

多次元特徴分布の可視化を実現した 教育用手書き数字認識システム

山本 健寛^{1,†} 竹下 鉄夫^{2,†} 江崎 信行^{3,†}

豊田工業高等専門学校 情報工学科[†]

1. はじめに

近年、文字認識や音声認識といったパターン認識システムが様々な分野に応用されている。これらのパターン認識システムはどれも、測定・採集した多次元の特徴量が大量に分布する特徴空間を演算処理に用いる。その特徴量の次元も、100次元や200次元といった非常に大きなものであることが多い。そのため、それら特徴量の分布や差異を具体的に思い浮かべることは難しく、このことがパターン認識を学習するうえでの妨げになっていると考えられる。そこで本研究では、先行研究[1]に倣い、パターン認識システムで扱う多次元ベクトルに対して多変量解析法の一つである主成分分析を行い、3次元に情報を集約することで、認識対象の特徴量の3次元空間上への可視化を試みる。そして、学習者が多次元パターンの分布の様子や広がりを感じやすく視認できるようなシステムの構築を目指す。またパターン認識システムの一例として、現在広く普及している手書き数字認識システムを取り上げ、研究を行う。

2. 手書き数字認識システム

2.1. システムの概要

本研究では、これまで手書き文字認識手法に対する学習支援システムの構築・評価を行ってきた。加重方向指数ヒストグラム法による文字認識画面を図1に示す。図1の4×4のマスキの中に縦・横・斜めの4つの半直線によって表現されたものが64次元の加重方向指数ヒストグラムを示している。このような多次元の特徴量の成分を一つ一つ比較するだけでも困難である上に、文字認識においては大量の辞書データを持つヒストグラムとユーザ入力を全て比較する必要がある。そこで、「0」から「9」の各数字のサンプルの集合(辞書データ)に対するユーザが入力した数字の類似度を直感的にわかりやすく視認できるように可視化する手法を考える。

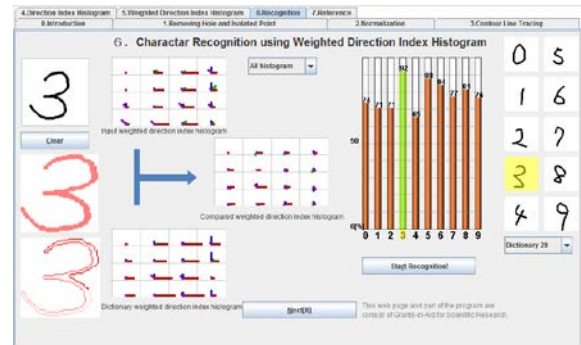


図1. 教育用手書き数字認識システム

2.2. 先行研究

先行研究[2]において、ヒストグラムに対し主成分分析を行い、得られた主成分スコアのうちの第3主成分までを抽出、3次元空間上に散布図としてプロットしたものを図2に示す。図2の画面は「3」の手書き数字の特徴空間を表示しており、赤い点が「3」の特徴分布、黄色い点が「2」の(「3」の空間上での)特徴分布である。また紫色の点がユーザ入力した文字を示している。これを見ればわかるように、同じ群に属するデータは空間上のある場所に固まって分布している。しかし、このようにただ点をプロットしただけでは、「2」と「3」の群の境界を視認しづらい。さらに、このシステムで同時にプロットできる手書き数字の分布は最大で2つであり、複数の群に属するデータの比較を行うことができない。また、すべての辞書データの主成分を座標として読み込む必要があるため、辞書データが大きくなると処理に時間がかかり動作しなくなってしまう。パターン認識の教育用システムとしてはこれらの問題点を改善する必要がある。

2.3. システムの改善

既存のシステムでは特徴量を散布図として表現していたため、それぞれの特徴量の比較はできても、前述したような問題が発生してしまった。そこで、各特徴量の分布の可視化を「点」ではなく、群ごとの等確率楕円体の表示によって試みる。これにより、ユーザが入力した手書き数字が全ての数字の群のうちどれに最も近い

Implementation of Visualizing Multidimensional Distribution for Educational Character Recognition System

1.Takehiro Yamamoto 2.Tetsuo Takeshita 3.Nobuyuki Esaki

†. Dept. of Information and Computer Engineering,

Toyota National College of Technology

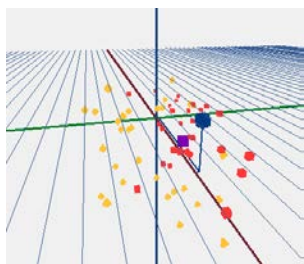


図 2. 先行研究 (主成分の散布図)

かをより視覚的にわかりやすく表現する。

3. 成果

産業技術総合研究所より提供されている ETL6[3]を用いて 1 つの群につき 1383 個のサンプルを用意し、それぞれの加重方向指数ヒストグラムを取得、これを特徴量として用いた。この特徴量から数字ごとに主成分分析を行い、特徴空間を得る。それらに対して全ての特徴量を射影し、第 1~3 成分のみを抽出することで、「0」から「9」の計 10 個の特徴空間に対する各数字の 3 次元分布を得た(これをそのまま 3 次元空間にプロットしたものが図 2 である)。新たに得られた 3 次元の特徴量について、各成分の度数分布を調べたところ、図 3 のようなヒストグラムとなり、これは正規分布に近いことがわかる(図は「0」の空間上の「0」の第一成分)。特徴量は多次元正規分布に基づくと仮定すると、楕円体の表面の軌跡は次の二次形式で表される。

$$(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (1)$$

ここで、 \mathbf{x} は楕円面上の任意のベクトル、 $\boldsymbol{\mu}$ は特徴量の平均ベクトルである。また、 $\boldsymbol{\Sigma}$ は 3 次元特徴量の分散共分散行列

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \quad (2)$$

である。式(1)を数字ごとに求め、特徴空間上に表示させた。これを図 4 に示す。図 4 は「0」の特徴空間における各数字の 95% 信頼楕円体を表示している。この図では全ての群の楕円体を表示しているが、ある特定の数字に限定して楕円体を表示し、異なる群間の形状を理解することもできる。図 4 の画面はマウスのドラッグによってユーザが自由に視点の回転や拡大・縮小を行うことができる。

4. 考察と今後の課題

等確率楕円体を 3 次元空間上に表示させる可視化方法により、異なる数字間の境界をはつき

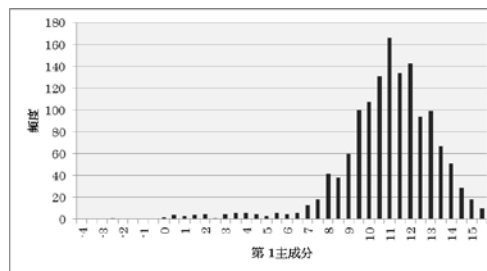


図 3. 「0」の第 1 主成分スコアヒストグラム

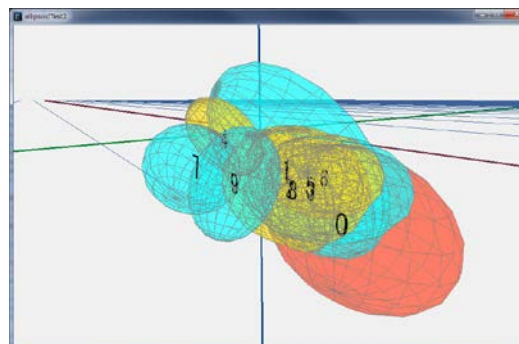


図 4. 等確率楕円体の表示

りと視認することができるようになった。さらに、各楕円体の描画自体に必要な値は式(1)を満たす行列 $\boldsymbol{\Sigma}^{-1}$ 及びベクトル $\boldsymbol{\mu}$ のみであるため、大量の辞書情報を反映しつつ全ての群を同時に可視化することに成功した。また、特徴量の分布を立体として表示することで、学習者にとって線形代数とパターン認識の橋渡しになるとも考えられる。

今後の課題としては次のようなことがある。

- 1) 多次元空間上の点からの文字の生成する逆問題の考察。
- 2) 2 つの群の境界付近の字形を提示し、文字の認識率をより高めるための知見を得る。
- 3) 高い認識率を競うゲーム性を考慮した学習プログラムの開発。

以上 3 つの課題を解消することで、パターン認識に対する理解度をより促進することのできるシステムとなる。

本研究の経費の一部は、科研費(23501194)によるものである。

参考文献

- [1] 山本康高, 吉川大弘, 古橋武: "判別分析を基準とするファジークラスタリングによる多次元データの可視化手法の提案", 電子情報通信学会論文誌, vol. J88-D-2, n. 6, p. 975-984 (2005)
- [2] 久田隆太郎, 竹下鉄夫, 江崎信行: "多次元パターン分布の可視化による理解促進教育用文字認識システム", 情報処理学会全国大会講演論文集, vol. 73rd, n. 4, 2011, p. 4487-4488 (2011)
- [3] 産業技術総合研究所: ETL 文字データベース (1984)