様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年 ５月２３日現在

機関番号：１７１０２
研究種目：挑戦的萌芽研究
研究期間：2011～2012
課題番号：23650089
研究課題名（和文）　事例稠密化による極限精度パターン認識

研究課題名（英文）　Toward ultimate pattern recognition by massive instances

研究代表者
内田　誠一（SEIICHI UCHIDA）
九州大学・システム情報科学研究院・教授
研究者番号：７０３１５１２５

研究成果の概要（和文）：
計算機の著しい性能向上によって，大量のデータを用いた認識が行われるようになった．しかし，パターンの分布については未だに明らかになっていない部分が多い．そこで本研究では，例として約 80 万個の手書き数字画像を登録した大規模データベースを用いて「パターンの真の分布解明」を目指した．そして，大規模文字認識や，パターンの近傍関係から作成した最小全域木の解析を通してパターン分布の解明に取り組んだ．また，クラスの隣接関係や，パターンの増加に伴うパターン空間の状況の変化を検証した．

研究成果の概要（英文）：
The ambitious goal of this research is to understand the real distribution of character patterns. Ideally, if we can collect all possible character patterns, we can totally understand how they are distributed in the image space. In addition, we also have the perfect character recognizer because we know the correct class for any character image. Of course, it is practically impossible to collect all those patterns — however, if we collect character patterns massively and analyze how the distribution changes according to the increase of patterns, we will be able to estimate the real distribution asymptotically. For this purpose, we use 822,714 manually ground-truthed handwritten digit patterns. The distribution of those patterns are observed by nearest neighbor analysis and network analysis, both of which do not make any approximation (such as low-dimensional representation) and thus do not corrupt the details of the distribution.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,100,000	930,000	4,030,000

研究分野：総合領域
科研費の分科・細目：情報学・ 知覚情報処理・知能ロボティクス
キーワード：パターン認識、ネットワーク解析、パターン分布解析

１．研究開始当初の背景

（１）一般にパターン認識とは、画像や波形などを予め定義された幾つかのクラスの一つに識別する問題である。例えば顔認識であれば、それが「A さんの顔」なのかそれ以外かを識別することになる。それでは「A さんの顔」認識の極限精度に到達するにはどうすればよいだろうか？答えは明確で、「A さんの顔の真の分布」、すなわちありとあらゆる「A さんの顔」があればよい。しかしそれは爆発的な数となる。このためパターン分布の全容は不可知なものと考えられた。結果的に、

何らかの数理統計モデルを仮定し、少数のパターンから具体的なモデルを構築し、認識に利用するのが通例であった。要するに、50年来のパターン認識研究は、まさにこの不可知さゆえに、分布の正規性などの大胆な仮定を置かざるを得なかった。

（２）要するに問題点は、全てのパターンが極めて膨大な数となる点に帰着する。しかし、逆に言えば、もし想定される全てのパターンもしくはそれに近いぐらいに大量のパターンがあれば、真の分布が得られ、仮定を一切伴わない極限精度を持った認識システムを実現できることになる。好都合なことに昨今の計算機は、数年前では不可能と思われたものと可能とするパワーを持っている。要するに、それら正確なラベル（正解）の付いた大量のパターン（事例）を準備できれば、極限の精度を持った認識システムを実現できる時代になりつつあると考えられる。

２．研究の目的

（１）実世界に存在するパターンを大量に蓄積・利用することで、「パターン分布の真の姿の解明」および「極限精度パターン認識」に挑戦したい。方針は徹底した正攻法である。すなわち、まず実世界で収集した大量のパターンをそれらの正確なクラスラベルと共に事例として計算機に蓄積する。次に、パターン空間内での位置関係を徹底的に解析し、クラス間の空隙・オーバーラップなどパターンの真の分布状況を解明する。さらにパターン空間内で稠密に分布するそれら超大量の事例を用いた最近傍法により極限精度認識を試みる。ただし数万画素からなる一般的な画像に対しては、パターン数が爆発するため、この正攻法では全く手が出せない。さらにクラスラベルも曖昧になる。そこで本研究課題では「極めて少画素」で「正確なラベル付けが可能」な文字画像を用いて、これまで不可知とされていたパターン分布の真の姿に、世界で初めて肉薄したい。

３．研究の方法

①既に準備が進んでいた目視によるラベル付き文字画像パターン約１００万個を事例として用い、それらの「分布の真の姿の解明」ならびに「極限精度パターン認識」の２課題に挑戦する。本課題では特に前者を重要視しており、多角的に取り組む。具体的には、大量事例の各クラスでの分布の定量的・定性的解析により、(i)外れパターン（アウトライヤ）や分布間のオーバーラップの様子、(ii)非文字領域（リジェクト領域）の連結性、について解明を図る。

②「分布の真の姿の解明」のために、事例集合を徹底的に吟味する。このために、パートナー企業(㈱オーリッド、主業務：データエントリ、本社：大分県別府市)の協力によって既に準備されている、100 万個オーダーの目視によるラベル付き手書き数字パターンを利用した（図１）。ところで、文字画像が小さいとはいえ、16×16=256 次元の直接の可視化は不可能である。可視化のためには主成分分析や多次元尺度構成法(MDS)が考えられるが、低次元に近似表現された分布を観察しても、「真の姿」と呼ぶには不十分である。そこで真の分布から直接定量化可能な方法として、「正規分布からの乖離分析」、「最近傍関係より構築した超巨大全域木による分布構造解析」などを考える。これらにより、特に分布境界の状況を定量化でき、具体的には、アウトライヤの存在、クラス間のオーバーラップの検証、無意味データの分布の把握、などが可能になる。

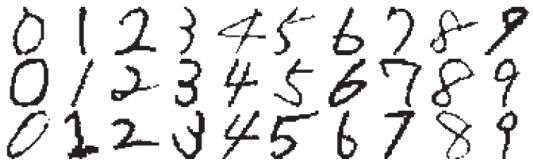

図 1　実験で用いた数字パターンの例

４．研究成果

（１）大量の学習パターンによる手書き数字認識

①学習パターン数が認識に及ぼす効果について定量的に検証するために，最近傍決定則による文字認識実験を行なった．具体的には，学習パターン数を変化させて実験を行ない，その結果から学習パターン数と認識率の関係性を見出す．同時に，その際のパターン空間における状況の変化を明らかにすることも目的とした．

②入力パターンと学習パターンとの距離尺度には，ハミング距離を用いた．本稿では，非常に多くの 2 値画像データを演算処理する必要がある．そこで，距離計算を高速かつ少ないメモリ領域で実現可能なハミング距離は好都合である．また，距離値が白黒の違う画素数に相当するため，解釈も容易である．

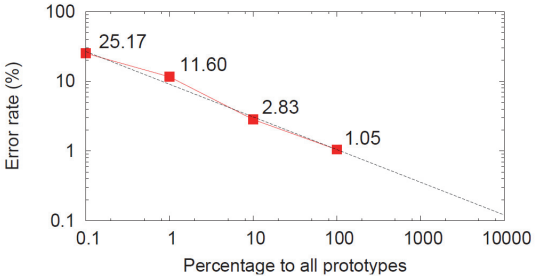

図 2 学習パターン数と認識率の関係

③図２にデータ数を変えながら測定した誤認識率を，対数グラフにプロットした結果を示す．図中の赤線は学習パターン数の増加による誤認識率の推移を示している．また，図中の破線は得られた誤認識率を直線近似したものである．学習パターン数の増加に伴い，誤認識率は低下する傾向にあった．すなわち，より多くの学習パターンを用いることで，より高精度な認識が可能であることが実証された．822,714 個全てのパターンを用いて認識を行なった際，最も低い誤認識率 1.05% が得られた．この結果は学習パターン数と認識率の間にある非常に興味深い関係を示している．すなわち，認識に用いる学習パターン数を 10 倍に増加させると，誤認識率が 65% 程度低下すると推察できる．以上より，学習パターン数の増加のみによって認識精度の向上を試みる場合は，学習パターン数を指数的に増加させる必要があることが明らかになった．

（２）最小全域木(MST) を用いた分布解析

①最小全域木 (Minimum Spanning Tree : MST) によってパターン分布をパターンのネットワークとして表現し，その構造的特徴からパターン分布の構造を解析する．具体的には，最小全域木のノードやエッジの分析，クラスタリング，経路解析などのネットワーク解析を通して，パターン分布の特徴について考察を行う．

最小全域木を用いる利点としては，グラフ理論やネットワーク理論の観点による解析手法を適用することができる点が挙げられる．グラフの辺の重みや，節点の次数を分析することで，比較的容易にその特徴を捉えることができる．また，膨大なデータを用いる上では，グラフ作成に高速なアルゴリズムが存在し，構造を記述するために必要な空間量が少ない点も魅力的である．また，最小全域木はしばしば階層型クラスタリングの手法として用いられるように，パターン分布の偏りを表現することにも有効である．たとえば医学分野では，疾患の感染地域や，遺伝子発現データのクラスタリングに最小全域木が利用され，有効性を示している．

②データベース内の画像パターンより MST を作成する手順は以下の通りである．まず，各パターンをノードとし，パターン間のハミング距離をエッジの重みとする完全グラフを考える．次に，この完全グラフから Prim のアルゴリズムによって MST を構築する．このように，パターン間距離を基に MST を作成することで，元のパターン空間における近傍関係を保存する．2 値画像において，ハミング距離は画像間で白黒が異なる画素の数となる．したがって，MST の性質により，隣り合うノードは視覚的に近いパターンとなる．

このようにして作成した MST は，次の 3 つの特性を持つ．まず，前述の通り，局所的に見ると類似画像が集まってクラスタを作る．また，クラスタ間のエッジも距離が小さいものを選択するため，クラスタ同士の近傍関係を保持する．つまり，大局的に見てパターン空間全体の構造を保持する．さらに，全体がひとつの木グラフになるため，MST の枝を辿ってあるノードと別のノードを結ぶパスが必ず存在し，かつ一意に定まる．

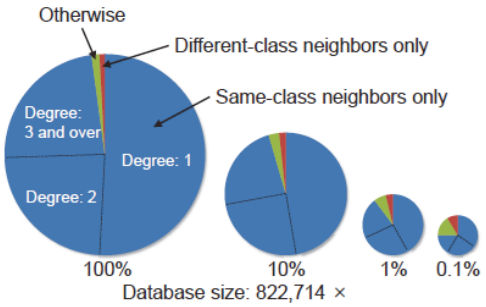

図 3 隣接ノードの分類結果

③以上で生成した MST のトポロジ等を解析することで，分布に関する様々な情報が得られたが，以下ではその一つの例として，クラスの隣接状況の定量化結果について述べる．図３は，MST の各エッジの両端にあるサンプル，すなわち隣接ノードのクラスによって，ノードすなわちパターンを分類した結果である．すなわちこの円グラフでは，自ノードと隣接するノードが，(i) すべて自ノードと同じクラスラベルを持つ，(ii) すべて自ノードと異なるクラスラベルを持つ，(iii) 一部のみ自ノードと同じクラスラベルを持つ，という 3 つの場合の割合を示す．ここで，(ii) の場合となったノードはアウトライヤのパターンであろう．そして，(iii) の場合となったノードは，クラス間の橋渡しをしているパターンである．図３より，ほとんどのノードが同じクラスのノードとのみ接していることがわかる．さらに，クラス間をつなぐノードが少ないことから，同じクラスのパターンが集まった領域があることが予想できる．

図４は，異なるクラスと隣接するノードの割合の推移である．MST のパターン数が増加するに伴い，異なるクラスとのみ隣接するノードも，どちらのクラスとも隣接するノードも，全ノードに対する割合が低下しているのがわかる．この傾向はべき乗則に従っており，MST のパターン数が 10 倍になると，割合はおよそ 60% 低下している．したがって，パタ

ーン数に対するおよそのアウトライヤ数の予測が可能である．また，この傾向は，（1）で述べた大規模文字認識における誤認識率の傾向と一致しており，両者に相関関係があると言える．MST の性質から，あるパターンが持つエッジのうち，少なくともひとつは最近傍パターンに接続されている．よって，異なるクラスと隣接するノードは，最近傍が異なるクラスのパターンである確率が高く，誤認識されやすい．したがって，異なるクラスと隣接するノードの数と誤認識率との間に，相関が現れたと考えられる．

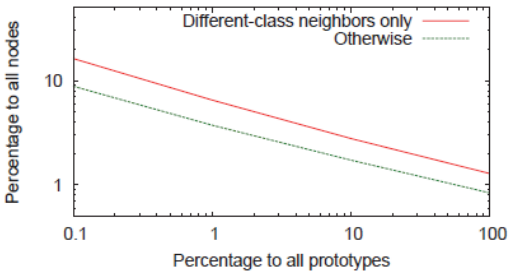

図 4 異なるクラスをつなぐエッジの割合

④最後に，MST を直接観察した結果について述べる．もちろん 80 万ものノードを全てプロットはできないので，ここでは「クラスタ木」という MST の粗視化法を導入する．これは MST において隣接する同じクラスのノードを併合することで生成される，新たな木グラフである．クラスタ木のエッジは，元の MST における異クラス間をつなぐエッジとなり，クラス間の隣接関係を保持している．この木グラフを用いることで，クラスごとのパターンの分布状況やクラス間の関係をより深くとらえられると期待できる．図 5 は実際に MST から作成したクラスタ木である．ひとつの円がひとつのクラスタを表し，数字がクラスを，円の大きさがそのクラスタに属するパターンの数，すなわちクラスタの大きさを表している．また，すべてのクラスタを描くには数が多いため，要素数 100 以下の微小なクラスタは図から省いた．微小なクラスタは本来の分布領域から外れたパターンと考えられ，省略してもクラスの分布や隣接関係の大勢に影響はないと判断した．ただし，大きなクラスタ間を結ぶように微小クラスタが存在している場合は，木構造が崩れないようにクラスタを残した．そのようなクラスタについては，例外として区別するために四角で表記した．別の言い方をすれば，省いた微小クラスタは，巨大クラスタが作る木構造から外れたクラスタである．したがって，その除外はクラスタ木の大局的構造に影響しない．

図 5 より，クラスごとに巨大なクラスタが存在することが見て取れる．これらの巨大クラスタは，各クラスの大部分のパターンを含んでおり，主となる分布領域を表している．一方，クラス 4 やクラス 7 は複数のクラスタが形成されており，分布領域が分断されている．こうした複数クラスタへの分断の一因は，MST の閉路を形成しないという制約であろう．しかし，そうした分断が起こってないクラスが他にあることも考えれば，分断化されたクラスは分断化されてないクラスに比べ，分布に粗密が大きいと見て間違いない．すなわち，1 つの標準形状とその連続的変形によりクラス全体が生成されているというよりは，異体字のような核となるパターンが複数存在していると考えられる．クラス 4 は 3 つの巨大クラスタを持ち，最大クラスタでも 40% 程度のパターンしか含まない．また，クラス 7 は中規模クラスタの数が多く，他クラスと比較してパターンが分散している．

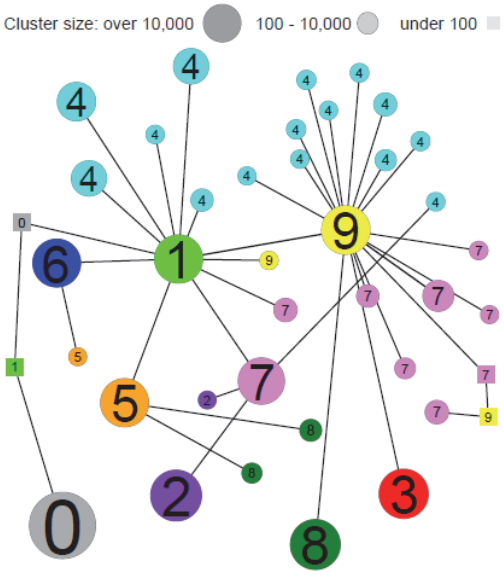

図 5 実際に得られたクラスタ木

図 5 のグラフからクラス間の位置関係を見ると，クラス 1 と 9 が多くのクラスタをつなぐハブの役割を果たしていることがわかる．特にクラス 1 は複数の巨大クラスタの架け橋となっている．このクラス 1 の特徴には，数字の“1”が，基本的に縦方向の単純なストロークのみで構成されることが関係している．全体的に，数字パターンは横方向よりも縦方向に長い形状をしている．そのため，他クラスのパターンの中でも縦方向に長いパターンがクラス 1 の近くに分布し，どのクラスもクラス 1 と近いという状況を作ったと考えられる．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　４件）
(1)　Masanori Goto, Ryosuke Ishida, Yaokai Feng and Seiichi Uchida, Analyzing the Distribution of a Large-scale Character Pattern Set Using Relative Neighborhood Graph, Proc. 12th International Conference on Document Analysis and Recognition, 査読有, 2013, Accepted.
(2)　Seiichi Uchida, Ryosuke Ishida, Akira Yoshida, Wenjie Cai and Yaokai Feng, Character Image Patterns as Big Data, Proc. 13th International Conference on Frontiers in Handwriting Recognition. 査読有, 2012, 477-482
(3)　Yutaro Iwakiri, Soma Shiraishi, Yaokai Feng, Seiichi Uchida, On the Possibility of Instance-Based Stroke Recovery, Proc. 13th International Conference on Frontiers in Handwriting Recognition, 査読有, 2012, 29-34
(4)　Seiichi Uchida, Wenjie Cai, Akira Yoshida, Yaokai Feng, Watching Pattern Distribution via Massive Character Recognition, Proc. IEEE International Workshop on Machine Learning for Signal Processing, 査読有, 2011, 1-6

〔学会発表〕（計　７件）
① 中本千尋，フォントネットワーク～大規模フォントセットの分布構造解析～，電子情報通信学会パターン認識・メディア理解研究会，2013 年 3 月 14 日，電気通信大学
② 佐藤洪太，大規模パターンを使った Self-Corrective Learning の挙動解析，電子情報通信学会パターン認識・メディア理解研究会，2013 年 3 月 14 日，電気通信大学
③ 柿迫良輔，分布構造を利用した半教師あり学習による文字認識，電子情報通信学会パターン認識・メディア理解研究会，2013 年 3 月 14 日，電気通信大学
④ 岩切裕太郎，大規模事例に基づく時系列推定の可能性―筆順復元問題を例として―，画像の認識・理解シンポジウム，2012 年 8 月 7 日，福岡国際会議場
⑤ 石田良介，吉田晃，蔡文傑，フォンヤオカイ，内田誠一，大規模数字画像データベースを用いたパターン分布解析，画像の認識・理解シンポジウム，2012 年 8 月 8 日，福岡国際会議場
⑥ 吉田 晃,大規模手書き文字認識－欠損部補完に見る文字パターン分布－，電子情報通信学会パターン認識・メディア理解研究会，2011 年 11 月 25 日，長崎大学
⑦ 石田良介,大規模手書き文字認識～ ネットワーク解析に見る文字パターン分布 ～，電子情報通信学会パターン認識・メディア理解研究会，2011 年 11 月 25 日，長崎大学

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　０件）
名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　０件）
名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
human.ait.kyushu-u.ac.jp/~uchida

６．研究組織
(1)研究代表者
内田　誠一（UCHIDA SEIICHI）
九州大学・大学院・システム情報科学研究院・教授
研究者番号：70315125

(2)研究分担者
金子　邦彦（KANEKO KUNIHIKO）
九州大学・大学院・システム情報科学研究院・准教授
研究者番号：50274494
馮　尭楷（FENG YAOKAI）
九州大学・大学院・システム情報科学研究院・助教
研究者番号：60363389

(3)連携研究者
該当無し