K_059

# 感情認識における画像情報と音声情報の統合

## Integration of visual and audio information on emotion recognition

青田 亨† 松本 哲也† 竹内 義則‡ 工藤 博章† 大西 昇†
Toru Aota Tetsuya Matsumoto Yoshinori Takeuchi Hiroaki Kudo Noboru Ohnishi

## 1. はじめに

近年、コンピュータやロボットなどの機械が人間の生活において担う役割は、ますます重要なものになりつつある。これに伴って、より柔軟な人間的マン・マシンインタフェースに対する要求が高まりつつある。

人間同士のコミュニケーションにおいては、非言語的情報の重要性は古くから知られており、その中でも顔表情や音声等に自然に表れる「感情」は非常に重要な役割を果たしている。人間がコミュニケーションにおいて、顔表情や音声の両者から相手の感情を認識しているにもかかわらず、従来の研究では画像情報単独[1]や音声情報単独[2][3]からの感情認識が行われているが、この両者を組み合わせた研究は報告されていない。そこで、本研究では画像情報による感情認識と音声情報による感情認識を統合することで、認識率の向上を行う。

## 2. 動画データ

人物に対して、平静・喜び・怒り・悲しみの 4 感情を表出して発話させ、発話の様子を正面からビデオカメラにて撮影し動画データを収録した。発話する単語には感情の表出しやすい単語を用い、単語に依存しないように人物に自由に決定させた。単語長は 2~12 文字からなり、以下が各感情に対する単語例である。

- 平静 ：「はい」「こんにちは」
- 喜び ：「やった」「ありがとう」
- 怒り ：「違うよ」「まだ決まってないんですか」
- 悲しみ：「できません」「すいません」

## 3. 画像情報による認識

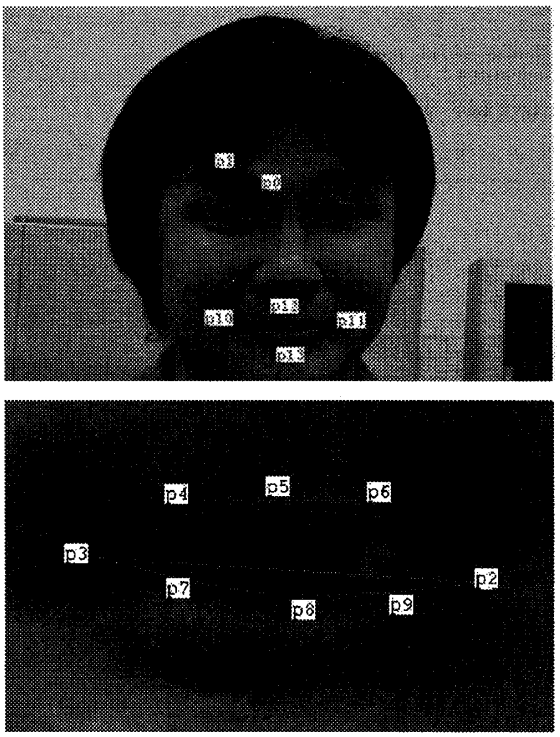
画像特徴量として動画データより抽出した顔画像より顔部位(眉・目・口)の 14 の特徴点 p0～p13 を抽出した。

次に、抽出した特徴点の外接矩形を抽出し、外接矩形の大きさが 1 となるよう正規化を行った。この正規化された特徴点の座標より得られた 28 次元の特徴ベクトルを最終的な画像特徴量とした。認識手法としては、入力サンプルと学習サンプルのユークリッド距離が小さい上位 k 個の学習サンプル中の割合から、各感情の尤度を出力した。

## 4. 音声情報による認識器

音声特徴量として、発話音声より得られたピッチの差分系列の平均、最大、最小、標準偏差を用いた。

動画データより抽出した音声を固定長フレームに分割し、各フレームに対して LPC 残差波形の自己相関値のピーク




p0：眉の内側の端点　p1：眉の中点
p2：目頭　p3：目尻
p4～p9：線分 p2p3 を 4 等分した点を通る垂線と瞼との交点
p10：口の左端　p11：口の右端
p12：口の上端　p13：口の下端

図 1 画像特徴量抽出例

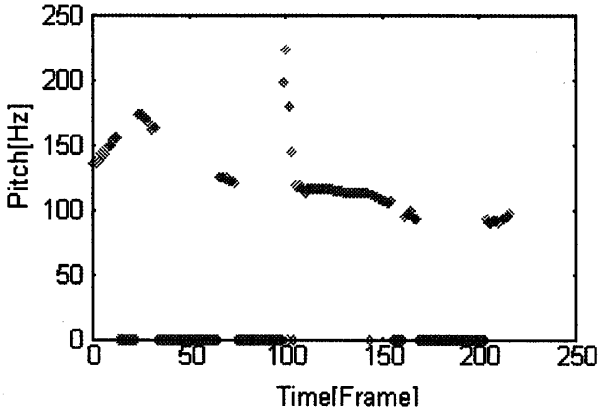



図 2 ピッチ系列抽出例

†名古屋大学大学院情報科学研究科
‡名古屋大学情報セキュリティ対策推進室, 理化学研究所 バイオミメティックコントロール研究センター

よりピッチを抽出することでピッチ系列を抽出した。用いた固定長フレームのフレーム長は 50ms、フレームシフトは 5ms とした。図 2 は抽出したピッチ系列の抽出例である。次に、このピッチ系列中の各点に対して、前点との差分をとることでピッチの差分系列を抽出する。その後、得られた差分系列より平均、最大、最小、標準偏差を抽出した。

認識手法は画像情報による認識器同様、ユークリッド距離の小さい上位 k 個の学習サンプル中の感情の割合より各感情の尤度を出力した。

## 5. 画像情報と音声情報の統合

上述した特徴量を用いて、画像情報単独、音声情報単独による認識を行い、得られた認識結果を統合した。

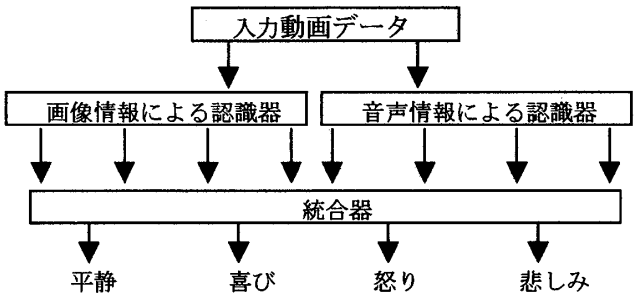



図 3 処理の流れ

入力動画データより顔画像と発話音声を抽出し、その後、それぞれを画像情報による認識器、音声情報による認識器に入力する。

統合器は 3 層パーセプトロンにより構成され、各認識器が出力した各感情の尤度を入力とする。学習サンプルを元に誤差逆伝播法によりユニット間の重みを更新することで学習を行い、入力サンプルに対しては各感情の尤度を出力する。教師データとしては学習サンプルにラベル付けされた感情を 100%、その他の感情を 0%として与えた。

入力動画データにラベル付けされた感情と、システムが出力した尤度最大の感情が一致した場合、システムは感情を認識したものと評価し、尤度最大の感情が複数現れる場合には認識しなかったものとした。

## 6. 実験

演劇経験のない 20 代前半男性 4 人の動画データを撮影し、人物ごとに 4 感情に対して 10 個、計 40 個の動画データを用意した。各動画データより最も感情が表出されていると判断した画像フレームを抽出し、このフレームを画像情報による認識器に入力する顔画像とした。顔画像のサイズは 720×480 であり、音声のサンプリングレートは 32kHz である。

各情報による認識器のパラメータ k=9、3 層パーセプトロンの入力層、中間層、出力層のユニット数はそれぞれ 8、50、4、誤差逆伝播法における学習率は 0.5 とした。

評価方法として Leave-one-out 法を用い、画像情報による認識率、音声情報による認識率、各情報の統合による認識率を調査・比較した。

表 1 各認識器に対する認識率(%)

| 人物 | 画像 | 音声 | 統合 |
|---|---|---|---|
| A | 55.0 | 45.0 | 90.0 |
| B | 72.5 | 12.5 | 100.0 |
| C | 47.5 | 47.5 | 87.5 |
| D | 52.5 | 40.0 | 77.5 |
| 平均 | 56.9 | 36.3 | 88.8 |

表 2 各感情に対する認識率(%)

| 人物 | 画像 | 音声 | 統合 |
|---|---|---|---|
| 平静 | 62.5 | 65.0 | 92.5 |
| 喜び | 67.5 | 40.0 | 95.0 |
| 怒り | 47.5 | 27.5 | 87.5 |
| 悲しみ | 50.0 | 10.0 | 80.0 |

表 1、表 2 は各情報を統合することにより画像情報単独、音声情報単独による認識と比較して、高い認識率が得られたことを示している。高い認識率が得られた原因として、各認識器の出力パターンに対する統合器の期待する出力が正しく学習できたことが挙げられる。

表 1 において人物 B の音声情報による認識率が他の人物と比較して低かった原因として、尤度最大の感情が複数現れる場合が多く見られたことが挙げられる。しかし、画像情報の付加によってこのような問題は解消され、高い認識率を実現している。

表 2 において悲しみの音声情報による認識結果が他の感情と比較して低いにも関わらず、統合による認識率が高かった原因として、音声情報による認識器の出力結果自体は低いものの、その出力パターンは類似しており、出力パターンに対する統合器の期待する出力を学習できたためであると考えられる。

全体的に見て、画像情報と音声情報の統合により、一方のみでは不完全であった情報が相互補間されることによって、高精度の感情認識が実現されることが確認された。

## 7. まとめ

感情認識における画像情報と音声情報の統合手法を提案し、画像情報単独、音声情報単独による認識と比較して、高い認識率が得られた。

現状では、学習サンプルに用いる人物と、入力サンプルに用いる人物は同一であり個人に特化しているため、今後の課題としては個人に特化しない認識手法の確立や、照明条件や雑音などにロバストな認識手法の確立が挙げられる。

## 参考文献

[1] 松野勝弘,李七雨,辻三郎:"ポテンシャルネットとKL展開を用いた顔表情の認識", 電子情報通信学会論文誌 D-II No.8 pp.1591-1600, 1994.

[2] 北原義典,東倉洋一:" 音声の韻律情報と感情表現", 信学技報SP88-158, 1988.

[3]Sherif Yacoub, Steve Simske, Xiaofan Lin, and John Burns:"Recognition of Emotions in Interactive Voice Response System", Proc. EuroSpeech 2003, 2003.