クラスタリングによる視点に不変なパターン認識の学習

井上, 光平
九州芸術工科大学

https://doi.org/10.11501/3168354

出版情報：九州芸術工科大学, 1999, 博士（工学）, 課程博士
バージョン：
権利関係：

クラスタリングによる視点に不変なパターン認識の学習

井　上　光　平

クラスタリングによる視点に不変な パターン認識の学習

Learning for View-Invariant Pattern Recognition by Clustering

1999年12月

井上 光平

Kohei INOUE

目 次

第1章

序論

ある物体を見るとき視点の位置を変えるとその物体の見え方は大きく変化するが，我々はそれを同一の物体として認識する．本論文は物体認識におけるこの視点不変性のモデル化に関する研究をまとめたものである．

本章では本研究の背景と目的を述べて本論文の構成と概要を示す．

1.1 研究の背景と目的

近年，電子計算機の情報処理能力が飛躍的に向上して情報化社会が急速な進展を見せる中で，様々な分野でより高度な情報処理が求められるようになってきている．その例としてはデータベース検索，メディア変換，自動翻訳，医療診断の支援，産業用ロボットなどが挙げられる．これらの高度な情報処理を実現するための基礎としてパターン認識の研究が行われている．

パターン認識は本来人間をはじめとする生体に備わった機能であるが，これは感覚器に与えられる刺激に対して何らかの応答を出す一種の情報処理として捉えられることから，電子計算機の登場と共に工学的に研究されるようになった．現在，パターン認識の研究内容は多岐にわたっており，それらを統一的に捉えることは容易ではないが，以下，アプローチによる分類，用いる数学的手法による分類，学習過程の違いによる分類などに基づいて本論文で提示する認識法の位置付けを説明していく．

まず認識手法を構成していくアプローチを大きく2つに分けると，個々の問題からのトップダウン的な開発法と生体の認識システムに基づくボトムアップ的な構成アプローチとがある．前者は文字や音声や図形など特定のパターンを対象としてそれを正しく安定に認識するためのアルゴリズムを追求する対象パターンを限定した認識技術で

ある．文字認識，音声認識，図形認識などがこれにあたり，適用範囲を制限して実用的な装置を開発する場合はこのアプローチは有効であり，文書OCR(Optical Character Reader)[1]などすでに実用化されている技術も多い．後者は人間などの生体を対象としてその認識メカニズムの解明に基づいて応用技術へと展開していくアプローチである．こちらは生体の振る舞いや構造のパターン認識器としての性質に着目するため，生物学，生理学，心理学など生体を研究対象とするいくつかの分野と深く関わっている．従来の研究としては生体の神経回路網の数理モデルであるニューラルネット[2]をはじめとする生体機能のモデル化に関する研究がある．パターン認識が生体に固有の機能であることからその機能を解明するためには後者のアプローチは妥当であり，またその成果は前者のアルゴリズム開発に対しても有用な示唆を与えるものと思われる．このような観点から本研究は後者の立場で進められている．

パターン認識の過程は学習の段階と識別の段階とに分けられる．学習は与えられたデータ(学習データ)にそれの所属するクラスを示す教師信号が付与されている教師付き学習(あるいは教師あり学習)と，教師信号が付与されない教師なし学習とに分けられる．教師付き学習ではクラスは予め形成されており，学習データに対してその教師信号にできるだけ近い応答をするようにパターン認識器が調整される．それに対して教師なし学習ではクラスは予め形成されておらず，学習データを用いてクラスを生成する必要がある．その方法としてはデータ間に定義される類似度あるいは距離に基づいたデータのクラスタリングが基本的である．クラスタリングはパターン認識においてだけでなく，多変量解析におけるクラスタ分析やグラフ理論におけるグラフ分割など多くの分野で研究されており[3]，応用範囲はパターン認識や画像処理などの工学的分野にとどまらず，社会調査や心理学などの人文科学や社会科学にも及んでいる[4]．パターン認識の分野で古くから用いられているクラスタリング法としてはmaximin-距離法，c-平均法(あるいはk-平均法)，ISODATA法などがある[5]．その中でc-平均法はクラスタ内偏差平方和の最小化問題として定式化される[6]．c-平均法をファジー化したファジーc-平均法[7]はファジークラスタリングの基礎になっているが，初期値依存性やノイズに対する弱さが指摘され，その後ノイズクラスタリング法，確率クラスタリング法，マウンテン法など多くのロバストなクラスタリング法が提案されている[8]．一方，グラフ理論においてはクラスタリングをデータ間の類似度を要素とする行列の固有値問題に帰着させる方法が知られている．これはグラフスペクトル法[9]と総

称される方法の一種であり，ノイズデータに対してロバストでありファジークラスタが解析的に求まる点が優れている．これらのクラスタリング法により学習データはいくつかのグループにクラスタリングされ，各クラスタは1つあるいは少数個の代表点で表される．これにより多数の学習データが少数の代表点で表されることになり記憶すべき情報が削減されると共に，データの分布の大局的な構造が抽出されるため識別の段階において未知データに対しても適切に応答するための汎化能力が期待される．

パターン認識に用いられる数学的手法は主に統計的手法と構造的手法とに分けられる．統計的パターン認識[10]はパターン認識研究の初期の段階から研究され，今日のパターン認識研究の基礎理論となっている．統計的パターン認識ではデータの変動が確率的に表現され，未知の入力データは事後確率が最大となるクラスに分類される．この方法はベイズ識別則と呼ばれ，誤識別率が最小になるという意味で最適な識別方法である．統計的パターン認識における学習は確率密度関数の推定として定式化される．推定の方法にはパラメトリックな方法とノンパラメトリックな方法とがある．パラメトリックな方法では少数個のパラメータで表現される関数形が与えられており，最尤推定やベイズ推定によってパラメータ値が決定される．ノンパラメトリックな方法にはパーゼン窓関数やk近傍推定などがある．パラメトリックな方法は推定が比較的容易であるが，関数形が固定されているためデータの分布を十分表現できない場合がある．一方，ノンパラメトリックな方法はデータの分布の自由な形状を表現可能であるが，一般に膨大な学習データを必要とする．そこでこれら2つの方法の中間に位置する方法としてセミパラメトリックな方法[11]があり，混合モデルがよく知られている．これはデータの分布をパラメトリックな関数の線形結合で表現したものであり，比較的少ないパラメータ数で複雑な分布形状を表現できる．ニューラルネットにおけるRBF(radial basis function)ネットは混合モデルの一種であり，関数近似やパターン認識のモデルとして近年盛んに研究されている[6]．そのパラメータ調整法としてはEM(expectation maximization)アルゴリズム[11]がある．EMアルゴリズムによるRBFネットの学習はクラスタリングとして捉えられる．すなわちRBFネットの基底関数の中心を適切に配置することはクラスタの中心を求めることに他ならない．

人間は五感によって外界の情報を得ているが，3次元空間の認識においては視覚が大きな役割を果たす．視覚は網膜に映った2次元の画像から3次元世界の構造あるいは状態を推定するという一種の逆問題を解いている[12]．視覚から得られる情報には明る

さ，色，形，テクスチャなどがあるが，視覚情報処理の主たる目的は物体の形状と物体相互の位置関係など外界の3次元的な構造を知ることであると考えられている[12, 13]．すなわち物体認識が視覚の大きな目的の1つである．我々が日常的に行っている物体認識の例として人の顔による個人識別を考えてみよう．ある人の顔を見る場合，自分と相手との相対的な位置関係が変わると視点の相対的な位置が動いて網膜に映る相手の顔の画像は変化する．しかし我々はそれを別の人の顔であると誤って識別することはない．すなわち個人識別には視点の変化によって生じる網膜上の顔の大きさ，向き，位置の変化に対する不変性が求められる．また一方で我々は相手に関係なく顔の大きさ，向き，位置を大まかに推定することができる．このように物体認識においては不変性の発見と変化量の推定とが並行して行われている．このような物体認識を実現するためにコンピュータビジョンにおいては物体の幾何学的な変換に対して不変に保たれる量すなわち幾何学的不変量を計算する研究が行われている[14]．画像から不変量が計算できれば不変量の照合により対象物を識別できるが，この方法では極端に大きく変形した物体でも同じ不変量で表されることがあり，人間の認識結果と一致しない場合がある．また不変量の計算には照合する物体間で点や線などの特徴の対応付けが必要な場合が多い．この対応問題[14, 15]はステレオ視，動きからの形状復元，モデルベースの物体認識やナビゲーションなどコンピュータビジョンの様々な場面で生じる重要な問題であるが，従来の研究の中にはこの対応は求まっているという仮定の下で進められているものが多い．別の物体認識の方法としては画像と予め記憶されたモデルとの照合により対象物を識別する方法がある．従来3次元モデルを用いる方法が多かったが，近年2次元モデルを用いる方法も多く研究されている[16, 17, 18]．Ullmanら[16]は少数枚の2次元画像(特徴点の2次元座標)の線形結合により3次元物体(特徴点の3次元座標)を表現できることを示した．それに対してPoggioら[18]のRBFネットモデルは2次元画像の非線形結合モデルとして捉えられる．Poggioらのモデルは心理実験結果とよく似た振る舞いを示す．また生理学においては顔の向きによらず顔に応答するニューロンと顔の向きに選択的に応答するニューロンとが観測されており[19]，Poggioらのモデルはそれらの動作を説明するモデルになっている．

Poggioらのモデルは教師付き学習によって調整される．しかし実際の物体認識では教師信号が与えられているとは考えにくく，むしろ外界の物理現象の中に教師を見出していると考えるほうが自然である．そこで物理現象の時間的な連続性に着目した教

師なし学習法が研究されている [20, 21, 22, 23]．前述の顔の例では視点の変化に伴って網膜像が変化している状況でも，ある一定の時間は同じ人物が網膜に映っており，従って現在見ている人物を次の時刻も見ている可能性が高いと考えられる．上の学習法ではこの時間的な文脈が教師信号として利用されるが，外部から明示的に教師信号が与えられるのではなくモデル内部で教師信号が作られるので全体として教師なし学習になっている．生理実験でも時間的な文脈に基づく学習が観測されている．酒井ら [24] は図形の対連合課題の実験によって時間的な文脈の影響を調べ，図形パターンのペアリングがパターンの類似度でなく提示時刻の近接性によって形成されることを示した．このように文脈情報は生体のパターン認識に大きく影響する要素であることが分かる．

文脈情報は入力データから得られるボトムアップ情報に関係なく与えられるトップダウン情報であり，文脈情報を取り入れたパターン認識器はそれら2つの情報をモードとするマルチモーダルパターン認識器として捉えられる．また人間は五感のそれぞれをモードとするマルチモーダルパターン認識器として捉えられる．そのためマルチモーダルパターン認識は心理学や生理学でも研究されている．その典型例としては心理学におけるマガーク効果 [25] が知られている．マガーク効果は視覚と聴覚をモードとする2モードのパターン認識において生じる心理学的錯覚である．例えば“ba”の音を聞きながら，“ga”の音を発音する唇の画像を見ると“da”の音を知覚する．このマガーク効果を説明するモデルがこれまでにいくつか提案されている [26, 27, 28, 29, 30] が，マガーク効果はノイズのある環境下でのみ生じるという性質が説明されていなかった．松永ら [31] はロバスト情報統合に基づく教師なし学習アルゴリズムを提案し，ノイズの影響を説明した．これらのモデルはフィードフォワードのニューラルネットであるが，心理実験においてトップダウンの効果が観測されており，また脳においてはフィードバック結合が見つかっており，これらを説明するモデルが望まれる．

以上のような背景から本研究では物体認識における視点不変性のモデル化を目的として，統計的パターン認識の理論に基づき，フィードバック結合を持つマルチモーダルパターン認識器を提案し，それに基づいて時間的あるいは空間的な文脈を取り入れたパターン認識器を構成し，更にそれを視点に不変なパターン認識に応用する．これらのパターン認識器の学習はクラスタリングとして捉えられる．またグラフスペクトル法に重みの概念を取り入れた新しいクラスタリング法を提案し，それに基づいて平面図形の視点に不変な認識を行う．

1.2 論文の構成と概要

本論文は7章からなる．以下に各章の概要を示す．

第1章では本研究の背景と目的を述べて本論文の構成と概要を示す．

第2章では視点に不変なパターン認識の基礎としてマルチモーダルパターン認識のニューラルネットモデルを提案する．提案モデルは松永ら[31]のモデルにメンバシップ値をフィードバックする機構を付加したものである．このモデルのベイズ識別則に基づく識別法とEMアルゴリズムによる教師なし学習法を示し，最尤推定によるマルチモーダルパターンの再構成法を示す．マガーク効果を説明する簡単なデータを用いてフィードバックの効果を調べ，心理学において報告されているいくつかのモードの入力パターンから別のモードパターンへの知覚誘導と生理学において観測されているモード情報が複数の感覚野からのフィードバックパスを通るトップダウン信号により誘導されるというニューロンの活動を説明する．

第3章では第2章で提案したマルチモーダルパターン認識器に基づき，認識器の出力であるメンバシップ値をフィードバックすることによって時間的あるいは空間的な文脈情報を取り入れたパターン認識器を提案し，ニューラルネットによる構成を示す．またそのパターン認識器の最尤推定に基づく教師なし学習法を提案する．時間的な文脈については1時刻前の識別結果を次の時刻にフィードバックする例を考え，簡単なデータを用いてそれらがパターンの類似度でなく提示時刻の近接性によってクラスタリングされることを示し，簡単な画像データを用いて位置不変なパターン認識への応用例を示す．空間的な文脈については空間的に1つ隣りにあるニューロンへメンバシップ値を伝搬する例を考え，空間データのノイズ平滑化やあいまいさの低減化やデータのない部分への充填現象などの空間的な整合化が行われることを示す．またデータの欠落と多重性を伴う空間データの例としてランダムドットステレオグラムを取り上げ，視差の計算を行う．

第4章では第3章で提案した時間的な文脈情報を取り入れたパターン認識器を視点に不変なパターン認識に応用する．時間的な文脈を伝搬するニューラルネットはいくつかの代表的な視点の2次元画像によって視点に不変な3次元物体の認識をするモデルであるRBFネットにメンバシップ値を事前情報としてフィードバックする機構を付加したモデルである．視点が時間的に変化する時系列データをパターン認識器に提示

することにより，明示的に教師信号を与えることなく時間的な文脈に基づき視点に不変なパターン認識器が学習できることを示す．視点に不変なパターン認識の例として顔画像を用いて顔の向きによらない個人識別を行い，このニューラルネットを構成するニューロンが生理学において観測されている顔の向きによらず顔に応答するニューロンと顔の向きに選択的に応答するニューロンとよく似た応答をすることを示す．またRBFネットの基底関数にロバストな分布を用いることによって学習時や識別時に混入する外れ値の画素を棄却できるようになり，時間的な予測による注視に似た処理が得られることを示す．例として3次元物体の画像からの注視領域の抽出を行う．

第5章ではグラフスペクトル法に重みの概念を取り入れて重み付きグラフで表されるデータから逐次にファジークラスタを抽出する方法を提案する．データは完全無向グラフで表され，各枝はデータ間の距離に基づく類似度を重みとして持つ．このグラフは枝の重みを要素とする隣接行列で表現され，第1クラスタはこの隣接行列の第1固有ベクトルとして求まる．またグラフの接点についても重みを考える．各接点の重みは隣接行列の対応する要素に乗じられる．接点の重みをすでに抽出したクラスタへのメンバシップ値を1から差し引いた値の積とすることによって抽出済みのクラスタを取り除きながら順にクラスタを抽出していく．抽出処理は抽出したクラスタの大きさの変化に基づいて重要なクラスタがなくなった時点で終了する．画像のセグメンテーションを例として本方法を津田ら[32]の方法と比較して性能を検証する．またカラー画像からの肌色領域の抽出への応用例を示す．

第6章では第5章で提案した逐次ファジークラスタ抽出法を用いた平面物体の視点に不変な認識法を提案する．ここでは3次元物体において生じる自己遮蔽の問題を避けるために対象を平面物体に限定している．平面物体は2次元平面上に分布する点の集合として表される．平面物体の透視射影像は非線形の変形を受けるが，視点の変化が小さいときは弱透視射影(weak perspective projection)[33]などの線形な射影で近似できることを利用して広範囲の視点から得られる多数の透視射影像を少数個の代表的な視点から得られる透視射影像で近似表現して視点に不変な認識を行う．この代表画像の選択に第5章のクラスタリング法を用いる．点パターンのクラスタリングを行うには点パターン間の類似度あるいは距離を定義する必要がある．ここでは点パターン間の点の対応は未知であるのでまず点パターンのマッチングを行う．そこでアフィン変換に不変な点パターンマッチング法を提案する．このマッチング法では2回の固有

値分解を行う．まず1回めでスケール係数を正規化し，2回めで正規化した点パターン同士のマッチングを行う．2回めの固有値分解法はShapiroとBrady[34]により提案されたものである．この方法は回転に不変であるため本方法は全体としてアフィン不変になっている．次に得られたマッチングに基づいて点パターン間の距離を測り，クラスタリングにより物体ごとに代表画像を求め，テスト画像は代表画像との距離に基づく最近傍識別により識別される．また相似変換についても同様の認識法を提案する．

第7章では本研究で得られた成果をまとめて今後の課題を述べる．

なお第7章の後ろに付録を付けて本論の補足をしている．

第2章

クラスメンバシップフィードバックをもつマルチモーダルパターン識別器

クラスメンバシップをフィードバックする機構をマルチモーダルパターン識別器に付加し，その教師なし学習アルゴリズムを提案する[35]．本モデルでは下位の識別決定がフィードバック情報によって修正され，その情報により下位のパターンの再構成が可能となる．簡単なモデルを用いてマガーク効果におけるフィードバックの効果を調べる．本章で提案するモデルは後の第3章と第4章の文脈伝搬ネットの基礎となるものである．

2.1 まえがき

マルチモーダルパターンの認識は心理学，生理学において研究されている．マガーク効果[25]は2モードすなわち聴覚，視覚信号からの音の認識において観測されるよく知られた心理学的錯覚である．これは例えば“ba”の音を聞き，“ga”を発音する唇の画像を見ると“da”の音を知覚するというものである．Massaro[26]はファジー理論に基づく知覚モデルFLMPを提案し，この視聴覚現象を説明した．この他にもこの2モードパターン認識の観測を説明する心理学的モデルがいくつか提案されている[27]が，FLMPを含むこれら全てのモデルは学習過程を導入するのが困難である．ニューラルネットによるモデルの実行には学習能力が必要である．そこで2モードの場合について教師なし学習アルゴリズムが提案され[28]，更に任意の数のモードへ拡張された[29]．またAkahoら[30]はEMアルゴリズムに基づく教師なし学習によりマルチモード情報からの概念獲得のモデルを提案した．しかしこれらのモデルは全てノイズの影響を考

慮していないためマガーク効果の説明としては不十分である．マガーク錯覚はノイズのある環境でのみ生じる[36]．松永ら[31]はロバスト情報統合に基づく教師なし学習アルゴリズムを提案し，マガーク実験におけるノイズの影響を説明した．しかしこれらのモデルは全てフィードフォワードである．心理実験においてはトップダウン効果が観測されており，また脳のニューラルネットワークにおいてはフィードバック結合も見つかっている．de Sa[37]はフィードバックに関する観測をまとめている．そこで本章では松永ら[31]のモデルにフィードバックを付加し，そのネットワークの教師なし学習アルゴリズムを導く．

2.2 マルチモーダル識別器

データdはマルチモーダルすなわち$d=[d_1,...,d_l]$とする．d_iは第iモードへの入力である．クラス数をn $(k=1,...,n)$，各モード成分データは次のような混合分布によりモデル化されるとする:

$$p(d_i)=\frac{1}{n}\sum_{k=1}^{n}p(d_i|k) \tag{2.1}$$

$p(d|k)$は第kクラスの成分データの密度である．混合は一様であるとする．松永ら[31]は成分密度のロバストな形状

$$p(d_i|k)=\epsilon_i+s_i e^{-a_i\|d_i-r_{ik}\|^2} \tag{2.2}$$

を仮定した．各モードは各クラスで互いに独立であるとすると，融合した密度は

$$p(d|k)=\prod_{i=1}^{l}p(d_i|k) \tag{2.3}$$

と分解される．データdは

$$\arg\max_{k} \quad p(d|k) \tag{2.4}$$

により決定されるクラスへ分類される．式(2.4)の“max”をファジー化して“softmax”にするとデータdの第kクラスへのメンバシップは

$$q_k=\frac{e^{\beta p(d|k)}}{\sum_{x=1}^{n}e^{\beta p(d|x)}} \tag{2.5}$$

により与えられる(付録A参照)．βは正の増幅パラメータである．q_kの計算は容易に実行される．これが先に松永らにより調べられたフィードフォワード識別器である[31]．式(2.2)の成分密度の形状はファジー多数決によるモードの統合を導く．

2.2.1 フィードバックをもつ識別器

ここでは高位の統合決定q_kが各モードにおいて下位の推定を修飾するとする．このトップダウン修飾はq_kを直接式(2.2)に乗じることにより実行されるとすると$p(d_i|k)$は

$$p(d_i|k) = q_k(\epsilon_i + s_i e^{-a_i \|d_i - r_{ik}\|^2}) \tag{2.6}$$

となる．この修飾は全てのモードにおいて各クラスへの応答の違いを拡大する，従ってポジティブフィードバックの効果を生じる．このフィードバックをもつ識別器は図2.1に示すニューラルネットワークにより実行される．ここでモード数lは2，クラス数nは3である．最下位の“R”で示される6個のニューロンは$\epsilon_i + s_i e^{-a\|d_i - r_{ik}\|^2}$を計算するRBFニューロンである．ここで$r_{ik}$はRBFニューロンの受容野の中心，$\epsilon_i$は背景ノイズにより生じるニューロンの自発応答，$s_i$は入力$d_i$の強度を表す．例えば入力信号がないときは$s_i = 0$である．RBFニューロンの上の“×”で示される6個のニューロンは2個の入力すなわちRBFニューロンの出力と最上位のニューロンからフィードバックされるq_kとの積を出力する乗算器であり，この乗算器ニューロンの出力は式(2.6)の$p(d_i|k)$である．この$p(d_i|k)$は次の3個の乗算器ニューロンで互いに乗算され式(2.3)の$p(d|k)$が出力される．最上位の3個のニューロンは式(2.5)のq_kを出力するファジーWTA(winner take all)ネットワークを構成する．すなわち本識別器はRBFニューロン，乗算器，WTAネットワークからなる．

各モードでの識別メンバシップスコアはp_{ik}すなわち$p(d_i|k)$によりカウントされる．これは主に入力信号d_iから計算され，一方同時にトップダウン情報q_kにより変調される．入力信号がq_kと呼応するときはスコアが上がり，q_kに反するときはスコアは抑えられる．従って識別情報は全モードが統合され同じ決定を出すようになるまでフィードバックループ内を走る．この統合過程は次の反復により表される．式(2.3)と(2.6)

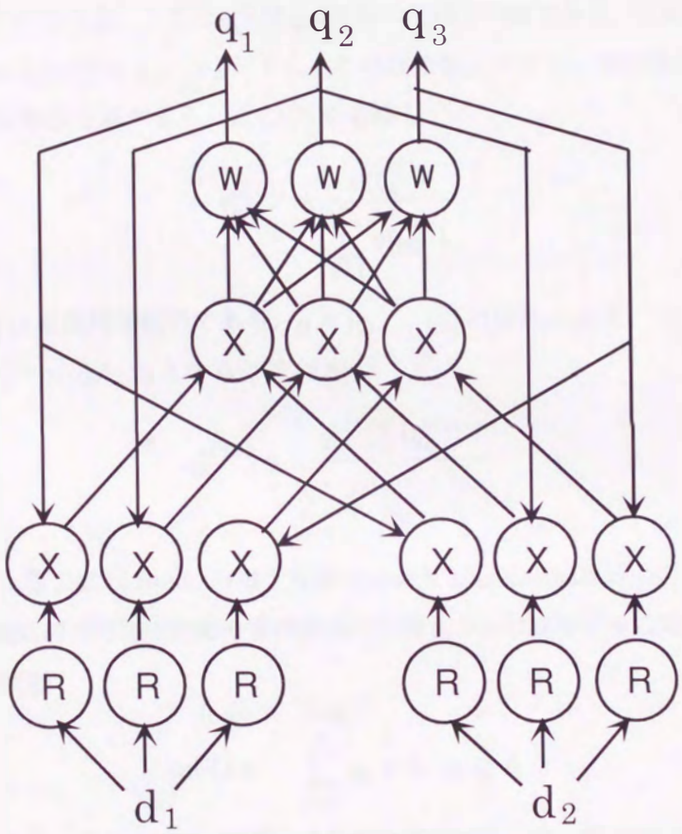


図 2.1: 2モードパターン認識のためのニューラルネットワーク

とを(2.5)に代入するとq_kに関する式が得られるが，これは反復

$$q_k^{(\xi+1)} = \frac{e^{\beta \prod_{i=1}^{l} q_k^{(\xi)}(\epsilon_i + s_i e^{-a\|\boldsymbol{d}_i - r_{ik}\|^2})}}{\sum_{x=1}^{n} e^{\beta \prod_{i=1}^{l} q_x^{(\xi)}(\epsilon_i + s_i e^{-a\|\boldsymbol{d}_i - r_{ix}\|^2})}} \tag{2.7}$$

により解くことができる．ここで$q_k^{(\xi)}$はq_kのξ回の反復での値である．反復の初期値$q_k^{(0)}$は一様に$q_k^{(0)} = 1/n$ $(k = 1, ..., n)$とする．この収束値はファジー識別決定を与える．

この反復の収束性を調べよう．式(2.7)を省略して

$$q_k^{(\xi+1)} = \frac{\phi(q_k^{(\xi)})}{\sum_{x=1}^{n} \phi(q_x^{(\xi)})} \tag{2.8}$$

と書く．$\phi(q_k)$は単調増加関数である．$q = [q_1, ..., q_n]$の関数$\psi(q)$を
$\psi(q) = \sum_{k=1}^{n} \int^{q_k} \phi(u)/u du$とすると式(2.8)は

$$q_k^{(\xi+1)} = \frac{q_k^{(\xi)} \frac{\partial \psi}{\partial q_k}(q_k^{(\xi)})}{\sum_{x=1}^{n} q_x^{(\xi)} \frac{\partial \psi}{\partial q_x}(q_k^{(\xi)})} \tag{2.9}$$

となる．この反復公式はBaumの増大変換(growth transformation)[38]と呼ばれるものであり，画像処理での確率緩和や音声認識での隠れマルコフモデルに現れる．式(2.9)は非線形計画問題

$$\begin{aligned} \max \quad & \psi(q) \\ \text{subj.to} \quad & \sum_{k=1}^{n} q_k = 1,\ q_k \geq 0 \end{aligned} \tag{2.10}$$

の反復解法であり，この反復で$\psi(q^{(\xi)})$は単調に増加する[39]．更に式(2.8)から
$q_k^{(\xi+1)}/q_{k'}^{(\xi+1)} = \phi(q_k^{(\xi)})/\phi(q_{k'}^{(\xi)})$を得，従って$q_k^{(\xi)} \geq q_{k'}^{(\xi)}$のとき$q_k^{(\xi+1)} \geq q_{k'}^{(\xi+1)}$となる．従って$q_k$は単調に収束する．

2.2.2 パターンの再構成

心理学においていくつかのモードの入力パターンから別のモードパターンの知覚への誘導が報告されており，また生理学においてはこのモード情報が複数の感覚野からのフィードバックパスを通るトップダウン信号により誘導されるというニューロンの

活動が観測されている[37]．提案モデルはこのようなパターン再構成活動を生じることができることを示す．第iモードの入力データをd_iとし，これが全モードでの再構成パターンf_i $(i=1,...,l)$を誘導するとする．各モードの確率密度は式(2.6)を用いて式(2.1)で表される．再構成の段階ではq_kは入力データd_iから計算される定数である．再構成パターンf_iは最尤推定

$$\arg\max_{f_i} \quad \sum_{k=1}^{n} q_k(\epsilon_i + s_i e^{-a_i\|f_i - r_{ik}\|^2}) \tag{2.11}$$

により計算されるとする．これは

$$\arg\max_{f_i} \quad \sum_{k=1}^{n} q_k e^{-a_i\|f_i - r_{ik}\|^2} \tag{2.12}$$

に簡単化される．式(2.12)をf_iについて微分しその導関数を0とおくと次のようなf_iの反復式を得る:

$$f_i^{(\xi+1)} = \frac{\sum_{k=1}^{n} q_k r_{ik} e^{-a_i\|f_i^{(\xi)} - r_{ik}\|^2}}{\sum_{k=1}^{n} q_k e^{-a_i\|f_i^{(\xi)} - r_{ik}\|^2}} \tag{2.13}$$

この反復公式の収束後のf_iの値が再構成パターンである．q_kは1つの入力モードのd_iから決定され，そのq_kは全モードに伝えられそこでパターンが再構成される．入力d_iと再構成f_iの両方が存在する第iモードではf_iは一般に元の入力d_iと異なる．

2.2.3 EMアルゴリズムによる教師なし学習

統合空間における混合密度は

$$\frac{1}{n}\sum_{k=1}^{n} p(d|k) = \frac{1}{n}\sum_{k=1}^{n}\prod_{i=1}^{l} p(d_i|k) \tag{2.14}$$

となる．m個のマルチモーダル学習データd_{ij} $(i=1,...,j;j=1,...,m)$を用いた受容野の中心r_{ik}の学習を調べよう．他のパラメータϵ_i, s_i, a_iは簡単のため適当な値に固定する．学習は次式に示す学習データの対数尤度の最大化により実行される．

$$\max_{r_{ik}} \quad \sum_{j=1}^{m} \ln \sum_{k=1}^{n}\prod_{i=1}^{l} p(d_{ij}|k) \tag{2.15}$$

この対数尤度のr_{ik}に対する導関数を0とおくとr_{ik}の反復公式

$$r_{ik}^{(\xi+1)} = \frac{\sum_{j=1}^{m} \lambda_{ijk}^{(\xi)} d_{ij} e^{-a_i \|d_{ij} - r_{ik}^{(\xi)}\|^2}}{\sum_{j=1}^{m} \lambda_{ijk}^{(\xi)} e^{-a_i \|d_{ij} - r_{ik}^{(\xi)}\|^2}} \tag{2.16}$$

を得る．ここで

$$\lambda_{ijk}^{(\xi)} = \frac{\prod_{i' \neq i}^{l} p(d_{i'j}|k)}{\sum_{x=1}^{n} \prod_{i=1}^{l} p(d_{ij}|x)} \tag{2.17}$$

であり，$p(d_{ij}|k)$は式(2.6)のd_iをd_{ij}に，r_{ik}を$r_{ik}^{(\xi)}$に置き換えたものである．学習の各段でq_kの値は反復(2.9)により計算される．

EMアルゴリズムについては付録BとCを参照されたい．

2.3 マガーク効果の実験

上述の性質を調べるために，典型例としてマガーク効果の簡単な実験を行った．学習に用いたデータを図2.2に示す．これはマガーク効果を説明できる最も簡単なデータであり，フィードバックなしのモデル[31]のシミュレーションで用いられたものである．聴覚の空間はmode1と書いた横軸の1次元で表され，視覚の空間もmode2と書いた縦軸の1次元で表される．音声のクラス数は3で“ga”,“da”,“ba”である．McGurk[25]はそれらの位置関係を図2.2のように推測した．左下の黒点が“ga”，中間の黒点が“da”，右上の黒点が“ba”を示す．実際のデータはこれらの点の付近に分布するがマガーク効果を説明するにはこの簡単なデータで十分である．この簡単なデータを用い，パラメータを$\epsilon_1 = 1, a_1 = 1, s_1 = 8, \epsilon_2 = 1, a_2 = 0.1, s_2 = 10, \beta = 0.5$として各クラスの中心$r_{ik}$を学習する．$a_1$と$a_2$の違いはデータの分散が聴覚よりも視覚のほうが大きいことを表す．すなわち音の単一モードでの知覚において聴覚は視覚よりも優れた性能をもつ．$\epsilon_1 = \epsilon_2 = 1$は大きなノイズレベルに対応するとする．学習後の$r_{ik}$は図2.2の白点で示される．これらの$r_{ik}$によりデータの3つのクラスを正しく識別できる．

マガーク錯覚では“ba”の値がmode1に入力され，“ga”の値がmode2に入力される．この入力から計算されるq_kの値を図2.3に示す．左図(a)がフィードバックなしの場合

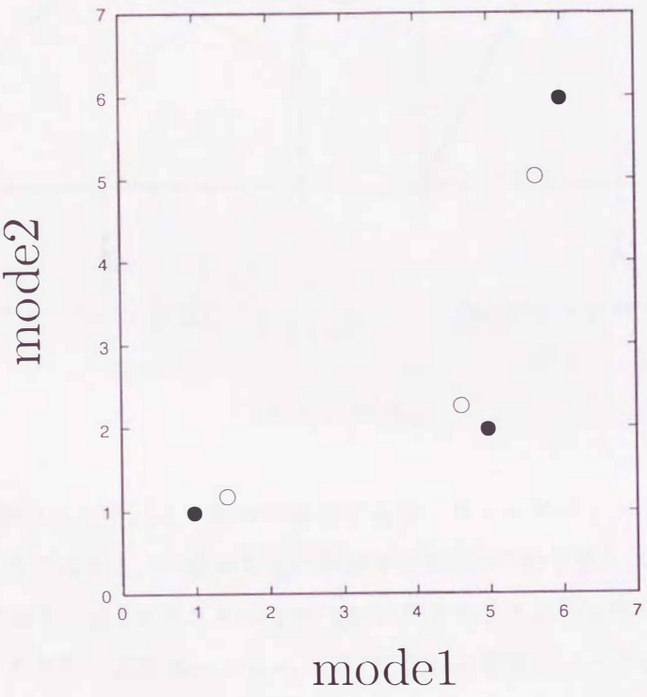
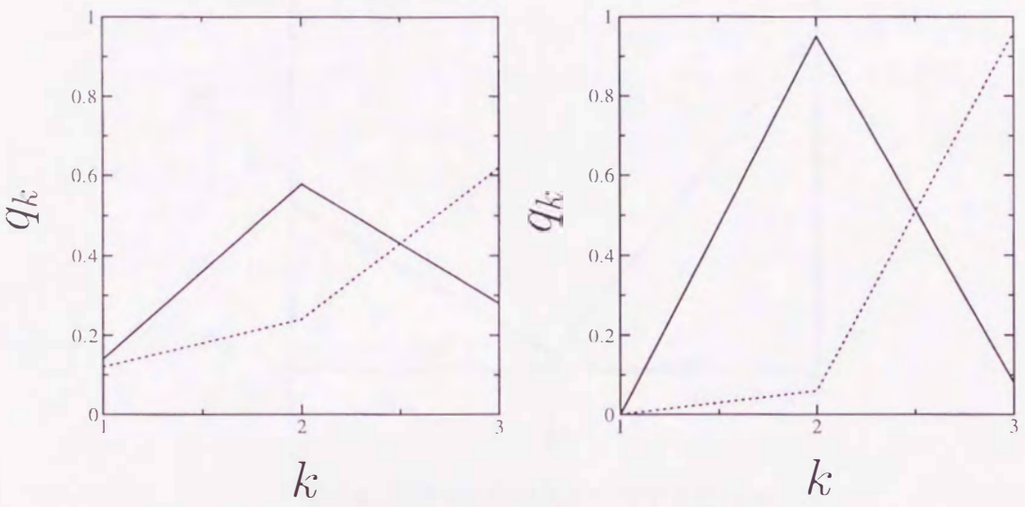


図 2.2: 入力データ (●) と代表点 (○)



(a) フィードバックなし　　(b) フィードバックあり

図 2.3: 出力q_k

で，右図(b)がフィードバックありの場合である．フィードバックによりデータの識別がよりクリスプになる．すなわちq_kの値が0または1に近くなる．図2.3の横軸はクラスの番号であり，第1クラス$k=1$が“ga”，第2クラス$k=2$が“da”，第3クラス$k=3$が“ba”である．実線は$\epsilon_1=1, \epsilon_2=0.1$すなわち聴覚がノイズを含み，視覚のノイズは小さいときの結果である．点線は$\epsilon_1=\epsilon_2=0.1$すなわち聴覚視覚共にノイズが小さいときの結果である．前者の場合はマガーク効果が生じ，視覚入力“ga”と聴覚入力“ba”の間の音“da”が知覚される．しかし後者の場合はマガーク効果は消え，聴覚入力“ba”が知覚される．

この結果は，マガーク錯覚は聴覚ノイズが大きいときに生じ，聴覚がはっきりしているときは生じないこと示す．これは心理学的な観測に一致する[40]．図2.4は聴覚入力が“ga”，視覚入力が“ba”すなわち図2.3の逆の場合の結果である．図2.4に示すように，この入力はマガーク効果を生じない．これも心理実験結果と一致する．

次にいくつかの音の他の組み合せを調べた．2次元の聴覚-視覚空間上の音の配置を

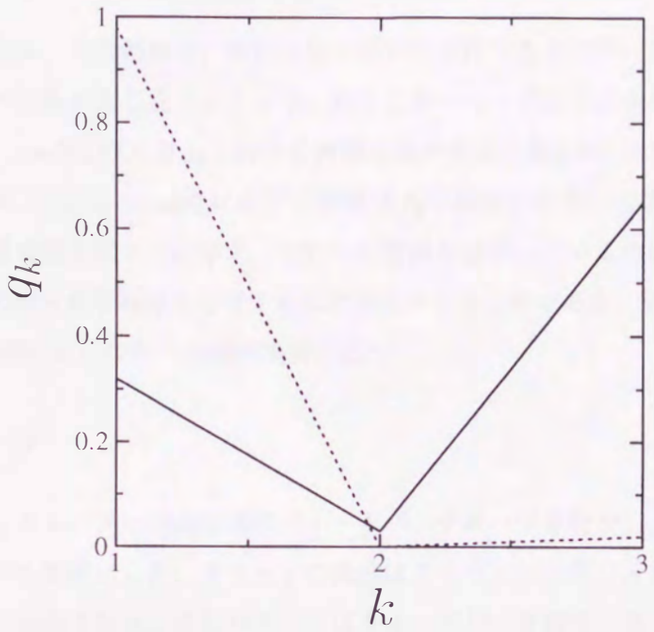


図 2.4: 図2.3と逆の入力に対する出力q_k

表 2.1: マガーク錯覚の例

聴覚	視覚	知覚
ba	ga	da
pa	ga	ta
ma	ga	na
pa	na	ma

図2.5に示すように仮定すると，表2.1に示すマガーク効果の例を再現できる．

2.2.1節で述べたように，q_kの値は識別の段階で時間と共に変化する．その初期値は$q_1 = q_2 = q_3 = 1/3$である．各モード$p(d_i|k)$の出力の初期値を図2.6に示す．$p(d_i|k)$はp_{ik}と略記した．左図はp_{1k} $(k = 1, 2, 3)$，右図はp_{2k} $(k = 1, 2, 3)$である．mode1の入力“ba”により左のグラフのp_{13}が最大となり，mode2の入力“ga”により右のグラフのp_{21}が最大となる．q_kの収束後のp_{ik}の値を図2.7に示す．両方のモードでp_{i2}が最大となっている．すなわちモードの統合によりp_{ik}が修飾されることが分かる．

図2.8にマガーク入力の各モードの再構成値を示す．右下の白い四角が入力であり，黒点は大きな聴覚ノイズのもとでの再構成値を示し，白点は小さいノイズのもとでの

再構成値である．この結果は，識別決定においてだけでなくパターンの再構成においてもマガーク効果が生じることを示す．最後に単一モードの入力からのパターン再構成を調べる．mode1の入力d_1に対する再構成値の変化を図2.9に示す．mode1における再構成f_1を左図に，mode2における再構成f_2を右図に示す．同様にmode2の入力d_2からの再構成値を図2.10に示す．これらの曲線が曲がっているのは，r_{ik}が3個だけでモード間の滑らかな写像を学習するには少なすぎるためである．学習データd_{ij}と中心r_{ik}の数が増えるにつれて曲線は直線に近づく．

2.4 むすび

マルチモーダルパターン識別器にフィードバックループを付加し，その教師なし学習アルゴリズムを導出した．各モードの識別はフィードバックパスを通して高位の統合決定により修飾される．またパターンはフィードバック信号に基づき各モードで再構成される．簡単なモデルを用いてマガーク錯覚に対するフィードバックの効果を調べた．フィードバックループを付加した提案法は，時系列あるいは空間的に分布したパターンにおいて各時点あるいは空間的位置をモードとみなすことによりパターン認識における文脈の影響を扱うことができる．次章ではこのような応用について述べる．

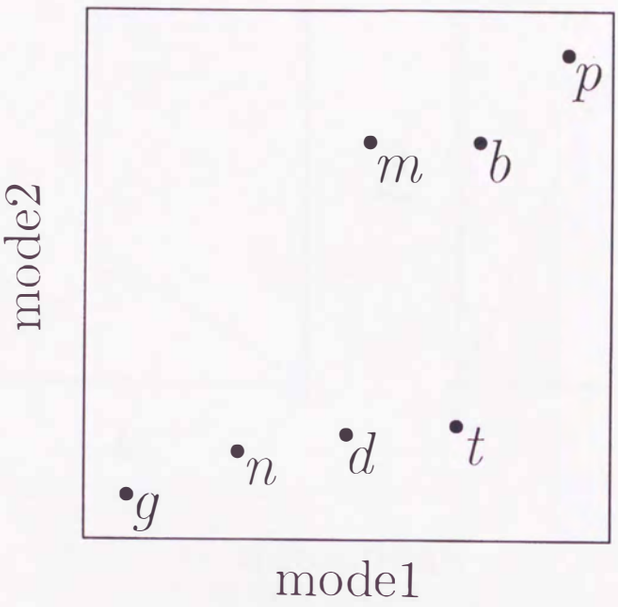


図 2.5: 聴覚-視覚空間における音の配置

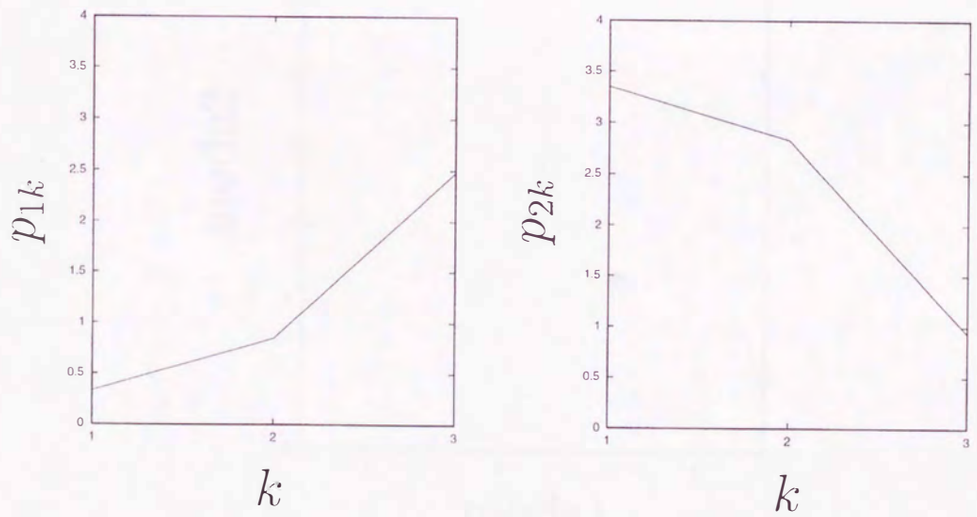


図 2.6: p_{ik}の初期値

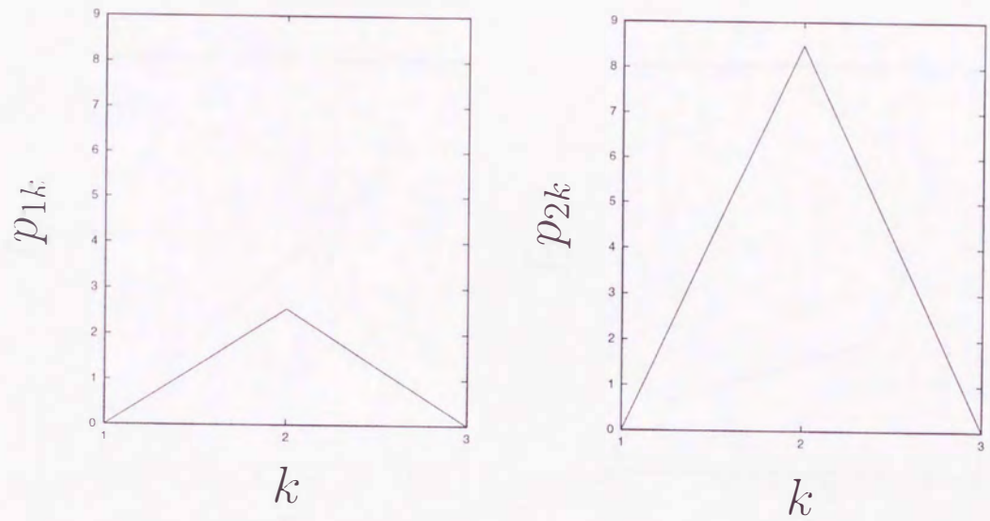


図 2.7: p_{ik}の収束値

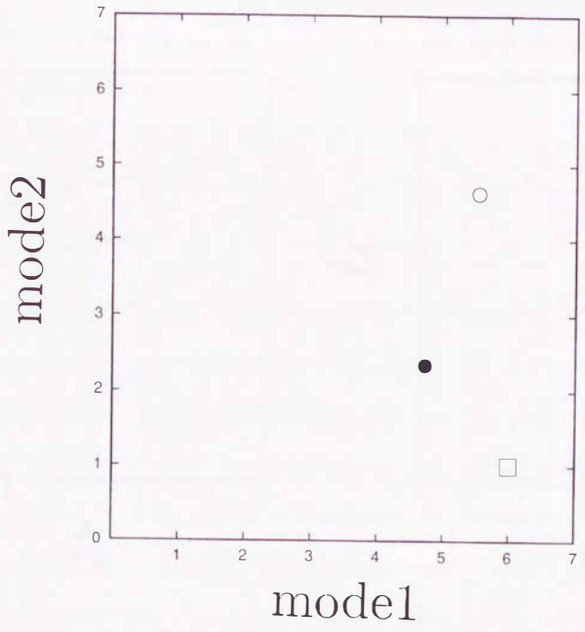


図 2.8: マガーク入力からの再構成値

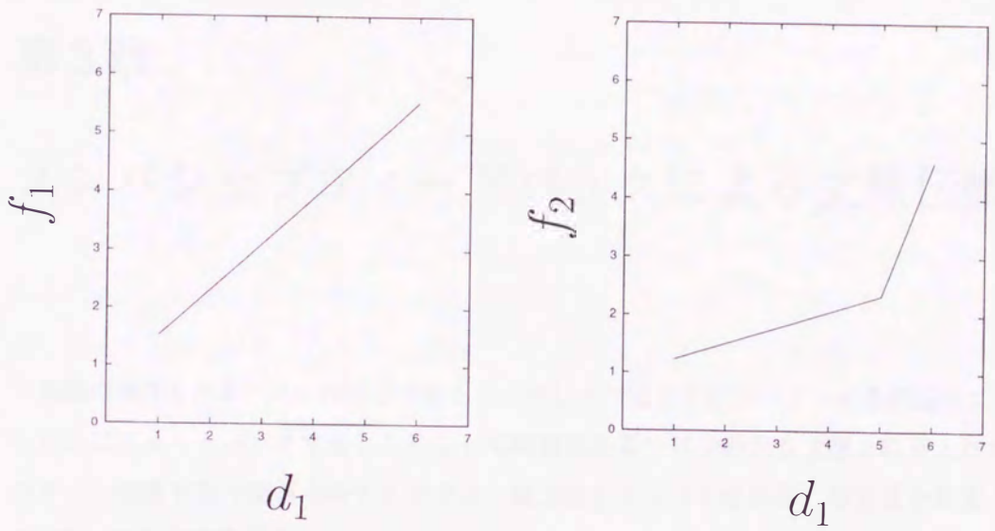


図 2.9: d_1からの再構成

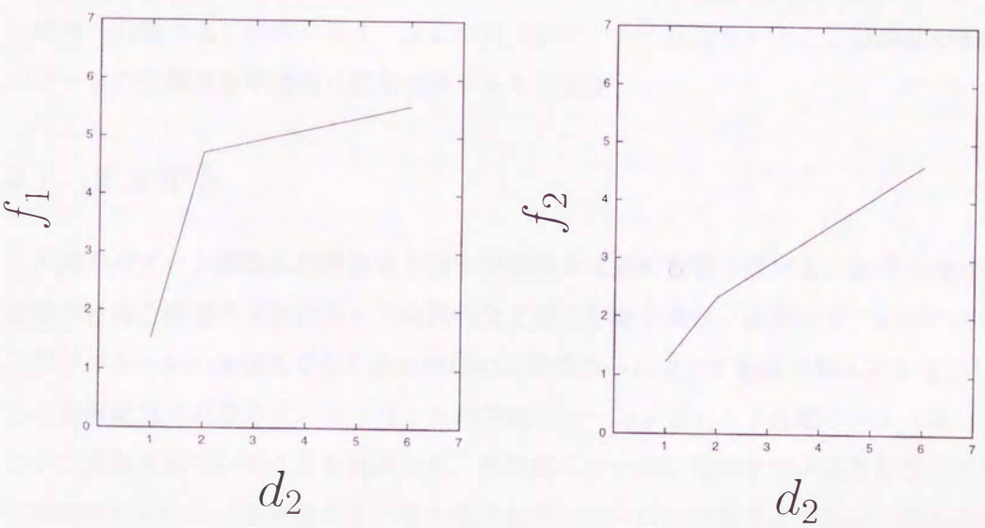


図 2.10: d_2からの再構成

第3章

メンバシップフィードバックによる文脈伝搬

高位のWTAニューロンの出力であるメンバシップ値を下位のパターン選択応答ニューロンにフィードバックすることによって時間的あるいは空間的な文脈を取り入れたパターン認識を行う簡単なモデルを考え，最尤推定に基づく教師なし学習法を提案する[42]．本章で提案するモデルはトップダウンで与えられる文脈情報と入力データから得られるボトムアップ情報とをそれぞれモードと考えると第2章のフィードバック付マルチモーダルパターン認識器の応用として捉えられる．まず時間文脈について1時刻前の識別結果が次の時刻にフィードバックされる例を考え，パターンの類似度でなく提示時刻の近接性によってクラスタリングされることを示し，位置不変なパターン認識への簡単な応用例を示す．次に空間文脈について同様なパターン認識法が画像パターンの空間的な平滑化に応用できることを示す．

3.1 まえがき

人間のパターン認識は時間的な文脈や空間的な文脈の影響を受ける．酒井ら[24]は図形の対連合課題の実験によって時間的な文脈の影響を調べ，図形パターンのペアリングがパターンの類似度でなく提示時刻の近接性のみによって形成されることを示し，この連合記憶は対符号化ニューロンと対想起ニューロンという2種類のニューロンによって表現されていることを見出した．対想起ニューロンはパターン入力がなくても文脈情報のみによって興奮する．また低次視覚ニューロンの応答は受容野の外の刺激にも影響され，空間的な文脈効果を示す[41]．充填(フィルイン)現象は入力がない場所でも周囲の文脈情報によって応答が生じることを示す．これらの生理及び心理学の知見に基づいて，時間文脈を活用した変形に不変なパターン認識器の学習がモデル化

されている [20, 21, 43].

本章でも文脈効果の簡単なモデルを考える．空間的な文脈はニューロン間の長距離ラテラル結合とフィードバック結合の両方で伝達されているようであるが，ここでは時間文脈と空間文脈の両方についてフィードバックの効果だけについて考える．すなわちパターン認識器にとって文脈情報はトップダウンで与えられ，入力パターンからのボトムアップ情報を修飾するとする．各情報を1種のモードと捉えれば，これはマルチモーダルパターン認識の1種と考えることもできる(トップダウンとボトムアップの2モード)．前章ではマルチモーダルな最近傍パターン認識を考え教師なし学習法を提案した．本章ではそれを応用して文脈を取り入れたパターン認識と学習法を提案する．まず時間文脈について1時刻前の識別結果が次の時刻にフィードバックされる例を考え，パターンの類似性でなく提示時刻の近接性によってクラスタリングされることを示し，場所不変なパターン認識への簡単な応用例を示す．次に空間文脈について同様なパターン認識法が画像パターンの空間的な平滑化(補間(フィルイン)を含む)に応用できることを示す．

3.2 時間近接性によるグルーピング

データ d の分布を混合密度

$$p(d) = \sum_{i=1}^{m} p(i)p(d|i) \tag{3.1}$$

で表す．$p(i)$ は第 i クラスタの事前確率であり，$p(d|i)$ は第 i クラスタでのデータ d の確率密度である．$p(i)$ はトップダウン情報，$p(d|i)$ は入力データによるボトムアップ情報であり，両者の積がとられることは両情報が互いに独立と仮定されていることに相当する．トップダウン(事前)情報がないときは $p(i)$ は一様分布 $p(i) = 1/m$ である．$p(d|i)$ は前報 [31] と同じく一様分布とガウス分布の和

$$p(d|i) = \epsilon + se^{-a\|d-r_i\|^2} \tag{3.2}$$

と仮定する．ϵ, a, s は正定数，r_i は第 i クラスタの代表点である．トップダウン情報 $p(i)$ だけでも識別出力が出るためには $\epsilon \neq 0$ が必要である．

次にこれら m 個のクラスタを $n(\leq m)$ 個のグループに分ける．すなわち2段の階層クラスタリングを行う(グループはクラスタのクラスタである)．第 j グループに含まれる

クラスタの集合をI_jと記す(例えば$m=5, n=3$で$I_1=\{1,2\}, I_2=\{3\}, I_3=\{4,5\}$など).

3.2.1 パターン識別

第iクラスタの事後確率$p(i|d)$は$p(i)p(d|i)/\sum_{k=1}^{m} p(k)p(d|k)$であるから，あるデータ$d$が所属するクラスタは

$$\arg\max_{i} \quad p(i)p(d|i) \tag{3.3}$$

で判定され，同様に第jグループの事後確率は$\sum_{i\in I_j} p(i)p(d|i)/\sum_{k=1}^{m} p(k)p(d|k)$であるから，データ$d$は

$$\arg\max_{j} \quad \sum_{i\in I_j} p(i)p(d|i) \tag{3.4}$$

のグループjへ所属すると識別される．このmaxをファジー化してsoftmaxにするとdの第jグループへの所属度(メンバシップ)は

$$q(j)=\frac{e^{b\sum_{i\in I_j} p(i)p(d|i)}}{\sum_{k=1}^{n} e^{b\sum_{i\in I_k} p(i)p(d|i)}} \tag{3.5}$$

と表される．bは正定数である．

以上には時間は入っていない．ここではデータdが1つずつ時系列として入力される場合を考える．時刻tでの入力を$d^{(t)}$と記す．そして時刻tでの事前確率$p(i)$として1時刻前のメンバシップ$q^{(t-1)}(j)$を使うことにする．すなわち第jグループに含まれる全てのクラスタ$i\in I_j$について$p^{(t)}(i)=q^{(t-1)}(j)$とする．そうすると式(3.5)は

$$q^{(t)}(j)=\frac{e^{b\sum_{i\in I_j} q^{(t-1)}(j)p(d^{(t)}|i)}}{\sum_{k=1}^{n} e^{b\sum_{i\in I_k} q^{(t-1)}(k)p(d^{(t)}|i)}} \tag{3.6}$$

となる．これが時刻tでの識別出力である．この式はマルコフモデルの状態遷移式とみれる．以上をニューラルネットで表すと，例えば$m=5, n=2$で$I_1=\{1,2,3\}, I_2=\{4,5\}$の場合図3.1のようになる．最下位の“R”と記したニューロンは式(3.2)を出力するRBF(radial basis function)ニューロンである．ϵは自発応答，sは入力刺激強度(すなわち入力がないときは$s=0$)，r_iは受容野の中心である．“×”と記したニューロンは2つの入力の積を出力する．同様にその上の“+”ニューロンは入力の和を出力する．

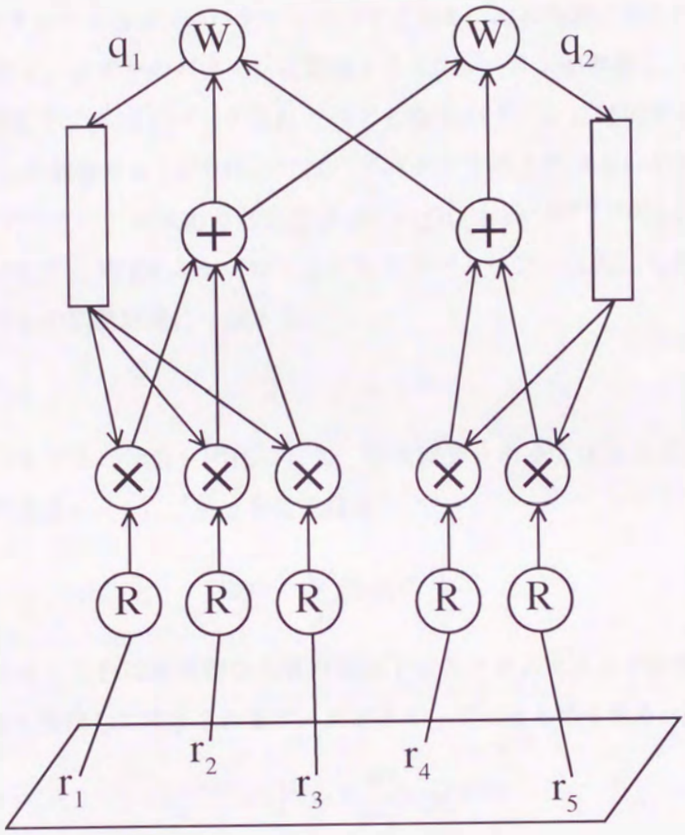


図 3.1: 時間文脈伝搬ネットの例

最上段の“W”と記したニューロンは式(3.5)を出力するファジーなWTA(winner take all)ニューロンである．フィードバックの長方形は1時刻の遅れを表す．

このニューラルネットは酒井ら[24]のパターン対の記憶の実験結果と似た振る舞いをする．図3.1の×印のニューロンが対想起ニューロンに対応し，最上段のWTAニューロンが対符号化ニューロンに対応するものとする．酒井らの実験ではグループの数は12個，各グループのクラスタの数は2個，各クラスタは1個のパターンだけからなる．すなわち各グループは2つのパターンのペアである．ある時刻にあるパターンが図3.1に入力されると，まずそのパターンに対応する×ニューロンが興奮し，$q(j)$が生じ，それが1時刻遅れてフィードバックされてペアとなるパターンに対応する×ニューロンが値$q^{(t-1)}(j)\epsilon$で興奮する(この時点ではペアパターンの入力はないので$s=0$である)．そしてペアパターンが入力され応答は$q^{(t-1)}(j)(\epsilon+se^{-a\|d^{(t)}-r_i\|^2})$に増える．またこのペアに対応するWTAニューロンはどちらのパターンの入力にも興奮する．以上の動作は酒井らの観測結果に一致する．

3.2.2 学習

隠れマルコフモデルの学習と同様にして，時系列データ$d^{(t)}$ $(t=1,2,...)$を使って各クラスタの代表点$r=[r_1,...,r_m]$を最尤推定

$$\max_{r} \quad \sum_{t} \ln p(d^{(t)}) \tag{3.7}$$

によって学習する．これは時間的な文脈の制約下でのクラスタリングの教師なし学習であり，時間的に隣接して提示されるデータがグループにまとめられる．r_iの学習則は

$$r_i^{(t+1)} = r_i^{(t)} + h\frac{\partial}{\partial r_i}\ln p(d^{(t)}) \tag{3.8}$$

とする．hは微小な正定数である．$p(d^{(t)})=\sum_{j=1}^{n}\sum_{i\in I_j} q^{(t-1)}(j)(\epsilon+se^{-a\|d^{(t)}-r_i^{(t)}\|^2})$であるから式(3.8)は

$$r_i^{(t+1)} = r_i^{(t)} + h\frac{q^{(t-1)}(j(i))}{p(d^{(t)})}e^{-a\|d^{(t)}-r_i^{(t)}\|^2}\cdot(d^{(t)}-r_i^{(t)}) \tag{3.9}$$

となる．ここで$j(i)$はこのiを含むグループすなわち$i\in I_j$であるjである．

以上のような学習を行うと，$q(j)$は時間的に近接したクラスタをグループにまとめるようになる．$p(d|i)$は式(3.2)であるから特徴値dが近いデータがクラスタを構成す

るが，それらのクラスタのグルーピングは時間近接性だけによってなされる．従って，いくつかの基本パターンがあってそれらが各々位置，大きさ，回転などの変形を受けたパターンが次々提示されるとき，変形が時間的に連続して生じるなら，ある時間区間ではある基本パターンが変形しつつ連続して提示され，引き続いて別の基本パターンが連続して提示されるということを繰り返すので，各基本パターンがグループを形成することになり，変形に不変なパターン認識器が学習できる．

3.2.3 実験例

まず最初に時間的な近接性がグルーピングに影響することを図3.2のような2次元データ(中央に点のある白丸)で検証した．左上部のデータと右上部のデータが交互に50回提示された後，左下部のデータと右下部のデータを交互に50回提示するというのを繰り返す．各部分内ではデータはランダムに選ばれる．学習ではクラスタは2個，各クラスタがそのままグループ，すなわちグループも2個とした．時間文脈を考えずデータの値だけでクラスタリングした場合，代表点は左右の黒長方形となり，データは左右2つに分割されるが，時間文脈を入れて学習すると上下の黒菱形の代表点となり，時間の近接性に従ってデータは上下2つに分割された．この例は，学習を2段階に分けて，まず代表点を文脈なしのクラスタリングで求めて次に文脈によって上層のグルーピングを行うという逐次法では最適解が得られない場合があることを示している．なおパラメータ値は次の例と同じである．

次にグルーピングが入り組んでいる例として図3.3の2次元データで学習してみた．グループは2個でそれぞれのグループは4個のクラスタからなる．白抜きの水平の正方形が第1グループのデータ，斜めの正方形が第2グループのデータで，各グループのデータを200個ずつ交互に提示した．データは提示のつど各グループのなかからランダムに選んだ．代表点は最初この範囲の中にランダムに配置した．黒の正方形が学習で得られたそれぞれのグループの代表点である．このようにデータはデータ値の近接度ではなく提示時刻の近接性によって2つのグループにまとめられた．パラメータ値は$\epsilon = 0.01, s = 1, b = 1, h = 0.1$とした．$a$はガウス分布の分散の逆数であり，最尤推定によって求めることもできるが，そのためには式(3.2)などでは省略している規格化定数を解析的に求める必要があり，今の場合ϵがあるので困難である．そこでここではアニーリングをした．最初aは0.015として各グループのデータ200個ずつ計400個を

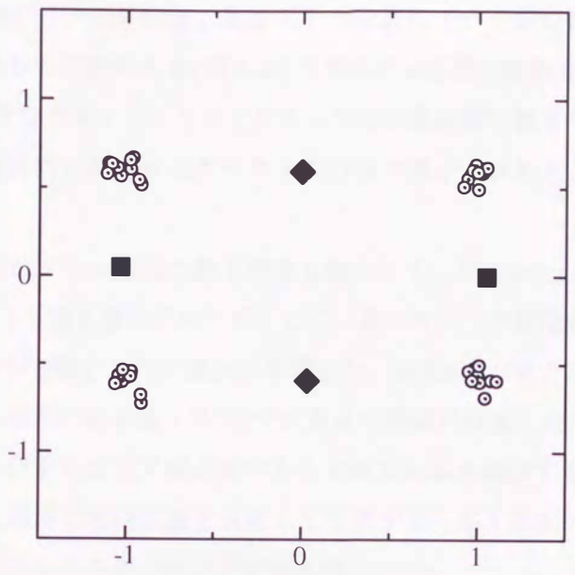


図 3.2: クラスタリングの例

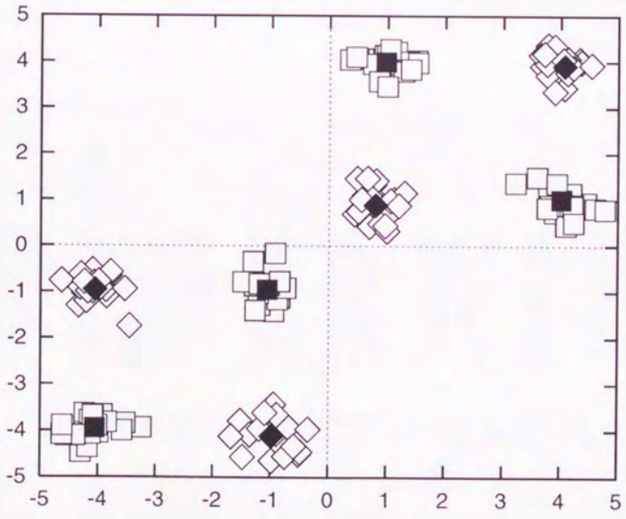


図 3.3: 時間文脈クラスタリング

提示する毎に1.1倍していった．このようにすると最初はaが小さいので図3.4のように代表点はまず全データの平均値に集まりその後別れていくので初期配置にほとんどよらない結果を得ることができる(図3.4は代表点のx座標の変化である．y座標も同様な動きをする)．すなわちアニーリングによって局所最適解に捕まりにくくなる．識別で使うaの値は最終的に得られたクラスタの分散の値から決めた．この例では$a = 5$となった．

次に位置不変なパターン認識の最も簡単な例として，図3.5の上5個の11×11画像を第1グループ，下5個を第2グループとして，各グループの画像を150個ずつ交互に提示して各グループ3個ずつの代表点を学習した．図3.6に示すランダムな代表点の画像から出発して，学習の結果図3.7に示す代表点の画像に収束した．この6個の代表点によって棒の提示位置によらず棒が縦であるか横であるか識別することができる．この場合も上の例と同様に時間文脈を無視して学習すると第1グループの代表点が横棒の画像に収束したりしてグルーピングが行えない．

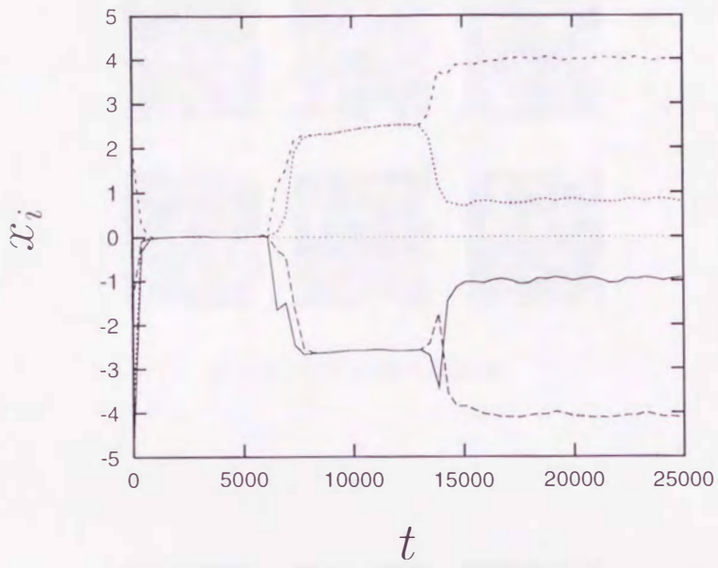


図 3.4: 第1グループの代表点の収束の様子

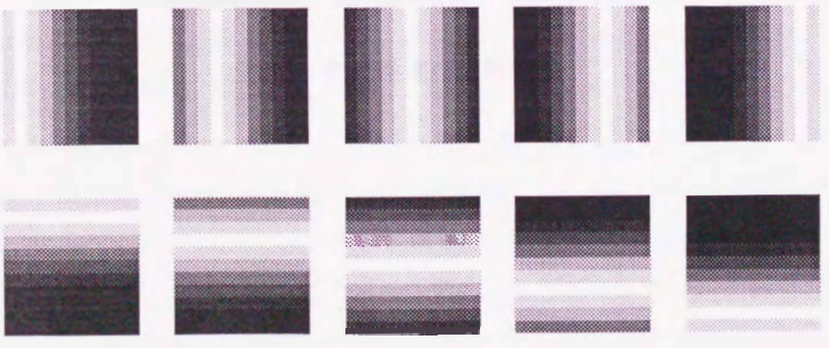


図 3.5: 学習用画像データ

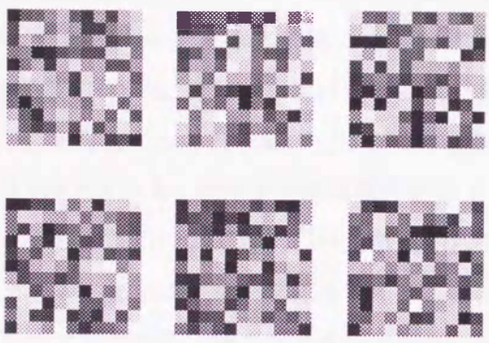


図 3.6: 代表画像の初期値

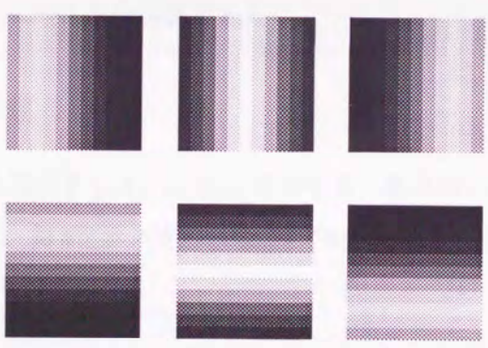


図 3.7: 代表画像の学習結果

3.3 空間伝搬による緩和整合化

以上では時間的な文脈の影響の例を考えた．今度は空間的な文脈の効果の例を考える．統計論的な基本は前節と同じである．前節では入力データは時間的に分布していた(すなわち時間の関数$d(t)$)のに対し，今度は空間的に分布する．m個のデータ$D_i = (x_i, d_i)$ $(i = 1, ..., m)$が与えられるとする．x_iは空間の場所，d_iは特徴値である．記述の簡単のため空間は1次元とする．データ全部の集合を$D = \{D_1, ..., D_m\}$とする．このとき各場所のデータそれぞれについて各データが所属するクラスタを推定することを考える．ここではクラスタを更にグループにまとめることはしない．すなわち各クラスタがそのままグループでもある．各クラスタの代表点r_iは前節のように学習で求めるべきであるが，ここでは単純に特徴空間の中に一様に等間隔にとる(一様分布のトレーニングデータで学習すればこのような代表点配置になる)．またクラスタを推定する場所は必ずしもデータが与えられた場所でなくてもよい．すなわちデータがない場所でもクラスタの推定を行うものとする(これも前節の時間文脈のときと同じであり，フィルインなどはこのような状況である)．

3.3.1 クラスタの推定

クラスタを推定する場所をy_j $(j = 1, ..., l)$とする．簡単のためy_jは等間隔とする．クラスタの数をnとし，第kクラスタの代表点の特徴値をf_kとする．そして場所y_jの第kニューロンの応答を

$$p_j(D|k) = \epsilon + \sum_{i=1}^{m} s_i e^{-a\|x_i - y_j\|^2 - b\|d_i - f_k\|^2} \tag{3.10}$$

とする．(y_j, f_k)はこのニューロンの受容野の中心であり，s_iはデータ$D_i = (x_i, d_i)$の強度である．前節との違いは受容野が特徴軸だけでなく空間方向にも広がっていることである．従ってこのニューロンは既に空間文脈をある程度取り入れている．しかしその範囲は受容野に限られ，受容野の外の文脈は取り入れられない．これを第1層のニューロンとし，次に第2層のニューロンの応答を事前確率$p_j(k)$と式(3.10)との積とし，これを

$$p_j(k)p_j(D|k) = \frac{q_{j-1}(k) + q_{j+1}(k)}{2} p_j(D|k) \tag{3.11}$$

とする．すなわち場所jでの第kクラスタの事前確率$p_j(k)$を隣接する場所$j-1$と$j+1$のメンバシップ$q_{j-1}(k)$と$q_{j+1}(k)$の平均値とする．このメンバシップは

$$q_j(k) = \frac{e^{c[q_{j-1}(k)+q_{j+1}(k)]p_j(D|k)}}{\sum_{s=1}^{n} e^{c[q_{j-1}(s)+q_{j+1}(s)]p_j(D|s)}} \tag{3.12}$$

で与えられる．cは正定数である．以上のニューラルネットを図示すると図3.8となる．ただしこの図の横軸は空間yである．これと直交して特徴軸fがあり，従って図3.8の構造が紙面に直交してn層重なっている．最下位のニューロンは式(3.10)を出力するRBFニューロンである．その上の+ニューロンは2つの入力の和を出力し，その上の×ニューロンは2つの入力の積を出力する．ここまでのニューロンには特徴軸方向の結合はない．最上段は式(3.12)の$q_j(k)$を出力するWTAニューロンであり，これは特徴軸方向に抑制性の結合をしている．図3.8を見て分かるようにこのフィードバックによりRBFニューロンの受容野の外へも情報が次々と伝搬していく．この伝搬は式(3.12)を次の反復法で解くことにより実現される．

$$q_j^{(t)}(k) = \frac{e^{c[q_{j-1}^{(t-1)}(k)+q_{j+1}^{(t-1)}(k)]p_j(D|k)}}{\sum_{s=1}^{n} e^{c[q_{j-1}^{(t-1)}(s)+q_{j+1}^{(t-1)}(s)]p_j(D|s)}} \tag{3.13}$$

ここでtは反復回数のカウントであり，時刻でもある．この反復が収束した$q_j(k)$が式(3.12)の解である．反復の初期値は一様な値$q_j^{(0)}(k) = 1/n$とする．この反復は画像処理で使われる確率緩和法[44]によく似ている．この反復の収束性について調べる．$q = [q_j(k)]$ $(j = 1, ..., l; k = 1, ..., n)$として$E(q) = \sum_j \sum_k e^{c[q_{j-1}(k)+q_{j+1}(k)]p_j(D|k)}$ $\ln q_j(k)$とすると$E(q)$は$q_j(k)$それぞれの単調増加関数であり，式(3.13)は

$$q_j^{(t)}(k) = \frac{q_j^{(t-1)}(k)\dfrac{\partial E}{\partial q_j(k)}(q^{(t-1)})}{\sum_{s=1}^{n} q_j^{(t-1)}(s)\dfrac{\partial E}{\partial q_j(s)}(q^{(t-1)})} \tag{3.14}$$

と書ける．式(3.14)の形の反復公式はBaum[38]の増大変換(growth transformation)と呼ばれるものであり，画像処理での確率緩和や音声認識の隠れマルコフモデルなどに現れる．式(3.14)は非線形計画問題

$$\begin{aligned} \max_{q} \quad & E(q) \\ \text{subj.to} \quad & \sum_{k=1}^{n} q_j(k) = 1,\ q_j(k) \geq 0 \end{aligned} \tag{3.15}$$

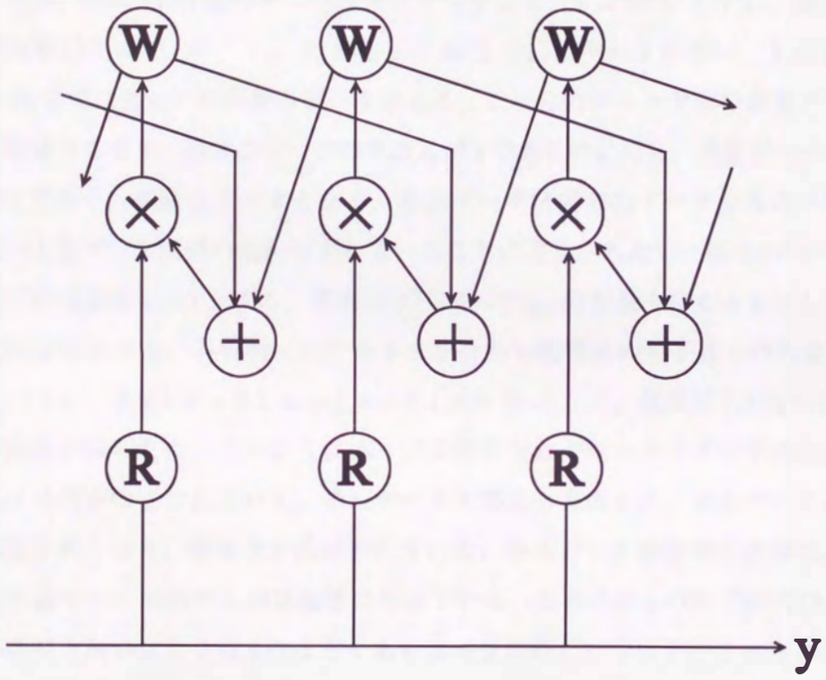


図 3.8: 空間文脈伝搬ネットの例

の反復解法であり，式(3.14)の反復では$E(q)$は単調に増加する[39]．従って式(3.14)の反復は式(3.15)の局所最適解に収束する．

3.3.2 実験例

以上のようにこの空間文脈効果は確率緩和法に似ており，平滑化による空間的整合化や曖昧さの低減化などの効果を示すと予想される．このことをまず図3.9に示す簡単な空間1次元，特徴値1次元のデータ例で実験してみた．このデータは全体的にガウスノイズを含み，$x=9$に1個のインパルスノイズがある．$x=20$および$x=50$ではデータが不連続にとんでおり，$x=30$から$x=40$までにはデータがない．また$x=60$から$x=70$までは各xで複数個のデータがある．これらのデータ欠落や多重データはデータの曖昧さを表す．通常のデータの重みs_iが1であるのに対し，多重データでは多重度分の1であり，曖昧なデータとなる．多重データ区間でのデータ分布の平均値はその左右の1重データ区間の値からずれていることに注意されたい．図3.9のデータについて上記の方法で$q_j(k)$を求め，各場所y_jにおいて$q_j(k)$が最大になるkのf_kをそこでの特徴値出力とする．各場所でのクラスタすなわち特徴値の量子化レベル数は100とした．パラメータ値は$\epsilon=0.1, a=1, b=0.1, c=30$とした．反復は7,8回で収束し，図3.10の結果が得られた．このようにエッジは保存されガウスノイズが平滑化されインパルスノイズが除去されている．またデータ欠落部も充填され，多重データ部分も一意に確定されており，曖昧さが低減されている．多重データ部分の平滑値は多重データの平均値でなく両側からの延長値になっている．なおこのaの値では式(3.10)の受容野の空間方向の広がりは4点ほどでありこの受容野だけではデータ欠落部はカバーできない．従って図3.10のようにこの欠落部が充填されたのは$q_j(k)$の伝搬のためである．多重データ部も同様である．従って欠落部分や多重データ部分が広くなると$q_j(k)$の収束時間も長くなる．

このような多重データによる曖昧さはアパーチャ問題として運動知覚やShape from X計算などで現れる．例えばバーバーポール錯視では線分の両端ではデータは一意に確定しているが内部では多重データとなっており平均は線分に垂直な運動方向であり，両端の運動方向とは異なる．図3.9の多重データも同様な分布をしている．次に多重データの別の例として図3.11のようなランダムドットステレオ視の実験をしてみた．この例は空間2次元で，特徴値は画素の左右のずれ，すなわち視差であり1次元である．

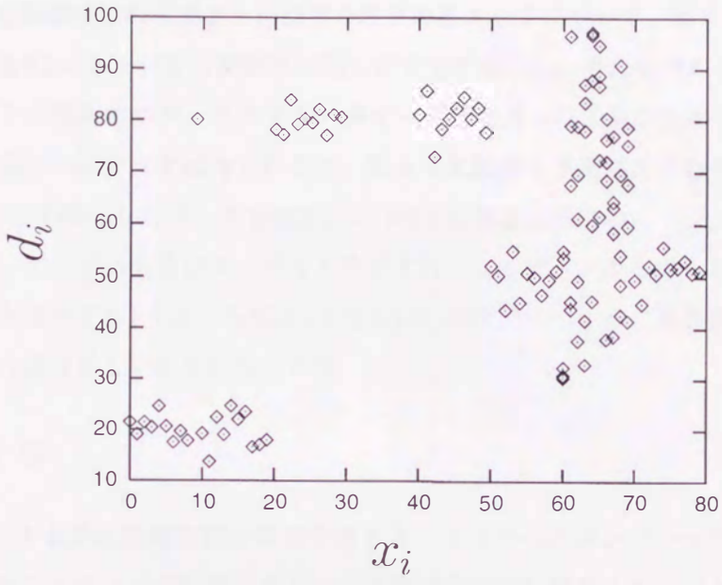


図 3.9: 空間データ例

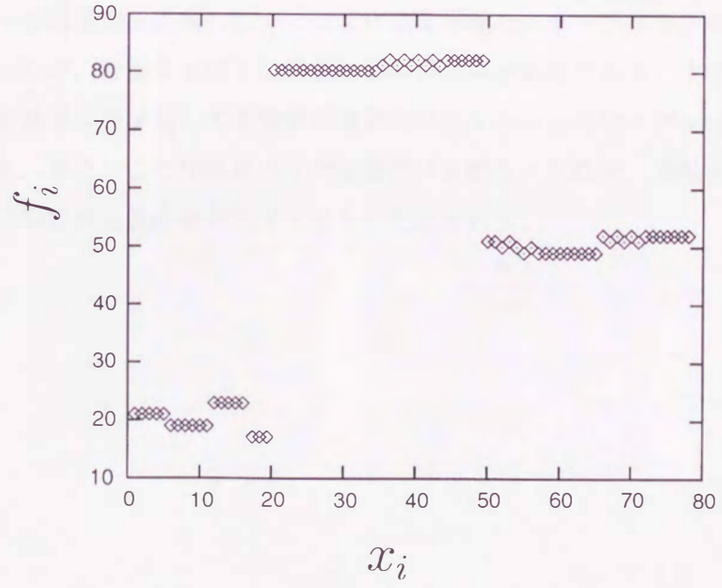


図 3.10: 収束値

まず最初にこの2枚の画像からデータd_{ij}を作る．(i,j)は画素の境界の垂直エッジの場所である．左眼画像での黒画素と白画素の境界の各エッジについて，同じタイプ(黒白か白黒の2通り)のエッジを右眼画像の同じ行で全て列挙し，それらのエッジの左右のずれを全てその場所でのデータとする．各データの重みs_iは多重度分の1とする．黒黒や白白の部分ではデータは作られない．従って欠落部と多重部を含むデータとなる(このようにして得られたデータを図3.12に示す(黒画素は黒白のエッジ，白画素は白黒のエッジ，グレイの画素はエッジなしを示す))．このデータから得られた各場所でのエッジの左右のずれ(すなわち視差)を図3.13に示す．このように境界部で少し乱れがあるもののほぼ正しい結果が得られた．

3.4 むすび

softmaxによる最近傍識別器の出力である各クラスタへのメンバシップ値をフィードバックすることによって時間的あるいは空間的な文脈を取り入れたパターン識別を行う簡単なモデルを考え，時間文脈に基づく教師なし学習によって位置不変なパターン認識が行えることを単純な実験例で示した．また空間的な文脈のフィードバックモデルによって確率緩和法によく似た空間整合化が行えることを例示し，ランダムドットステレオへの応用例を実験した．ここでは最も単純なフィードバックモデルでしか実験しなかったが，性能を上げるには更に多くの拡張が必要である．まず変形に不変なパターン認識では多層化して受容野が階層的に広くなるようなモデルにするのが有効と思われる．またここでは隣接点との文脈だけを取り入れたが，文脈が影響を及ぼす範囲を広げた方が文脈効果が増大するものと思われる．

図 3.11: ランダムドットステレオグラム

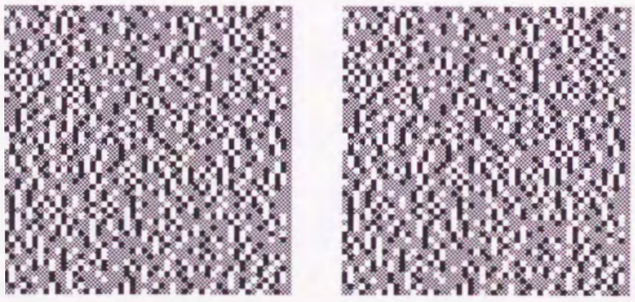


図 3.12: データの欠落部と多重部

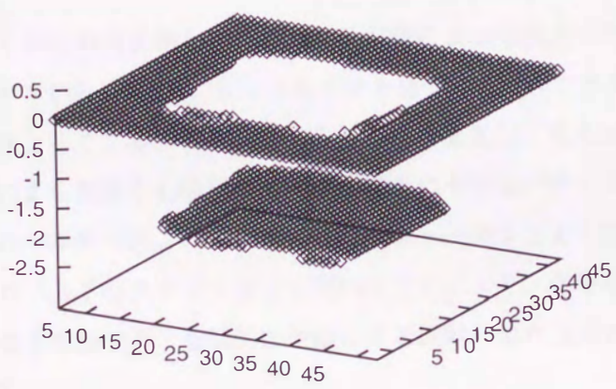


図 3.13: 得られた視差

第4章

時間的な文脈に基づく視点に不変なパターン認識器の学習

前章では，いくつかの代表的な視点の2次元画像によって視点に不変な3次元物体の認識をするモデルであるRBFニューラルネットについて，各クラスへのメンバシップ出力を事前情報としてフィードバックするモデルを提案し，視点が時間的に変化する時系列データによる教師なし学習によって，視点に不変なパターン認識器が構成できることを示した．本章ではこのモデルが顔認識ニューロンとよく似た応答をすることを示し，またロバストなクラスタ分布を用いることにより，学習時や認識時に混入する外れ値の画素を棄却でき，時間的な予測による注視に似た認識処理が得られることを示す[45, 46].

4.1 まえがき

視点に不変なパターン認識の方法として，いくつかの視点での2次元画像の結合によって3次元物体をモデル化する方法がある．Ullmanら[16]は線形結合によるモデルを提示したが，Poggioら[18]のRBF(Radial Basis Function)ニューラルネットによるモデルは一般的な非線形結合モデルである．

RBFネットを図4.1に例示する．この例では物体は2個で第1の物体はr_1, r_2, r_3という3個の代表的な視点の2次元画像の結合でモデル化され，第2の物体はr_4, r_5という2個の代表画像でモデル化される．1番下の平面は入力画像dの空間である．Rと記したニューロンは入力画像dに対し$e^{-a\|d-r_i\|^2}$という応答を出力するRBFニューロンである．その上の+印のニューロンはRニューロンの応答の線形和を出力する．最後のW

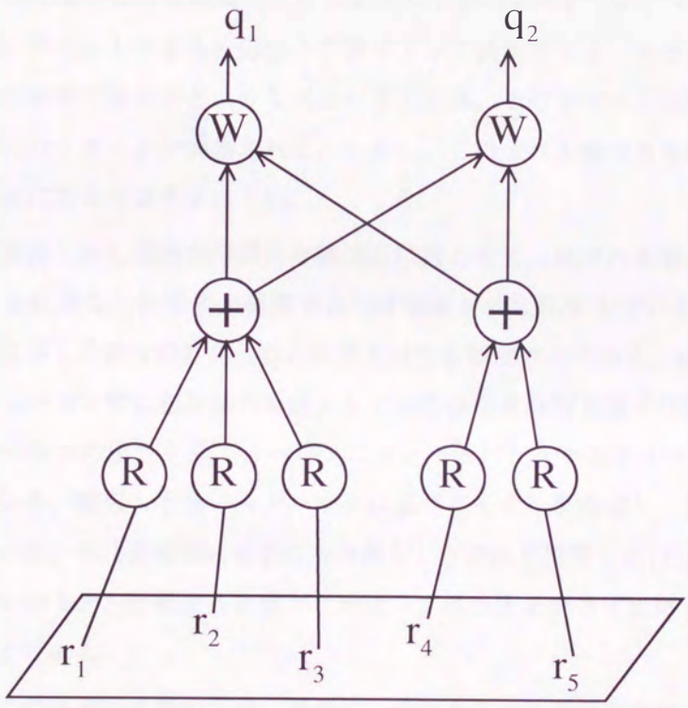


図 4.1: RBF ネット

と記したニューロンはWTA(Winner Take All)ニューロンであり，その下の“+”ニューロンの応答が大きい方のq_jが1になり，もう1方のq_jは0となる．このようなネットワークは心理実験結果とよく似た振る舞いを示し，生理学的にもRニューロンに対応する視点中心ニューロンと，“+”(あるいはW)ニューロンに対応する物体中心ニューロンとが観測されている[19, 47]．図4.2に生理実験データ[19]を示す．横軸は顔の向き(略図に示す8方向)であり，縦軸はニューロンのスパイク頻度である．左が視点中心ニューロン(これは右向きの顔に最も強く反応する)であり，右が物体中心ニューロンの応答である．

Poggioらのモデルでは代表画像r_iとRニューロンから+ニューロンへの結合係数が入力画像のデータセットによる教師付き学習によって調整される．すなわち各入力画像がどの物体の画像であるかというラベルが与えられ，そのラベルに対応するq_jが大きくなるようにパラメータが調整される．しかし，このラベル情報を生成する部分が脳内の別の場所にあるとは考えにくい．

このラベル情報と同じ役割を時間的な連続性に持たせて，視点に不変なパターン認識器を構成する教師なし学習法が提案され[20]発展させられてきている[21, 22]．時間的な文脈に基づく学習は酒井ら[24]の生理実験でも観測されている．RBFネットについてもRニューロン間に相互結合を導入して結合係数を時間文脈で学習するモデル[48]やWニューロンの出力をRニューロンにフィードバックするタイプのモデル[43]が提案されている．筆者らもフィードバックに基づくモデルを提案し，最尤推定に基づいてRニューロンの代表画像r_iを求める教師なし学習法を提案した[45, 46]．本モデルがBecker[43]のモデルと異なる点はゲイティングネットを介さず直接q_jをフィードバックすることである．

本章では，このモデルを顔の認識に適用し，またクラスタの分布をロバストにすることにより注視に似た効果を持たせられることを示す．なお本章の主目的はフィードバックの効果を調べることであるので，画像からの特徴抽出は特には行わず，各画素のグレイレベルをそのまま並べてデータベクトルとする．

また4.2と4.3節の式の導出は第3章の3.2節のそれとほぼ同じであるがPoggioら[18]の標準正則化に基づく関数近似法としてRBFネットとの比較という観点からあらためて議論している．図についても図4.3の時間文脈伝搬ネットは図3.1と同じであるが図4.1のRBFネットとの比較のためあらためて本章にも掲載している．

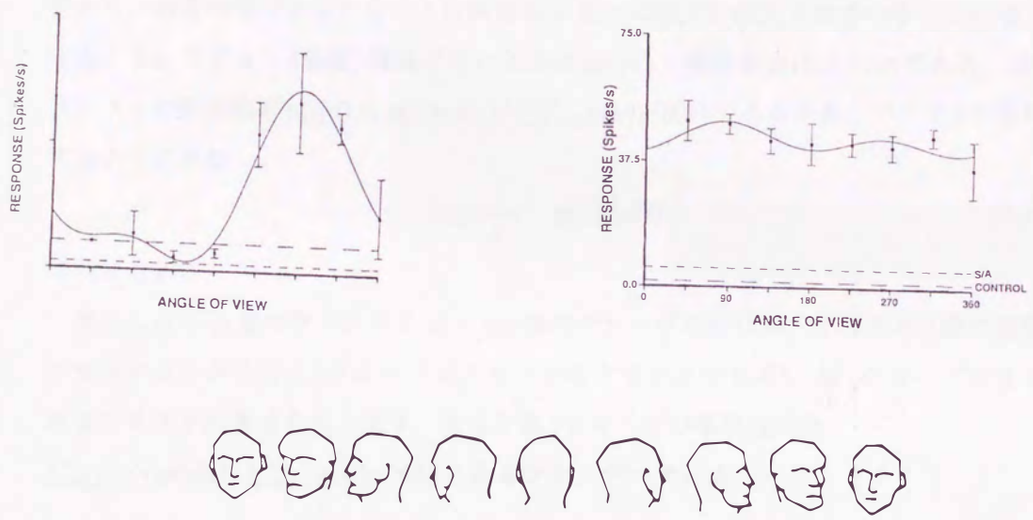


図 4.2: 顔ニューロンの応答

4.2 RBFネット

Poggioら[18]は標準正則化に基づく関数近似法としてRBFネットを導出し，それをパターン識別にも応用したが，ここでは混合密度に基づくベイズ識別則として直接RBFネットを導く．

データdの分布を混合密度

$$p(d) = \sum_{i=1}^{m} p(i)p(d|i) \tag{4.1}$$

で表す．$p(i)$は第iクラスタの事前確率であり，$p(d|i)$は第iクラスタでのデータdの確率密度である．$p(i)$はトップダウン情報，$p(d|i)$は入力データによるボトムアップ情報であり，両者の積がとられることは両情報が互いに独立と仮定されていることに相当する．トップダウン(事前)情報がないときは$p(i)$は一様分布$p(i) = 1/m$である．第iクラスタの事後確率$p(i|d)$は$p(i)p(d|i)/\sum_{k=1}^{m} p(k)p(d|k)$であるから，データ$d$が所属するクラスタは

$$\arg\max_{i} \quad p(i)p(d|i) \tag{4.2}$$

で判定される．

次にこれらm個のクラスタを$n(\leq m)$個のグループに分ける．すなわち2段の階層クラスタリングを行う(グループはクラスタのクラスタである)．第jグループに含まれるクラスタの集合をI_jと記す．すると第jグループの事後確率は
$\sum_{i\in I_j} p(i)p(d|i)/\sum_{k=1}^{m} p(k)p(d|k)$となるから，データ$d$は

$$\arg\max_{j} \quad \sum_{i\in I_j} p(i)p(d|i) \tag{4.3}$$

のグループjへ所属すると識別されることになる．このmaxをファジー化してsoftmaxにするとdの第jグループへの所属度(メンバシップ)は

$$q(j) = \frac{e^{b\sum_{i\in I_j} p(i)p(d|i)}}{\sum_{k=1}^{n} e^{b\sum_{i\in I_k} p(i)p(d|i)}} \tag{4.4}$$

と表される．bは正定数である．例えば$m = 5, n = 2$で$I_1 = \{1,2,3\}, I_2 = \{4,5\}$で，各クラスタのデータ分布がガウス分布$p(d|i) \propto e^{-a\|d-r_i\|^2}$のとき$q_j$は先程の図4.1のRBFネットで計算される．

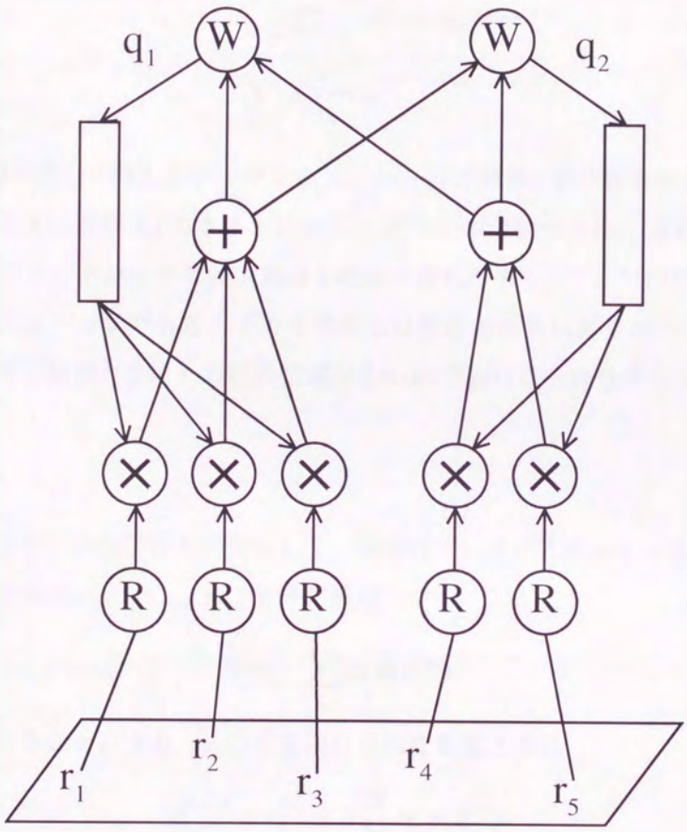


図 4.3: 時間文脈伝搬ネット

これがPoggioらのRBFネットであり，時間は入っていない．すなわち提示データは履歴に無関係にそのときのデータの値のみによって識別される．我々のモデルでは識別がデータの提示履歴にもよるとし，1時刻前の識別結果が現時刻のトップダウン情報となると仮定する．すなわち時刻tでの事前確率$p(i)$として1時刻前のメンバシップ$q^{(t-1)}(j)$を使うことにする．つまり第jグループに含まれる全てのクラスタ$i \in I_j$について$p^{(t)}(i) = q^{(t-1)}(j)$とする．そうすると式(4.4)は

$$q^{(t)}(j) = \frac{e^{b\sum_{i\in I_j} q^{(t-1)}(j)p(d^{(t)}|i)}}{\sum_{k=1}^{n} e^{b\sum_{i\in I_k} q^{(t-1)}(k)p(d^{(t)}|i)}} \tag{4.5}$$

となる．$d^{(t)}$は時刻tでの入力データである．$q^{(t)}(j)$が時刻tでの識別出力となる．この識別則を図示すると図4.1のネットにフィードバックが付加された図4.3のネットワークになる．フィードバックの長方形は1時刻の遅れを表し，“$\times$”印は2つの入力の積を出力するニューロンである．このモデルでは各時刻の識別結果が次の時刻の識別に影響を及ぼす．時刻0ではその前の情報はないので$p(i)$は一様分布とする．

4.3 学習

隠れマルコフモデルの学習と同様にして，時系列データ$d^{(t)}$ $(t = 0, 1, 2, ...)$を使って各クラスタの代表点$r = [r_1, ..., r_m]$を最尤推定

$$\max_{r} \quad \sum_{t} \ln p(d^{(t)}) \tag{4.6}$$

によって学習することにする．r_iの学習則は単純な最急上昇法

$$r_i^{(t+1)} = r_i^{(t)} + h\frac{\partial}{\partial r_i} \ln p(d^{(t)}) \tag{4.7}$$

とする．hは微小な正定数である．$p(d^{(t)}) = \sum_{j=1}^{n} \sum_{i\in I_j} q^{(t-1)}(j) e^{-a\|d^{(t)} - r_i^{(t)}\|^2}$であるから式(4.7)は

$$r_i^{(t+1)} = r_i^{(t)} + h\frac{q^{(t-1)}(j(i))}{p(d^{(t)})} e^{-a\|d^{(t)} - r_i^{(t)}\|^2} \cdot (d^{(t)} - r_i^{(t)}) \tag{4.8}$$

となる．ここで$j(i)$はこのiを含むグループすなわち$i \in I_j$であるjである．

パラメータについては，まずhは厳密には確率近似法に従って徐々に小さくすべきだが，ここでは簡単のため固定する(以下の実験では$h = 0.01$)．次にbは大きすぎて

も小さすぎても悪く，適度な値(すなわち式(4.5)のWTAが適度にファジー)でなくてはならない．大きすぎるとq_jが1か0に近づき文脈が強すぎてパターンが切替っても識別が切替らなくなり，逆に小さすぎると全てのq_jが$1/n$に近づき文脈の影響がなくなってしまう．以下では$b=2$とした．代表点に関する事前情報はないので最初代表点はランダムに配置する．そして局所最適解へのトラップを防ぐためにaの値をアニーリングする．すなわち最初aは十分小さな値にして最終的に十分大きな値になるまで徐々に大きくする．このようにすると最初はaが小さいので代表点はまず全データの平均値に集まりその後別れていくので初期配置にほとんどよらない結果を得ることができる．このアニーリングの性質については付録Dを参照されたい．識別で使うaの値は最終的に得られたクラスタの分散の値から決める．

この学習法では，各時刻でのステップは1時刻前の識別結果が現時刻の教師信号となる教師付き学習とも言えるが，ネットワークの外からは教師信号が与えられないので全体的には教師なし学習である．この学習で得られるのは階層的なクラスタリングであり，2層めのグルーピングが時間的な文脈に基づいて行われ，時間的に隣接して提示されるクラスタがグループにまとめられることになる．すなわち各クラスタはdの値の近接性だけによって構成され，それらのクラスタのグルーピングは時間近接性だけによってなされる．従って，いくつかの3次元物体があってそれらのいろいろな視点からの2次元画像が次々に提示されるとき，ある時間区間ではある物体の画像が連続して提示され，引き続いて別の物体が連続して提示されるということが繰り返されれば(各時間区間のなかでは視点の変化は必ずしも連続していなくてもよい)，各物体の各視点の画像がクラスタを構成し，同じ物体のクラスタがグループを形成することになり，視点に不変なパターン認識器が学習できる．

4.4 実験例

本モデルが位置不変な画像認識を学習できることを簡単な例で前にも示した(3.2節)が，ここでは図4.4に示す3人の向きの異なる顔画像について実験を行った．最初の人の画像をサイクリックに30回提示した後，2番めの人の画像を同様に提示し，次に3番目の人を同様に提示した後，再び最初の人に戻って同じデータ入力を繰り返した．図4.5に示すランダムな代表点の画像から出発して，学習の結果図4.6に示す代表点の画像に収束した．この代表画像により，向きによらない顔の認識ができた．図4.4のデ

図 4.4: 顔画像の例

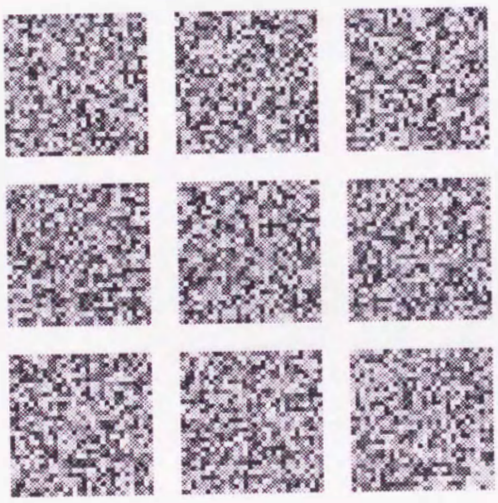


図 4.5: 代表画像の初期値

図 4.6: 代表画像

図 4.7: 文脈なし学習の代表画像

ータでは，同じ人の向きの異なる顔画像間の距離が向きの同じ他人の画像間の距離より大きいので，時間文脈なしでは各人へのグルーピングが行えない．図4.7に文脈なし，すなわち通常のクラスタリングで得られた代表画像を示す．このように文脈なしではグルーピングに無関係に図4.4のデータから9個の代表画像が選ばれることになり，人の識別が行えない．

図4.6の代表画像で図4.4の画像を識別したときの$e^{-\|d-r_2\|^2}$と$q(j)$の値を図4.8に示す．横軸は図4.4と同じdの並びすなわち顔の向きであり，左図の破線，実線，点線はそれぞれ図4.4の上5個，中5個，下5個である．右図の破線，実線，点線はそれぞれ$q(1), q(2), q(3)$である．r_2は図4.6上中央の画像である．図4.8は図4.2とよく似ている．

4.5 ロバスト化

以上では各クラスタの確率密度はガウス分布とした．次に学習データに外れ値が含まれる場合を扱ったり注視に似た機能を持たせるためにロバストな分布を使うことにする．まず外れ値の影響を見るためにガウス分布を使って図4.9のような画像データで学習してみた．右端の画像は左から2番めの画像にノイズが乗ったものであり，データ提示の際に3回のうち1回，2番めの画像を右端の画像に置き換えて提示し学習した．この結果図4.10のように外れ値も入った代表画像が得られた．これはガウス分布では代表画像はクラスタに含まれる画像の平均の画像になるからである．

データ画像dと代表画像r_iの第k画素をそれぞれd_k, r_{ik}とすると，ガウス分布では

$$p(d|i) \propto e^{-a\|d-r_i\|^2} = \prod_k e^{-a(d_k-r_{ik})^2} \tag{4.9}$$

である．これは各画素の識別結果のANDが画像全体の識別結果となることを表している．すなわち全画素で$d_k \cong r_{ik}$のときにだけ$p(d|i) \cong 1$となりその他の場合$p(d|i) \cong 0$となる．従って例えば図4.9の左3列の画像だけで学習してこれらと同じ代表画像が得られたとして，それによって右端の画像を識別するとどちらのグループに属すか分からない(どちらのグループにも0.5ずつ属す)という結果となる．これは左から2番めの画像と一致する部分(中央の水平線)があるにもかかわらず違う部分(斜めの線)があるためにANDの結果違うと判定されるためである．

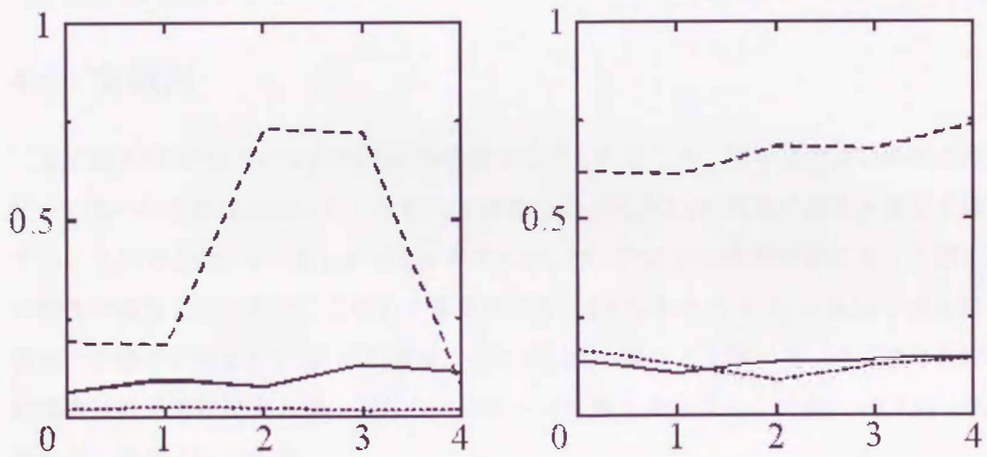


図 4.8: Rニューロンと Wニューロンの応答

以上のような欠点を無くすためにガウス分布をロバストな分布

$$p(d|i) \propto \prod_k e^{-a[1-e^{-\alpha(d_k-r_{ik})^2}]} \tag{4.10}$$

に変えることにする．ガウス分布が$|d_k - r_{ik}| \uparrow \infty$のとき0になるのに対し，式(4.10)の分布は$e^{-a}$という正の値に漸近する．すなわちこの分布はガウス分布に一様分布を加えたものとほぼ同じである．一様分布が外れ値を表す．$|d_k - r_{ik}|$の値が大きい(すなわち外れ値の)画素の総数をMとすると，式(4.10)の値はe^{-Ma}となり，外れ画素(逆に言えば一致する画素)の数を数えることができ，ガウス分布のときのような全画素一致ではなく，画素の多数決による識別ができる．アニーリングは今度はaは固定してαを徐々に大きくしていく．

4.6 実験例

まず図4.9の例について式(4.10)の分布で学習したところ，図4.10の薄い斜めの線が完全に無い代表画像が得られ，その代表画像によって図4.9の右端の画像も正しく識別することができた．$a = 0.1, b = 2, \alpha = 0.1$とした．次にこの代表画像によって図4.11の画像の識別をしてみた．このとき事前情報なし(すなわち$q_1 = q_2 = 0.5$)で図4.11を提示する場合と図4.9のどれかの画像の後に提示する場合とを調べた．まずガウス分布の場合は提示方法によらず，どちらのグループに属すか分からない(q_1 =0.5,q_2 =0.5)となり，式(4.10)の場合

1) いきなり(すなわち$q_1 = q_2 = 0.5$の後)提示した場合q_1 =0.55,q_2 =0.45となり，僅かな差ではあるが横棒のグループに属すとなり，

2-1) 図4.9の左上の画像を提示した(q_1 =0.9,q_2 =0.1)後，図4.11を提示した場合は$q_1 = 0.73, q_2 = 0.27$となり，横棒のグループに属すとなるが，

2-2) 図4.9の左下の画像を提示した(q_1 =0.1,q_2 =0.9)後，図4.11を提示した場合はq_1 =0.29,q_2 =0.71となり，縦棒のグループに属すとなった．

このように1枚の画像に二つのグループの画像が重ね描きされているような場合，直前に提示された画像のグループと同じグループに識別された．これは予測による注意に似た現象である．同様な例として，図4.12の画像データで各物体に3個ずつの代表画像(よって図4.12の画像がそのまま代表画像となる)を求めた後，図4.12の右下の画像を提示した後，図4.13の画像を識別すると自動車として識別された．このように直

図 4.9: 外れ値を含む画像データ例

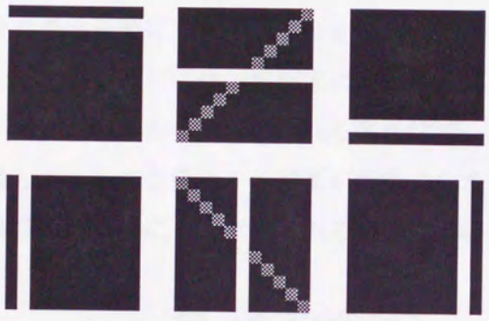


図 4.10: 代表画像

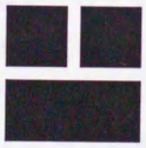


図 4.11: テスト画像

前に提示された物体の向きが変わったり，他の物体によって1部隠蔽されているときでも認識することができた．隠蔽されていても大丈夫なのは画素の多数決によって識別されるからである．なお通常のように重ね描きがなく，あるグループに含まれる画像だけが提示されるときは直前の画像が別のグループの画像であっても事前情報よりも現在のデータが勝って正しく認識される．

複数の物体が提示されたときに注意に似た振る舞いを示すとはいっても本識別法は画像全体を扱っており，画像のなかの物体を切り出して識別してはいない．しかし，次のようにして画像のなかの注視部位を求めることができる．まず各グループについて背景に相当する画像を求める．それには，各画素について背景に相当する画素値を次のようにして求める．各画素はグループに含まれる各画像でいろいろな画素値をとるが，多くの画像で共通してとる画素値はその画素が背景に含まれるときにとる値である．この背景の画素値は次のようにして求められる．第jグループでの各画素の値d_kの分布は

$$p(d_k) \propto \sum_{i \in I_j} e^{-a[1-e^{-\alpha(d_k - r_{ik})^2}]} \tag{4.11}$$

という混合密度となる．これの最尤値

$$\arg\max_{d_k} \quad p(d_k) \tag{4.12}$$

がこの画素が最も頻繁にとる値であり，この画素の背景での値である．例えば図4.12の下3個の画像の背景画像は図4.14となる．このようにここでの背景とは，グループの全画像に共通して現れるので個々の画像の識別では有益な情報をもたらさない部分という意味であり，厳密な意味での背景とは違うが，簡略のため背景と呼ぶことにする．各グループについてこの背景画像を求めておくと，次のようにして注視部位を特定できる．まずq_jの値により入力画像がどのグループに所属するかが分かり，更に所属するグループの各クラスタの代表画像について式(4.10)を計算すると，式(4.10)の値が最大となるクラスタが入力画像が所属するクラスタであることが分かる．そこでまず最初に，入力画像が所属するグループの背景画像と入力画像との共通部を求め，その部分を入力画像から削除する．次に残った各画素について，入力画像が所属するクラスタへのメンバシップ$e^{-a[1-e^{-\alpha(d_k - r_{ik})^2}]}$を計算し，この値が1に近い画素を取り出せば，そこがこの入力画像の識別に重要な部分であり，注視部分である．すなわち注視部分は背景や他の物体が写っている部分を取り除いた残りである．例えば上記の

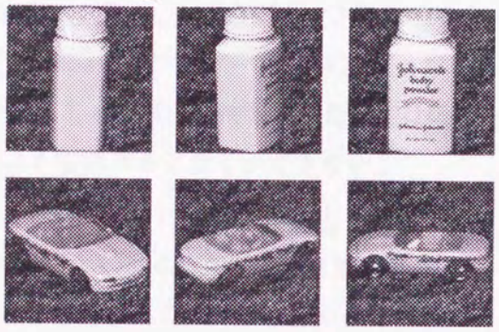


図 4.12: 3次元物体の画像例

図 4.13: テスト画像

図4.11についての2-2の場合の注視部分は図4.11の上半分の縦棒となり，図4.12右下画像の後の図4.13では図4.15の白い部分，すなわち自動車の見えている部分となった．

最後にトップダウン情報による予測におけるロバスト分布の役割について考える．まず以上のモデルには入力dの強度が入っておらず，いつでも強度は1とされている．ニューラルネットにおいては入力刺激は値dに付随して強度sをもつ．入力がないときは$s=0$である．この強度も入れると，$p(d|i)$がガウス分布の場合は$p(d|i) \propto se^{-a\|d-r_i\|^2}$となり，従って$p(i)p(d|i) \propto sp(i)e^{-a\|d-r_i\|^2}$となる．従って入力がない$(s=0)$場合には応答は生じず，すなわちトップダウン情報$p(i)$だけでは応答を生じることはできず，$p(i)$は入力$d$による応答を変調することしかできない．

次に$p(d|i)$がロバスト分布の場合は，$p(d|i) \propto \prod_i e^{-a[1-e^{-a\|d-r_i\|^2}]}$は$\| d-r_i \| \uparrow \infty$のとき$e^{-Na}$($N$は画素の総数)という$d$によらない定数になるので，$p(d|i) \propto e^{-Na} + f(\| d-r_i \|)$と表すと$f(\| d-r_i \|)$はガウス分布によく似た関数となる．第1項は入力$d$にはよらないから，入力の強度を入れると$p(d|i) \propto e^{-Na} + sf(\| d-r_i \|)$となる．すると$p(i)p(d|i) \propto p(i)[e^{-Na} + sf(\| d-r_i \|)]$となり，トップダウン情報$p(i)$は入力による応答$sf(\| d-r_i \|)$の大きさを変調もするが，入力がない$(s=0)$ときも$p(i)e^{-Na}$という応答を生じることができる．$e^{-Na}$はRニューロンの自発応答である．すなわち自発応答がない場合がガウス分布のRニューロンである．

以上のようにガウス分布ではトップダウン情報だけでは応答は生じ得ないが，ロバスト分布なら生じることができる．トップダウン情報によって生じる視知覚はメンタルイメージと呼ばれ，そのときV1ニューロンも応答しているという報告もある[49]．また酒井ら[24]は側頭連合野のニューロンが，現時刻のデータ入力がなくても直前の入力画像に応じた応答を生じることを観測している．これらの生理実験結果はロバスト分布モデルを支持しているように思われる．

4.7 むすび

RBFニューラルネットにフィードバックを付加したモデルを使って顔画像の学習をし，生理実験データとよく似た応答が得られることを示した．またロバストな分布を使うことにより，画像中の外れ値の画素を棄却する認識が行えることを例示した．この棄却機能を使うことにより，物体のセグメンテーションを陽に行うことなく，注視に似た処理が行えることも述べた．しかし一方，物体の切り出しや動き検出による物

体追跡を伴うように本モデルを拡張して，実際の注視に近い機構をモデル化することも考えられる．このためには，ここでのように画素値をそのままデータベクトルとするのでなく，画像中の位置も含めた特徴抽出をする必要があると思われる．またここでのニューラルネットには位相がなく，クラスタの代表ニューロンの順序には意味はないが，生理実験では視点の変化につれて活動する場所も連続的に移動するので，空間的なラテラル結合などを付け加えて位相保存性を持たせるのも今後の課題である．

図 4.14: 図 4.12 下列の画像の背景

図 4.15: 図 4.13 での注視領域