ロボットの挙動選択のための 顔情報と音声情報を統合した感情判別

松本 祥平 † 山口 健 ‡ 駒谷 和範 ‡ 尾形 哲也 ‡ 奥乃 博 ‡
† 京都大学工学部情報学科 ‡ 京都大学大学院情報学研究科知能情報学専攻

1 はじめに

近年、人間とインタラクションを行うロボットの研究が盛んになりつつある。人間・ロボット間のインタラクションの高度化、あるいはソーシャルインタラションのためには、感情認識は重要である。感情は多様な形で現われるので、センサー情報を用いたボトムアッププロセスと言語情報、話し方、心理状態などを用いたトップダウンプロセスを併用する必要がある。

本研究ではトップダウンプロセスは今後の課題として、ボトムアップによる感情認識に焦点を当てる。ただし、単一モダリティでは曖昧性が多いので、顔情報、音声情報から話者の感情を推定する。顔情報からは快・不快情報と驚きの強度を、音声情報からは快・不快情報を取得し、3 次元の感情空間の中で話者がどこに位置しているのかを判断する。

2 個別モダリティによる感情認識

2.1 顔表情からの 2 次元の情報の抽出法

本研究では FACS[1] に基づいて表情認識を行う。[3] FACS は表情記述法の 1 つで、AU(Action Unit) と呼ばれる表情動作の最小単位を組み合わせて表情を記述する。AU は Ekman らによって提案された基本 6 感情 [2](驚き、恐怖、嫌悪、怒り、幸福、悲しみ) との対応が明らかになっている。基本 6 感情の認識に必要な AU のうち、本研究で認識対象とする AU を表 1 に示す。ただし、話者の発話区間においては、表情を作るための動きではなく、単に発話のための動きである可能性が高いので AU20, AU25, AU26 は認識対象としない。

AU1	眉の内側を上げる	AU10	上唇を上げる
AU2	眉の外側を上げる	AU12	唇端を引っ張り上げる
AU4	眉を下げる	AU15	唇端を下げる
AU5	上瞼を上げる	AU20	唇を横に引っ張る
AU7	瞼を緊張させる	AU25	顎を下げずに唇を開く
AU26	あごを下げて唇を開く		

表 1: 認識対象とする Action Unit (AU)

AU の認識は図 1 に示した特徴点を抽出し、特徴点間の距離を特徴量としてその変動から AU の認識を行う。顔検出は、ロボットと人間が対面しているという仮定をおき、入力画像中で最大の肌色領域が顔であると想定して行なっている。唇の検出は上唇と下唇の間の輝度値が低いこと利用してまず唇の間を検出する。さらに唇領域は顔領域の他の部分と比べて YIQ 表色系の Q 成分が高いことを利用して端点の抽出を行う。目と眉の検出では、顔領域の上半分をグレイ化して値が小さくかつ面積が大きな領域を目と眉の候補とし、顔領域における位置情報、前フレームにおける目と眉の位置情報などを用いて目と眉の領域を決定する。以上のようにして抽出した特徴点間の距離の変動を AU と対応づけ、AU_i の発現度 $\mathcal{A}_i$ を次式で定義する。

$$\mathcal{A}_i = 100\frac{v_i - v_i^{(c)}}{v_i^{(m)} - v_i^{(c)}}$$

v_i : AU_i と対応する特徴量
$v_i^{(c)}$: 平静時の v_i の値
$v_i^{(m)}$: v_i の最大値

この $\mathcal{A}_i$ を算出することで AU の認識を行う。

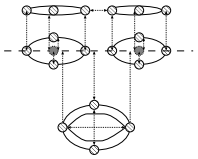

図 1: 抽出する特徴点と特徴量

平静時の値はインタラクションの開始直後に取得した値の中央値で代用し、最大値はインタラクションの中]で更新していくことで事前学習なしでの認識を可能にしている。

基本 6 感情を肯定的感情 (幸福) $\mathcal{F}_p$、否定的感情 (恐怖、嫌悪、怒り、悲しみ) $\mathcal{F}_n$、驚き $\mathcal{F}_s$ の 3 グループに分ける。したがって、顔情報から快・不快情報と驚きの強度 C、S を次の式により求める。

$$\mathcal{C}_{\mathrm{def}} = \mathcal{A}_p - \mathcal{A}_n$$

$\mathcal{A}_p$: 肯定的感情を記述する AU の発現度の和
$\mathcal{A}_n$: 否定的感情を記述する AU の発現度の和

$$\mathcal{S}_{\mathrm{def}} = \mathcal{A}_s$$

$\mathcal{A}_s$: 驚きを記述する AU の発現度の和

2.2 音声情報からの 1 次元情報の抽出法

音声においては言語的情報、韻律的情報が話者の感情を推定するための有力な手がかりとなる。本研究では韻律的情報に焦点を当て、一発話毎に基本周波数の最大値、初期値、平均値最大値と最小値の差、パワーの最大値、平均値、発話時間の特徴量を抽出する。感情表現の個人差に対応するために、前発話との差分値やその差分値を現在の値で正規化した

Emotion Recogntion based on Audio-Visual Integration for action selection systems of robots *Shohei Matsumoto, Takeshi Yamaguchi, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno (Kyoto University)*

値、さらに第一発話で正規化した値も用いて SVM (Support VectorMachine) で喜び、困惑の有無を判別する [4]。

音声情報から得られる、喜び判別用の SVM の境界面からの距離 $\mathcal{J}$ と困惑判別用の SVM の境界面からの距離 $\mathcal{P}$ とから快・不快情報 $\mathcal{V}$ を次の式で求める。

$$\mathcal{V} = \mathcal{J} - \mathcal{P}$$

3 システムの設計

本研究では、顔画像から得た感情、音声から得た感情はそれぞれに意味があると考え、両方の情報をロボットの挙動選択部へ送る。3 次元の話者の感情空間を定義し、話者の状態がその空間の中のどこにあたるかを推定し、その 3 次元ベクトルを挙動選択部へ渡す。

第 1 軸に音声情報から得た話者の快、不快情報 ($\mathcal{V}$) を、第 2 軸に顔画像から得られる快、不快情報 ($\mathcal{C}$) を、第 3 軸に驚きの強さ ($\mathcal{S}$) を反映させる。

$\mathcal{V}$ に関しては、発話毎に算出される SVM の境界面からの距離を用いて定義通りに計算した値を新たな $\mathcal{V}$ の値とする。

$\mathcal{C}, \mathcal{S}$ に関しては、表情は比較的短時間しか表われないが、表情が消えた途端に感情も消え去るわけではないので減衰関数を導入する。まず $\mathcal{C}_{\mathrm{decay}}$ を以下のように定義する。

$$\mathcal{C}_{\mathrm{decay}} = \sum_{i=1} \mathcal{C}_{t-i} e^{-\alpha i}$$

ただし、α は感情によって減衰する速度が異なることを考慮するための係数である。この $\mathcal{C}_{\mathrm{decay}}$ と $\mathcal{C}_{\mathrm{def}}$ を用いて

$$\mathcal{C} = \mathcal{C}_{def} + \mathcal{C}_{decay}$$

を新たな $\mathcal{C}$ の値とする。$\mathcal{S}$ も同様である。減衰関数を導入したことによって、小さな感情の積み重ねを大きな感情だと判断する、同じ表情でも時系列情報によって異なる結果を出すといったことが可能となる。

単なる感情の判別結果だけではなく、以上のようにして求めた 3 次元ベクトルをロボットの挙動選択部に渡すことで、ロボットは話者の感情の強度を反映したモーター速度や話速で対応したり、タスクによって挙動決定の閾値を変動させたり、発話直後は音声情報からの感情を重視し、それ以外のときは顔情報からの感情を重視するなどといったことができるようになる。

4 実験

システムの動作例を示すための実験を行った。被験者の正面にカメラを置き、マイクを持って話してもらう。被験者は「(1) 快の表情→ (2) 快の発話→ (3) 不快の表情→ (4) 不快の発話→ (1) 快の表情→ (2) 快の発話」を意図的に行う。

4.1 結果・考察

$\mathcal{V}, \mathcal{C}, \mathcal{S}$ の値を図 2,3,4 に示す。また、発話区間を矢印で、被験者の意図的に出した感情との対応を円と番号で示した。発話区間が終了してから $\mathcal{V}$ の値が変わるまでの 2,3 秒のずれがはピッチ抽出などの処理にかかる時間である。その点に注意すれば被験者の意図と一致するかたちで感情認識ができていることがわかる。

図 3 に関してはノイズや影の影響で正確に特徴量を抽出できず、誤認識を起しているところも見受けられる。時系列情報を重視して感情認識を行うなどの対応が考えられる。

図 4 の (*) に注目すると、発話区間付近で意図せず $\mathcal{S}$ の値が上昇している。これは発話区間では顔の動きが大きく、顔器官の抽出が困難になったために誤認識が起っているからである。発話区間においては AU20, 25, 26 を認識対象から外しているが、他の AU の認識についても検討が必要である。

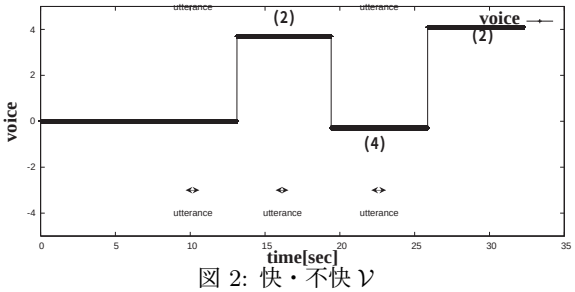

図 2: 快・不快 $\mathcal{V}$

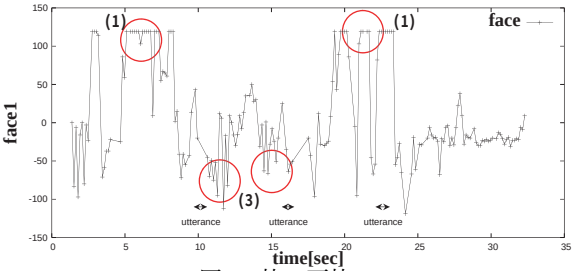

図 3: 快・不快 $\mathcal{C}$

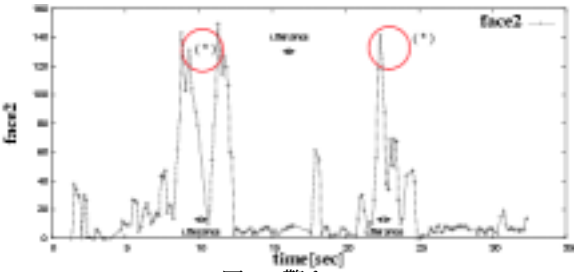

図 4: 驚き $\mathcal{S}$

5 まとめ

カメラ、マイクの入力信号から話者の感情状態の推定を行った。推定結果を表示する GUI も作成したのでこのシステムによってロボットは話者の感情に応じた挙動選択を GUI でロボットの内部状態を示しながら行うことが可能となる。今後、システムの評価実験、ロボットへの実装、人間とインタラクションを行なっている被験者の感情状態を推定した場合の結果の考察を行う予定である。本研究の一部は、科研費、21 世紀 COE、SCAT、栢森財団の支援を受けた。

参考文献

[1] P.Ekman and W.V.Friesen. *The Facial Action Coding System.* Consulting Psychologists Press, 1978.

[2] P.Ekman, W.V.Friesen, 工藤訳. 表情分析入門. 誠信書房, 1987.

[3] 下田宏, 國弘威, 吉川榮和. 動的顔画像からのリアルタイム表情認識システムの試作. ヒューマンインタフェース学会論文誌, Vol.1, No.2, 1999.

[4] 伊藤亮介, 駒谷和範, 河原達也, 奥乃博. ロボットとの音声対話におけるユーザの心的状態の分析. 情報処理学会研究報告, 2003-SLP-45-18, 2003.