K-063

ライフログのための対話検出

Dialogue Detection for Lifelog

熊谷 怜史† 全 炳東‡
Satoshi Kumagai Heitoh Zen

1. はじめに

個人的な体験をライフログとして取得・活用するための研究が盛んである．この研究は身体に装着したセンサのデータから対話区間を推定し，さらに対話の相手，内容，状況等によるインデクシングをもとにした活用を目指している．この報告では頭部に装着したカメラとマイクによる画像・音信号から，対面での対話区間を検出する手法について述べる．カメラからの顔検出とマイクからの発話検出のみでは対話区間の正確な推定は困難だが，両者を組み合わせることにより，複雑な背景や騒音の大きな環境下でも良好な対話区間検出が可能なことを示す．

2. ライフログと人／対話の検出

ライフログとは人の生活や行動(Life)をデジタルデータとして記録(Log)することである．例えばカメラや加速度計などの異種センサを装着しやすい小型パッケージにまとめた，ライフログに役立てようとする試みがある [1]．また加藤らは，肩に装着したアクティブ(パン-チルト)カメラを用いて周囲の人物の顔を追跡し，出会った人を記録する研究を行っている[2]．同じように高松らの研究でも顔検出によって会った人を記録している [3]．

顔の検出だけではなく，音（声）による会話の検出，記録を試みる研究もある。Choudhury らは，Sociometer と呼ばれるウェアラブルデバイスを，対話を交わす人物それぞれに装着させることで対話の検出を行い，さらに人間関係の解析を試みている [4]．また [5]や[6]では，複数人に装着したデバイスとユビキタスセンサを統合することで，会話などのインタラクションを検出することを提案している．

この研究でもカメラとマイクを用いた対話（対面での会話）の検出を行う．装着したカメラやマイクは独立して動作するので，対話の相手が同じシステムを装着する必要はない．得られたデータからケプストラムによる母音検出と，Haar-Like 特徴による顔の検出をそれぞれ行い，両者を統合することで対話区間を決定する．Tancharoen らも，カメラとマイクを用いている．まず音圧，周波数などの分析によりデータを区間に分割し，会話を行っている区間のみを抽出する．つぎに会話区間のビデオから顔が検出されるか否かで区間の重要性を分類している [7]．

3. 対話区間の検出

図 1 に実験に使用した機材（カメラ，マイク）と装着の様子を示す．3.1 以降に述べるように，カメラのビデオ画像から顔検出を，またマイクの音から発話（母音）区間の検出を行い，両者が重なる区間を対話区間として検出する．図 2 にその様子を模式的に示した．次節以降で顔検出，発話検出それぞれの手法について述べる．

3.1 顔検出

顔検出手法は Lienhart[8]らの手法を用いる．この手法は，Haar-Like 特徴を 1 つの弱識別器として用い，それらを複数用いた強識別器を構成し，さらにこれらを複数直列に接続した検出器を用いる．実装には OpenCV に付属する正面顔・横顔検出用の検出器をそれぞれ用いた．

3.2 発話検出

人の声の母音に含まれるピッチ(基本周波数)に着目し，発話検出を行う．ピッチ検出にはケプストラムを用い，高ケフレンシー部の最大値と平均値を比較することで判定を行う．

図 1　カメラとマイク

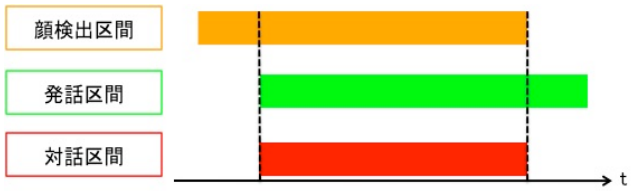

図 2　対話区間の検出

3.3 区間検出

以上の手法で顔と声を検出するが，実際には時間の短い検出漏れや誤検出がしばしば発生する．

そこで前後の区間における検出の割合に着目する．発話検出の場合，図 3 のようにピッチが検出された区間を 1，それ以外の区間を 0 とし，これを図 4 のように移動平均をとり閾値以上の区間を発話区間とする．顔の検出の場合は，

† 千葉大学大学院工学研究科
Graduate School of Engineering Chiba Univ.
‡ 千葉大学総合メディア基盤センター
Inst. Media and Info. Tech, Chiba Univ.

ある時刻に対し，その前後の区間における顔が検出された画像フレームの枚数を求め閾値以上の区間を顔検出区間とする．

4. 実験

4.1 実験方法

USB カメラとマイクを図 1 のように頭部に装着し，実際に対話を交わした画像と音信号を取得した．画像は解像度 640×480 でフレームレート 15fps，音信号はサンプリングレート 22050[Hz]で取得し後処理にて対話区間の検出を行った．

4.2 実験結果

図 5 に上から顔検出区間，発話検出区間，検出された対話区間，実際の対話区間を示す．また，表 1 に対話区間と実際の対話区間の開始時刻と終了時刻を示す．

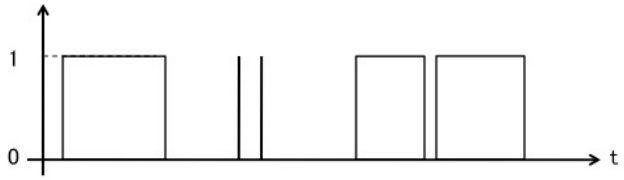

図 3　ピッチが検出された区間

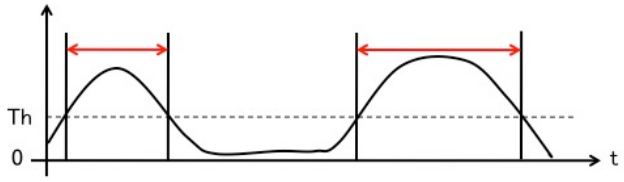

図 4　移動平均

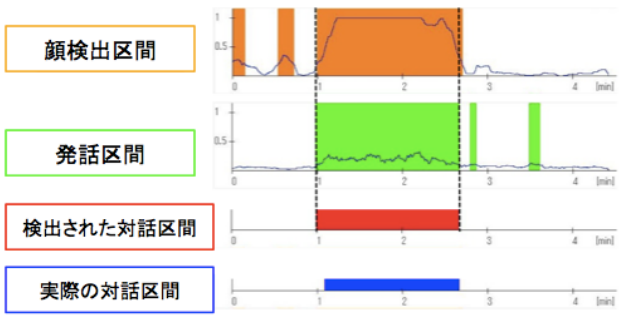

図 5　実験結果

表 1　検出された対話区間と実際の対話区間

	開始時刻	終了時刻
検出された対話区間	59.56[sec]	160.08[sec]
実際の対話区間	63.51[sec]	160.00[sec]

5. 考察

顔検出区間では，開始 0 秒と 30 秒付近で誤って検出されてしまっている．これは，人の顔でない背景を誤検出したためである．また発話検出に関しても同様に周囲の騒音，主に他人の話し声による誤検出がある．しかし，顔と声の検出両方を組み合わせることによってそれらの誤検出の影響を抑えることができた．

6. おわりに

本研究では，頭部に装着したカメラ，マイクを用いて対話を検出する手法を提案した．顔検出，発話検出それぞれでは誤検出となるような場面においても両方を用いることによってその影響を抑えられることが実験で示された．

今後はカメラ，マイクを正面だけでなく横方向や後方に装着し，横に並んでの会話など様々な状況に対応させていきたい．また，さらにセンサの種類を増やすことで，対話検出の精度を向上させていきたい．例えば，骨伝導マイクを用いれば，自分と相手の声を識別することによって，発言の割合が推定できるので周囲の騒音に頑強に対話の検出や状況の推定を行うことができると考えられる．

参考文献

[1]J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell, ”Passive Capture and Ensuing Issues for a Personal Lifetime Store”, ACM CARPE2004, pp.48-55, 2004.

[2]加藤 丈和 , 蔵田 武志 , 坂上 勝彦, ”VizWear-Active -記憶補助のための顔画像検出,追跡,登録-”, ITE Technical Report Vol.25, No.85, pp.41-46, VIS2001 - 102, 2001.

[3]高松創介, HAYASHI Oribe,西村 邦裕, 谷川 智洋, 廣瀬 通孝, ”顔情報を用いたライフログ利用に関する研究”, 映像情報メディア学会技術報告, 33(21), pp.73-78, 2009.

[4]Tanzeem Choudhury, Alex Pentland, ”The So- ciometer: A Wearable Device for Understanding Human Networks”, CSCW’02 Workshop: Ad hoc Communications, 2002.

[5]間瀬 健二他, ”インタラクションに基づく体験共有コミュニケーション”, 情報処理学会論文 誌, コンピュータビジョンとイメージメディア 48(SIG 1(CVIM 17)), pp.53-64, 2007.

[6]中田 篤志, 角 康之, 西田 豊明, ”非言語情報の出現パターンによる会話状況の特徴抽出”, 情報処理学会研究報告, UBI, [ユビキタスコンピューティングシステム], 2009-UBI-24(13), pp. 1-8, 2009.

[7]Tancharoen Datchakorn, 河崎 晋也, 山崎 俊彦, 相澤 清晴, “個人的なビデオのための会話の検出”, 電子情報通信学会総合大会講演論文集 2005 年_情報・システム(2), 171, 2005-03-07

[8]Rainer Lienhart, Jochen Maydt “An Extended Set of Haar-like Features for Rapid Object Detection”, IEEE ICIP 2002, Vol. 1, pp.900-903, (2002).