2J1-04

対話型ロボットのための口領域動画像に基づく発話推定

Speech Activity Detection Using Mouth Image Sequences for an Interactive Robot

元吉大介[*1] 嶋田和孝[*2] 榎田修一[*2] 江島俊朗[*2] 遠藤勉[*2]
Daisuke Motoyoshi Kazutaka Shimada Shuichi Enokida Toshiaki Ejima Tsutomu Endo

[*1]九州工業大学大学院情報工学研究科
Graduate School of Computer Science And Systems Engineering, Kyushu Institute of Technology

[*2]九州工業大学情報工学部
Faculty of Computer Science And Systems Engineering, Kyushu Institute of Technology

In this paper, we describe a method of speech activity detection for an interactive robot. The method detects the speech events by using mouth image sequences captured from a USB camera. We calculate the sum of absolute difference and optical flow from the mouth image sequences. Finally, the method classifies an image into two states (speech activity or non-activity) by using a machine learning algorithm.

1. はじめに

近年，生活支援ロボットや受付ロボットなど，人間と自然なコミュニケーションをとるロボットに関する研究が盛んに行われている．これらのロボットと人間のコミュニケーションにおいては，音声発話によってコマンドを入力することが多いが，ロボットが複数の人間に囲まれている場合，人間同士の会話を誤ってコマンド入力として受け取り，誤動作を起こす可能性がある．そこで，ロボットはコマンド入力者の発話にのみ応じる必要があり，発話推定の技術が必要となる．Matsumoto ら [Matsumoto 04] は，顔方向や注視方向を測定し，人間がロボットを見ている間だけ発話に応答する手法を提案した．しかし，この手法では，複数の人間がロボットの周りに存在する場合，コマンド入力者以外の発話に誤って応答してしまう可能性がある．増田ら [増田 06, 増田 07] は，唇領域の動静判定を行うことで発話区間の推定を行った．簡単な判別手法を採用しているにも関わらず高精度であることを報告しているが，論文 [増田 06] では，同じサイズの唇領域しか検出できないという問題点がある．また，論文 [増田 07] では，唇の詳細な形状を検出するために EBGM という複雑な手法を用いている．

本研究では，高速かつ頑健に口領域を検出し，検出した口領域の動画像から発話推定を行う手法を提案する．口領域を検出する際，正面顔領域内の特定の領域に制限した画像を高解像度化及びヒストグラムの均一化をした画像で検出処理を行うことで，高速かつ頑健に口領域を検出する．発話推定では，フレーム間で口の動静判定を行うことで発話の有無を検出する．口の動静判定には，1 つ前のフレームと現フレームの口領域画像から求めたオプティカルフローと絶対値差分和を特徴量として用いる．この 2 つの特徴量について数フレーム前から現フレームまでを素性とし，決定木による判別手法である C4.5 [Quinlan 93] を利用して発話推定を行う．本研究では，上記の手法と単純に現フレームの各特徴量のみを用いた手法とを比較することで，2 つの特徴量を統合的に扱いかつ数フレームの情報を利用する手法の有効性を実証する．

2. システムの概要

作成するシステムは，正面顔及び顔部品検出部と発話推定部で構成される．本章では，それぞれについて詳しく説明する．

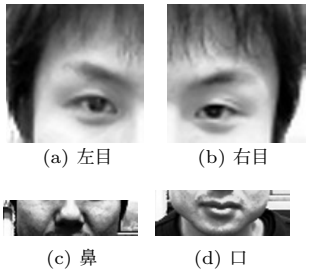
連絡先: 元吉大介，九州工業大学，〒 820-8502 福岡県飯塚市川津 680-4，d_motoyoshi@pluto.ai.kyutech.ac.jp

2.1 正面顔及び顔部品検出部

正面顔及び顔部品検出部では，Viola ら [Viola 01] が提案し，Rainer ら [Rainer 02] によって改良された物体検出器を用いて正面顔と顔部品 (両目，鼻，口) を検出する[*1]．利用する物体検出器は，Intel がオープンソースで公開しているコンピュータビジョン関連のライブラリである OpenCV に実装されているため，容易に利用が可能である．正面顔については，画像全体で検出処理を行っても高速かつ頑健に検出可能であるが，顔部品については，背景や服装の一部を誤検出することが多く，処理速度も遅いという問題点がある．そこで，顔部品検出に関しては，検出精度と処理速度の向上のため，以下に示す追加処理を行う．

- 検出処理領域について，左目及び右目は顔領域の左上半分及び右上半分，鼻は顔領域の目より下の上半分，口は顔領域の鼻より下に制限する．
- 制限した領域を高さと幅ともに 2 倍に高解像度化して，更にヒストグラムの均一化した画像で検出処理を行う．

検出処理領域の制限により，背景や服装などの誤検出の解消や処理領域の削減に繋がり，検出精度と処理速度ともに向上すると考えられる．高解像度化とヒストグラムの均一化を行う理由は，検出処理領域の画像サイズは大きいほうが検出精度が良いことと，逆光時に画像全体が暗くなることで検出精度が低くなることが実験的に分かっているためである．図 1 に各顔部品の検出処理領域を，図 2 に顔部品検出結果例を示す．



(a) 左目 (b) 右目

(c) 鼻 (d) 口

図 1: 顔部品検出処理領域

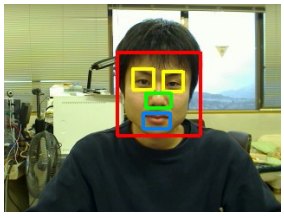


図 2: 顔部品検出結果例

[*1] 口以外の顔部品は，今後顔の方向推定や人物認証などに応用利用する予定である．

2.2 発話推定部

発話推定部では，正面顔及び顔部品検出部で検出された口領域の動静判定を行うことで，現フレームが発話中か否かの判別を行う．本節では，発話推定部について詳しく説明する．

口の動きを測定するための特徴量として，武田ら [武田 03] が読唇に用いたオプティカルフローと，増田ら [増田 06] が唇の動静判定に用いた絶対値差分和のそれぞれを特徴量として使用する．オプティカルフローを用いた特徴量としては，1 つ前のフレームと現フレームの口領域画像からブロックマッチング法によりオプティカルフローを求め，その大きさの総和を画像サイズによって正規化した値を用いる．画像サイズで正規化するのは，ブロックマッチング法によるオプティカルフローは，画像サイズの大小に比例するため，そのままオプティカルフローの総和を特徴量として用いると，分類器が生成時の画像サイズに依存してしまうという問題点があり，これを防ぐためである．絶対値差分和とは，1 つ前のフレームと現フレームの対応する全画素値の差の絶対値和である．絶対値差分和はオプティカルフローと同様，画像サイズの大小に比例するため，画像サイズで正規化した値を特徴量とする．また，2 つの特徴量算出の際，1 つ前のフレームと現フレームの口領域の画像サイズを比較し，小さい方の画像サイズに合わせる．これは，オプティカルフローと絶対値差分和ともに，1 つ前のフレームと現フレームの画像サイズは同じでなければ算出はできず，フレーム毎に検出される口領域のサイズは異なる場合が多いためである．画像のリサイズ手法には，バイキュービック法を採用する．

発話推定手法として，この 2 つの特徴量について数フレーム前から現フレームまでを素性として，決定木による判別手法である C4.5 を利用することで発話中か否かの判定を行う手法を採用する．

3. 実験

3.1 実験環境

動画像を撮影する USB カメラには，Logicool の Qcam Pro 9000 を使用した．撮影された画像のサイズは 320×240 である．PC のスペックについては，CPU が Intel Core2 Duo 3GHz，メモリが 3GB である．

3.2 顔部品検出実験

顔部品検出の精度を求めるため，正面顔を含む 100 枚の画像と背景のみの 50 枚の画像の計 150 枚の実験画像データを用意した．実験については，追加処理無しと追加処理有りの比較実験を行った．追加処理無しと有りそれぞれの顔部品検出実験結果を表 1 に示す．ここで，追加処理無しに関しては，左目と右目の区別を行っていないため，両目ともに検出できた場合を正解としている．表 1 の結果より，追加処理により顔部品全ての検出精度は向上しており，追加処理は有効であるといえる．また，処理速度については，実験画像データ全体の平均で追加処理無しの場合は 1 フレームあたり 225.7ms(4.43fps) だったのに対して，追加処理有りの場合は 1 フレームあたり 50.94ms(19.63fps) と大幅に向上した．

3.3 発話推定評価実験

発話推定評価実験として，2 つの特徴量について 2 つ前のフレームから現フレームまでの計 6 つを素性として，C4.5 で学習し，発話推定を行った．実験データとして，443 発話フレームと 513 非発話フレームからなる 956 フレームの動画像データを用意した．実験方法には，10 分割交差検定を採用した．2 つの特徴量を統合的に扱いかつ数フレームの情報を用いる手法の有効性を確認するための比較対象として，2 つの特徴量を単体で用い，C4.5 により決定された閾値により口の動静判定をする発話推定実験も行った．オプティカルフロー (OF) と絶対値差分和 (SAD) それぞれ単体による実験結果と，2 つの特徴量について 2 つ前のフレームから現フレームまでを C4.5 の素性とした実験結果 (PREV2) を表 2 に示す．表 2 より，OF と SAD に比べて PREV2 の結果が良いことから，2 つの特徴量を統合的に扱いかつ数フレームの情報を用いることは有効であるといえる．

表 1: 顔部品検出実験結果

	追加処理無し			追加処理有り			
	目	鼻	口	左目	右目	鼻	口
再現率	0.32	0.03	0.72	0.90	0.83	0.40	0.90
適合率	0.88	0.09	0.22	1.00	1.00	0.98	0.99
F 値	0.47	0.05	0.33	0.95	0.91	0.57	0.94

表 2: 評価実験結果

	OF	SAD	PREV2
再現率	0.77	0.66	**0.88**
適合率	0.71	**0.87**	0.86
F 値	0.74	0.75	**0.87**

4. おわりに

本研究では，音声発話をコマンドとして入力する対話型ロボットを想定し，コマンド入力者の発話区間推定手法を提案した．実験結果より，特徴量単体による閾値判別手法に比べて，2 つの特徴量を統合的に扱いかつ数フレームの情報を用いる手法の方が有効であることが分かった．今後は，特徴量の追加や C4.5 以外の分類器により実験を行い，更なる精度向上を目指す．また，今回の実験データに含まれる人物は全て同一であったため，複数人の実験データを用いて実験することで，手法の汎用性についても調査する予定である．

謝辞

本研究は，次世代ロボット知能化技術開発プロジェクト（独立行政法人新エネルギー・産業技術総合開発機構）における「施設内生活支援ロボット知能の研究開発」の成果の一部である．

参考文献

[Matsumoto 04] Y, Matsumoto., J, Ido., K, Takemura., M, Koeda., and T, Ogasawara.: Portable Facial Information Measurement System and Its Application to Human Modeling and Human Interfaces, The Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04), (2004).

[増田 06] 増田 健, 松田 博義, 井上 淳一, 有木 康雄, 滝口 哲也: 唇領域の動静判定と音声・雑音判定の統合に基づく発話区間の検出, 画像の認識・理解シンポジウム (MIRU2006), (2006).

[増田 07] 増田 健, 青木 政樹, 松田 博義, 滝口 哲也, 有木 康雄: EBGM を用いた唇の形状抽出による発話区間の検出, 画像の認識・理解シンポジウム (MIRU2007), (2007).

[Quinlan 93] Quinlan, J. R.: C4.5 Programs for Machine Learning, Morgan Kaufmann Publishers, (1993).

[Viola 01] P, Viola., M, Jones.: Robust Real-time Object Detection, Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, (2001).

[Rainer 02] Rainer, L., Alexander, K., Vadim, P.: Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection, MRL Technical Report, (2002).

[武田 03] 武田 和大, 重留 美穂, 小野 智司, 中山 茂: オプティカルフローによる読唇の研究, 2003 PC Conference, (2003).