

研究速報

デーヴァナーガリ文字における類似した文字の識別法

准員 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[†]

A New Discrimination Method for Similar Fonts of Devanagari Character Set

Akihiro SUZUKI[†], Associate Member, Hiroshi KANAIE^{††}, Yoshiyuki KAWAZOE^{††}, Shozo MAKINO[†] and Ken'iti KIDO[†], Members

[†] 東北大学応用情報学研究所センター, 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

あらまし 本論文では、正準判別分析を用いて類似した形状をもつ文字の識別に有効な特徴を自動的に抽出する手法を提案する。正準判別分析の適用における問題点を指摘し、その解決法を示す。極めて類似した形状をもつデーヴァナーガリ文字9字種についての認識実験で、99.9%の認識率を得て、本手法の有効性が示された。

1. まえがき

デーヴァナーガリ文字においては、図1に示すように非常に類似した文字(以下、類似文字)が存在する。現在まで発表されているデーヴァナーガリ文字の認識法^{(1),(2)}においては、構造解析的な手法によりこうした文字の識別を試みている。そこにおいては、前処理として細線化処理⁽³⁾を入力パターンに施すのが一般的である。しかし、本論文で扱っている活字デーヴァナーガリ文字のフォントには、図2に示す“pa”と“ya”のように、細線化後のパターンでは両者の識別が困難になるものもある。従って、構造解析的な手法だけで類似文字を完全に識別するのは困難であり、細線化していない文字のパターンそのものを比較して認識を行う必要がある。そのため、我々は、パターンマッチング手法を用いたデーヴァナーガリ文字の認識法の開発を行っている^{(4),(5)}が、これまでは類似文字の識別に関しては必ずしも良好な認識率は得られていなかった。そこで、本論文では、文字の識別に有効であると考えられる特徴を、正準判別分析を用いて自動的に抽出し、類似文字を高精度に識別するための手法を提案する。また、認識実験によりその有効性を示す。



図1 類似文字の例
 Fig. 1 Examples of similar fonts in printed Devanagari.



図2 “ya”と“pa”の細線化結果
 Fig. 2 Thinned patterns of “ya” and “pa”.

2. 正準判別分析を用いた類似文字の高精度な識別法

パターンマッチング手法により文字の識別を行う場合、単純なパターンの重ね合せでは、やはり類似文字の高精度な認識は難しい。例えば、図2の“pa”と“ya”において、二つの文字が共通にもつ右側縦棒および上側横棒の部分には、この2文字の識別に有効な情報はほとんどなく、むしろこれらの棒の輪郭線に出現する標準化雑音等が悪影響を与える。従って、それ以外の領域からできるだけ多くの文字の識別に有効な特徴を、統計的に抽出することが必要となる。

識別に有効な特徴を抽出する分析法として、正準判別分析^{(6),(7)}が知られている。この分析法では、 n 次元の特徴ベクトル $s(x)$ ($x=1, \dots, n$) に対して、次式で与えられる n 個の係数 $m(x)$ によって射影をとり、新たな特徴量 z を求める。

$$z = \sum_{x=1}^n m(x)s(x) \tag{1}$$

その際に、 z の群間平方和 S_B と群内平方和 S_W の比 λ

$$\lambda = \frac{S_B}{S_W} \tag{2}$$

を最大とするように、 $m(x)$ を定める。実際に $m(x)$ を求める計算は、次式で表される一般固有値問題を解くことに帰着する。

$$(B - \lambda W)m(x) = 0 \tag{3}$$

ここで、 B および W は $s(x)$ の群間平方和積和行列、群内平方和積和行列 (n 行 n 列)である。こうして求められた正準変量 z を用いることにより、各文字の特徴がより強調され、類似文字の認識が可能となる。

しかし、この手法を実際にデーヴァナーガリ文字の

認識に適用する場合、次の二つの問題点がある。

(1) ここで扱う文字パターンの特徴ベクトルの次元数 n は、 $96 \times 96 = 9216$ と非常に高く、しかも現状では文字サンプルを大量に収集するのが困難であるため、式(3)の群間平方和積和行列 B 、群内平方和積和行列 W が安定に推定できない。

(2) 文字の背景部や線の中央部の画素においては、常にその値が0または1となって、標準偏差すなわち式(2)の分母が0となるために、 λ が発散、または不定となる。

この問題点を解決するために、本論文では、次の二つの処理を導入する。

(a) 次元数の圧縮による近似計算 96×96 の画素を 8×8 ごとのマスクに区切り、そのマスク中の画素の値の平均値を新たに1画素とすることにより、次元を $12 \times 12 = 144$ に圧縮する。

(b) 標準偏差がしきい値以下の画素の除去 辞書作成用の全文字サンプルパターンから、各画素ごとの標準偏差を求め、それが0または非常に小さい値となる画素は特徴ベクトルとして意味をもたないと考え



図3 認識対象として用いたデーヴァナーガリの類似9字種
Fig. 3 9 similar fonts of Devanagari characters for recognition experiments.

て除去する。

3. 認識実験

本手法の有効性を確認するため、認識実験を行った。認識対象としたのは、図3に示す類似した9字種である。

図4に9字種について各字種20サンプル、計180サンプルにおける各画素の標準偏差を示す。この場合、上記(b)におけるしきい値=0として、すなわち標準偏差0となった画素は $12 \times 12 (=144)$ 画素のうち111画素であり、残り33画素により正準判別分析を行って求めた正準変量に関して、ユークリッド距離による認識を行う。

図5に、認識実験の結果を示す。正準変量を用いた認識においては、認識に用いる正準変量の個数を変えて実験を行った。また、比較のために、正準変量を用いずに $96 \times 96 (=9216)$ 次元の文字パターン特徴ベクトルをそのまま用いたユークリッド距離による認識も行った。認識結果には、認識率を示すと共に、次式で定義する入力文字と、その文字に対応する辞書との距離と、

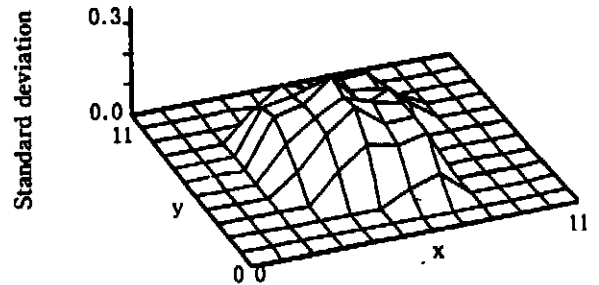


図4 9字種180サンプルにおける各画素の標準偏差
Fig. 4 Standard deviation of each pixel for the 180 samples of 9 Devanagari characters.

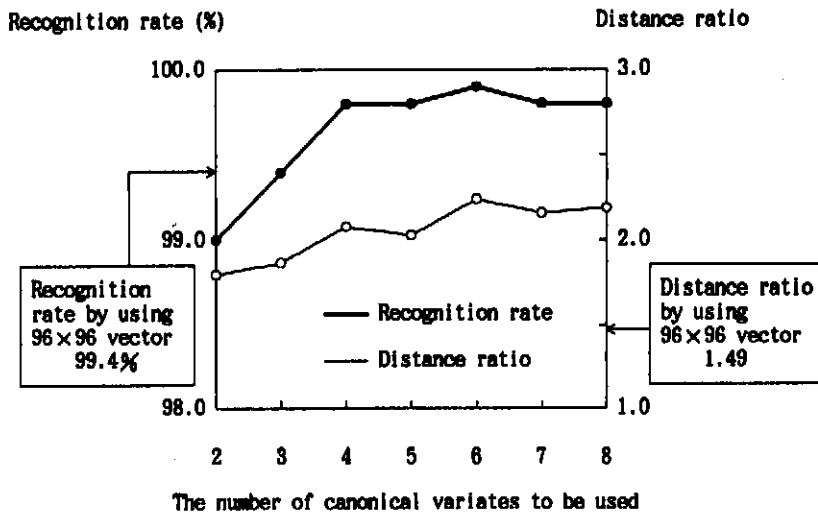
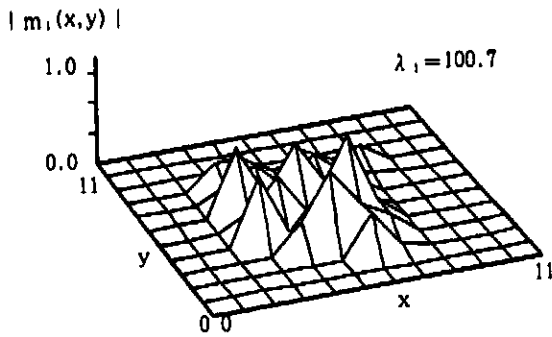
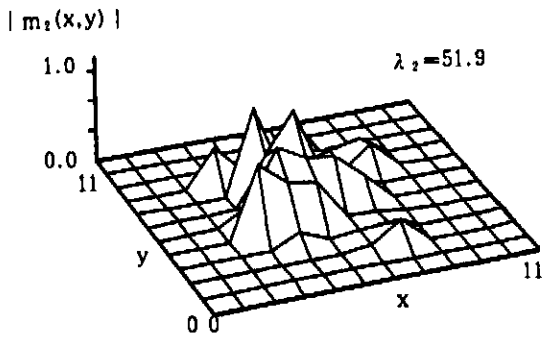


図5 認識実験結果(しきい値=0)
Fig. 5 Results of recognition experiments (threshold value=0).



(a) Coefficients of 1st canonical variates m_1



(b) Coefficients of 2nd canonical variates m_2

図6 正準変量の係数 m_1 と m_2

Fig. 6 Coefficients of canonical variates m_1 and m_2 .

それ以外の辞書との距離の最小値の比(距離比)を示した。

$$\text{距離比} = \frac{(\text{入力字種以外の辞書との距離の最小値})}{(\text{入力字種の辞書との距離})}$$

(4)

この距離比は、識別の分離度の良さを示す。実験結果より、正準変量を用いた認識においては、通常のユークリッド距離による認識よりも良い認識率が得られていることがわかる。また、第6正準変量までを用いた場合が最も高い認識率が得られていることから、第7、第8正準変量には文字の識別に有効な情報が少なく、ノイズや識別に有効でない成分が多く含まれていると考えられる。距離比についても同様の結果が得られており、本手法の有効性が確かめられている。また、図6に正準変量を構成する係数を示す。この係数を見ると、文字の横棒、縦棒の部分以外の文字の判別に有効な特徴をもつと思われる部分で大きな値をもっている様子がわかる。以上から、本手法によって雑音の影響を受けやすく、識別に寄与しない特徴が除去され、識別に有効な特徴が自動的に抽出されていることがわかる。

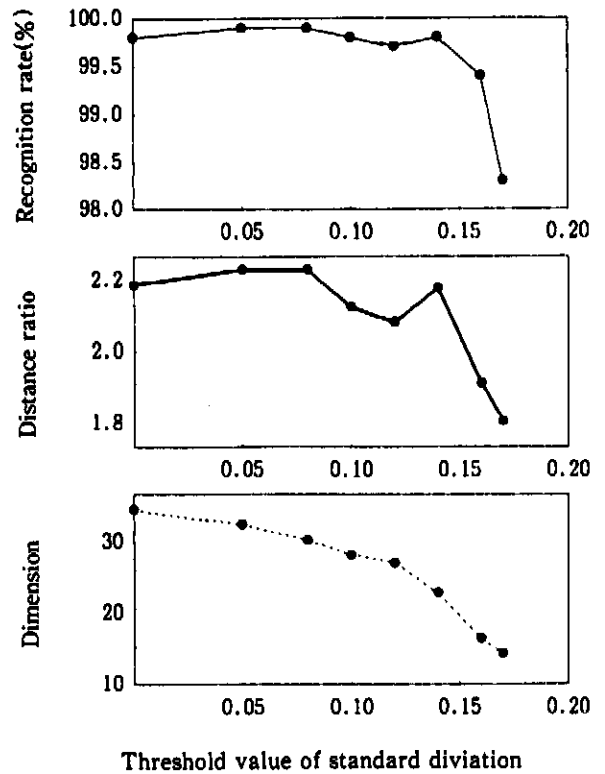


図7 標準偏差に対するしきい値と次元数、距離比および認識率の関係

Fig. 7 Relations between threshold value and results of recognition; recognition rate, distance ratio, and the dimension.

図7に標準偏差に対するしきい値を変化させ、第1から第6の正準変量を用いた認識実験を行った結果を示す。認識率および距離比を見るとしきい値が0.05から0.08のときに最も識別能力が高いことがわかる。

4. 結論

デーヴァナーガリ文字における非常に類似した文字の識別に有効な特徴を、正準判別分析によって自動的に抽出するための手法を提案した。従来の正準判別分析の手法における問題点として、特徴ベクトルの次元数の問題と、群内分散が0または非常に小さくなる軸が存在するときの計算における不安定性を挙げ、その解決法の一例を述べた。類似文字9字種の識別実験では、通常ユークリッド距離を用いたパターンマッチング法では、誤認識率が0.6%であったが、本手法を用いた認識では0.1%と大幅に低下し、その有効性が示された。

文 献

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

-
- (2) ジャヤンティ クリシナマチャリ, 鈴木昭浩, 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) C. M. Holt, A. Stewart, M. Clint and R. H. Perrott: "An improved parallel thinning algorithm", Commun. ACM, 30, 2, pp. 156-160 (1987).
- (4) 鈴木昭浩, 金井 浩, 牧野正三, 川添良幸, 城戸健一: "切り出しと認識を同時に行う活字デーヴァナーガリ文字の認識法", 信学技報, PRU88-81 (1988).
- (5) 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群書画像の推定—活字デーヴァナーガリ文字を例として", 信学技報, PRU88-84 (1988).
- (6) 奥野忠一, 久米 均, 芳賀敏郎, 古澤 正: "多変量解析法<改訂版>", 昭科技速 (昭56).
- (7) 奥野忠一, 芳賀敏郎, 矢島敏二, 奥野千恵子, 橋本茂司, 古河陽子: "統多変量解析法", 日科技速 (昭51).
(平成元年3月28日受付)
-