2013-Interaction (2EXB-21)
2013/3/1

マルチモーダルデータを用いた映像閲覧者の興味推定

倉野 大二郎[1,a)]　松村 耕平[1,b)]　角 康之[1,c)]

概要：タブレット型 PC やスマートフォンなどのモバイルデバイスの普及により，映像コンテンツを閲覧する形態が変わりつつある．その事により，外で歩きながらや家で寝転びながら端末を持って映像を閲覧する事を可能とし，映像を見る際のデバイスと映像閲覧者との身体的関わりが深くなったといえる．本論文では，モバイルデバイスを用いて，映像閲覧者の無意識的な行動や発話，しぐさなどからその映像における閲覧者の興味の変化を推定するシステムについて述べる．

Estimating Video Viewer's Interests by Multimodal Data

Kurano Daijiro[1,a)]　Matsumura Kohei[1,b)]　Sumi Yasuyuki[1,c)]

Abstract: With the spread of mobile devices such as smart phones or tablet PCs, a method of viewing a video content has been changed. Thus, we can view the video while walking at out, and while lying down at home. In other words, the physical interaction of the viewer and a mobile device became deep. In this paper, we describe a system that estimates the change in the interest of a viewer from a gestures and speech when a viewer is viewing a video using a mobile device.

1. はじめに

近年では，YouTube やニコニコ動画など，映像クリエーターが自由に映像を制作・公開し，それを手軽に視聴することができる投稿型映像閲覧サービスが普及している．さらに，Ustream やニコニコ生放送のように，遠隔からイベントを「横目」で眺めるような状況が増えてきた．また，タブレット型 PC やスマートフォンの普及によって，そのようなサービスの閲覧方法も従来のテレビを閲覧するような形態から，閲覧者が端末を手に持って閲覧する形態へと変化しつつある．この変化によって，他の作業をしながら（移動しながら，家事をしながら，仕事の合間に）映像閲覧することが可能になり，「ながら」的に映像を閲覧することが多くなってきた．

また，近年ではユーザの発話やしぐさなどの非言語行動を計測し，人の行為や会話の意味を推測する試みが活発に行われている．その技術的なアプローチとして，マイクやビデオカメラ，モーショントラッキングなどの多様なセンサを複合して用いるマルチモーダルセンシングが注目されている．

本研究では，多様なセンサを持ち，映像閲覧のデバイスとしても普及しつつあるタブレット型 PC を用いて，マルチモーダルセンシングによって取得したデータを利用し，端末を持っている状態での映像閲覧者の発話やしぐさ，映像閲覧時の状況などから，閲覧している映像に対して閲覧者がどのように興味を変化させるかを推測する．

本研究によって，CM クリエーターなどの映像制作者が意図して制作したビジュアルアテンションに対して映像閲覧者がどのような興味の変化を起こすか，即座にフィードバックすることが期待できる．

2. 関連研究

2.1 マルチモーダルデータの利用

人は会話において言語的な発話の他に，視線やジェス

[1] 公立はこだて未来大学
Future University of Hakodate
[a)] d-kurano@sumilab.org
[b)] matsumur@acm.org
[c)] sumi@acm.org

チャ，相槌などの非言語的な行動によって様々な意図を表現している．これらの非言語行動には一定の時間的・空間的なパターンがある [1]．このような人同士のコミュニケーションで生じる非言語的な行動を計測することをマルチモーダルセンシングと呼ぶ．マルチモーダルセンシングによって取得したデータは主に発言内容を構造化したデータと発言場面の映像データ，音声データから構成されており，各々のデータをリンクさせることで発話時の表情や状況などを振り返ることができる [2]．

本研究ではマルチモーダルセンシングによって取得したデータを利用し，映像閲覧者が映像に対して反応した時の表情や状況を振り返ることによって映像閲覧者の興味推定を行う．

2.1.1 ユーザの意図・興味を探り情報検索・提示を行うシステム

平山らは，対話的コミュニケーションを通してユーザの意図や興味を推定し，情報を検索・提示するシステムの実現を目指している [3]．これを実現するためにロボットを含む多様なインタフェースを用いてユーザに対して気の利いたタイミングで対話を積極的に行っている．また，宮崎らは移動中の個々のユーザに対して，それぞれの状況を検知・認識し，適切な情報を提供できるキオスク端末型情報システムを実現するために，エージェント指向処理モデルの提案を行っている [4]．

本研究では，ロボットなどのユーザとは異なるエージェントの存在によって意図された対話を用いずに，映像を閲覧しているユーザの反応や状態変化によって映像閲覧者の興味や意図を推定する．

2.1.2 マルチモーダルなセンシングよるユーザの状態把握

河原は，学会などで一般に行われているポスター会話においての，相槌，頷き，視線配布などの聴衆の反応に着目し，そのような情報からいつ誰がどのような質問をするか予測したり，聴衆の興味度を推定することを検討している [5]．また，Bohus らは多人数で行われる会話において，ユーザの視線や身振りから連続された会話の中で，会話のターンを読み取り，適切なタイミングでその会話にエージェントが介入する研究を行なっている [6]．

これらの研究は実験を行うために大きな空間と多数のデバイスを必要とし，ユーザがデータを取得するデバイスとの身体的な接触がない事から，デバイスにとってデータを取得できる位置にユーザが寄ることで初めてシステムが成り立つ．本研究では端末を手に持った状態で情報コンテンツを閲覧するため，閲覧者がどういう状態なのかをいつでもダイレクトに知ることができる．また，画像認識や音声認識などの高度な分野に踏み込まず，抽象度の高いデータから閲覧者の興味を推定する事ができる．

2.2 映像閲覧者の興味推定

映像を閲覧している際の視線や顔向きから映像閲覧者の興味の変化や注目の度合いを図る試みは電子広告などに応用されている．

2.2.1 閲覧者の顔向きによる興味推定

プラズマディスプレイや，プロジェクタなどを設置し，広告などの情報の提示を行うデジタルサイネージは，その利便性から広告提示媒体として注目を集めている．南竹らは，広告が映されている画面を歩行者が通過した際の顔向きを解析し，広告への注目状況を取得するツールキット SignageGazer と SignageTracker の実装を行った [7]．

これに対して，本研究では携帯端末であるモバイル PC を用いることによって場所を選ばずに，また，テレビで放映されている CM のようなデジタルサイネージ以外の分野でも興味推定を行える．

2.2.2 閲覧者の視線による興味推定

映像コンテンツを閲覧する際に，映像に集中していないときは外部に注意をそらす事がある．Conneor らは，閲覧者の視線によって映像に集中しているかどうかの状況を，映像を閲覧する端末に LED カメラを設置して分析している [8]．

この研究では視線のみに着目し，映像閲覧者の集中度合いを分析しているが，本研究では，マルチモーダルデータを用いて，視線のトラッキングのような高度な処理を用いずに映像閲覧者の興味を推定することができる．

3. タブレット PC を使用した映像閲覧者のマルチモーダルデータの収集

3.1 システム概要

今回作成したシステムは，映像を閲覧しながらマルチモーダルデータを取得する事ができるデータロガーと，取得したマルチモーダルデータを用いて分析を行うためのマルチモーダルデータ分析ツールの 2 つで構成されている．

データロガーは多様なセンサを持ち，映像閲覧者の表情や発話，持っている端末の揺れの度合いを取得することができるタブレット PC を対象とし，今回は iPad2 を利用した．マルチモーダルデータ分析ツールは様々なデータを同期状態で閲覧し，多人数会話の分析を行う事ができるインタラクション統合分析環境 iCorpusStudio[2] を拡張して利用した．拡張は，閲覧している映像のタイムラインに追従した動画プレーヤーの作成，データロガーで取得した三軸加速度のグラフ化を行った．

3.2 マルチモーダルデータとして利用するセンサ群

本研究で利用するマルチモーダルデータは以下の 3 つのセンサから取得している．

- マイクロフォン：映像閲覧者の発話を取得するために，iPad2 のモノクロマイクロフォンを利用する．

- フロントカメラ：映像閲覧時の状態や，発話状況を分析するために iPad2 のフロントカメラを利用し，閲覧の開始から終了までの様子を記録する．
- 三軸加速度：映像閲覧者の端末の持ち換えや，寝返りなどの閲覧状態の変化，笑った時の端末の揺れを認識するために，iPad2 の三軸加速度を利用する．また，取得する周波数は 100Hz とする．

図 1 記録の様子
Fig. 1 The state of the record.

3.3 閲覧動作記録

データロガーを用いて取得したデータに加えて，閲覧している映像の再生位置や閲覧中に行う一時停止や巻き戻しなどの映像の再生状態を CSV ファイル形式に出力し，マルチモーダルデータ分析ツールを用いて分析する．

また，タブレット型 PC である iPad2 を用いることによって，様々なシチュエーションでの映像閲覧者の閲覧動作を記録することができる．例として，1 人で映像を閲覧 (図 2 左上) する他に 2 人で映像を閲覧 (図 2 右上)，家で寝転がりながら映像を閲覧 (図 2 左下)，食べ物を食べながら映像を閲覧 (図 2 右下) するシチュエーションがある．

図 2 様々なシチュエーションでの映像閲覧
Fig. 2 Video viewing in various situations.

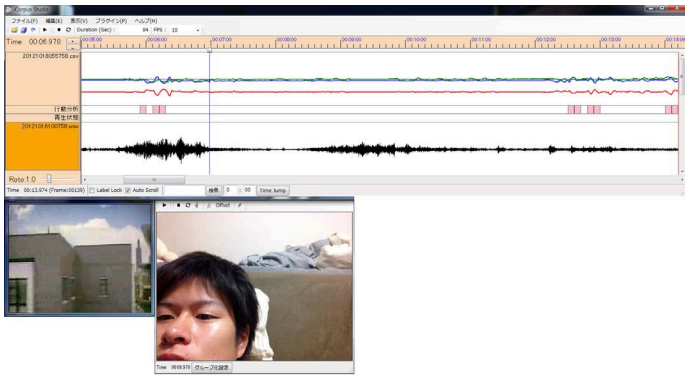

図 3 マルチモーダルデータ分析ツールを用いて分析を行う様子*1
Fig. 3 Analyzed using the Tool to analyze multi-modal data.

3.4 閲覧動作の分析

マルチモーダルデータ分析ツールを用いて，取得したセンサ群を可視化し，時系列を統一させて分析を行う．

図 3 の通り，分析には閲覧している映像コンテンツ，フロントカメラからの閲覧者の状態，マイクロフォンからの発話の音声を変換した波形データ，取得した三軸加速度と映像閲覧時の再生状態などのログを可視化したデータを用いる．

図中上の部分は時系列情報（音声，加速度のグラフ），左下は閲覧中のビデオコンテンツ，右下はフロントカメラの映像の映像を表している．

また，三軸加速度がある閾値を超えて変化した際にラベルを追加し，映像閲覧者の興味変化の推定を行う手がかりとする事ができる．現在は，三軸加速度のみに注目した閾値計算でラベリング処理を行なっているが，そこに音声データなどを組み合わせ，and 演算して計算するなど，様々な計算を用いてラベリング処理する事で，より詳しい事象での興味変化の推定を自動で行うことが期待できる．

4. マルチモーダルデータによる映像閲覧者の興味変化の推定

4.1 興味推定の仮説

本システムによって，映像閲覧者の興味を推定するために以下のような仮説を立てた．

(1) 映像に対して閲覧者が何か反応した際，三軸加速度と音声データに変化が生じる．
(2) 複数人で映像を閲覧する際，その発話や行動によって映像に対して興味を持ったポイントを推定できる．
(3) 映像を閲覧するシチュエーションによって，閲覧者の反応の仕方に変化が生じる．

4.2 実験による検証

仮説の妥当性を検証するために，10 名の被験者に本システムを利用して映像を閲覧してもらった．また，映像閲覧時の状態は，我々が場所と時間を指定するシチュエーショ

ンと，タブレット PC を被験者に貸し出して，被験者が映像を閲覧する場所と時間を自由に選択できるシチュエーションの 2 つのシチュエーションを設定し，実験を行った．

4.2.1 閲覧者の笑い声や笑う時に生じる端末の揺れによる興味変化の推定

1 人で映像を閲覧している際に，面白いシーンに対して笑っている閲覧者がいた．この時の音声と三軸加速度データは笑う前の状態と比べて大きく変化していることがわかる (図 4)．この事から三軸加速度や音声の変化によって興味変化の推定が行え，仮説 (1) が妥当であると言える．

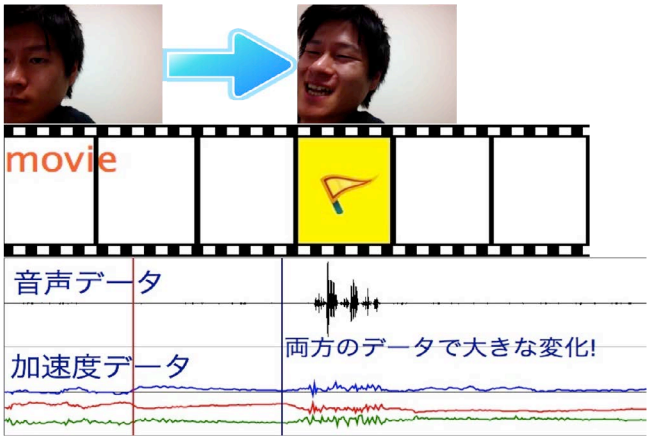

図 4 三軸加速度と音声の変化によって興味が推定できるケース
Fig. 4 To estimated of viewer's interests by changes in three-axis accelerometer and audio.

4.2.2 閲覧者の寝返り動作での興味変化の推定

閲覧者が寝転がりながら映像を見ている際に，映像のシーンの切れ目や緊張感のない場面で，閲覧者が寝返りを打つ動作が見られた．この時三軸加速度は大きく変化していることがわかる．また，寝返りを打つことによって閲覧している体勢が変わることから，その後の三軸加速度が寝返り前と異なる値になることがわかる (図 5)．この事から閲覧者の動作によって映像の切り替わりや閲覧者の注目度が低いポイントなどが推定できる．

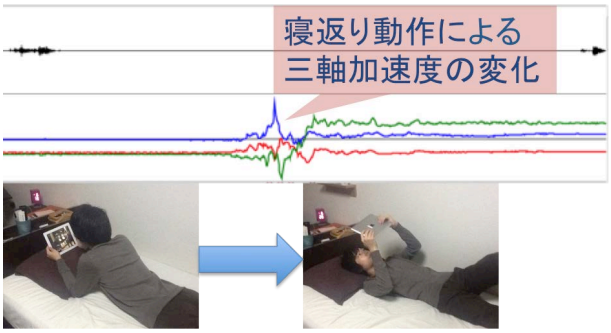

図 5 閲覧者の寝返り動作によって興味変化を推定できるケース
Fig. 5 To estimated of viewer's interests by rolling over.

4.2.3 ２人の発話，再生時間軸の変化による興味推定

タブレット PC を被験者に貸し出し，被験者の家で映像を閲覧してもらった際に，閲覧者が面白いと感じたシーンを共有しようと，たまたま一緒に居た別の被験者を呼び出し，映像を巻き戻して再び再生しているケースがあった (図 6)．この時，発話状態から映像を一時停止したポイントが，閲覧者の興味の変化が生じた終点となり，映像の巻き戻しを行い再び閲覧者が再生を行ったポイントが，その閲覧者の興味の変化が生じた始点であると推定できる．この事から，仮説 (2) が妥当であると言える．

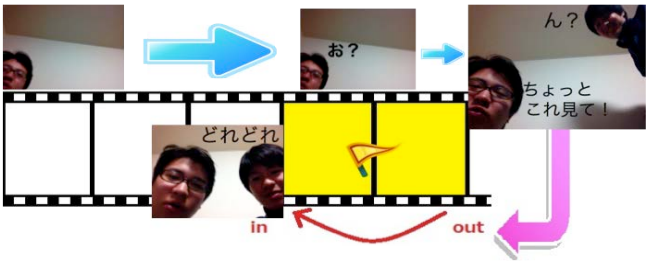

図 6 閲覧者の発話，行動から興味が推定できるケース
Fig. 6 To estimate of viewer's interest by talk and behavior with other viewer.

5. おわりに

現在，本研究はマルチモーダルデータを収集，分析するためのデータロガーとマルチモーダルデータ分析ツールの作成を行い，三軸加速度による自動的な閲覧者の反応判定を手がかりに，手動によって状況を分析して閲覧者の興味の変化を推定できる簡易的なシーンを発見した段階である．今後の予定として，三軸加速度だけでなく各モダリティの基本的な反応判定を自動化し，これらの組み合わせのパターンから自動的に閲覧者の興味推定を行いたい．また，4.2 節で仮説の妥当性を証明する特徴的な事例を紹介したが，実際には三軸加速度や音声データが変化していない場面でも閲覧者が笑顔になるなど，興味の変化が生じるシーンが見られた．このように同じ映像を閲覧していても閲覧者によって反応の度合いが異なる [9] ので，多くのユーザのデータから機械学習的に興味推定の構造を発見したい．

参考文献

[1] 角康之：マルチモーダルデータを用いた会話的インタラクションの構造理解，人工知能学会誌，Vol.27，No.4，pp.405-410，2011.
[2] 角康之，西田豊明，坊農真弓，來嶋宏幸：IMADE：会話の構造理解とコンテンツ化のための実世界インタラクション研究基盤，情報処理学会誌，Vol.49，No.8，pp.945-949，2008.
[3] Takatsugu Hirayama, Yasuyuki Sumi, Tatsuya Kawahara, and Takashi Matsuyama: *Info-concierge: Proactive multi-modal interaction through mind probing*, The 2011 APSIPA Annual Summit and Conference, 2011.

[4] 宮崎泰彦，藤本憲司：個人状況適応型キオスク・システムにおけるエージェントモデルの提案，情報処理学会研究報告（マルチメディア通信と分散処理研究会報告），Vol.98，No.84，pp.25-30，1998.
[5] 河原達也：スマートポスターボード：ポスター会話のマルチモーダルなセンシングと認識，電子情報通信学会誌，Vol.112，No.141，pp.7-12，2012.
[6] Dan Bohus, and Eric Horvitz: *Multiparty Turn Taking in Situated Dialog:Study, Lessons, and Directions*, Proceedings of the SIGDIAL 2011 Conference, pp.98-109, 2011.
[7] 南竹俊介，高橋伸，田中二郎：歩行者の顔向き情報と移動軌跡を利用したデジタル広告の視聴率測定，情報処理学会全国大会講演論文集，vol.72，No.3，pp.323-324，2010.
[8] Conneor Dickie, Roel Vertegaal, Changuk Sohn, and Daniel Cheng: *eyeLook: Using Attention to Facilitate Mobile Media Consumption*, UIST '05 Proceedings of the 18th annual ACM symposium on User interface software and technology, pp.103-106, 2005.
[9] 中田篤志，角康之，西田豊明：非言語行動の出現パターンによる会話構造抽出，電子情報通信学会論文誌，Vol.J94-D，No.1，pp.113-123，2011.