字に添えられる．典型的なデーヴァナーガリ文字の構成例を図1に示す．デーヴァナーガリの基本的な文字は，母音独立体9字種，子音文字33字種，母音半体文字12字種，数字10字種の計64字種より構成される．加えて複数の子音文字が複雑に結合した数多くの結合文字が使われる．

デーヴァナーガリによる文章は，文字を左から右に並べることにより構成されるが，一つの文節中の文字は上部の横棒によって連結されている（図2(a)）．また，活字漢字と比較したときの活字デーヴァナーガリの文字形状に関する特徴として，次の2点が挙げられる．

（i） 一つの文字が文字列の上段又は下段に存在する文字（通常は母音半体文字）の影響を受けて一つの方形で囲めない場合がある（図2(b)）．

（ii） 字種によって文字幅にかなりの違いがある（図2(c)）．

文字列画像データは，上述の写本の1ページから28ページまでの計397行分を対象とした．画像の入力には解像度が8本/mmのイメージスキャナを用い，1行分を2176×96画素からなる2値画像データに変換している．このデータのうち1ページから5ページ分までを辞書作成に，6ページから28ページ分までを認識実験に用いた．

辞書に登録した文字は，1ページから5ページまでの間に2回以上出現した，子音文字29字種，母音独立体文字1字種，母音半体文字1字種，数字10字種，特殊記号4字種，結合文字44字種の合わせて89字種である．

3. 従来法による文字の切出し

手書き日本語文字列については，文字列や文字に関する一般的知識に基づき文字の切出しを行う手法[6]が利用されている．すなわち，文字の大きさや形についての一般的な知識を用いて文字の切出しを行うわけであるが，前章で述べたようにデーヴァナーガリ文字は，文字幅や文字の形が字種により異なり，しかも文字間が連結しているので，この手法をデーヴァナーガリ文字に直接適用することはできない．

文字形状に関する知識を用いずに文字の切出しを行う手法として，図2のような文字列の縦方向の投影を取りヒストグラムがあるしきい値以下になった点で文字を切り出す方法が考えられる．

前章で述べたように一般にデーヴァナーガリ文字列は，文字同士が横線部分で接している場合が多いが，印刷されたデーヴァナーガリ文字は各活字間に若干のすき間があるため，その情報による切出しの可能性がある．しかし，横線の部分は量子化誤差等のノイズや印刷時の濃度のばらつき等により太さが一定でない．そのため，図3に示すように1文字中のヒストグラムの最小値が文字間の値よりも小さな値をとる場合がしばしばある．従って，すべての文字を完全な形で切り出すしきい値の設定は困難である．しきい値をいくつか設定してそれぞれの場合に関して縦投影により文字の切出しを行い，その結果をもとに認識実験を行った結果，最適なしきい値（この実験では5画素）を用いた場合でも，主として文字の切出しの失敗のために70.7％

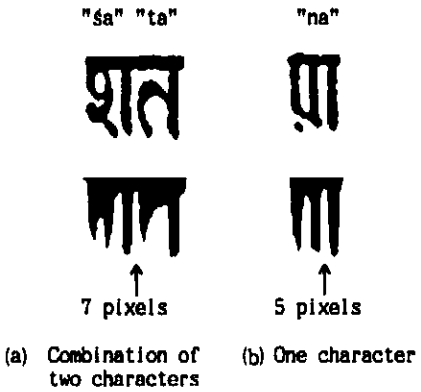


図3　文字列とその縦投影ヒストグラム
Fig. 3　Histograms on vertical projection of the characters.

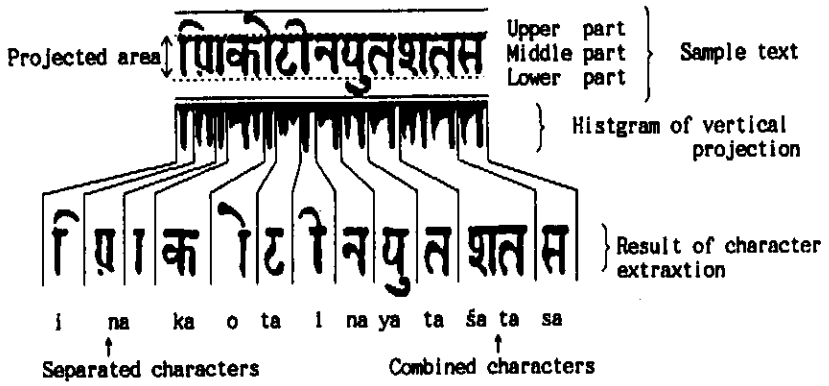


図4　縦投影による文字の切出し（中段のみ使用）
Fig. 4　Character extraction using histogram of vertical projection. Only middle part is projected.

の認識率しか得られなかった．

縦投影による文字の切出し例を図4に示す．ここで縦投影の範囲は図2(b)のような一つの方形で囲めない文字の存在を考慮し，図に示すように文字列の中段だけとした．この例より，文字中のヒストグラムの値が文字間の値よりも小さい場合には，一つの文字が2文字に分離して切り出されることと，2文字が結合して1文字として切り出されることがわかる．

オンライン文字認識については，各文字の切出しと認識を同時に実現し，高精度の認識を実現する候補文字ラティス法[6]が提案されている．この手法の適用を考えた場合，印刷されているデーヴァナーガリ文字からストロークの抽出を行うことは，デーヴァナーガリ文字の重要な部分（横棒と縦棒以外）が曲線主体で構成されていること，更にその部分のストローク数が小さいことにより困難であるため，この点を考慮した候補文字の抽出法の考案が必要である．

4. 文字の切出しと認識の同時処理

前章で述べたようにデーヴァナーガリ文字は，従来の手法を用いて，機械的に切出しを行うことが困難である．そこで，デーヴァナーガリ文字に適した新しい認識技法を提案する．

本手法は，文字列からの文字候補の抽出，最適な文字系列の決定の二つの処理より構成される．切出しと認識を同時に行う手法は既に候補文字ラティス法[6]で，提案されている．本論文では，その中の基本セグメント抽出に当る部分に，活字デーヴァナーガリ文字の性質を考え，文字列を1画素ずつずらしながらパターンマッチングを行う手法を適用し，個々のストロークが分離できない文字列からの文字候補の抽出を実現する．また，この文字候補抽出法の適用に伴い，新たに最適文字列の評価関数を定義し，導入した．

以下に，各処理について概要を述べる．

4.1 文字列からの文字候補の抽出

文字列に含まれると考えられる字種とその位置の検出を行う．処理には，パターンマッチング法[7]を用い，次のように文字列から直接文字候補の抽出を行う．

（1） 文字の仮の切出し　一つの文節内の文字列で，左から右に1画素ずつ位置をずらしながら文字の仮の切出しを行う．切出しを行う際の文字幅は，整合しようとする字種の文字幅と等しく設定する．ここで整合する字種の文字幅は，辞書作成用パターンから求めた文字幅の平均値とし，あらかじめ字種ごとに辞書に登録しておく．

（2） 整合する字種との距離計算　仮に切り出された文字と辞書文字との距離計算を行う．ここでは，距離尺度としてユークリッド距離を用いた．また，整合する際，縦方向の位置ずれの影響を除くため重心による縦方向の位置合せを行った．辞書文字は，各字種ごとに20サンプルの平均パターンから作成した．但し，特に出現頻度の低い文字については20サンプル以下の場合もある．

（3） 文字候補抽出　整合により得られた距離があるしきい値以下のとき，その字種，位置および距離を出力し，文字候補とする．ここで，字種ごとに文字幅が異なることによってパターンの次元数も異なるため，しきい値は文字幅で正規化した距離に関して設定した．

(1)から(3)の処理をすべての認識対象字種について行うことにより，個々のストロークが分離できない文字列からの文字候補の抽出を実現する．

図5に文字候補抽出例を示す．文字列に含まれる正

しい文字はすべて文字候補として抽出されているが，それ以外の文字も候補として抽出されている．そこで，この文字候補抽出結果をもとにして文字列中に実際に存在する文字の決定を行う処理を次に述べる．

4.2 最適な文字系列の決定

文字列から抽出された文字候補について，最適と思われる字種およびその位置の組合せを決定する．この決定のための評価関数として次の2種の評価値の和の荷重平均値を定義し，その最小値をとる文字列を最適候補とする．

（i） 入力文字と候補辞書文字間の距離の和

（ii） 候補文字により覆われていない文字列の余白部分

具体的には，

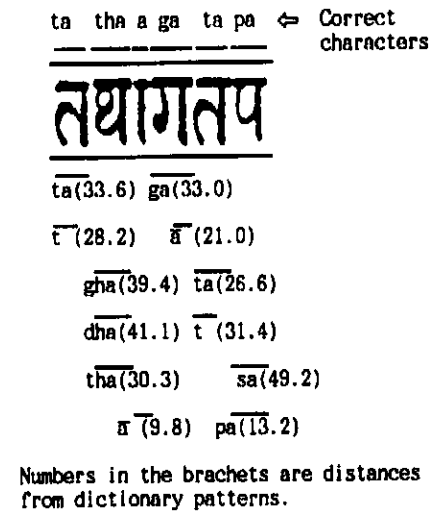
$F = W_1$（候補文字の分布する部分の距離）^{2}
$+ W_2$（余白の部分の距離）^{2}

で定義する評価関数 F を最小にする文字列の組合せを選ぶ．ここで，W_1 と W_2 は重み係数であるが，本論文では $W_1 = W_2 = 1$ と設定した．また，余白の部分の距離とは，文字候補により覆いつくされていない部分の文字列と，非黒点から構成される「空白」の辞書文字とのユークリッド距離とする．図6に示すような文字候補抽出結果の組合せに対する具体的な評価関数の値は，

$F = D_A^2 + D_B^2 + D_C^2 + D_{SP}^2$

で与えられる．

この評価関数は，認識の対象とする文字列と文字候補の組合せにより構成した文字列との，ユークリッド

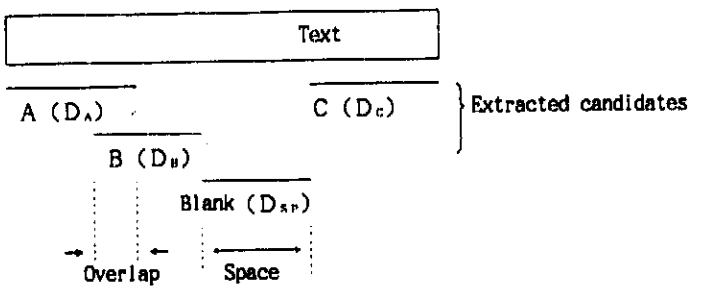


(a) Result of character extraction

Character	Width	Threshold
त ta	37	48.1
थ tha	36	46.8
ग ga	38	49.5
घ gha	39	50.7
प pa	35	45.5
स sa	38	49.5
ा ā	17	22.1
ृ t	29	37.7
ध dha	40	52.0

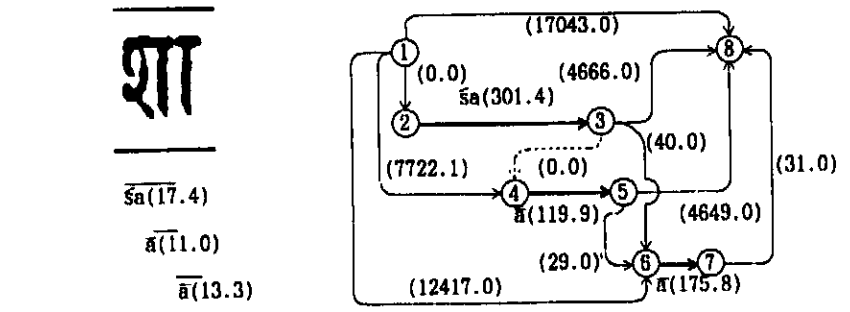
(b) Character widths and the thresholds for extraction

図5 文字候補抽出例
Fig. 5 An example of candidate character extraction.



A, B, C ・・・Candidates
D_A, D_B, D_C ・・・Distances between dictionary patterns and A, B, and C
D_{SP} ・・・Diatance between the space and the "blank" dictionary patterns

図6 文字候補同士の重なりと余白
Fig. 6 An example of overlap of candidate characters and blank part.



(a) Result of character extraction　(b) Graph extracted from the result indicated as (a)

図7　文字候補抽出結果に基づくグラフ表現
Fig. 7　Graph for representing candidate characters.

距離に関する厳密な整合を考えたものである．この評価関数においては文字列中の位置の絶対座標を用いているため，候補文字同士の間を接続する「空白」を導入している．これにより，文字列全体を覆いつくしていない候補文字の組合せに関しても，残った文字列の部分を「空白」とマッチングすることにより，文字列全体としての整合性の良さを評価することが可能となる．また，「空白」とともに，「画素上の候補文字同士の重なり」という概念も導入することにより，互いに重なり合う部分をもつ候補文字の組合せに関しても評価を与えることを可能とする．

以上述べたような評価関数 F を文字候補のすべての組合せについて計算し，F が最小となる組合せを最適文字系列とする．しかし，実際にすべての組合せを考慮した場合演算量が膨大となるため，本論文では処理を効率的に行うために文字候補抽出結果よりグラフを作成し，動的計画法[8]を用いて F を最小にするパスを選択する．この手法は，候補文字ラティス法[6]で用いられている方法と同様であるが，ブランチの重みのとり方を以下に述べるように活字デーヴァナーガリ文字の特徴に合わせて設定した．

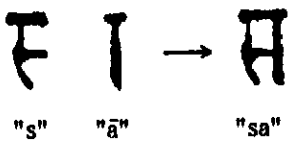
まず，文字候補抽出結果をグラフで表現する．文字列の左端および右端，候補文字の始点終点に対応するノードを設定し（図7中の①～⑧），ノードとノードを接続するブランチを，

（i）候補文字の分布する部分に対するブランチ（図7中の"——→"）．ブランチの重みは，マッチングにより得られた辞書文字との距離の2乗．

（ii）候補文字同士の接続および候補文字により覆われつくされていない部分の文字列に対するブランチ（図7中の"·····»"）．ブランチの重みは，前述の「余白部分の距離」の2乗．

（iii）候補文字同士の重なりに対するブランチ（図7中の"·····»"）．ブランチの重みは，0．

の三つの条件で設定する．ここで，デーヴァナーガリ文字においては，図8に示すように1文字が他の2文字の本来存在しない組合せにより完全に構成される場合があるため，このような文字系列を構成するブランチは設定しないこととした．



Since separated "s" and "a" are not expected to exist, single character of "sa" is selected.

図8　文字の結合に関する知識の利用
Fig. 8　Utilization of information concerning character connection.

次に，生成されたグラフについて，ブランチの重みの和を最小とするパスを，動的計画法を用いて探索する．ここでは，終点のノードから始点のノードに向かって経路を探索する逐次近似法を用いた．これにより求められたパスを構成する候補文字系列を認識結果とする．

なお，ブランチの重みを距離の2乗で定義したことにより，ブランチの重みの和はパスを構成する候補文字の組合せに対する前述の評価関数 F に対応する．

5. 認識実験

本手法をデーヴァナーガリ文書に適用し，認識実験を行った．認識対象は2節で述べた法華経経典の校訂本であり，1ページから5ページまでの文字データより辞書パターンを作成し，6ページ以降の文字列について認識を行った．認識対象字種は文字列の上部および下部に付加する母音半体文字を除いた89字種である．この文字列の上部と下部に付加する母音半体文字については，個々の文字が独立に存在しており，従来の文字切出し法で容易に処理できるので，実験対象から除外した．実験に用いた全文字列中に認識対象字種は合わせて9431字含まれている．

5.1 実験結果

表1に本手法による認識実験の結果を示す．99.4％の認識率が得られ，縦投影により文字を機械的に切り出して認識を行う場合と比較して格段の認識率向上が認められる．ここでの認識率は，正しく認識された文字数と文字列中の認識対象文字数との比の値で定義した．

表2に誤認識された文字についてその原因を分類して示した．文字認識誤りは，図9に示すような極めて類似した文字（類似文字）間の誤認識がほとんどであった．文字候補抽出の誤りは，図10に示すように文字ノイズが付加した場合に生じた．図11に実際の認識結果の出力例を示した．ここでは，文字列の上部および下部に付加する母音半体文字の認識も行っており，すべて正しい結果が得られている．

5.2 考　　察

実験結果より，本論文で提案した文字の切出しと認識を同時に行う文字認識法が，デーヴァナーガリ文字の認識に極めて有効であることがわかった．しかし，今後更に認識率の向上を目指すためには，次の点を検討する必要がある．

表1　認識実験結果

	縦方向写影	本手法
認識率	70.7％	99.4％

表2　誤認識文字の分類

誤認識率	0.39％（37文字）
候補誤り	0.19％（18文字）

全認識対象文字数＝9431文字

（1）類似文字の識別　　表2からもわかるように，本方法による誤認識の原因のほとんどは類似文字に関するものである．今回用いたユークリッド距離を用いた単純なパターンマッチング法では，活字とはいえ印刷時に発生するノイズがあるためこうした類似文字を高精度に識別するのは困難と考えられる．従って，類似文字に関しては更に詳細な認識法の導入が必要と考えられる．

（2）少数サンプル字種に対する辞書の作成　　今回扱ったデーヴァナーガリ文字に限らず，過去においてあまり認識の対象とされずデータの蓄積が少ない文字種の認識を行おうとする場合，辞書作成のためのサン

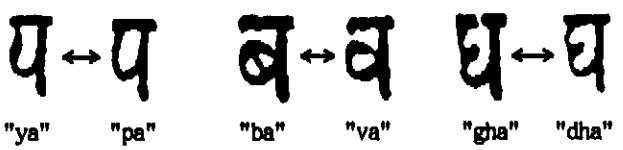


図9　類似文字の例
Fig. 9　Examples of similar characters.

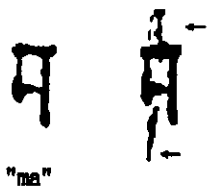


図10　ノイズの影響
Fig. 10　Influence of noise.

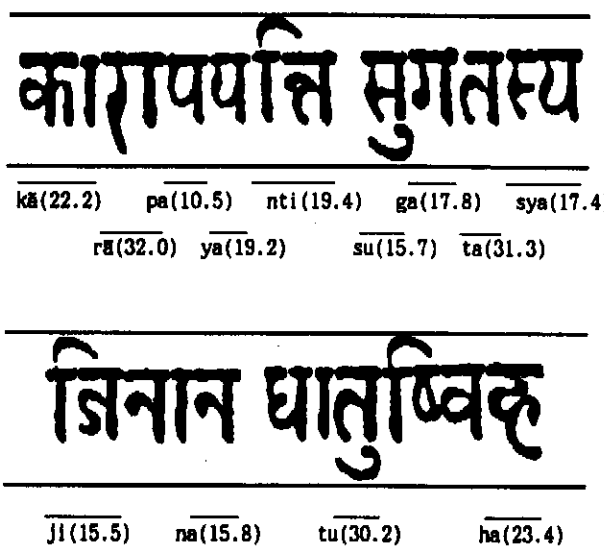


図11　認識結果出力例
Fig. 11　Results of character recognition.

ブルが十分ではないという問題は常に考えられる．この問題に対しては，周波数領域における伝達関数を利用して最適なフィルタを設計する手法を研究中である[9]．

6. むすび

活字デーヴァナーガリ文字認識のための文字の切出しと認識を同時に行う文字認識法を提案した．本手法を用いることにより，直接文字列から幅の異なる文字候補の抽出が可能となった．

また，法華経の活字印刷本中の89字種を認識対象とした認識実験においては99.4％という高い認識率が得られ，本手法の有効性が示された．

今後は，5.2の考察で述べたような問題点について更に検討を行う予定である．

文　献

(1) R. M. K. Sinha and H. N. Mahabala : "Machine recognition of devanagari script", IEEE Trans. Syst., Man. & Cybern., SMC-9, pp. 435-441 (1979).

(2) K. ジャヤンティ，鈴本昭浩，金井　浩，牧野正三，川添良幸，木村正行，城戸健一："An approach to Devanagari character recognition using outstanding features"，信学技報，PRU87-103 (1988).

(3) 辻直四郎："サンスクリット文法"，岩波文庫 (昭49).

(4) H. Kern and B. Nanjo : "Saddharmapuṇḍarīka", Bibliotheca Buddhia X (1912).

(5) 馬場口登，塚本正敏，相原恒博："手書き文字列における文字切り出しの基礎的考察"，信学論(D)，J68-D, 12, pp. 2123-2131 (昭60-12).

(6) 村瀬　洋，若原　徹，梅田三千雄："候補文字ラティス法による枠無し筆記文字列のオンライン認識"，信学論(D)，J68-D, 4, pp. 765-772 (昭60-04).

(7) 長尾　真："パターン情報処理"，コロナ社 (昭58).

(8) 鍋島一郎："動的計画法"，森北出版 (昭43).

(9) 鈴木昭浩，金井　浩，川添良幸，牧野正三，城戸健一："文字認識における少数サンプル辞書画像の推定―活字デーヴァナーガリ文字を例として"，信学技報，PRU88-84 (1988).

(平成元年2月3日受付，5月15日再受付)

鈴木　昭浩

昭62東北大・工・電気卒．平成元年，同大大学院修士課程了．文字の自動認識に関する研究に従事．現在，NTT勤務．

金井　浩

昭56東北大・工・通信卒．昭61同大大学院博士課程了．同年同大・情報処理教育センター助手．昭64同大・工・助手．現在に至る．音響・振動信号等のディジタル信号処理と機械系診断への応用に関する研究に従事．64年度石川賞受賞．IEEE，日本音響学会，日本機械学会各会員．工博．

川添　良幸

昭45東北大・理・物理第二学科卒．昭50同大大学院博士課程了．理博．同東北大学教養部助手を経て，昭56同大学情報処理教育センター助教授．その間昭56マックスプランク研究所員として西ドイツ在住，61年西オーストラリアWACAE客員教授．情報処理教育法，文字認識，並列計算機，原子核理論の研究に従事．日本物理学会，情報処理学会，日本印度学仏教学会等各会員．著書「コンピュータ概説」，「文科系のための計算機とプログラム」，「SASへの招待」(共著)など．

牧野　正三

昭44東北大・工・電子卒．昭49同大大学院博士課程了．同年同大電気通信研究所助手，昭56同大応用情報学研究センター助手，現在同所助教授．昭59～61米国STL客員研究員．言語情報を利用した音声認識の研究，音響信号処理，文字，画像信号処理の研究に従事．工博．日本音響学会，情報処理学会各会員．

城戸　健一

昭23東北大・工・電気卒．同大学電気通信研究所助手，同工学部助教授を経て，昭38同大電気通信研究所教授，昭51応用情報学研究センター教授，センター長，現在に至る．音響機器，建築音響，騒音制御，心理音響の研究から始まり，現在は音声自動認識，ディジタル信号処理，特にその音響工学への応用に関する研究に従事．著書「音響工学」(本会編，コロナ社)，「ディジタル信号処理入門」(丸善)，「電子計算機原論　上・下」(丸善)，「過渡現象論」(朝倉書店)等．工博．日本音響学会，電気学会，計測自動制御学会，情報処理学会，韓国音響学会，IEEE，AES等各会員．アメリカ音響学会フェロー．