カーネル PCA の手書き平仮名識別への応用

坂井佳帆　(指導教員：吉田裕亮)

1　はじめに

パターン認識とは, いくつかの概念に分類できる観測データが存在する時, 観測されたデータをそれらの概念のうちの一つに対応させることである.

パターン認識において, 正規化された情報から特徴抽出を行う際に何が重要かは明示的には分かりづらい. そこで PCA(主成分分析) などの統計的な情報圧縮手法を用いる. それにより情報を次元圧縮された特徴ベクトルへと変換して認識に必要な情報を抽出することが可能となる.

本研究では, 手書き平仮名文字において, カーネル PCA を用いて, 識別器を構成し, 書き手の識別の可能性についての研究を行った.

2　先行研究

研究 [2] においては, 描かれた浮世絵の役者の顔の形状の顔の部位間の角度情報を用いて PCA(線形) を行い, 9 人の絵師の識別を行っている.

角度情報を用いることで, スケーリングフリーな識別を行うことができるという利点を応用して, 本研究では, 文字の特徴点間の角度情報を用いた人の識別を行う.

3　カーネル PCA

3.1　PCA(主成分分析)

PCA とは, 分散の大きい方向にデータを射影することで, 多次元データの情報を, 特性を保ちながらより低い次元に縮約させる方法である.

しかしスケーリングフリーな識別ができる一方, 線形データ解析手法のため, 非線形なデータの構造が捉えにくいという欠点がある.

3.2　カーネル法

一般にカーネル法では, 非線形変換を介して x のいろいろな特徴量を取り出している.

$\phi_1, ..., \phi_d$ という非線形関数で特徴抽出された特徴ベクトルを $\vec{\phi}(x) = (\phi_1(x), ..., \phi_d(x))^T$ と書く. すると, カーネル関数は特徴抽出の内積に基づき, 以下のように定義できる.

$$k(x, x') = \vec{\phi}(x)^T \vec{\phi}(x') = \sum_{m=1}^{d} \phi_m(x)\phi_m(x')$$

したがって, 非線形に写像した空間での $\vec{\phi}(x)$ と $\vec{\phi}(x')$ の内積が, 入力特徴 x と x' のみで計算でき, $k(x, x')$ から最適な非線形写像を構成することができる. このような関数 k をカーネルと呼び, このように高次元に写像しながらカーネルの計算のみで最適な識別関数を構成することを, 一般に, カーネルトリックという.

3.3　カーネル PCA

カーネル PCA とは, 高次元の特徴ベクトルに変換してから, 通常の PCA を行い, 低次元の線形部分空間を求める多変量解析手法である. 以下がアルゴリズムである.

1) 中心化されたデータ点の集合 $x^{(1)}, x^{(2)}, ..., x^{(n)}$ から, グラム行列 $K = \left(k(x^{(i)}, x^{(j)})\right)$ を作る.

2) $K\alpha = \lambda\alpha$ という固有値問題を解く.

3) 上から M 個の固有値 $\lambda_1, \ldots, \lambda_M$, 固有ベクトル $\alpha_1, \ldots, \alpha_M$ を用いて, M 次元 PCA プロットを行う.
本研究においては, M を 2 として, 第 2 主成分まで用いて 2 次元 PCA プロットを行なった.

4　提案手法

本研究では, カーネル PCA を用いて, 手書き平仮名文字の識別を以下のように行うことを提案する.

1) 手書き平仮名文字の書かれた紙をスキャン.

2) 各文字の特徴点 (9 or 10 点), それぞれの座標を取得.

3) 文字ごとに座標のセンタリングを行い, そのデータにカーネル PCA を施す.

4) 第 1,2 主成分を用いた主成分プロットを行い, 識別可能性を調べる.

5　文字データ

実データとして, 6 人分のひらがな文字「す」「そ」「み」「さ」の文字データを用いる.
データの集計は 4 日間に分けて行い, 1 日に各 8 文字ずつとした.

一人につき「す」「そ」「み」「さ」各 32 文字のデータを集めた. また,「す」「そ」「み」は各 10 点,「さ」は 9 点の特徴点をとり, 計 7488 点の座標データを用いる.
「す」「そ」「み」「さ」各 32 文字のうち, 各 24 文字を学習データ, 各 8 文字をテストデータとする.

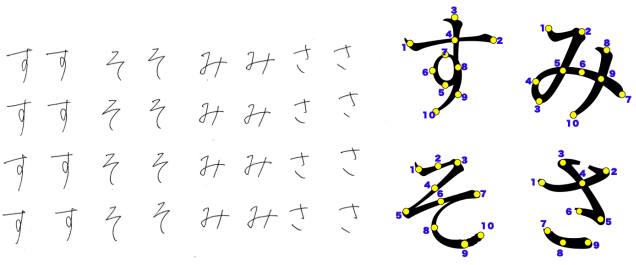

図 1: (左) 手書き文字データ例, (右) 特徴点

6　実験概要

6.1　実験 1

目的　「す」「そ」「み」「さ」4 文字をカーネル PCA を用いて人の識別を行う際の, 重要な点を調べる.

手法 識別に用いる座標データの数を, 6 人が識別できなくなる限界まで減らしていく.

結果 図 2 の重要な点のデータのみで 6 人を識別できた. また, 識別を行なった際に似たような位置に分類されているグループの重要な点の座標の平均を図形として表してみたところ, 目視では大きな違いを見つけることはできなかった.

考察 識別に用いる特徴点を半分程度にまで減らしても, カーネル PCA を用いて 6 人を識別することは可能である. また, 人間の目では見つけることの難しい特徴を用いて, 識別を行なっている.

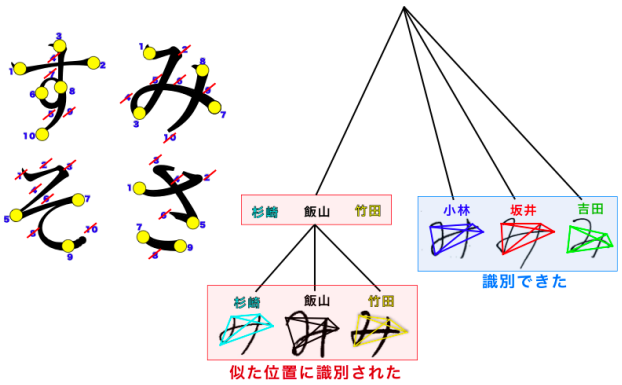

図 2: (左) 重要な点, (右) 識別の階層図

6.2 実験 2

目的 識別器を構成し, 精度を調べる.

手法 1) 学習データを用いて, 各文字 3 つずつ識別器を構成する.

2) テストデータを用いて, 識別器の精度を調べる. 各文字 3 つの識別器での識別率を実験した後, 1 文字のみの場合の識別率を用いて, 複数文字を組み合わせた識別率を計算する.

識別率) $P_X(A)$ を"文字 X が A さんの書いた文字であると正しく識別された確率"とするとき, 複数文字を合わせた場合の識別率と誤識別率は以下のように表される.

$$P_{XY}(A) = 1-(1-P_X(A))(1-P_Y(A))$$
$$P_X(\overline{A}) = 1-P_X(A)$$

結果 図 3 の例で示すように,「す」「そ」「み」「さ」4 文字全てにおいて, 3 つの識別器を構成することができた. しかし, 3 日分の学習データを一度に用いた識別器では、人の識別に重なりが多い識別器しか構成することができなかったため, 日ごとに分けて 3 つの識別器しか構成できなかった. 理由としては, 同じ人でも日によって書く字が大きく違っていることがあったためである. また, 図 4 に示すように, 1 文字の場合の識別率は 0.30 であったが, 4 文字まで増やすと 0.72 に増加した.

考察 カーネル PCA を用いれば, 6 人を一度で識別する識別器の構成は可能であった. しかし, 同じ人でも日によって異なる特徴を持った文字を書いており, 全ての学習データを用いた識別器を構成することはできなかったため, 色々な日に色々な状態で書いた文字を用いると良い. また, 学習データの量と種類を増やすことが, 精度の高い識別器の構成には必要であると考える.

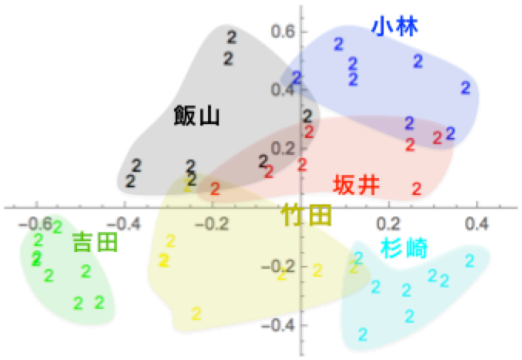

図 3: 「み」の識別器 2

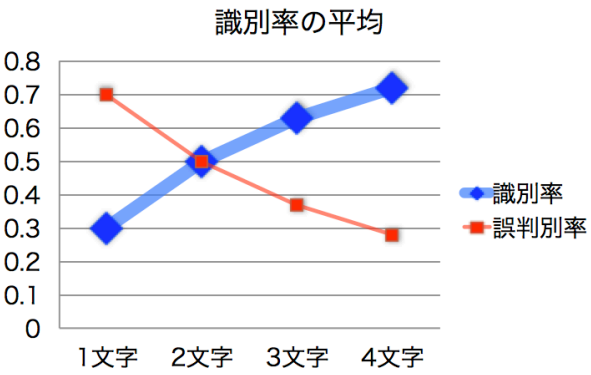

図 4: 文字の識別率

7 本研究のまとめと今後の課題

カーネル PCA は, 手書きひらがなのパターン識別の手法の一つとして有効であると考えられる.

しかし, 平仮名は普段書き慣れていないこともあり, 本人でも書く字が大きく違うことがあった. そのため今後は, 本人の名前を漢字で書いてもらうなど, 書き方に比較的強い癖がついているであろう文字をデータとして用いたい. また, 今回の研究では「す」「そ」「み」「さ」4 文字合わせても, 識別率が 0.72 までしか上がらなかったため, 学習に用いるデータを増やすことで識別器の精度を上げたい.

参考文献

[1] 赤穂昭太朗 「カーネル多変量解析 -非線形データ解析の新しい展開-」, 岩波書店 (2008)

[2] 村上征勝, 浦部治一郎, 「浮世絵における役者の顔の描画法に関する数量分析」, 統計数理, 55(2), 223-223 (2007)