修士論文要旨

温度画像を用いた発声時の表情認識

（知能情報システム学）池添史隆

1．序論

人間の感情は顔の表情によくあらわれるため、顔の表情は、日常でのコミュニケーションにおいて重要な役割を果たしている。そこで将来、人間とコンピュータが平和に自然な協調作業をしていく為には、コンピュータが人間の表情を理解する必要がある。これまでは主に、可視光画像を用いた表情認識の研究が行われてきた。しかし、可視光画像を用いた表情認識では、異なる照明条件下での正確な認識が一般的に困難である。そこで、赤外線による温度画像を用いる表情認識手法の研究が行われてきた[1]。

本研究では、既報の研究[1]においては手動で行われてきた、音声波形を基にした動画像からの静止画像抽出の自動化を実現した。また、動きが伴う、典型的で実用上重要な場面として発声時を選び、自動閾値法[2]による 2 値化やテンプレートマッチング等を用いて温度画像から顔部品領域の抽出を行った。そして、特徴ベクトル空間での最小距離識別法による表情認識を行い、その性能を評価した。

2．処理概要

本研究における処理手順の流れを図 1 に示す。

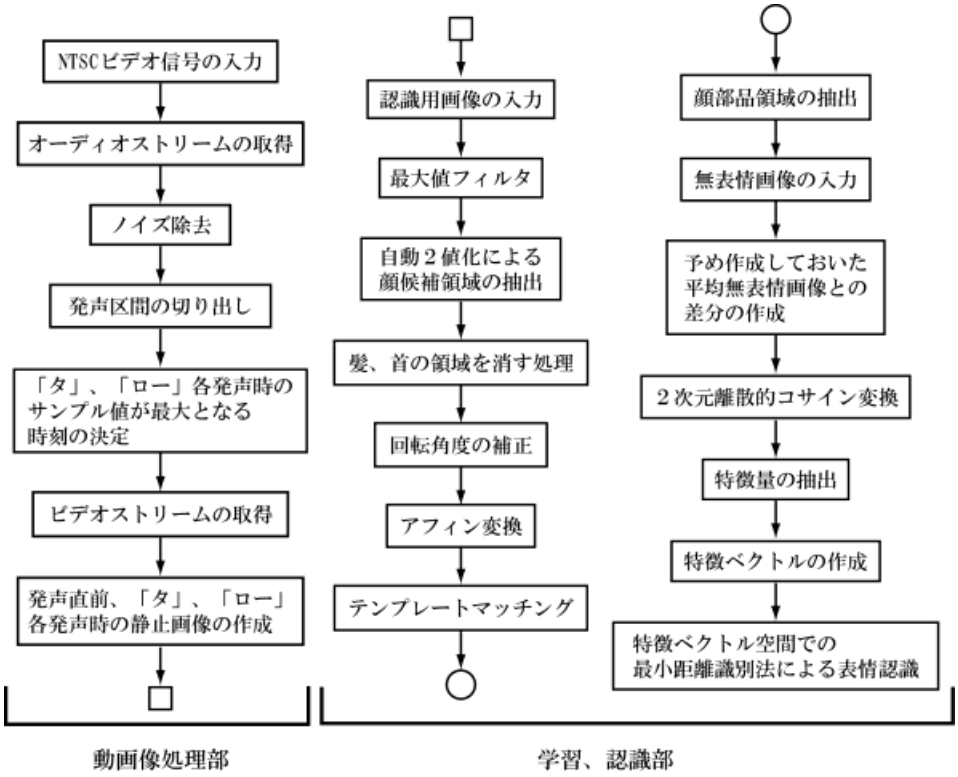

図 1　処理フロー

動画像処理部では、入力された NTSC ビデオ信号（AVI ファイル）からオーディオストリームを取り出し、得られた音声データを基に発声区間の切り出しを行う。そして、切り出された発声区間ごとに「タ」、「ロー」各発声区間に分割し、サンプルの値の絶対値が最大の値をとる時刻を各々決定する。そして、それらの時刻と発声直前におけるフレームを切り出し、静止画像（ビットマップ形式）を得る。

学習、認識部では、動画像処理部で作成された顔温度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、自動閾値法[2]による 2 値化を行い、髪、首の領域を消して顔候補領域を抽出する。得られた顔候補領域を基に、回転角度補正、及びアフィン変換による顔のサイズ、位置の規格化を行う。更にテンプレートマッチングによる上下方向の補正を行った後、顔部品領域を抽出する。次に、顔部品領域において平均無表情画像との差分を作成し、各顔部品領域に対して 8×8 画素毎に 2 次元離散的コサイン変換（2D-DCT）を施す。得られた各 2D-DCT 係数の絶対値ごとに、各顔部品領域内での平均値（以下、「DCT 特徴量」と表記）をとる。得られた DCT 特徴量から更にルールを用いて選出し、特徴ベクトルを作成する。そして、特徴ベクトル空間での最小距離識別法による表情認識を行う。学習用画像については、予め同様にして特徴ベクトルを作成しておく。

3．動画像処理部

AVI ファイルは Video For Windows の API 関数群によって操作する。ここで、AVI ファイルは Type2 DV-AVI 形式であり、音声データはサンプリングレート 48kHz、ビットレート 16bit のステレオ音源であった。

発声中の値を決める、無音の値（ノイズ）との閾値は、開始 1 秒間（1 秒以内に発声が始まるものは 0.5 秒間）の無音区間におけるサンプル値から求める。図 2 のように、無音区間におけるサンプル値のヒストグラムは正規分布に従うとみなす事ができる。平均値を m、標準偏差をσとすると、無音区間における全サンプル値がほぼ全て（1 個未満）無音の値の範囲に含まれるには、m±5σを閾値とすればよい。僅かに残った発声中の値とみなされるノイズを除去した後、閾値を基にして発声区間の切り出しを行う。次に、学習用データから発声区間内における「タ」、「ロー」各音素の時間比を算出する。これを基に、各発声区間を更に、各音素の発声区間に分割する。そして各音素の発声区間におけるサンプル値の絶対値が最大となる時刻を求め、その時刻の静止画像を採取する。切り出された発声区間の例を、図 3 に示す。

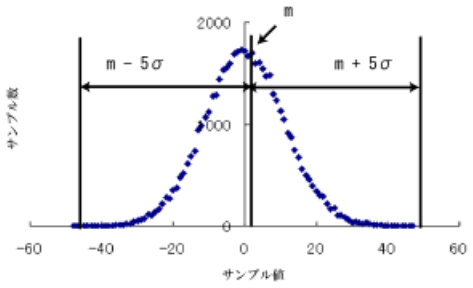

図 2　開始 1 秒間の無音区間おける
サンプル値のヒストグラム

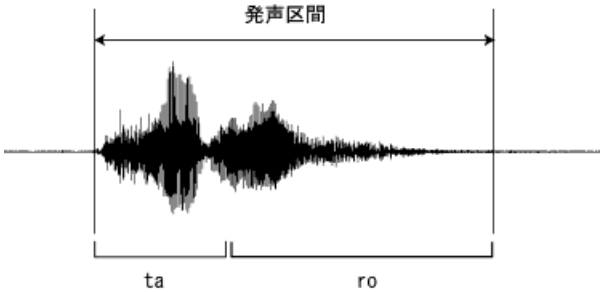

図 3　切り出された発声区間

4．学習、認識部

図 4(左図)に示す顔温度画像が入力されると、まず最大値フィルタを施した後に自動閾値法[2]による 2 値化を行い(図 4(右図))、髪、首の領域を消して顔候補領域を抽出する。得られた顔候補領域を基に、顔の回転角度、サイズ、位置等の補正を行った後、更にテンプレートマッチングによる上下方向の補正を行う。ここで、これまでの研究[3]では、顔候補領域における上辺と下辺の各中点を結んだ直線から顔の回転補正角度を求めていた。しかしこの場合、上辺と下辺における顔候補領域の抽出精度の影響が大きい。そこで、顔候補領域内における各水平方向の中点座標を求め(図 5(左図))、その回帰直線の傾きから顔候補領域の面内回転を表す角度θを算出した（図 5(右図))。θ=90°を回転無しとして 90°からのずれを補正角度とし、顔候補領域と温度画像に対する回転の補正を行った。各補正後に顔部品領域を抽出し、平均無表情画像との差分を作成する。差分における各顔部品領域に対して 8×8 画素毎に 2D-DCT を施し、各 2D-DCT 係数から計 90 個の DCT 特徴量を得る。この 90 個の DCT 特徴量から、予め定めた実験的パラメータを基に作成したルールに従い、学習画像のデータを表情毎に分類しやすいものを選ぶ。選出された DCT 特徴量から特徴ベクトルを作成し、特徴ベクトル空間での最小距離識別法による表情認識を行う。

5．実験

中立的な言葉（日本名「taro」）を女性アナウンサーに発声してもらい、その時の「怒り」、「喜び」、「無表情」、「悲しみ」、「驚き」の 5 つの表情画像を採取した。1 表情に対して 30 データを採取し、学習用画像 20 サンプル、認識用画像 10 サンプルとした。前述の各処理から 17 次元の特徴ベクトルを作成し、表情認識を行った。また、自動的に採取した静止画像に対しても同様の処理を行い、12 次元の特徴ベクトルによる表情認識を行った。ここで、各 DCT 特徴量から特徴ベクトルの候補を選出する基準には実験的に定めたパラメータが 1 つあり、その値は画像のマニュアル採取の場合と自動採取の場合で異なったものを用いた。また、自動的な静止画像採取の対象とした動画像は、既報の研究[1]と共通である。

6．結果と考察

既報の手法[1]における表情認識の結果を表 1 に、本法での認識結果を表 2 に示す。対象とした画像は共通である。本法を用いた事で平均 92%の認識率が得られ、既報の手法[1](平均 56%)に比べて高い認識率を得る事ができた。また、本法で画像を自動採取した場合でも、平均 90%の高い認識率を得る事ができた(表 3)。

図 4　左図：入力画像
　　右図：2 値化後の画像

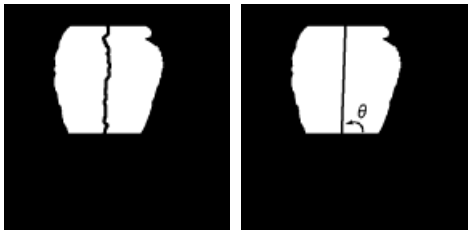

図 5　左図：水平方向における各中点座標
　　右図：回帰直線による回転角度θの算出

表 1　既報の手法[1]による認識結果

		入力表情				
		怒り	喜び	無表情	悲しみ	驚き
認識	怒り	**0**				
	喜び		**90**	20	10	20
	無表情			**80**		50
	悲しみ	70	10		**90**	10
	驚き	30				**20**

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	**100**	10		30	
	喜び		**90**			
	無表情			**100**		
	悲しみ				**70**	
	驚き					**100**

表 3　自動採取した画像での認識結果

		入力表情				
		怒り	喜び	無表情	悲しみ	驚き
認識	怒り	**70**	10			
	喜び		**90**			
	無表情			**90**		
	悲しみ	30		10	**100**	
	驚き					**100**

学習用に用いた画像を目視観察した結果、「怒り」の表情には種々のバリエーションがあり、再現性が低い傾向が見られた。このような「怒り」の感情の表出の際の変動が、特徴ベクトルの分散の要因となり、未知画像においても他の表情が「怒り」の表情に認識される傾向がでたものと考えられる。このような特異なクラスが存在する場合の学習と認識の方法については、今後とも更に検討を進める必要がある。また、特徴ベクトルの作成方法の再検討や、特徴ベクトルの最適次元数の決定において、主成分分析等の多変量解析法の適用を検討する必要がある。

7．結論

自動閾値法による 2 値化、回転角度の補正、そしてテンプレートマッチングを用いた上下移動の補正によって、より正確な顔部品領域の抽出が可能となった。また、特徴ベクトル空間での最小距離識別法による認識を行うことで、5 つの表情に対して平均 92%の認識率を得た。そして、AVI ファイルから音声データを取得し、そこからのノイズの除去、及び発声区間の切り出し、静止画像の抽出が可能となった。また、本法で行った動画像からの自動的な画像採取によって得られた静止画像を用いた表情認識においても、平均 90%の認識率を得た。

参考文献

[1] Yoshitomi, Y., Kim, S., Kawano, T. and Kitazoe, T.: Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face; Proc. of 9th IEEE Int. Workshop on Robot and Human Interactive Communication, pp.178-183, 2000.

[2] 吉冨康成，谷尻豊寿：画像の 2 値化装置および方法;特願 2002-344185，2002.

[3] 胡玲琴：発声時の温度顔画像からの顔部品抽出に関する研究;京都府立大学大学院人間環境科学研究科環境情報学専攻 修士論文，2003.