

ニューラルネットを用いた顔画像の識別と特徴抽出

小杉 信

NTT ヒューマンインタフェース研究所

顔画像の中心部を正方に切り取り、これを粗いモザイク表現に変換後、モザイク片の代表値をニューラルネット(3層BPN)に入力して顔の識別を試みた。この方法により多人数の学習が可能であり、学習の結果、中間層の各ユニットは、モザイク画像の各所から濃淡情報をきめ細かく集めることにより、互いに共通の特徴をもつ顔画像、例えば、男女や、のつぺり顔とふっくら顔などを自動的に分類した。さらに、この方法は、多人数の顔の識別が可能であるだけでなく、表情の変化や年齢の違いによる顔画像の変形、ならびに焦点はけ、照明の変化、ノイズ等に対し非常にロバストであることがわかった。また、顔画像の切り出し時における位置やサイズの変動に対する許容の程度を示した。

Human-Face Feature Extraction and Identification by Neural Network

Makoto KOSUGI

NTT Human Interface Laboratories

1-2356 Take, Yokosuka-shi, Kanagawa-ken, 238-03 Japan

The back propagation network(BPN) is applied to human-face identification. A mosaic pattern transformed from central part of human-face image, is put into BPN. This combination succeeds in identification of hundreds of people with robustness not only to defocused or noisy image but also to image of different face expression or different age. Hidden units of the BPN extract peculiar and delicate features of human-face, which can not be obtained from existing statistical methods. A few of hidden units can especially select only men or women. Moreover, marginal region to keep correct answer against shift deviation or size change of a central part of human-face image is shown.

1 まえがき

顔画像は個人同定の重要な対象であることから、コンピュータによる顔画像のパターン認識に関する研究が古くよりなされている。最近、人にやさしいセキュリティシステムや顔の識別を利用した知的通信への適用などのため、こうした研究への関心が高まっている。しかしながら、未だ実用に耐え得る成果は得られてない。

従来、パターン認識は画像からエッジを抽出し、その形状や互いの位置関係を見い出して識別を行う。顔画像の場合、顔の輪郭や目・鼻・口などの形状、ならびにこれらの位置関係が顔の重要な特徴であるとし、これらを画像から抽出して識別に利用する。

しかしながら、この方法では、弱い照明や焦点ぼけ、ノイズなどにより、エッジの抽出自体が困難である場合が多い。文書や図面など2値画像でさえ、このような前処理には手をやいており、自然画像においてはなおさらである。とくに、顔の場合、鼻や口などは濃淡や色に差が少なく、正確にエッジを抽出することは非常に困難である。そこで、画素レベルの濃淡情報を用い、統計的手法で顔画像を認識する試みも進められている[1],[2],[3]。

これに対し、人間は、線分形状のみでなく、色やテクスチャが認識に重要な役割を果たしていることが分かっている。例えば、リンカーンの顔のモザイクはよく知られている[4]。従来の線分形状を用いるものが、画像の高周波成分を用いたのに対し、これは極く低周波の成分を用いるものである。人間の視覚には複数のバンドパスフィルタがあり、これらの出力が統合化されて処理されることが分かっている[5]。但し、これが認識にどのように関与しているのかは殆んど分かっていない。

ここでは、低周波成分だけでも人間は認識が可能であることを手がかりに、モザイクを認識に利用する。ただし、従来の統計的手法ではなくニューラルネットを適用する。

まず、顔画像を少数のブロックに分割し、各ブロックを代表値で表わし、これをニューラルネットに入力する。とくに多人数の顔の識別を試みる。こうした例はまだ非常に少なく、文献[6]などに若干見られるが、いずれもブロックサイズが小さくかつ少人数への適用に留まっている。

既に、ニューラルネットによる学習と収束については報告したので[7],[8]、ここでは、中間層で抽出された特徴や、学習後のニューラルネットの特性、とくに未知画像入力時の変動要因に対する耐性などについて述べる。

2 顔画像のモザイク化とニューラルネット

顔画像の領域としては、髪を除いた顔の部分とし、顔の中心を固定して、相対的に顔幅を一定となるよう手で切り出す。

即ち、図1に示すように、水平方向の中心は両目の目頭を結ぶ線、両端は顔幅、垂直方向にはほぼ眉から顎までとする濃淡画像の顔部分を $n \times m$ のブロックに分割する。さらに各ブロック内の画素の平均値をブロックの代表値とする。本検討では、 $n, m = 12$ 、階調は256とした。

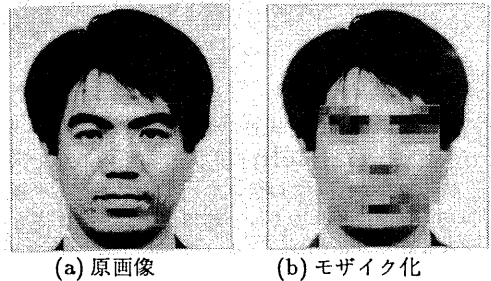


図1: 入力画像データ

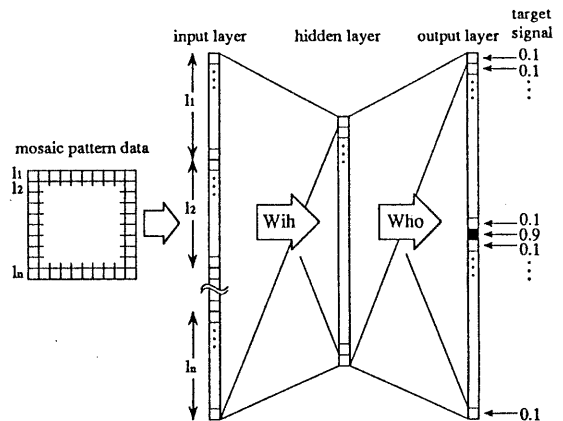


図2: モザイク画像識別に用いたニューラルネット

即ち、 $ND = 12 \times 12 = 144$ コの代表値を図2のような3層のBPNに入力した。入力層の数を上記ブロック数 ND 、中間層の数を NH (可変)、出力層の数を人数分 NP とし、入力値を I_i 、入力層から中間層への重みを W_{hi} とすると、中間層のユニット H_j の値は次式で表わされる。

$$H_j = f_j \left(\sum_{i=1}^{ND} W_{hi} I_i \right) \quad (1)$$

ここで f_j は次のようなロジスティック関数である。

$$f_j(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

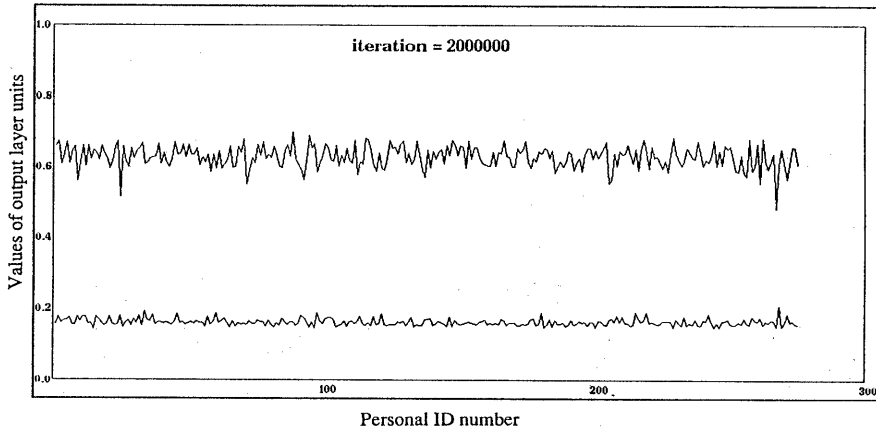


図 3: 学習による収束状況

同様に、中間層から出力層への重みを W_{oh} とすると、出力層のユニット O_j は次式で表わされる。

$$O_j = f_j \left(\sum_{i=1}^{NH} W_{ohji} H_i \right) \quad (3)$$

全ての重みは初めランダムとし、上記のモザイクデータに対応する個人識別信号を教師として、対象出力ユニットに 0.9、その他のユニットに 0.1 を与え、これを繰り返すことにより学習させた。全人数に対し、出力ユニットにおける出力値と正解値との平均自乗誤差が一定値内となった時、学習が終了したものとした。

この結果、学習用画像に対しては、多人数の場合も非常にうまく収束した。図 3 は、約 300 人を対象としたときの収束状況であり、横軸は個人識別番号、縦軸は各番号の個人データを入力したとき、その番号の出力ユニットの出力値（実線）と、この番号を除く他のすべての出力ユニットの中の最高値（破線）を示す。

3 中間層で抽出された特徴

3 層 BPN では、それぞれの層間でデータ表現の変換がなされる。入力層から中間層への重み W_{hi} によりモザイク画像から何らかの特徴空間への変換、中間層から出力層への重み W_{oh} により特徴空間に基づく個人識別がなされる。即ち、中間層の各ユニットはそれぞれ異なる特徴を表現することになる。しかし、どのような特徴が得られるかを予想することはできない。

約 80 人（男女半々）を対象としたときの、入力画像、中間層、出力層の結合関係を図 4 に示す。入力層はモザイクデータが 1 次元に配列され、これらが中間層の各ユニットと重み W_{hi} でリンク（図示は割愛）している。中間層からある出力層ユニットへの重み W_{oh} を

リンクとともに示し、黒は正重み、グレイは負重み、線分の幅は重みの大きさを示す。中間層、出力層とも各ユニットは 0 から 1 の値を取り、塗りつぶされた部分がその値を示す。

この場合、入力と結合されている中間層のユニット数は 35 であり、36 番目（# 36）はしきい値を与える。図のように、いずれの入力にも反応しないもの、逆に全ての入力に反応するものがある。これは中間層のユニット数が過多のため不要のユニットがでたことによるものであり、真に有効なユニットは 2/3 程度である。どの中間層ユニットが効いているかは画像ごとに異なるが、常に 1 となるユニットを除き、ユニット値とこのユニットから注目する出力層ユニットへの重み W_{oh} を乗じた値がある程度以上となるものがあてはまる。図の入力画像の場合、# 4、5、9、14、16、17、28、32 などである。

そこで、各ユニットごとに、入力データに対する値を要素とするベクトルで表わし、各ユニット間のなすユークリッド距離を多次元尺度法で 2 次元平面に近似展開し、これを図 5 に示す。多次元尺度法はベクトル間の距離の順序を保って、低次元空間にベクトル間の距離関係をマッピングする方法である。図 5 中の数字は図 4 における中間層と同一のユニットを示す。この図において、左端は全てに無反応、右端は全てに反応するユニットがある（省略）。また、互いに近距離にあるユニットは特徴が類似し、座標軸に対称的位置にあるものは何らかの対称的な特徴がある。

そこで # 16 の中間層ユニットの値を大きくする入力画像を調べる。即ち、入力モザイクからユニット # 16 への重み W_{hi16i} により、何らかの共通する特徴により # 16 が自動的に選別する画像を調べる。

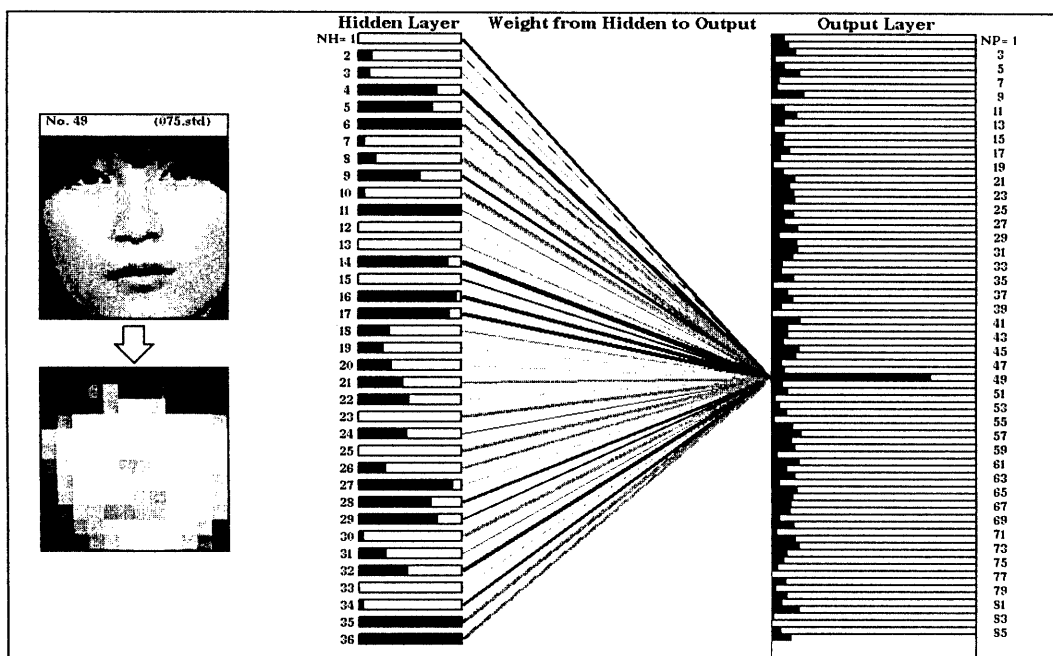


図 4: ある入力画像と中間層・出力層の結合関係

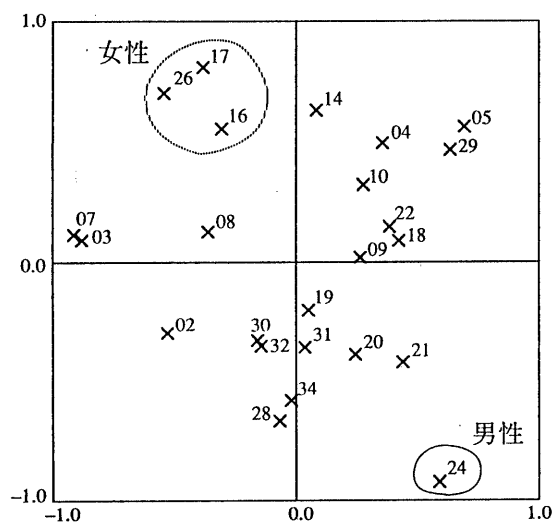


図 5: 多次元尺度法による中間層のユニット間距離

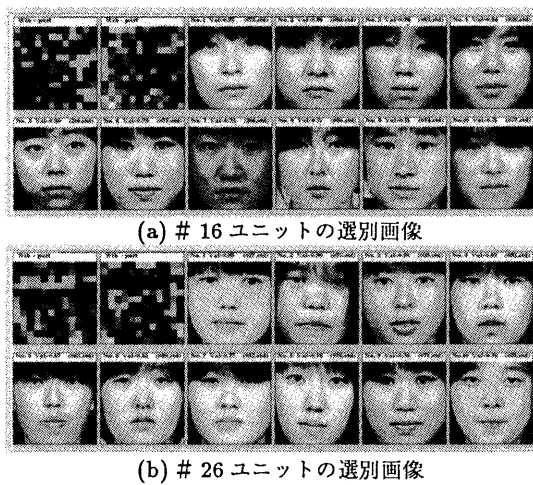


図 6: 中間層のユニットで選別された入力画像

この結果を図6(a)に示す。この図のように、#16はなんと女性のみを選別する。図6の初めの2ブロックは入力層から中間層ユニットへの正負の重みを示す。これを解析すると、額に髪がかかっていたり、額の両側に髪があることなどが大きな要因となり、この結果女性だけが選別される。

さらに、図5において隣接している#26について同図(b)に示す。これも同様の理由で女性のみが選別される。ところで、#16と#26は、図5において若干の距離があり、それぞれは何らかの違いがあることを示唆している。確かに、両者は互いに異なるタイプの顔を選別しており、重み W_{hi} を解析すると、両者は目の下の明るさの差が特徴的な違いとなっており、結果的に、#16はふくら顔、#26は逆にのっぺり顔を選別している。

これについては、さらに次のように言える。肌色の部分の濃淡はある方向からの光に対する反射率を表わし、光に直角であれば最も明るく、傾斜が大きくなるほど暗くなる。一方、3次元物体における表面は連続であるから、この傾斜は表面の凹凸など奥行きの情報を表わす。#16は目の直下は暗く、その下の頬は明るいことから、目の下に膨らみがあったり、頬骨がはった顔となる。さらに、鼻の両脇の暗部の負重みは、頬のでっぱりにより鼻の両脇との高低差によって生じる陰影を抽出している。こうして総体的には、凹凸のある顔をした女性を選別されている。

一方、#26のように目の下が明るいことは、目の下に膨らみやへこみがなく表面が平坦であることを示し、結果的には、他の特徴とも合わせ全体的にのっぺりした顔の女性を選別される。この結果、#16と#26では上位10名が1名を除きオーバーラップしない。

一方、これらのユニットとは点対称の位置に#24がある。このユニットの値を大きくする画像を調べると、0.8以上の値をとる22名のうち86%が男性、逆に0.2以下となる20名のうち80%が女性となり、図5から予測される事実が確かめられた。

以上のように、ニューラルネットは顔全体から実にきめ細かく特徴を捉え、識別に必要な情報を捕まえる。これは統計的手法などには見られないニューラルネットの大きな特徴である。

ところで、このような特徴をもつ中間層が常に現われるわけではない。中間層のユニット数が少ないと、一つのユニットに多くの特徴が縮退され、直観ではユニット間の差がわからない。逆に、ユニット数が多くなると、一つの特徴が分散表現されこれも直観では見分け難くなる。

4 顔の変形に対する耐力

今までは、顔画像識別の仕組みを考察したが、ここでは、学習で用いた顔画像ではなく、異なる表情や異なる時期の顔画像に対するニューラルネットの応答を調べる。

(1) 表情の変化

普通の表情、即ち真顔を用いて学習したネットワークに、表情の変化した画像を入力した。図7は笑顔の場合を示したもので、含み笑いと開口した笑顔の例である。当然、口を開けた方が表情の変化は大きく、それだけ真顔との差は大きくなる。これらの影響で出力値は幾分下がるが、識別は問題ない。

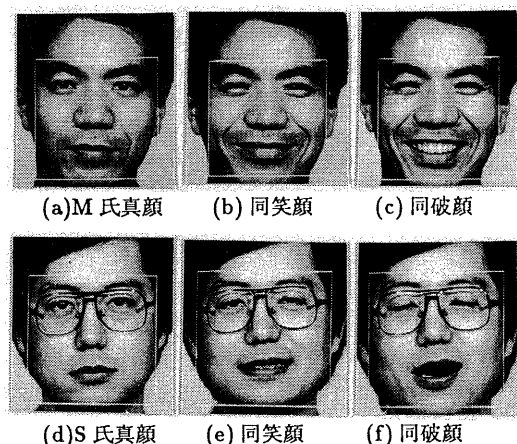


図7: 表情の変化画像の例

(2) 時間差 (年齢変化)

同じ人物でも、学習時と未知画像入力時で時間差がある場合、両画像間には当然、差を生じる。時間差が5～10年となると顔つきもかなり変わることが予想される。図8は学習時の3年後、5～10年前の顔について検証し、同図(c)を除きいずれも正解をだした。中には人間でさえ判断の困難なものがある。

これらの画像でノイズがのっているのは、印刷顔画像を用いたため、スキャナで読みとる際、濃淡をサイズの大きさに表わす画素が拡大されたことによる。それにもかかわらず正解が得られたのは、ノイズがモザイクより小さく、モザイク化によりこれらのノイズが平均化されたためである。ノイズへの優れた耐性もさることながら、10年というようなロングレンジでも同一の顔が選ばれうることは、モザイク自体が多くの情報量をもっていることの証左であろう。



(a)M 氏現在 (b) 同 5 年前 (c) 同 10 年前



(d)S 氏現在 (e) 同 3 年前 (学習用) (f) 同 10 年前

図 8: 時間差のある顔画像の例 (学習画像は 3 年前のもので、M 氏の学習画像は図 1(a) と同じ)

(3) 顔の向き

学習時には、通常、正面を向いた画像を用いる。この場合、未知画像取り込み時に学習時との顔の向きの違いが問題となる。これは、種々の変動要因の中で最も厳しいものである。本ニューラルネットはある程度、正解をだすことができるが、現在検討中であり、別途報告したい。

以上のような種々の変形に対するニューラルネットの出力値の変化を図 9 に示す。

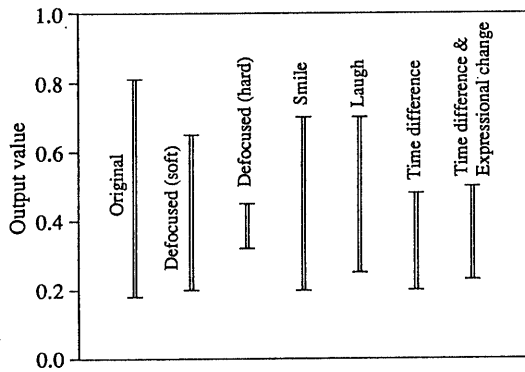


図 9: 種々の変形に対する出力値の例 (対象人数 70、中間層ユニット数 35)

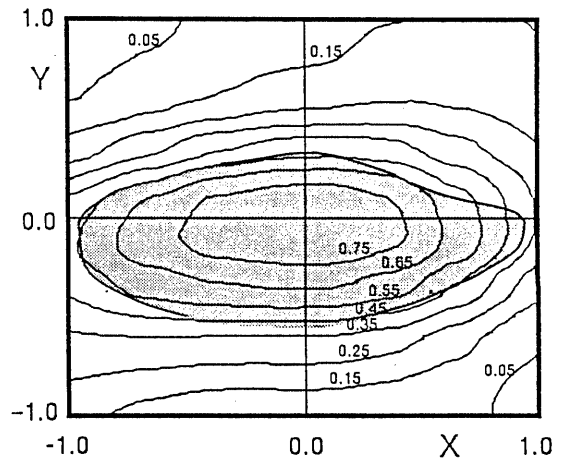
5 画像入力時の変形要因に対する耐力

未知画像入力時に中心とサイズを指定するが、手入力であるためずれを生じうる。この場合、入力画像の変形は平行移動と拡大・縮小の混在したものとなる。この原因は、前述と異なり、対象となる人間のほうでなく、画像を取り込む側にあり、また自動化のうちに

はある程度解決されうる問題である。なお、焦点ぼけ (図 9) や濃淡の変化については非常にロバストであることをすでにのべた [7],[8])。

(1) 平行移動

学習用原画を用いて新たに顔部分を切り出す時、学習時の中心位置からのずれにしがたってニューラルネットの出力値は減少し、ついには第 2 位以下に落ち込む。図 10(a) は、モザイク 1 個分の長さを単位として、位置ずれがおきたときのニューラルネットの出力値を示す。第 1 位を保持する範囲 (グレイで示す) は、横方向には、左右はモザイク 1 コ分、上方向は 0.3、下方向は 0.5 コ分が限界となる。これらに対応する顔画像を同図 (b) ~ (e) に示す。上下が非対称なのは、入力時の中心を目の位置にしているためである。但し、これらの値は顔ごとにより異なる。



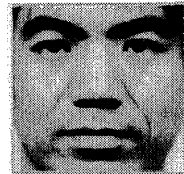
(a) 出力層の値 (グレイ部は第 1 位の範囲)



(b) 右シフト 0.9blk



(c) 左シフト 0.9blk



(d) 上シフト 0.3blk



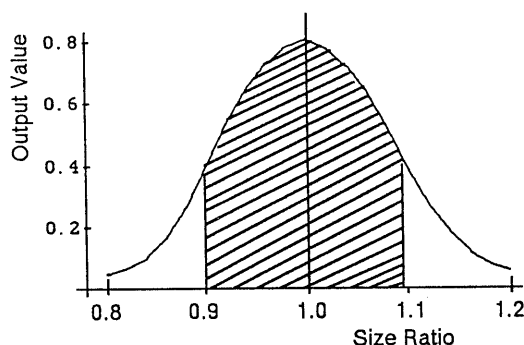
(e) 下シフト 0.5blk

図 10: 平行移動によりずれた画像入力時の出力値と第 1 位を維持する画像の例

このように、学習時の画像と入力画像の間に大きな変化がなければ、平行移動に対するマージンは十分あり、手操作による位置ずれの影響は少ない。これは、もともと正しい位置におけるニューラルネットの出力値が大きく、且つ第2位候補との差が大きかったためである。これに対し、表情が変わったり、時期が大きすぎた場合、これらの程度に応じて、許容される位置ずれは小さくなる。

(2) サイズ変形

上記と同様に、学習用原画からの切り出し時にサイズがずれたときのニューラルネットの出力値を図11に示す。この画像の場合、学習時の約0.9～1.1倍が第1位を維持できる許容範囲であり、この値は、他の画像に対してはほぼ当てはまる。また、同図の各顔画像から分かるように、この範囲はかなり広い。しかし、通常は、上記の平行移動とサイズ変形が同時に起こるため、許容範囲はこれを考慮する必要がある。さらに、表情や時期が学習時とずれば、それだけ許容範囲が小さくなるのは平行移動の場合と同様である。



(a) 出力層の値 (斜線部は第1位の範囲)



(b) SR=0.90

(c) SR=1.09

図11: 切り出しサイズのずれた画像入力時の出力と第1位を維持する画像の例

6 未学習画像に対する応答

今までは、学習済みの顔について、種々の変化・変形に対するニューラルネットの応答をのべた。ここでは、ニューラルネットで学習していない人物の顔が入力された場合について述べる。

通常は、未学習の顔が入力されると、どの出力ユニットの値も小さく、しかも第1位と第2位、さらには多くのユニット間で出力値に差を生じない。そこで、これを利用し、第1位の値が小さく、かつ連続する上位いくつかの値との差が小さければ、該当する顔はないという解を出せば良い。しかしながら、未学習の顔が学習済みの顔のいずれとも絶対似ていないという保証はなく、むしろ人数が多くなるほど問題となる。文字などと異なり、全体の集合が決められないため、認識率は学習した顔と未学習顔の係に依存する。

例えば、100人を学習し、ある未学習の50人を入力した。この結果、図12に示すように、前記の条件により38人が該当なしとされた。即ち、まず10人は第1位の値が小さい(0.4以下)という条件で該当なしと判定され、28人は1位と2位の差が小さい、具体的には、 $(maz1 - maz2) / maz1 < 0.2$ という条件で該当なしとなった。しかしながら、残りの12人に対しては、こうした判定基準にかからずエラーとなった。これは150人を母集団として8%のエラーとなる。

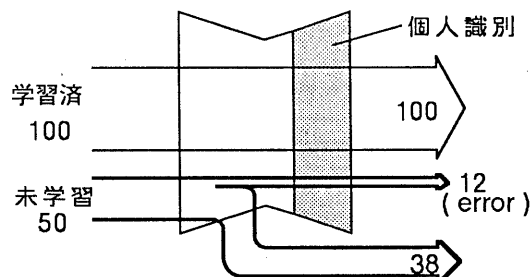


図12: 条件導入による未学習顔の判別

そこで、1つの試みとして、例えば100人の顔画像の集合を半分に分け、50人を個人識別、他の50人を1つのグループとして学習させる。具体的には、一方には個人ごとに出力層の1ユニットを割当て、他方の50人分については、どの顔も同一のユニットを割り当てる。即ち、一方は個人ごとに識別するのに対し、他方は、「その他」の人とする。

この場合も学習時の収束の問題は全くなく、学習終了時の結果を図13に示す。図の右半分は同一のユニットを割り当てたものであるが、個人識別に比べ第1位、即ち、「その他」の出力値は0.1ほど高くなり、第2位の出力値は0.02ほど下がった。本来であれば、異なった顔を受け入れねばならないため、重みの調整が厳しく、収束しなかったり、収束しても第1位の出力値は小さいと予想していた。

これらの事実は、このニューラルネットは意図的なグルーピングが容易に可能であることを示している。

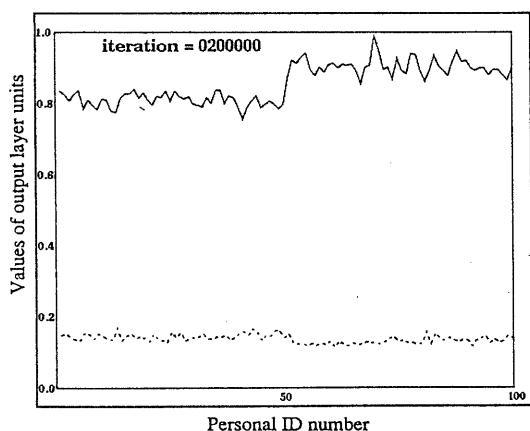


図 13: 「その他」のユニットを含む学習の結果

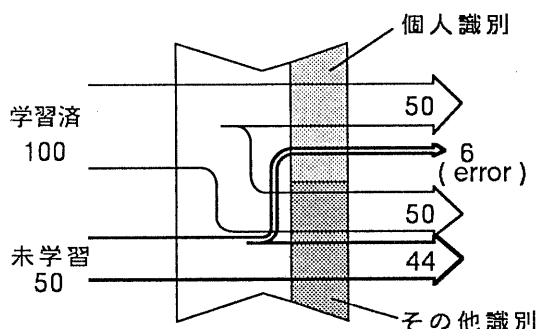


図 14: 「その他」を学習したニューラルネットワークにおける対未学習顔の判別

このニューラルネットに、個人識別により学習した画像を入力すると、確かに対応する出力ユニットが反応する。また、「その他」として学習したどの画像を入力しても「その他」に対応する唯一の出力ユニットが反応する。そこで、学習には用いなかった未知の画像を入力したときの反応を調べる。前記と同様、未知の 50 人を入力した。この結果、図 14 に示すように、44 人は「その他」のユニットに反応した。即ち 50 人のうちの 88 % の画像は「その他」の出力ユニット値を第 1 位にし、それも殆ど (29 人) が 0.95 以上という高い値であった。これに対し 6 人は、個人識別した 50 ユニットのいずれかを示し、エラーとなった。これは、母集団の 150 人の 4 % である。

結果的に、この方法は前者に比べ、エラー率は半減した。ニューラルネットを認識に利用するためには、未学習の顔への対応は重要な課題であり、こうした方法は一つの方向を示すものといえよう。

7 むすび

ニューラルネットを用いて、モザイク化した顔画像の認識を試み、非常に良好な結果を得た。本手法の特徴は、線分形状ではなく濃淡情報を用いたこと、抽出すべき特徴を事前に指定するのではなく、系が自動的に獲得することである。

当初、顔画像とくに髪を除いた部分は視覚的に良く似ており、多人数を対象とする個人識別は困難なことが予想された。しかし、上述のとおり、本手法は、個人識別に十分有用であるのみでなく、ノイズや焦点ぼけ、さらに表情の変化、年齢による変化にも強いなど、大きな成果が得られた。

これらの好結果は、当然、モザイク自体が識別に必要な情報を十分担っており、ニューラルネットが特徴をうまく引き出しかつ分類できたためである。モザイクは顔の概形のみでなく、濃淡による奥行き情報も有しており、他の手法にも有用である。

しかし一方で、本手法は、自動的に顔を切り出す方法が必要である。その際、本手法は、位置ずれやサイズの変化に対してもある程度のマージンがあることがわかった。このため、自動切り出しも精度に余裕をもつことができる。

本研究の機会を戴いたヒューマンインタフェース研究所釜江尚彦所長、小林幸雄視覚部長、ならびに御討論戴いた視覚部の方々に感謝する。

参考文献

- [1] L.Sirovich,M.Kirby:Low-dimensional procedure for the characterization of human faces,J.Opt. Soc.Am.A,4,3(1987).
- [2] M.Turk and A.Pentland:Face recognition without features,MVA'90 IAPR Workshop on Machine Vision Applications,Nov.(1990).
- [3] 赤松、他:KL 展開によるボタン記述法の顔画像識別への応用の評価,信学技報 PRU90-152,p55 (1991).
- [4] L.D.Harmon,B.Julesz:Masking in visual recognition:Effects of two-dimensional filtered noise, Science,180,15 June,p1194(1973).
- [5] H.R.Wilson et al;Spacial frequency tuning of orientation selective units estimated by oblique masking,Vision Res. 23,9,p873(1983).
- [6] J.L.Perry:Human face recognition using a multi-layer perceptron,IJCNN90-WASH,II-416(1990).
- [7] 小杉:ニューラルネットを用いた顔画像識別の一検討,テレビジョン学会技術報告,VAI'90-30,14, 50,p7(1990).
- [8] M.Kosugi:Human-face identification using mosaic pattern and BPN,ACNN'91,p111(1991).