ロボットとの対話のための発話推定に関する事例研究

元吉大介・嶋田和孝・榎田修一・江島俊朗・遠藤勉（九州工業大学）

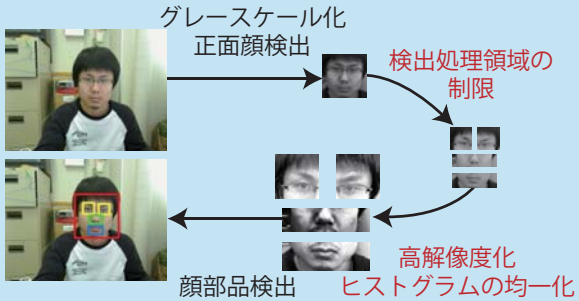
発表番号
IS3-26

研究背景・目的

- 対話型ロボット
  - 人間同士の会話に誤って応答
    - 話者の特定・発話区間推定
- 提案する発話推定システム
  - 現フレームが発話か非発話か
  - 口領域の検出とその動きによる発話推定

口領域検出

- Viola & Jones の物体検出器の利用
  - 口検出の精度と処理速度の問題
- 追加処理
  - 検出領域の制限
  - 高解像度化
  - ヒストグラムの均一化

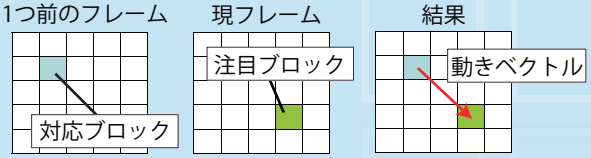


発話推定

- ２つの特徴量を利用
  - オプティカルフロー（OF）と絶対値差分和（SAD）

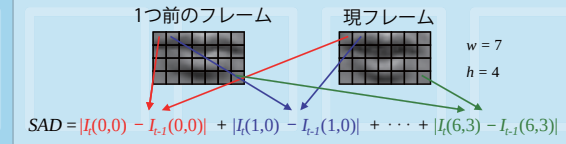
オプティカルフロー

- ブロックマッチング法



- 特徴量
  - 全ブロックのフローの大きさの総和
  - 画像サイズで正規化

絶対値差分和

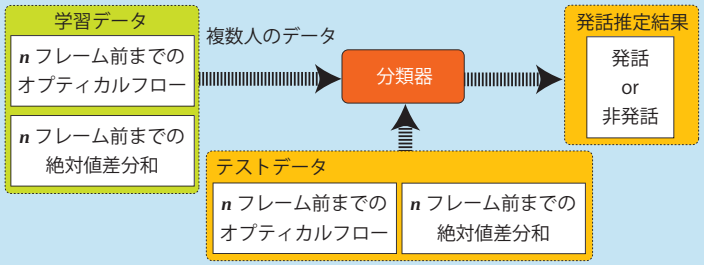


$$SAD = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} |I_t(i,j) - I_{t-1}(i,j)|$$

- 特徴量
  - 画像サイズで正規化

発話推定手法と分類器

- 数フレーム前までの情報を利用



- 分類器：様々な種類を比較
  - C4.5，Random Forest，SVM，Naive Bayes，k-NN，AdaBoost

実験

- 実験環境
  - -USB カメラ：Logicool Qcam Pro 9000
    - 画像サイズ：320×240
  - -PC スペック
    - Intel Core2 Duo 3GHz & メモリ 3GB
- 実験 1：顔部品検出
  - –実験データ：被験者 1 名
    - 画像データ 150 枚
      - 正面顔＋背景：100 枚
      - 背景のみ：50 枚
- 実験 2：発話推定
  - 実験データ：被験者 4 名
    - 3694 フレームの動画データ
      - 発話区間：1827 フレーム
        - 「あいうえお」など母音の異なる言葉が続くように発話
      - 非発話区間：1867 フレーム

実験結果：顔部品検出

	追加処理なし			追加処理あり			
	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
処理速度	4.43fps			19.63fps			

実験結果：発話推定

分類器	*n*	再現率	適合率	F 値
C4.5	0	0.759	0.663	0.708
	1	0.747	0.739	0.743
	2	0.784	0.755	0.769
	3	0.787	0.761	0.774
	4	0.764	0.768	0.766
	5	0.766	0.752	0.759
	6	0.774	0.752	0.763
	7	0.782	0.743	0.762
	8	0.768	0.755	0.761
	9	0.766	0.752	0.759
	10	0.765	0.756	0.761
Random Forest	0	0.598	0.665	0.625
	1	0.683	0.734	0.707
	2	0.742	0.777	0.759
	3	0.757	0.792	0.774
	4	0.774	0.815	0.794
	5	0.783	0.808	0.795
	6	0.783	0.811	0.797
	7	0.778	0.811	0.794
	8	0.785	0.817	0.801
	9	0.771	0.827	0.798
	10	0.775	0.815	0.795
SVM	0	0.624	0.723	0.670
	1	0.727	0.760	0.743
	2	0.796	0.786	0.791
	3	0.816	0.793	0.804
	4	0.824	0.807	0.816
	5	0.825	0.808	0.817
	6	0.819	0.820	0.819
	7	0.825	0.822	0.824
	8	0.829	0.819	0.824
	9	0.828	0.823	0.825
	10	0.832	0.825	0.828

分類器	*n*	再現率	適合率	F 値
Naive Bayes	0	0.621	0.721	0.667
	1	0.713	0.760	0.736
	2	0.783	0.792	0.787
	3	0.812	0.800	0.806
	4	0.826	0.805	0.816
	5	0.819	0.810	0.814
	6	0.828	0.818	0.823
	7	0.836	0.817	0.827
	8	0.837	0.820	**0.829**
	9	**0.839**	0.820	**0.829**
	10	**0.839**	0.816	0.827
k-NN (k=9)	0	0.670	0.671	0.670
	1	0.750	0.739	0.745
	2	0.797	0.767	0.782
	3	0.819	0.788	0.803
	4	0.831	0.801	0.816
	5	0.824	0.803	0.813
	6	0.825	0.809	0.817
	7	0.830	0.807	0.818
	8	0.826	0.806	0.816
	9	0.819	0.807	0.813
	10	0.808	0.808	0.808
AdaBoost	0	0.640	0.604	0.622
	1	0.697	0.740	0.718
	2	0.757	0.776	0.766
	3	0.773	0.787	0.780
	4	0.795	0.808	0.801
	5	0.808	0.823	0.815
	6	0.793	0.830	0.811
	7	0.809	0.822	0.816
	8	0.804	**0.836**	0.820
	9	0.810	0.826	0.818
	10	0.800	0.825	0.812

オプティカルフロー (OF) 単体　再現率：0.700　適合率：0.646　F 値：0.627
絶対値差分和 (SAD) 単体　再現率：0.785　適合率：0.662　F 値：0.718

- 分類器
  - Naive Bayes がベスト
  - SVM もほぼ同値
- 時系列情報
  - –時系列情報なし (***n***=0)
    - 精度の低下
  - –過去の情報は有効
- 特徴量の組み合わせ
  - OF 単体と SAD 単体
    - 低精度
  - –組み合わせは有効

まとめ

- 顔部品検出
  - 高精度かつ高速な検出が可能
- 発話推定
  - 複数の特徴量と時系列の有効性
- 今後の課題
  - 被験者の追加
  - 精度向上へのアプローチ
    - 新たな手法や特徴量の調査

D. Motoyoshi, K. Shimada, S. Enokida, T. Ejima and T. Endo

A Case Study of Speech Activity Detection for an Interactive Robot