平成30年度 卒業論文

顔と手の動きを特徴量レベルで統合した個人識別の評価

宮崎大学 工学部 情報システム工学科

池田 拓矢

指導教員　椋木 雅之 教授

目 次

1　はじめに

近年、個人認証には、顔、指紋、虹彩などの生体情報が使用されることが増加してきている。しかし、生体情報を偽装する手口も進化しているため、単体の生体情報では偽装に耐えられないことが多々ある。そのため、近年注目されているのが複数の生体情報を用いて個人認証を行うマルチモーダル認証であり、盛んに研究が行われている。

本研究では、顔と手の動きを組み合わせた個人識別を扱う。顔を用いて個人識別をすることは、人間が他人の顔を見て誰なのかを判断するように、普段誰しもが行っていることであることから、指紋などのほかの生体情報に比べ対象者の心理的な抵抗が少ないという利点がある。一方で、対象が双子、サングラスやマスクで顔の一部が分からないといった場合に判別が難しいという欠点が存在する。これに対して手の動きを生体情報として用いることは、手の動きをはじめとする仕草は、各個人によって癖が出やすく、顔の情報が分かりにくい状態でも取得しやすいという利点がある。これらを組み合わせることで、識別結果の向上が期待できると考えられる。

従来から、顔と動きを組み合わせた個人識別に関する研究は、盛んに行われている。例えば[1]では、顔や動きの特徴をそれぞれ取得し、それらの相違度をDPマッチングやユークリッド距離により独立に評価し、その結果をAND/OR演算またはファジィ推論により統合し、更に評価している。[2]では、顔や動きなどの特徴を、それぞれ取得し、これらの相違度を求め、事後確率を求めることにより統合し、評価している。しかし、これらの研究はいずれもそれぞれの特徴を単独で評価し、相違度を統計学的に統合したもの、いわゆるスコアレベルで統合した研究である。

しかし、それぞれの特徴を特徴量レベルで統合し、評価しているものは少ない。異なるモーダルの特徴量は、異なる性質を持つことから、単純に統合することは難しく、単純に統合したとしても良い結果が得られるとは考えられていないからである。

一方で、近年の深層学習の進化により、特徴量の優劣が重要になっていることが明らかになってきている。そのため、各特徴量に適切に優劣をつければ、特徴量レベルで統合しても、十分な結果が得られる可能性がある。

そこで、本研究では顔と手の動きというマルチモーダル情報を特徴量レベルで統合 (図 1) して個人識別に利用する手法を評価し、スコアレベルでの統合による評価との比較実験を行う。

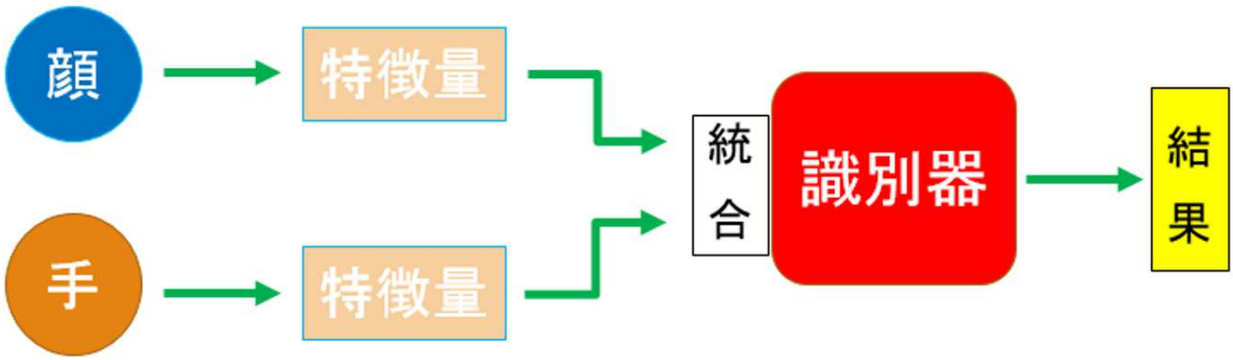

図 1 特徴量レベルでの特徴量の統合の流れ

2　特徴量レベルでの統合評価手順

本研究における個人識別の特徴量レベルでの統合評価は、以下の手順により行う。

(1) 顔の特徴量の取得

(2) 手の動きの特徴量の取得

(3) 特徴量の統合

(4) 識別器による判定

2.1　顔の特徴量の取得

顔の特徴量取得に用いる特徴抽出器として、davidsandberg/FaceNet[3] を用いる。特徴量取得は、図 2 のような流れで行われる。

davidsandberg/FaceNet[3] は、最初に Google 社から提案された顔認証で使うことを想定されたニューラルネットワーク FaceNet[4] をもとに、davidsandberg 氏が開発したオープンソースである。davidsandberg/FaceNet[3] では、図 1 のような顔画像から、Multi-Task-Convolutional Neural Network(MTCNN)[5] と呼ばれる顔検出に特化したニューラルネットワークを用いた顔検出器を用いて顔部分を切り出し、この切り出した顔を、Convolutional Neural Network(CNN) を介して 512 次元の特徴量に変換する。得られた特徴量は、L_2 ノルムが 1 となるように正規化されている。

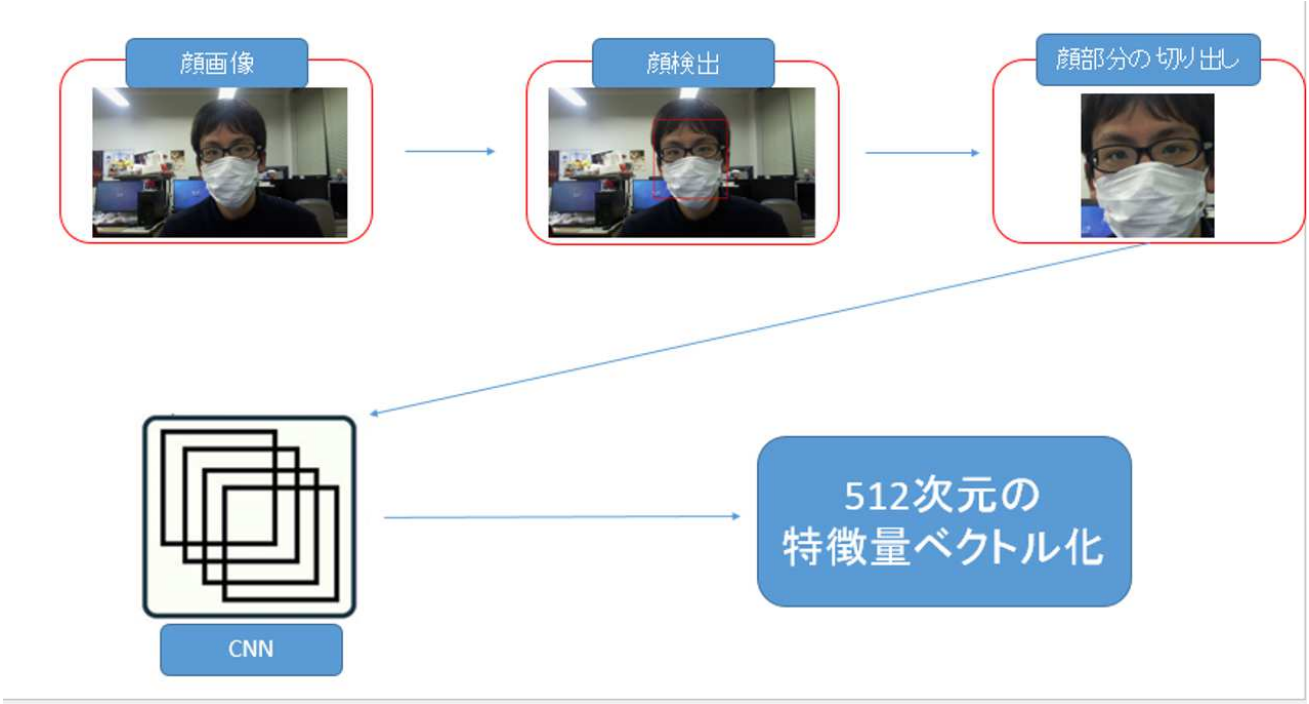

図 2 顔の特徴量の取得の流れ

2.2　手の動きの特徴量の取得

手の動きの特徴量は、以下の手順により取得する。

(1) 手の位置の座標の取得

(2) 手の位置の座標の時間合わせ

(3) 手の位置の座標の位置合わせ

(4) 特徴量ベクトルの L_2 ノルムの正規化

2.2.1　手の位置の座標の取得

本研究では、手の動きの特徴量として、被験者にカメラの前で空中で三角形を描いてもらい、これを1周して得られた座標列を用いる。その際、被験者に緑色の軍手を装着してもらい、その色を追跡することで手の位置とする。

手の位置の取得には、カラートラッキングの手法を用いる。まず、図3に示すような画像から得られる RGB 値を以下の式により HSV 値に変換する (図 4)。ここで、各画素の RGB 値のうち、一番大きいものを MAX、一番小さいものを MIN とする。

$$H = \begin{cases} 60 \times \frac{G-R}{MAX-MIN} + 60 & (MIN = B \text{ のとき}) \\ 60 \times \frac{B-G}{MAX-MIN} + 180 & (MIN = R \text{ のとき}) \\ 60 \times \frac{R-B}{MAX-MIN} + 300 & (MIN = G \text{ のとき}) \end{cases} \tag{1}$$

$$S = \begin{cases} \frac{MAX-MIN}{MAX} & (MAX \neq 0 \text{ のとき}) \\ 0 & (\text{それ以外}) \end{cases} \tag{2}$$

$$V = MAX \tag{3}$$

HSV変換した各画素に対して、追跡の対象としたい緑色の軍手が含まれるHSV値の範囲を指定し、その範囲に入っている画素を白，範囲外の画素を黒で表すように二値化処

理を行う（図5)。ある画素のHSV値を$h(H,S,V)$、各画素の画素値を$f(x,y)$とし、式で表すと以下のようになる。

$$f(x,y) \quad = \quad \begin{cases} 1(\text{白色}) \quad (\theta_{min}(H,S,V) \leqq h(H,S,V) \leqq \theta_{max}(H,S,V)) \\ 0(\text{黒色}) \quad (\text{それ以外}) \end{cases} \tag{4}$$

そして、白の領域の重心座標を求める。重心座標は、以下の式で求める。

画像中のp,q次の画像モーメントは、以下の式により求まる。

$$m_{p,q} = \sum_x \sum_y x^p y^q f(x,y) (\text{二値画像なので} f(x,y) = 1) \tag{5}$$

0次の画像モーメントが面積、1次のモーメントがx軸方向およびy軸方向の平均値を表すので、

$$0\text{次画像モーメント} \quad = \quad m_{0,0} = \sum\sum x^0 y^0 (\text{画素数}) \tag{6}$$

$$1\text{次画像モーメント} \quad = \quad m_{1,0} = \sum\sum x^1 y^0 (x\text{座標の和}) \tag{7}$$

$$= \quad m_{0,1} = \sum\sum x^0 y^1 (y\text{座標の和}) \tag{8}$$

となるので、白の領域の重心座標(x_g, y_g)は、

$$(x_g, y_g) = (\frac{m_{1,0}}{m_{0,0}}, \frac{m_{0,1}}{m_{0,0}}) \tag{9}$$

と求まる。この重心座標を手の位置とみなす(図6中の赤い点)。

図 3 手の動きの入力画像

図 4 HSV への変換

図 5 二値化

図 6 重心座標 (図中の赤い点)

2.2.2　手の位置の座標の時間合わせ

被験者が軌跡を描くまでの時間は、毎回変動する。しかし、これをこのまま特徴量ベクトルとすると、特徴量ベクトルの次元数がバラバラになってしまい、判定ができない。そのため、判定に用いる手の位置の座標を補間により揃える (図 7)。

被験者が手を動かし始めた瞬間を始点 (時刻 t_0)，動かし終わる瞬間を終点 (時刻 t_N) とし，始点から終点までのフレーム数を $N+1$ フレーム (始点，終点を含む) とする。始点 (時刻 0) から終点までの間を P 等分した時刻 P までの各時刻の手の位置の座標を求める。等分した地点の座標は，時刻 i の前後のフレーム N_{t_i},$N_{t_{i+1}}$(i=0$\cdots P$、t_i は i の値によって $0 \leqq t_i \leqq N$ の範囲をとる整数) における手の位置の内分（線形補間）によって求める。フレーム番号 t_i，内分比 s_i を次式で求める。

$$t_i = \left[\frac{N}{P_i}\right] ([] : \text{ガウス記号}) \tag{10}$$

$$s_i = \frac{N}{P_i} - t_i \tag{11}$$

内分比 s_i と t_i 番目のフレームでの x,y 座標を用い、時刻 i における手の位置の座標 (x_i,y_i) を次式で求める。

$$x_i = x_{t_i}(1-s_i) + x_{t_{i+1}} s_i \tag{12}$$

$$y_i = y_{t_i}(1-s_i) + y_{t_{i+1}} s_i \tag{13}$$

本研究では P=100 の座標列とする。時間合わせ前の軌跡の例を図 8 に、時間合わせ後の軌跡の例を図 9 に示す。

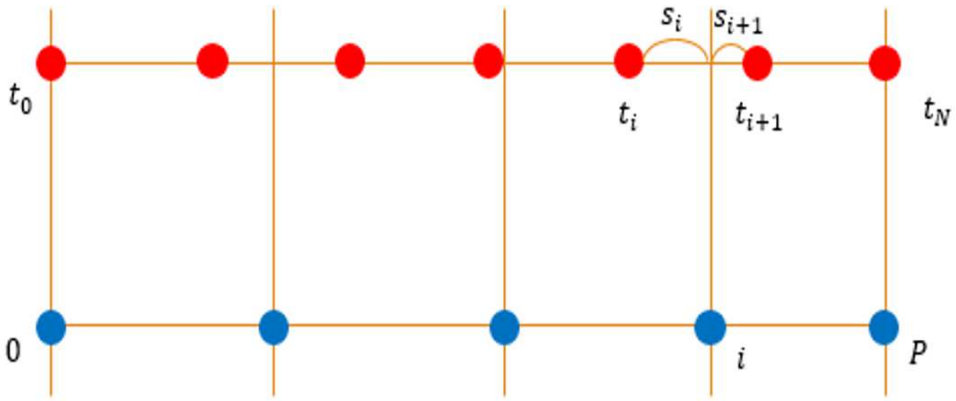

図 7 座標の時間合わせの図

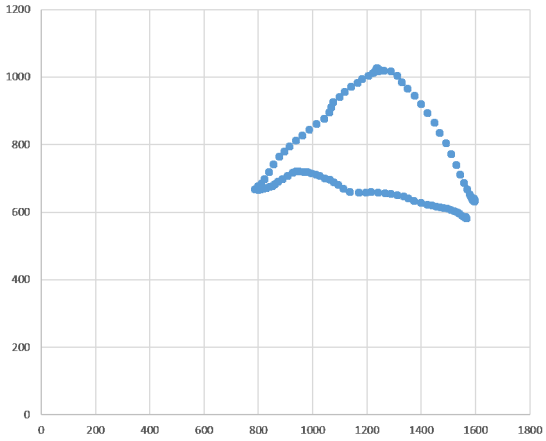

図 8 時間合わせ前の軌跡の例

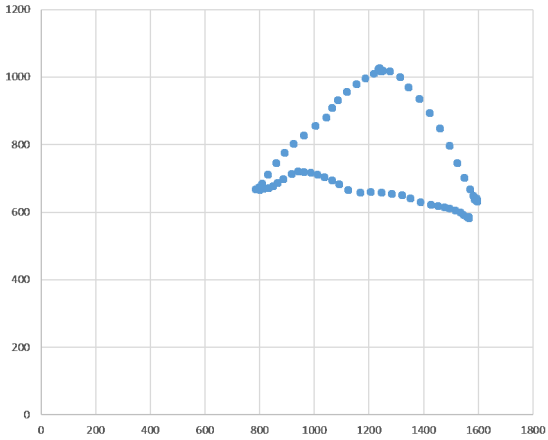

図 9 時間合わせ後の軌跡の例

2.2.3　手の位置の座標の位置合わせ

軌跡を描く位置は、被験者によって異なる。時間合わせをした手の位置の座標列をそのまま特徴量として使って識別を行うと、位置ずれの影響のため、正しく判定ができない。そのため、すべての軌跡の重心座標を $(x,y)=(0,0)$ に位置合わせする。重心座標は、N を座標数、x_i、y_i を軌跡を構成する点の座標 $(0 \leqq i \leqq N\text{-}1)$ とすると、軌跡の重心座標を $(\overline{x},\overline{y})$ とすると、

$$\overline{x} = \frac{1}{N}\sum_{i=0}^{N-1} x_i \tag{14}$$

$$\overline{y} = \frac{1}{N}\sum_{i=0}^{N-1} y_i \tag{15}$$

とすることにより求まる。更に、

$$x_i' = x_i - \overline{x} \tag{16}$$

$$y_i' = y_i - \overline{y} \tag{17}$$

とすることで、すべての軌跡の重心座標を $(\overline{x},\overline{y})=(0,0)$ に位置合わせすることができる。位置合わせ前の軌跡の例を図10に、位置合わせ後の軌跡の例を図11に示す。

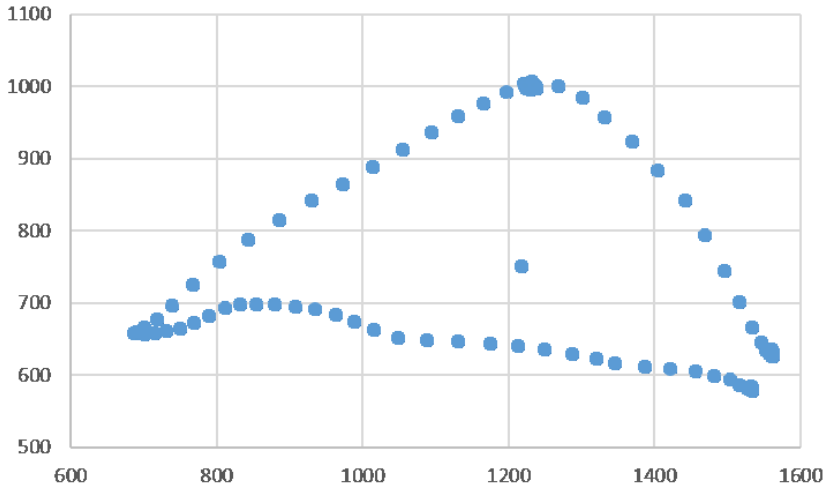

図 10 位置合わせ前の軌跡の例

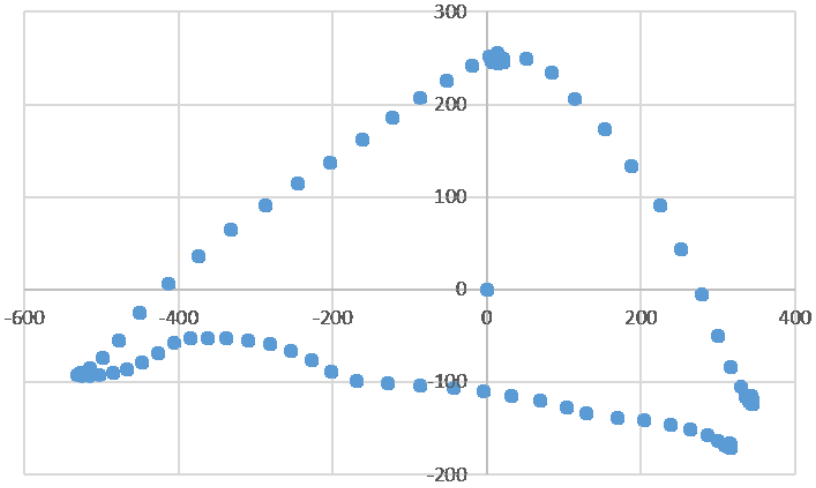

図 11 位置合わせ後の軌跡の例

2.2.4 特徴量ベクトルの L_2 ノルムの正規化

顔の特徴量ベクトルの L_2 ノルムが 1 になっていることから、手の動きから取得した座標列を 1 つのベクトルとし、このベクトルの L_2 ノルムを 1 にする。手の動きから取得した座標列からなるベクトルを $\boldsymbol{x}$=$(x_0, y_0, ...x_{N-1}, y_{N-1})$(本研究では N=100) とし、次式により $\boldsymbol{f}$ を求める。

$$\boldsymbol{f} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||} \tag{18}$$

$$||\boldsymbol{x}|| = \sqrt{\sum_{i=0}^{N-1} x_i^2 + \sum_{i=0}^{N-1} y_i^2} \tag{19}$$

こうして得られた 100 点の座標列からなるベクトル $\boldsymbol{f}$ を、200 次元の手の動きの特徴ベクトルとする。正規化前の軌跡の例を図 12 に、正規化後の軌跡の例を図 13 に示す。

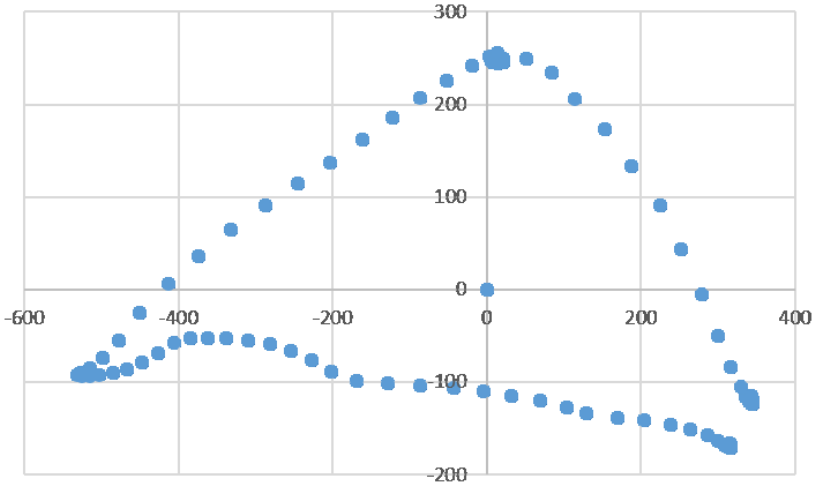

図 12 正規化前の軌跡の例

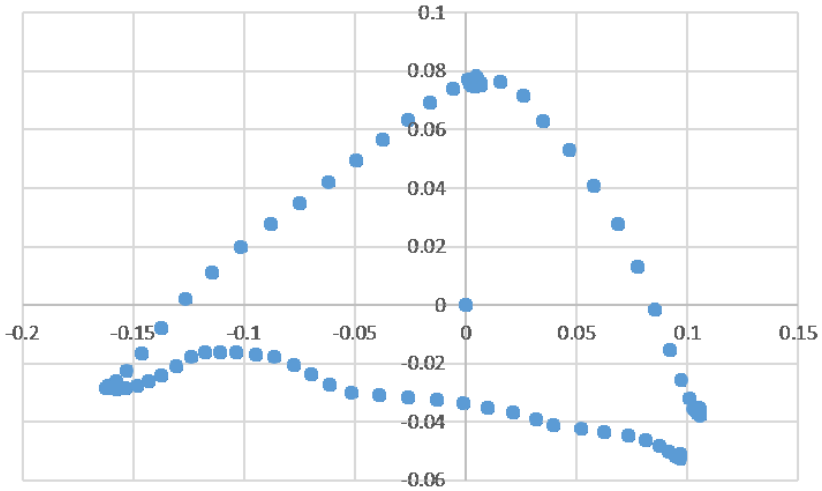

図 13 正規化後の軌跡の例

2.3　特徴量の統合

2.1 節で取得した 512 次元の顔の特徴量ベクトルと、2.2 節で取得した 200 次元の手の動きの特徴量ベクトルを統合し、712 次元の特徴量ベクトルとする。統合の際には、どちらの特徴量をどの程度重視するか、重みづけをおこなう。重みづけは、顔の特徴量ベクトルを $\boldsymbol{f_{face}}$、手の動きの特徴量ベクトルを $\boldsymbol{f_{hand}}$、統合後の特徴量ベクトルを $\boldsymbol{f_{int}}$ とし、以下のように行う。

$$\boldsymbol{f_{int}} \quad = \quad (\alpha \boldsymbol{f_{face}}, (1-\alpha)\boldsymbol{f_{hand}})(0 \leqq \alpha \leqq 1) \tag{20}$$

α =0 の場合は手の動きの特徴量ベクトルのみ、α =1 の場合は顔の特徴量ベクトルのみで識別を行っていることを意味する。

2.4　識別器による判定

得られた特徴量を用いて、識別器により個人識別を行う。識別器には、Collaborative Mean Attraction(CMA) 法 [6] を用いる。CMA 法は、未知のテストデータを複数の既知のカテゴリのどれかに分類する手法である。テストデータである特徴ベクトルを、学習用の特徴ベクトルのクラスのいずれかに分類し、誰であるかを分類する。

CMA 法 [7] では、まず、テストデータの特徴量ベクトルを $\boldsymbol{y} \in R^m$(m:特徴量ベクトルの次元数)、学習用データの特徴量ベクトルを並べた行列を $\boldsymbol{X} \in R^{m \times N_x}$ (N_x:学習データ数) とし、以下の式を最小化する係数ベクトル $f(\boldsymbol{\beta})$ を求める。

$$f(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_2||\boldsymbol{\beta} - \frac{\mathbf{1}_{N_x}}{N_x}||^2 \tag{21}$$

$||\cdot||$:ベクトルの L_2 ノルム、λ_2:重みパラメータ、$\mathbf{1}_{N_x}$:各要素が 1 の N 次元ベクトル

こうして求まった係数 $\boldsymbol{\beta}$ を用いて、各カテゴリ i について式 (22) により評価値を求め (図 14)、その値が最小になったカテゴリにテストデータを分類する (図 15)。

$$d^i = ||\boldsymbol{X}||_* \cdot ||\boldsymbol{y} - \boldsymbol{X_i}\boldsymbol{\beta_i}||^2||\boldsymbol{\beta}||/||\boldsymbol{\beta_i}|| \tag{22}$$

$\boldsymbol{X_i}$:カテゴリ i の学習データの特徴量ベクトルを並べた行列、$\boldsymbol{\beta_i}$:カテゴリ i の学習用データの係数

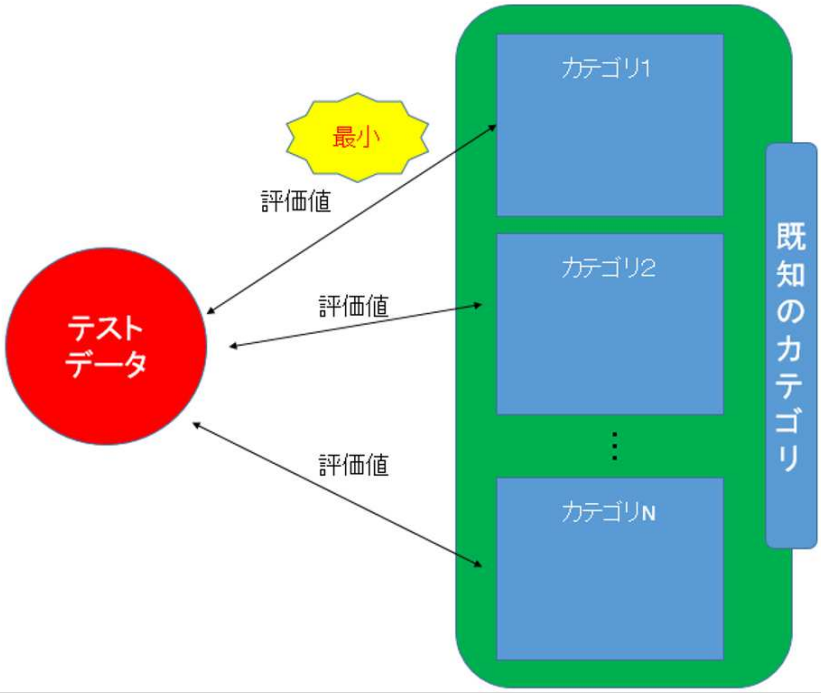

図 14 各カテゴリ間の評価値の算出および比較

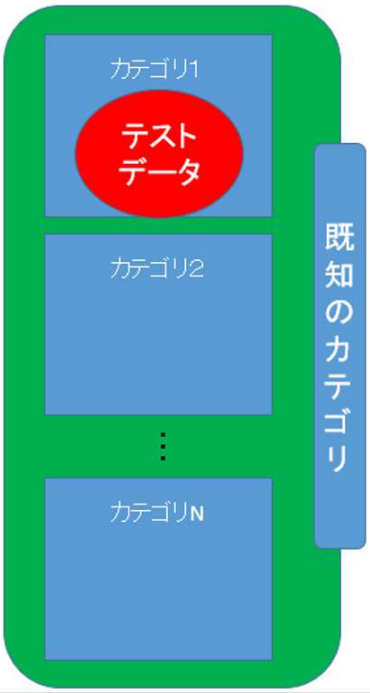

図 15 既知のカテゴリへの分類

3　実験

実験では、各特徴量ベクトルに対し、重みづけを行った上で識別を行い、識別率がどのように変化していくかを確かめる。

3.1　実験データ

実験では、顔画像と手の動きの動画を用いる。

顔画像については、学習用として、何も装着していない状態 (図 16) で被験者を撮影した動画のうちの 40 フレーム× 3(異なる場所で撮影したもの) × 4(人)=480 枚を、テスト用として、サングラスをした画像 (図 17)、マスクをした画像 (図 18) それぞれについて、50 × 3 × 4=600 枚を用意した。これらの画像の解像度は、1920 × 1080 ピクセルのものである。

手の動きについては、学習用に三角形の軌跡を描いた動画 2（一人当たり）× 4（人）=8 個を、テスト用に、8 × 4=32 個を用意した。これらの動画は、フレームレートが 30fps、解像度が 1920 × 1080 ピクセルのものである。

これらはすべて、正面から撮影したものである。顔と手の動きの特徴量ベクトルは、同じ被験者のもの同士でランダムに組み合わせて用いる。

図 16 なにも装着していない画像の例

図 17 サングラス画像の例

図 18 マスク画像の例

3.2　比較手法

比較手法として、顔と手の動きをスコアレベルで統合した手法を用いる (図 19)。顔と手の動きそれぞれの特徴量を別々に識別器に与え、評価値を求める。求まった評価値について、同じカテゴリ同士で足し合わせ、統合後に最小値となったカテゴリにテストデータを分類する。この比較実験では、手の動きの評価値 (d^i_{hand}) および顔の評価値 (d^i_{face}) に対し、式 (23) のように重みづけを行った。

$$d^i_{int} = \alpha\ d^i_{face} + (1 - \alpha)d^i_{hand}(0 \leqq \alpha \leqq 1) \tag{23}$$

α =0 の場合は手の動きの評価値のみ、α =1 の場合、顔の評価値のみで評価を行っていることを意味する。手の動きの評価値と、顔の評価値の足し合わせの組み合わせ方は、特徴量を統合したときと同じ組み合わせ方である。

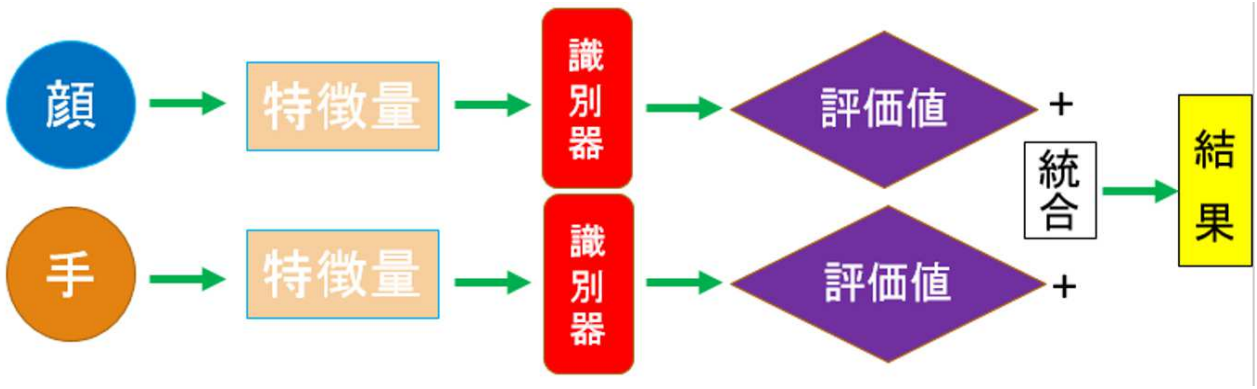

図 19 特徴量のスコアレベルでの統合の流れ

3.3 結果

表1から表4に、結果を示す。表1、表2は、被験者がサングラスを装着した顔画像での結果、表3、表4は、被験者がマスクをした顔画像での結果である。それぞれ、表1、表3が特徴量レベルでの統合、表2、表4がスコアレベルでの統合の結果である。

表 1 特徴量レベルでの識別結果 (サングラス画像)

α ＼被験者	1	2	3	4	総合
0.0	87.3%	99.3%	100%	76.0%	90.7%
0.1	87.3%	100%	100%	82.7%	92.5%
0.2	87.3%	100%	100%	88.0%	93.8%
0.3	87.3%	100%	100%	88.0%	93.8%
0.4	95.3%	100%	100%	88.0%	95.8%
0.5	100%	100%	100%	92.6%	98.2%
0.6	100%	100%	100%	97.3%	99.3%
0.7	100%	100%	100%	97.3%	99.3%
0.8	100%	100%	100%	97.3%	99.3%
0.9	100%	100%	100%	81.3%	95.0%
1.0	100%	100%	100%	72.0%	93.0%

表 2 スコアレベルでの識別結果 (サングラス画像)

α ＼被験者	1	2	3	4	総合
0.0	87.3%	99.3%	100%	76.0%	90.7%
0.1	87.3%	100%	100%	96.0%	95.8%
0.2	87.3%	100%	100%	100%	96.8%
0.3	87.3%	100%	100%	100%	96.8%
0.4	87.3%	100%	100%	100%	96.8%
0.5	87.3%	100%	100%	100%	96.8%
0.6	87.3%	100%	100%	100%	96.8%
0.7	87.3%	100%	100%	100%	96.8%
0.8	87.3%	100%	100%	100%	96.8%
0.9	87.3%	100%	100%	98.0%	96.3%
1.0	100%	100%	100%	72.0%	93.0%

表 3 特徴量レベルでの識別結果(マスク画像)

α ＼被験者	1	2	3	4	総合
0.0	87.3%	99.3%	100%	76.0%	90.7%
0.1	87.3%	99.3%	100%	76.0%	90.8%
0.2	87.3%	99.3%	100%	88.0%	93.6%
0.3	87.3%	99.3%	100%	88.0%	93.6%
0.4	87.3%	100%	98.0%	91.3%	94.1%
0.5	90.0%	100%	91.3%	98.6%	95.0%
0.6	91.8%	100%	77.3%	98.6%	91.8%
0.7	94.6%	100%	52.0%	98.0%	86.1%
0.8	95.3%	96.0%	40.0%	97.3%	82.1%
0.9	82.6%	84.6%	34.6%	96.0%	74.5%
1.0	72.6%	78.0%	20.7%	95.3%	67.8%

表 4 スコアレベルでの識別結果(マスク画像)

α ＼被験者	1	2	3	4	総合
0.0	87.3%	99.3%	100%	76.0%	90.7%
0.1	87.3%	99.3%	100%	96.0%	95.7%
0.2	87.3%	99.3%	100%	100%	96.7%
0.3	86.7%	99.3%	98.7%	100%	96.1%
0.4	86.0%	99.3%	98.0%	100%	95.8%
0.5	86.0%	99.3%	96.7%	100%	95.5%
0.6	86.0%	99.3%	92.0%	99.3%	94.1%
0.7	86.0%	99.3%	87.3%	99.3%	93.0%
0.8	86.0%	99.3%	70.0%	98.7%	88.5%
0.9	85.3%	98.0%	45.3%	98.0%	81.7%
1.0	77.3%	78.0%	20.7%	95.3%	67.8%

また、図20に、サングラス画像での識別結果、図21に、マスク画像での識別結果をグラフで示す。いずれについても、単独の特徴量で識別を行った場合より精度の向上がみられる。サングラス画像の場合は、特徴量レベルで識別を行った場合の識別率は最大99.3%、スコアレベルで識別を行った場合は最大96.8%であり、特徴量レベルで統合した場合の方が2.5ポイント良いという結果であった。一方で、マスク画像の場合は、特徴量レベルで識別を行った場合の識別率は最大95.0%、スコアレベルで識別を行った場合は最大96.7%であり、スコアレベルで統合した場合の方が1.7ポイント良いという結果であった。

これらのことから、特徴量レベルで統合したほうがスコアレベルで統合する場合より良い結果を得られる場合があるが、常に特徴量レベルでの統合のほうが良いわけではないということが分かった。

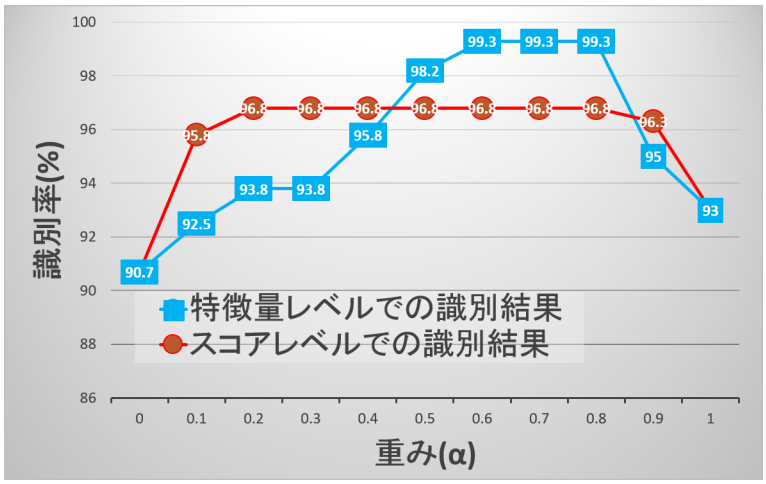

図 20 サングラス画像での識別結果

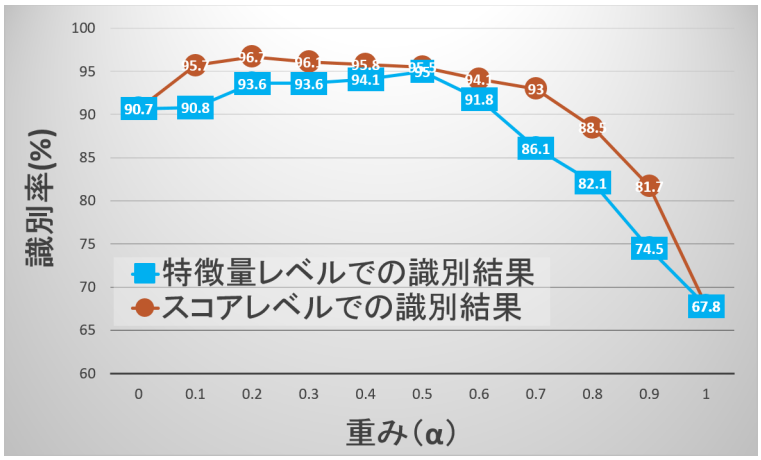

図 21 マスク画像での識別結果

4　おわりに

本研究では、顔と手の動きを使ったマルチモーダル個人識別において、特徴量レベルとスコアレベルそれぞれで情報の統合を行い、評価を行った。従来のマルチモーダル認証では、スコアレベルで情報を統合することが多かったが、特徴量レベルで情報を統合した場合のほうが良い結果が得られる可能性があるということが分かった。

顔の特徴量は、CNN を介して得られた質の高い特徴量であったのに対し、手の動きの特徴量は、手の座標から直接取得したものであったため、質があまり良いとはいえない。そのため、質の高い特徴量同士で統合を行った場合にどうなるは、検討するべきである。また、被験者数が少数であったこと、顔画像がすべて正面画像であったことから、被験者が多数の場合や顔画像が横顔であった場合には、特徴量レベルで統合しても良い結果が得られるとは限らないため、それらについても検討することが今後の課題として挙げられる。

謝辞

本研究を進めるにあたり、指導教員である椋木雅之教授には、大変お忙しい中、手法のアイデアやプログラムの実装、論文に関する助言など、大変丁寧にご指導いただいたこと、そして特に私に多くの時間を割いていただいたことを大変感謝いたします。また、貴重な時間を割いてお付き合いくださった実験の被験者の方々、何度も挫折しそうになった私に励ましの言葉をくださった椋木研究室の皆様にも感謝いたします。

また、顔の特徴抽出器として用いた davidsandberg/Facenet を開発した davidsandberg 氏にも感謝いたします。

参考文献

[1] 真部雄介, 齋藤隆輝, 嶋田弦, 菅原研次," 歩行・顔・身体のソフトバイオメトリック特徴を用いた正面観測個人認証", 知能と情報 (日本知能ファジィ学会誌),Vol.24,No.5,pp.988-1001,2012.

[2] 村松太吾, 岩間晴之, 木村卓弘, 槇原靖, 八木康史," 一歩行映像から取得される複数特徴を用いた個人認証", 電子情報通信学会論文誌 A バイオメトリクス小特集，Vol. J97-A, No. 12, pp. 735-748,2014.

[3] https://github.com/davidsandberg/facenet,2019 年 1 月 29 日閲覧.

[4] Florian Schroff, Dmitry Kalenichenko, James Philbin,"FaceNet: A Unified Embedding for Face Recognition and Clustering",IEEE Computer Society Conf,Computer Vision and Pattern Recognition(CVPR),pp.815-823,2015.

[5] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks"IEEE Signal Processing Letters (SPL), vol. 23, no. 10, pp. 1499-1503, 2016.

[6] 荻原弘樹, 椋木雅之,"Collaborative Mean Attraction 法による画像分類", 電子情報通信学会技術研究報告,Vol.116,No.259,PRMU2016-109,pp103-106,2016.

[7] Yang Wu,Masayuki Mukunoki,Michihiko Minoh,"Collaborative Mean Attraction for Set Based Recognition",MIRU2014,OS3-3,2014.