修士論文要旨

温度画像を用いた発声時の個人差にロバストな表情認識

(知能情報システム学)　中野真里

1. はじめに

人間同士のコミュニケーションにおいて顔の表情は重要な役割を果たしている。人とコンピュータの関係においても、コンピュータが人の表情を理解できればより円滑なコミュニケーションを実現することができる。そのための表情認識技術は実用可能なレベルにまで高める必要があり、現在まで様々な研究が行われてきた。

可視光画像を用いる場合、照明環境の変化が認識率に影響を及ぼす。顔の識別においてはそれを解決するための研究が行われている。しかし表情認識においては顔識別よりも画像におけるわずかな濃度変化を正確に抽出する必要性が高いので、より照明環境の変化の影響を受けやすいと考えられる。そこで、原理的に照明環境の変化に影響を受けない赤外線温度画像を用いた表情認識の研究が行われ、既報の研究ではその有効性が示されている[1],[2]。しかし同手法が多くの人物へ適用できるかどうかについては十分に検討されていなかった。

本研究では被験者を 7 人に増やし、既報の手法[1],[2]を用いて５表情（怒り・喜び・無表情・悲しみ・驚き）の認識実験を行った。複数の被験者が同じ手法で表情認識を行う場合、その個人差や画像撮影環境が認識に影響を及ぼすが、本研究では既報の手法[1],[2]で行われていた画像補正処理や音声処理に改良を加えることによりこの課題に対処し、個人差にロバストな表情認識を実現している。また、個人差が表情認識に及ぼす影響とその要因・度合についても検討した。

2. 処理概要

本研究における処理フローを図 1 に示す。

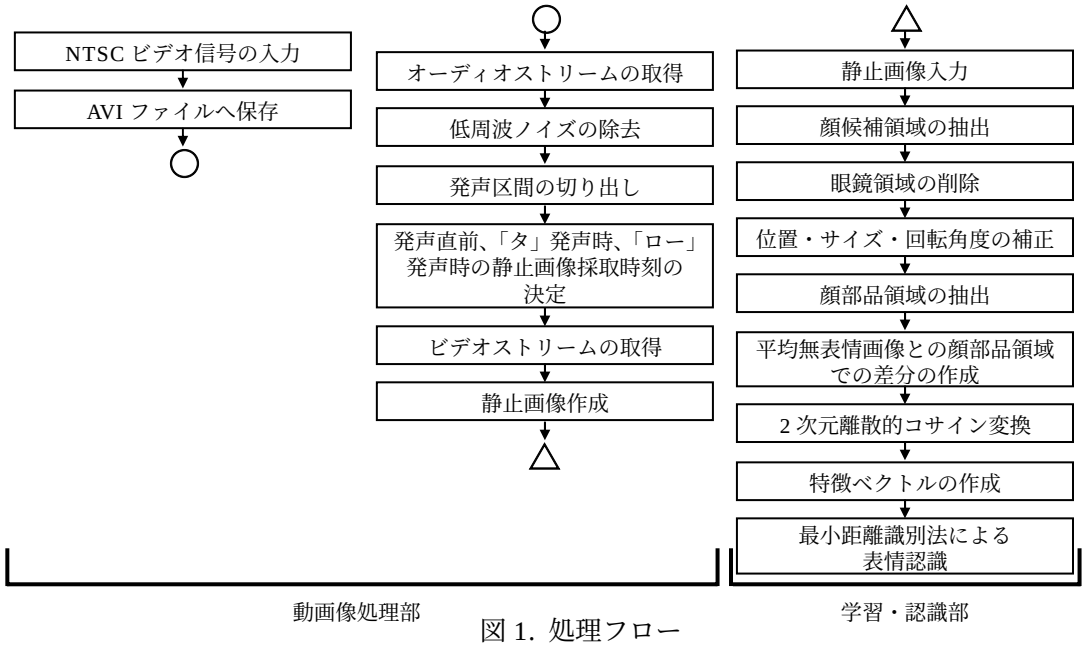

図 1. 処理フロー

本法における処理は、表情認識に用いる静止温度顔画像を採取する「動画像処理部」と、採取した画像を用いて表情認識を行う「学習・認識部」で構成されている。DirectShow プログラミングを用いて「NTSC ビデオ信号の入力～AVI ファイルの保存」を自動的に行うことで、これに続く「オーディオストリームの取得～最小距離識別法による表情認識」までのスレッド処理を可能にし、処理の効率化とオンライン処理を実現している。

3. 動画像処理部

被験者に「怒り」、「喜び」、「無表情」、「悲しみ」、「驚き」の各表情で中立的な言葉(日本名「taro」)を発声してもらったときの赤外線動画像を AVI ファイルに保存し、各 AVI ファイルから VFW の API 関数群を用いてオーディオストリームを取得して音声データを得る。被験者の音声データには既報の実験[1],[2]では見られなかったノイズが現れた。これは発声時またはその前後の息遣いの変動が原因であり、発声の音声データに比べて低周波であったため、ハイパスフィルタ処理を全音声データに施して除去した。遮断周波数は 3 人の被験者のノイズデータの周波数解析を行い、280Hz とした。図 2 にノイズ除去前後の音声データの例を示す。

(a) (b)

図 2. 低周波ノイズ除去前の音声データ(a)、低周波ノイズ除去後の音声データ(b)

続いて低周波除去後の音声データを基に、発声直前、「タ」発声時、「ロー」発声時の計 3 枚の静止温度顔画像を採取する。

4. 学習・認識部

動画像処理部で採取した静止温度顔画像に対して最大値フィルタを施し、頭部髪毛領域の濃度変動を極力吸収する。そして、人物が写っていない背景領域をあらかじめ決定し、「背景領域内の最大濃度値+1」を閾値として画像全体を 2 値化し、顔の横幅の最大値(フェレ径)を求める。温度画像においては顔画像内部に眼鏡領域が黒く残るので、顔の左右両端内の画素を白く埋めることで眼鏡領域を削除する。次いで顔の横幅がフェレ径の a %未満の部分を削除することで髪の毛領域を、さらに b %未満の部分を削除することで首領域を削除し、顔候補領域を抽出する。a、b は被験者ごとに設定するパラメータである。得られた顔候補領域を基に、回転角度補正、およびアフィン変換による顔のサイズ、位置の規格化を行った後、顔部品領域を抽出する。顔部品領域における平均無表情画像との差分をとり、純粋な表情変化のみを抽出した差分部分画像に、二次元離散的コサイン変換(2D-DCT)を施す。3 枚の画像から得た DCT 特徴量から、特徴ベクトル作成ルール[1]に基づいて、高い確率で表情を分類できる DCT 特徴量を選び、特徴ベクトルとする。特徴ベクトル作成ルール[1]には実験的に決定するパラメータ P が含まれており、被験者ごと

に設定している。そして学習画像から得た特徴ベクトル空間に未知画像の特徴ベクトルを入力し、最小距離識別法による表情認識を行う。

6. 実験方法

7 人の被験者 A～G に、「怒り」、「喜び」、「無表情」、「悲しみ」、「驚き」の各表情で中立的な言葉(日本名「taro」)を発声してもらい、そのときの AVI ファイルを作成し、各タイミングにおける静止温度顔画像を採取した。被験者は成人男性 5 名(うち眼鏡着用者 2 名)、成人女性 2 名(眼鏡着用者なし)である。それぞれの被験者において 1 表情につき 30 回の発声データを採取し、そのうち 20 データを学習用、10 データを認識用に用いた。赤外線カメラで設定した検出温度幅は 5℃、表情認識に用いる静止温度顔画像は 256 階調であるので、画像の濃度 1 階調は 1.95×10^{-2}℃に相当する。採取した静止温度顔画像を用いて、前述の各処理によって被験者 A～G の温度顔画像それぞれから特徴ベクトルを作成し、表情認識を行った。

7. 結果と考察

既報の手法による認識結果[2]を表 1 に、本法での認識結果を表 2 に示す。表情ごとの認識率を 7 人の被験者の平均値で表している。5 表情の平均認識率は 88 %と、既報の研究[2]における結果(84 %)とほぼ同等の高い認識率が得られた。なお、既報[2]および本法とも、最良の認識率を与える特徴ベクトル作成ルールのパラメータ P での認識結果を表記している。

「怒り」の表情は他の表情より認識率が低い。これは「怒り」の表情が他と比べて動きが大きく、様々なバリエーションが存在することが原因である。また、誤認識を起こした認識用画像の特徴ベクトルと 2 番目にベクトル距離の近い学習用画像を調べたところ、2 番目に近い画像が「怒り」であるものがあった。このことから、誤認識した一因は最小距離識別法によって認識していることにあると考えられる。

表 1. 既報の手法による認識結果[2]

		入力表情				
		怒り	喜び	無表情	悲しみ	驚き
認識	怒り	**70.0**	10.0		10.0	
	喜び	10.0	**80.0**		10.0	
	無表情			**100.0**	10.0	
	悲しみ	10.0			**70.0**	
	驚き	10.0	10.0			**100.0**

表 2. 本法での認識結果

		入力表情				
		怒り	喜び	無表情	悲しみ	驚き
認識	怒り	**77.1**	4.3			1.4
	喜び	14.3	**90.0**			2.9
	無表情	8.6		**97.1**		3.1
	悲しみ			2.9	**89.0**	7.1
	驚き		5.7		11.0	**85.4**

続いて表情認識に影響を及ぼす個人差の要因とその度合を検討した。表 3 は各被験者の特徴ベクトル次元数と、その要素の顔部品領域ごとの割合を示している。灰色に網掛けされた数値はその被験者の特徴ベクトル要素数が最も多かった顔部品領域とその割合を示している。

特徴ベクトルの次元数は被験者により異なっている。また、同じ被験者でもパラメータ P の与え方により異なるものである。表 3 から、特徴ベクトルの要素として選出される顔部品領域の割合には偏りが見られ、被験者ごとに異なっていることがわかる。このことから、個人によっ

て表情の特徴が出る顔部品領域には相違があると思われる。しかし、特徴が出てくる領域は実験環境や体調などとも関係して、個人によってもその時々で変化する可能性があるため、今後も検証が必要である。

本法で用いた特徴ベクトル作成ルール[1]には、実験的に決まるパラメータ P が含まれていて、この設定値によって選出される特徴ベクトルの要素数は異なる。本実験では、P の値を 60～83 %の範囲で検討した。P がこの範囲内のとき、平均認識率は被験者 A～G で最低 80 %、90 %、82 %、86 %、90 %、40 %、76 %と被験者 F を除いて極端に低下することはなかった。被験者 F においても P が 60～76 %の間であれば最低でも 78 %の平均認識率を得ることができている。このことから、P は 60～76 %の間であればどの値に設定しようとも高い認識率を得られることがわかる。このパラメータは個人差が表情認識に及ぼす影響を緩和するために与えたものであるが、今後は統一した値を用いることも可能であると言える。

表 3. 特徴ベクトルの要素として選出された顔部品領域の割合

	被験者 A	被験者 B	被験者 C	被験者 D	被験者 E	被験者 F	被験者 G
特徴ベクトル次元数	13	22	16	26	5	11	16
右頬	23.1%	18.2%	31.3%	3.8%	20.0%	36.4%	6.3%
鼻頭部	38.5%	13.6%	6.3%	0.0%	20.0%	0.0%	18.8%
左頬	7.7%	0.0%	37.5%	15.4%	40.0%	0.0%	25.0%
口右側	0.0%	27.3%	0.0%	19.2%	0.0%	0.0%	25.0%
口・顎	7.7%	9.1%	6.3%	34.6%	20.0%	36.4%	0.0%
口左側	23.1%	31.8%	18.8%	26.9%	0.0%	27.3%	25.0%

8. 結論

既報の手法[1],[2]を基にして、ハイパスフィルタによる音声のノイズ処理や眼鏡領域の削除処理、顔領域抽出のためのパラメータの設定を行い、個人差や撮影環境にロバストな表情認識手法を提案した。本手法を用いて 7 人の被験者で表情認識を行った結果、7 人の平均で 88 %という高い認識率が得られた。

今後はさらに被験者数を増やしたり、同じ被験者で撮影環境(季節や場所など)を変えて実験を行ったりして、実用化の検討を進めていく必要がある。また、現在は発声区間の切り出しや補正処理に時間がかかるため実時間処理は実現できていないが、今後は音声と画像の同時取得などによって実時間処理を行うことができる手法を検討する予定である。

参考文献

[1] 池添史隆，胡玲琴，谷尻豊寿，吉冨康成，「温度画像を用いた発声時の表情認識」，ヒューマンインターフェース学会論文誌，6(2004)，19-27.

[2] 池添史隆，中野真里，吉冨康成，田伏正佳，「発声時の温度顔画像の自動取得と表情認識」，ヒューマンインターフェースシンポジウム 2005 論文集，(2005)，7-12.