

パターン認識のための 特徴抽出法の評価に関する基礎的考察

瀬脇 康宏*・政水 克典†・浜本 義彦‡
金岡 泰保‡・富田 眞吾‡

A Fundamental Study on the Evaluation of a Feature Extraction Method for Pattern Recognition

Yasuhiro SEWAKI, Katsunori MASAMIZU, Yoshihiko HAMAMOTO
Taiho KANAOKA and Shingo TOMITA

Abstract

Feature extraction can be considered as a problem of finding a transformation that maps an n -dimensional pattern space down to a d -dimensional feature space without significantly increasing the degree of overlap between different class distributions. In a feature space, a classifier is generally designed. In order for the classifier to be reliable in predicting the future performance of a classifier, the feature space provided by a feature extraction method must have two properties. The first property is a statistical one that the pattern distribution in the feature space is a normal distribution. The second property is a topological one that the above transformation is a continuous function.

In this paper, we investigate the evaluation of a feature extraction method in terms of both of the statistical and topological properties using the real data set.

1. まえがき

一般にパターン認識系は、Fig1に示すような、観測系、前処理系、特徴抽出系、識別系の4つの処理系からモデル化されている [1].

外界に存在する認識対象は、観測系と前処理系を経てパターン空間の1点として記述される。特徴抽出系では識別に有効な特徴が抽出され、特徴空間の1点としてパターンが記述される。特徴空間におけるパターンは、特に特徴パターンと呼ばれる。次の識別系では特徴パターンが既知クラスのいずれかに対応づけられ

る。これらの処理系の中で、観測系、前処理系、特徴抽出系は、いずれも認識対象の性質に強く依存する [1]。このためパターン認識系の設計に関して一般論を展開することは困難となり、認識対象の性質を反映した個別的手法が数多く提案されている。このように一般パターン認識問題は認識対象の性質に基づいて解かれなければならない。このため従来の研究では、直観的に認識対象から特徴を取り、それら特徴の組で特徴空間を構成するというアプローチがとられてきた。直観的であるがゆえに、得られる特徴空間の性質は明らかにされていない。しかし、実際の認識問題を

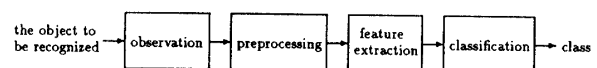


Fig. 1 Pattern recognition system

*日産自動車株式会社

†大学院知能情報システム工学専攻

‡知能情報システム工学科

考える上で、特徴空間には正規性及び位相性という二つの性質が必要とされる。

まず、正規性の必要性について述べる。多くの統計的識別理論は、パターン分布が正規分布であるという仮定の下に構築されている。もし、正規分布の仮定が満たされないと、そのような理論に基づいて設計された識別器の識別性能は著しく劣化する、という問題が生じる。このため、パターン認識においてこれまで得られた多くの成果が活用されるためには、パターン分布が正規分布であることが望ましい、と考えられる。また、正規性を仮定しないノンパラメトリック手法は、多数の訓練パターンを必要とするために、訓練パターン数に限りがある実際の認識問題に対しては、その識別性能がパラメトリックな手法のものに比べ高くない、という問題を含んでいる。それゆえ、特徴空間の正規性が問題となる。

次に位相性の必要性について述べる。一般に特徴パターンのなす統計的構造は未知であるため、学習が導入される。学習により未学習の特徴パターンに対しても認識が可能となるためには、特徴空間が認識対象間の類似の程度に応じた相対位置関係を保つような位相空間でなければならない、ということが飯島 [1] により指摘されている。すなわち、特徴空間に適当な位相性が存在するときのみ、パターン認識における学習が可能である。同様な指摘を上坂 [2] も行なっている。このように、学習可能性の問題が、特徴空間の位相性の問題として考えられている。しかし、実際には特徴空間について、その位相性の有無は明確に検討されていない。このため、学習が可能ではない状況にもかかわらず、パターン認識系を構築している場合がある。

前述したように、特徴抽出法は認識対象の性質に依存している。本研究は、手書き漢字から得られた二値文字を認識対象とし、特徴抽出法の評価を特徴空間の正規性及び位相性の観点から論じるものである。

まず第2章で特徴空間の正規性に関する評価法を提案する。次に第3章で特徴空間の位相性に関する評価法を提案する。第4章では、計算機シミュレーションを通して、提案された手法により特徴抽出法の評価を行う。

2. 正規性に基づく評価法

特徴パターンの分布が正規分布をなしているか否かという問題は、識別系の多くが正規分布の仮定の下に設計されるために、パターン認識において重要である。まず、2.1節で分布の歪度および尖度なる統計量 [3] [4] を述べ、これらを基に正規性の評価関数を2.2節

で定義する。

2.1 分布の歪度および尖度

特徴パターンの各特徴量に対し歪度と尖度を用いて、周辺分布の正規性から特徴空間の正規性を評価する。本章の議論はクラス毎になされるため、クラス名を表す添字は省いている。

いま、 n 次元特徴空間上に N 個の特徴パターンが分布しているとする。このとき第 i 番目の特徴パターン x_i は、以下のように表現できる。

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{in})^t \quad (1)$$

ここで、 x_{ij} は第 i 番目のパターンの第 j 番目の特徴量を表す。また、平均 \bar{x} は次のように求まる。

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_j, \dots, \bar{x}_n)^t \quad (2)$$

$$\text{ただし、} \bar{x}_j = \frac{1}{N} \sum_{i=1}^N x_{ij} \quad 1 \leq j \leq n$$

ここで第 j 番目の特徴量に関して、分布の非対称度を示す歪度 b_{1j} 、分布の裾の長さを示す尖度 b_{2j} はそれぞれ、

$$b_{1j} = \sqrt{N} \frac{\sum_{i=1}^N (x_{ij} - \bar{x}_j)^3}{\left\{ \sum_{i=1}^N (x_{ij} - \bar{x}_j)^2 \right\}^{\frac{3}{2}}} \quad (3)$$

$$b_{2j} = N \frac{\sum_{i=1}^N (x_{ij} - \bar{x}_j)^4}{\left\{ \sum_{i=1}^N (x_{ij} - \bar{x}_j)^2 \right\}^2} \quad (4)$$

で与えられる [3] [4]。この歪度 b_{1j} は $b_{1j} = 0$ のとき正規分布を、 $b_{1j} > 0$ のとき右裾広がり分布を、 $b_{1j} < 0$ のとき左裾広がり分布を示している。また、尖度 b_{2j} は $b_{2j} = 3$ のとき正規分布、 $b_{2j} > 3$ のときは t 分布などの裾広がり強い分布、 $b_{2j} < 3$ のときは裾が途切れた分布を表している。

2.2 正規性の評価関数

統計量 b_{1j} 、 b_{2j} を用いて、正規性の評価関数 $J(n)$ を以下のように定義する。

$$J(n) = \frac{1}{n} \sum_{j=1}^n \{ b_{1j}^2 + (b_{2j} - 3)^2 \} \quad (5)$$

の $J(n)$ の値が小さい程、特徴空間内の分布が正規分布に近い、と考えられる。

3. 位相性に基づく評価法

前述したように、学習が可能であるためには、特徴空間が、認識対象である二値文字間の類似の程度に応じた相対位置関係を保つような位相空間でなければならない。すなわち、類似している二値文字同士は、それらに対応する特徴パターンが特徴空間上で互いの近傍に存在しなければならない。もしそうでなければ、多数の訓練パターンを用いて学習を行なっても、将来入力されるパターンの認識は困難となる、と考えられるからである。

本研究では、まず原二値文字を連続的にかつ微小に変化させた二値文字を人工的に生成する。次に、それらの類似二値文字に対し、特徴抽出法により特徴抽出を行なう。得られた特徴空間上において、類似二値文字にそれぞれ対応する特徴パターン間の遠近関係を、階層的クラスタリングにより調べる。これにより、外界において互いに類似している二値文字同士が、特徴空間上においても互いの近傍に存在しているか否かを調べることができる。

基本的な準備として、3.1節で階層的クラスタリングを、また3.2節で類似二値文字を生成するための手法である線密度イコライゼーション法を述べる。これらの準備のもとに、3.3節で位相性の評価関数を定義する。

3.1 階層的クラスタリング [5]

特徴パターンを分類する際、似たもの同士を集めるために、似ている度合すなわち類似度、もしくは似ていない度合すなわち距離をどのように測るかを決めなければならない。本論文ではユークリッド距離を用いる。

3.1.1 クラスターの形成

次にクラスターの階層的形成法を述べる。いま、 N 個の特徴パターンに1から N までの番号を与えておく。そして最初に各特徴パターンをそれぞれ1つのクラスターと考え、この分類を C_0 とする。

$$C_0 = \{1, 2, \dots, N\}$$

次の段階で1と2、3と4と5がクラスターを形成したとする。この段階を C_1 とし

$$C_1 = \{1-2, 3-4-5, 6, 7, \dots, N\}$$

とする。ここで1-2とか3-4-5をそれぞれ分類における要素と呼ぶことにする。こうして、 m 回の結合を得

て $C_0, C_1, C_2, \dots, C_m$ という分類系列が得られるのであるが、階層的というのは C_{i-1} から C_i が C_{i-1} の要素によってできている場合である。

この階層的な手法のうち次の場合を考える。まず N 個の特徴パターンを N 個のクラスターと考える。次に $d_{\alpha\beta}$ をクラスター α とクラスター β の距離を定義する。いま、すべての α, β について最小の $d_{\alpha\beta}$ を d_{hi} とすると、この最小の距離をもつクラスター h とクラスター l の2つを融合する。このような操作を $N-2$ 回繰り返せば、最後に2つのクラスターが残る。

クラスター h とクラスター l を結合してクラスター g ができるときに、クラスター g に属さないクラスター f とクラスター g との距離 d_{fg} が結合する前の距離 d_{fh}, d_{fl}, d_{hl} のみから得られれば、このクラスター分析を組合せ的手法と定義する (Fig2参照)。

距離としてユークリッド距離を用いる場合 d_{fg} は、次式のように表される。

$$d_{fg} = \alpha_h d_{fh} + \alpha_l d_{fl} + \beta d_{hl} + \gamma |d_{fh} - d_{fl}| \quad (6)$$

ただし、 $\alpha_h, \alpha_l, \beta, \gamma$ は定数である。

以下、式(6)の定数の異なる3種類の手法を述べる。

(a) 最短距離法

$$\alpha_h = \alpha_l = \frac{1}{2}, \quad \beta = 0, \quad \gamma = -\frac{1}{2}$$

(b) 最長距離法

$$\alpha_h = \alpha_l = \frac{1}{2}, \quad \beta = 0, \quad \gamma = \frac{1}{2}$$

(c) メジアン法

$$\alpha_h = \alpha_l = \frac{1}{2}, \quad \beta = -\frac{1}{4}, \quad \gamma = 0$$

3.1.2 デンドログラム

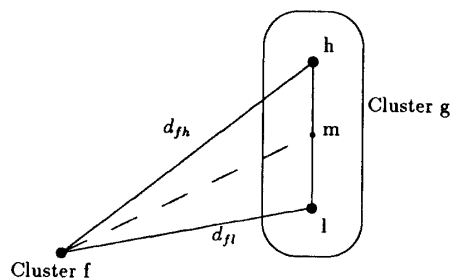


Fig. 2 Generation of cluster

デンドログラム [6] は、形成されたクラスター間の遠近関係を図に表したものである。デンドログラムの一例を Fig3 に示す。

ユークリッド距離を縦軸にとると、クラスター間の距離は図中でクラスターとクラスターが継っている位置を見ることで分かる。

3.2 線密度イコライゼーション法による類似二値文字作成

本節では、手書き漢字文字の類似二値文字を作成するための一手法である線密度イコライゼーション法 [7] を説明する。

3.2.1 標本点の算出

未知文字を、一定の標本化間隔 δ で標本化された二値文字を

$$f(x_i, y_j) \tag{7}$$

とする。ただし $i=1, 2, \dots, I, j=1, 2, \dots, J$ 。ここで、 $x=x_i$ の垂直線と二値文字 f との交差数を

$$h_X(x_i) = \sum_{j=1}^J f(x_i, y_j) \cdot \bar{f}(x_i, y_{j-1}) + \alpha \tag{8}$$

同様に $y=y_j$ の垂直線と二値文字 f との交差数を

$$h_Y(y_j) = \sum_{i=1}^I f(x_i, y_j) \cdot \bar{f}(x_{i-1}, y_j) + \alpha \tag{9}$$

とする。 α は、パラメータ、である。次に、それぞれの線の総数を

$$N_x = \sum_{i=1}^I h_X(x_i) \tag{10}$$

$$N_y = \sum_{j=1}^J h_Y(y_j) \tag{11}$$

とする。ここで標本化間隔と線密度が一定であると言

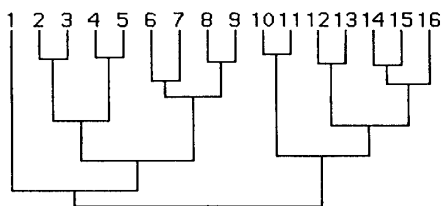


Fig. 3 Dendrogram

うことは、

$$\epsilon_X(i) \cdot h_X(x'_i) = constant \tag{12}$$

$$\epsilon_Y(j) \cdot h_Y(y'_j) = constant \tag{13}$$

であることを意味する。なお、 $\epsilon_X(i), \epsilon_Y(j)$ は、新しい標本点の間隔で、 $h_X(x'_i), h_Y(y'_j)$ は、その間隔での f との交差数である。また、変形後のメッシュ数も変形前と同じ ($I \times J$) とすると、上式の定数は、それぞれ $\delta \cdot N_X / I, \delta \cdot N_Y / J$ となる。

以上の条件により、新しい標本化文字のための標本点 (x'_i, y'_j) は、

$$\begin{aligned} x'_i &= \left\{ x \mid H_X(x) = i \cdot \frac{N_X}{I} \right\} \\ y'_j &= \left\{ y \mid H_Y(y) = j \cdot \frac{N_Y}{J} \right\} \end{aligned} \tag{14}$$

から求まる。ただし、

$$\begin{aligned} H_X(x) &= \int_0^x \frac{h_X(x')}{\delta} dx' \\ H_Y(y) &= \int_0^y \frac{h_Y(y')}{\delta} dy' \end{aligned} \tag{15}$$

であり、 $H_X(x)$ は、変形前の図形の 0 から x までの f との交差数の和である。

また式 (14) の x, y を離散的な値をとらせたい場合には、

$$\begin{aligned} i' &= \min \left\{ l \mid \sum_{k=1}^l h_X(x_k) \geq (i-0.5) \cdot \frac{N_X}{I} \right\} \\ j' &= \min \left\{ m \mid \sum_{k=1}^m h_Y(y_k) \geq (j-0.5) \cdot \frac{N_Y}{J} \right\} \end{aligned} \tag{16}$$

なる i', j' に対し $x'_i = x_{i'}, y'_j = y_{j'}$ として変換すればよい。Fig4 に生成過程を示す。

3.2.2 パラメータ α の意義

例えば「い」という文字の場合、縦方向の中央部分は何もないので、 x_i における X 軸と二値文字の交差数がゼロになるところができる。このためその標本化間隔が無限大になり、変形後は中央部分の空白は無くなり左右の部分が引っついてしまう。これを防ぐためにあらかじめ交差数 h にある程度値 α を与えておくこと変形後も何も存在しない部分に空間を保つことがで

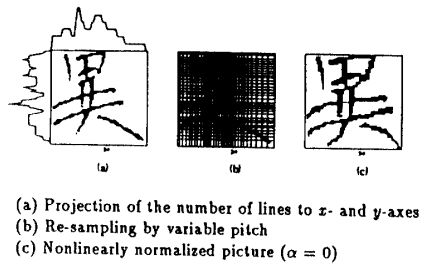


Fig. 4 Illustration of Line density equalization method



Fig. 5 Example of similar binary patterns

きる。また $\alpha \rightarrow \infty$ にすると標本化間隔は一定となる。このことから α の値の大きさを変えることで、変形強度を変化させることが可能となり、様々な類似二値文字を生成することができる。

3.3 位相性の評価関数

前節の線密度イコライゼーション法で原二値文字を、連続的にかつ微小に変形させることにより類似二値文字を生成する。生成された類似二値文字からなる類似二値文字系列上で、隣接する二値文字同士は、主観的に類似しているように α の値を変えて生成されている。これらの類似二値文字に対して特徴抽出を行なう。もし特徴空間上においても位相性が保たれているならば、特徴パターンの遠近関係をデンドログラムに表した場合、類似二値文字系列上で隣接する類似二値文字にそれぞれ対応する特徴パターンは、特徴空間上でクラスターを形成するはずである。このとき、位相ずれが生じていないと呼ぶ。そのようになっていないならば、位相のずれが生じていることになる。このことを図を用いて説明する。いま、Fig5に示すような類似二値文字に対し、それぞれ特徴抽出を行なうことで複数個の特徴パターンが得られる。これらの特徴パターンに対し特徴空間上で階層的クラスタリングを行なう。これにより得られた遠近関係をFig6に表す。図中の番号は類似二値文字の識別番号で、その値が小さい程、原二値文字に似ている類似二値文字であることを意味する。ここで、Fig6は、位相ずれが生じていない例を、またFig7は位相ずれが生じている例を示す。

以上の準備の下、次の評価関数 $\rho(n)$ を定義する。

$$\rho(n) = \frac{\text{位相ずれが生じた類似二値文字系列数}}{\text{類似二値文字系列の総数}} \times 100 (\%) \quad (17)$$

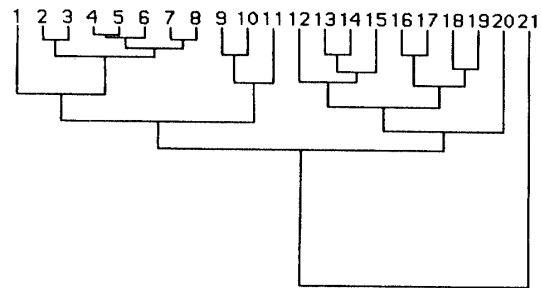


Fig. 6 Example of the dendrogram without the topological gap

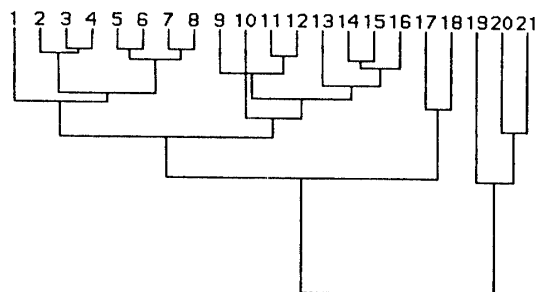


Fig. 7 Example of the dendrogram with the topological gap

この $\rho(n)$ の値が小さい程、特徴空間は位相性を有していると考えられる。

4. 計算機シミュレーション

本章では、提案された二つの評価法により特徴抽出法を評価する。4.1節で実験について述べ、4.2節では得られた実験結果を示す。最後に4.3節で検討を述べる。

4.1 実験データ

今回実験に使用したデータは、電子技術総合研究所で作成された手書き教育漢字データベースETL-8である。ETL-8の手書き漢字データは、64×63の二値文字であり、この二値文字に対し、代表的な特徴抽出法である外郭方向寄与度法 [8]、加重方向指数ヒストグラム法 [9] によりそれぞれ特徴を抽出し、二種類の特徴空間を構成した。評価の対象となった外郭方向寄与度法、加重方向指数ヒストグラム法は、いずれも手書き漢字のための有効な特徴抽出法としてこれまで広く用いられているものである。

4.2 シミュレーション結果

まず、4.2.1節で正規性の観点からの評価結果を示し、次に、4.2.2節で位相性の観点からの評価結果を示す。

4.2.1 正規性の観点からの評価結果

Table1の10クラス(各クラス160文字)に対して、外郭方向寄与度法と加重方向指数ヒストグラム法によりそれぞれ特徴空間を構成した。各特徴抽出法のパラメータを変えることで、Table2に示す特徴数を有する特徴空間が得られる。

得られた特徴パターンに対し、式(3)、(4)により b_{1j} 、 b_{2j} を求め、評価関数 $J(n)$ の値を計算した。結果をTable3, 4に示す。参考までに、二つの方法で字種“愛”を1024次元で特徴抽出したときの b_{1j} 、 b_{2j} ($j=1, 2, \dots, 1024$)の分布をFig8, 9に示す。

4.2.2 位相性の観点からの評価結果

ETL-8の全881クラスの中の59クラスから、1クラスあたり一つの特徴文字を任意に選び出した。原二値文字毎に線密度イコライゼーション法により、 α の値を徐々に減少させ類似二値文字を20個生成し、原二値文字(ETL-8のデータ)一つと類似二値文字20個の計21個の二値文字からなる類似二値文字系列を得た。各クラスより一つの特徴文字を選び、類似二値文字系列の総数を59とした。類似二値文字生成の各種パラメータの詳細については、文献[10]を参照されたい。

以下の処理は、類似二値文字系列毎になされる。系列内の21個の類似二値文字に対し、外郭方向寄与度法、加重方向指数ヒストグラム法でそれぞれ特徴抽出した後、得られた21個の特徴パターンに対し最短距離法、最長距離法、メジアン法により階層的クラスタリングを行なった。この処理を59個の類似二値文字系列全てについて行なった。次に、デンドログラムにより位相ずれが生じている類似二値文字系列数を調べた

Table 1 Examples of characters

characters				
愛	悪	暗	案	以
困	委	医	因	飲

Table 2 Feature extraction methods to be evaluated and the number of features

Feature extraction methods	Number of features				
PDC	512,	1024,	1536,	2048,	2560
WDIH	64,	100,	324,	484,	1024

PDC : Peripheral direction contributivity method
WDIH : Weighted direction index histogram method

Table 3 Normality for PDC (values of $J(n)$)

Class	Number of features				
	512	1024	1536	2048	2560
愛	5.99	11.44	9.18	9.17	8.64
悪	5.50	4.01	17.46	14.72	20.70
暗	3.49	8.99	7.01	9.44	9.08
案	13.14	8.62	6.85	6.16	6.38
以	8.09	14.15	13.03	13.38	13.91
困	12.92	10.65	8.75	7.82	11.36
委	7.21	5.48	5.76	5.25	8.00
医	6.18	6.42	10.84	10.24	11.68
因	7.76	6.45	8.02	8.49	7.73
飲	4.79	4.41	4.38	4.23	4.40
Mean	7.51	8.07	9.13	8.90	10.19
Deviation	9.30	9.75	13.26	9.99	19.04

Table 4 Normality for WDIH (values of $J(n)$)

Class	Number of features				
	64	100	324	484	1024
愛	22.9	88.4	1441.6	1792.2	3033.4
悪	4.3	14.1	652.7	547.0	1000.1
暗	2.7	10.9	666.5	989.5	1252.2
案	7.6	15.5	929.6	1291.9	2386.6
以	53.9	80.9	1108.5	1194.9	2123.8
困	3.1	18.7	190.4	281.4	656.4
委	11.7	116.4	1664.5	1487.5	2275.7
医	25.2	38.5	546.5	683.5	1374.8
因	3.7	18.3	281.6	424.9	901.9
飲	12.6	20.4	504.1	918.2	1637.5
Mean	14.8	42.2	798.6	961.1	1664.2
Deviation	2.28×10^3	1.32×10^4	2.10×10^6	2.13×10^6	5.27×10^6

(Table5, Table6). さらに、3種類のいずれの方法に対しても位相ずれが生じた数をTable7(a), (b)にそ

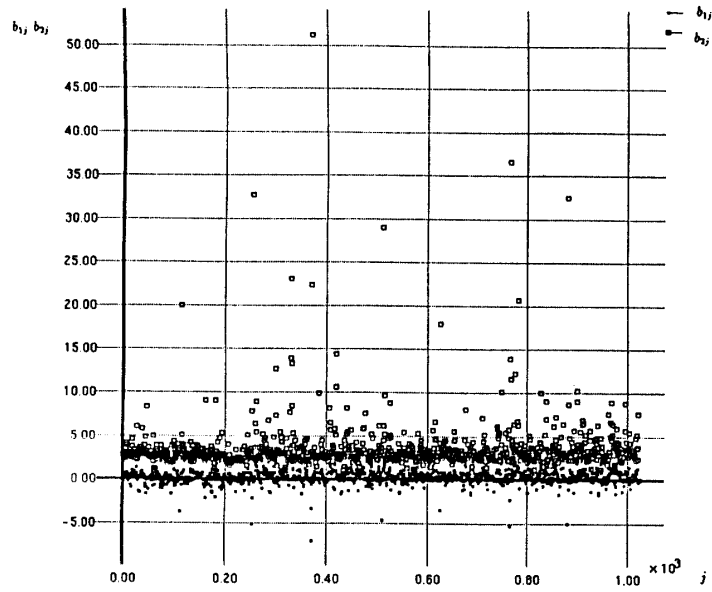


Fig. 8 Distribution of $b_{1,j}$ and $b_{2,j}$ for PDC

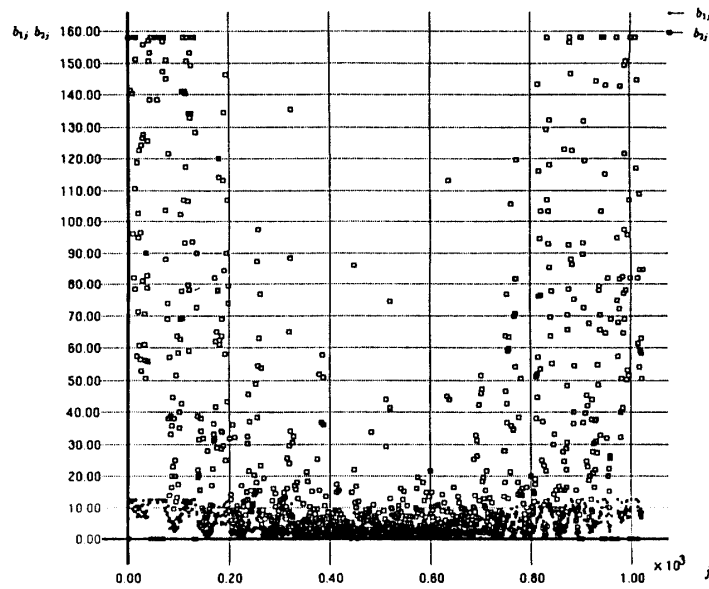


Fig. 9 Distribution of $b_{1,j}$ and $b_{2,j}$ for WDIH

Table 5 Topological gap for PDC

Number of features	512	1024	1536	2048	2560
Nearest neighbour method	13	6	8	8	5
Furthest neighbour method	13	6	8	8	4
Median method	33	31	29	25	25

Table 6 Topological gap for WDIH

Number of features	16	64	100	196	324	484	1024
Nearest neighbour method	37	34	26	13	11	6	3
Furthest neighbour method	37	35	26	13	10	6	3
Median method	44	44	35	23	26	21	9

Table 7 (a) Topological gap for PDC

Number of features	512	1024	1536	2048	2560
Number of sequences with the topological gap	12	6	8	8	4

Table 7 (b) Topological gap for WDIH

Number of features	16	64	100	196	324	484	1024
Number of sequences with the topological gap	34	34	25	13	10	5	3

れぞれ示す。これは特定のクラスタリング手法（類似度）に関係なく、位相ずれを表しているものと解釈される。

次に3種類すべての手法で位相ずれが生じた文字列に対して、式(17)により $\rho(n)$ を求めた。結果をTable8, 9に示す。

4.3 検 討

特徴空間の正規性に関しては4.2.1節の結果を見ると、いずれの特徴抽出法も特徴数が増えていくにつれ、 $J(n)$ の値が増えていく傾向にある。これは、特徴数を増加していくと、特徴空間の正規性が崩れていくことを示している。また、外郭方向寄与度法と加重方向指数ヒストグラム法を比較してみると、外郭方向寄与度法の $J(n)$ の値が特徴数の増加に対してほとんど変化しないのに対して、加重方向指数ヒストグラム法の $J(n)$ の値は、特徴数の増加に伴って増加していく。この結果から、高次元空間では外郭方向寄与度法より加重方向指数ヒストグラム法の方が、特徴空間の正規性を欠いていることがわかる。

次に特徴空間の位相性に関して4.2.2節の結果を見ると、いずれの特徴抽出法も特徴数が増えていくにつれ、位相ずれが生じた類似二値文字系列数が減少している。これは、特徴数を増加させることによって、よりきめの細かい特徴が得られ、それにより外界における二値文字の微小な変形に特徴パターンが対応することができるからである、と考えられる。Table7から加重方向指数ヒストグラム法の方が、外郭方向寄与度法よりも、特徴数の増加による位相性の改善が大きいことがわかる。

5. むすび

本論文では、特徴空間の正規性及び位相性の観点から特徴抽出法の評価法を提案した。

まず、正規性については、特徴パターンに対して歪

Table 8 Topological property for PDC

Number of features	512	1024	1536	2048	2560
$\rho(n)$ (%)	20.0	10.2	13.6	13.6	6.8

Table 9 Topological property for WDIH

Number of features	16	64	100	196	324	484	1024
$\rho(n)$ (%)	57.6	57.6	42.2	22.0	16.9	8.5	5.1

度、尖度なる統計量を用いて周辺分布の正規性から特徴空間の正規性を評価した。その結果、加重方向指数ヒストグラム法は外郭方向寄与度法に比べ特徴数を増やしていくと正規性の失われる程度が大きいことがわかった。

また、位相性については、原二値文字を微小にかつ連続的に変形させた類似二値文字を生成し、それらの類似二値文字に対し特徴抽出法により得られた特徴パターンの遠近関係を、階層的クラスタリングにより調べた。実験結果から、加重方向指数ヒストグラム法、外郭方向寄与度法、いずれの特徴抽出法も特徴数を増加させることによって位相性の改善が見られた。

今後の課題としては、まず周辺分布の正規性と同時分布の正規性との関連に関する問題がある。統計論によれば、周辺分布が正規分布でも同時分布が必ずしも正規分布であるとは限らない、ということが知られている。実際の立場では、このことがほぼ満たされるといふ報告[4]がある。しかし、さらに検討が必要である。

次に、位相性については二値文字変形のアルゴリズムの評価がある。今回二値文字変形の基準となったのは、人間が見て主観的に似ている文字でかつ微小に変化している、というものであった。しかし本論文で用いた手法が最適な二値文字変形手法か否かの検証はなされていない。これについては課題として残されている。

謝 辞

本研究に際し、御協力頂いた本講座の渡辺政典技官に深く感謝致します。また、実験に用いた手書き教育漢字データベースETL-8を作成、提供して頂いた電子技術総合研究所の関係諸氏に深く感謝致します。

参考文献

- 1) 飯島泰蔵：“パターン認識理論”，森北出版(1989)

- 2) 上坂吉則：“パターン認識と学習の理論”，総合図書 (1971)
- 3) 柴田義貞：“正規分布特性と応用”，東京大学出版会 (1981)
- 4) 若林哲史，鶴岡信治，木村文隆，三宅康二：“文字認識における特徴量の次元数と変数変換に関する考察”，信学技報，PRU92-115，pp.99-106 (1993-01)
- 5) 奥野忠一，芳賀敏郎，矢島敬二，奥野千恵子，橋本茂司，古河陽子：“統多変量解析法”，日科技連 (1976)
- 6) “多変量統計解析法見方・考え方”，数学セミナー (1990)
- 7) 山田博三，齊藤泰一，山本和彦：“線密度イコライゼーション-相関法のための非線形正規化法-”，信学論 (D)，J67-D，11，pp.1379-1383 (1984-11)
- 8) 萩田紀博，内藤誠一郎，増田功：“外郭方向寄与度特徴による手書き漢字の識別”，信学論 (D)，J66-D，10，pp.1185-1192 (1983-10)
- 9) 鶴岡信治，栗田昌徳，原田智夫，木村文隆，三宅康二：“加重方向指数ヒストグラム法による手書き漢字・ひらがな認識”，信学論 (D)，J70-D，7，pp.1390-1397 (1987-07)
- 10) 瀬脇康宏：“パターン認識のための特徴抽出法の評価に関する研究”，山口大学大学院知能情報システム工学専攻修士論文 (1993)
(平成5年4月15日受理)