Vol.2012-MBL-61 No.22
Vol.2012-UBI-33 No.22
2012/3/13

タブレット PC を利用した映像閲覧者のマルチモーダルな興味度推定

駒木定生† 角康之†

スマートフォンやタブレット PC の普及に伴い，映像の閲覧形態が大きく変わりつつある．つまり，いつでもどこでも映像を見られるようになったため，短い時間に断片的に見たり，何かしながら気が向いたものにだけ目を向けたりする傾向が高くなってきた．また，受動的に与えられる映像を見るばかりでなく，検索あるいは探索的に好みの映像を選んで見たり，それに対して積極的にコメントを残して他の人に推薦したりする社会的な側面が強くなりつつある．さらに興味深いことは，映像デバイスと閲覧者との身体的関係の大きな変化である．つまり，お茶の間でくつろぎながらテレビモニターを眺めるのと，手で把持して歩いたり寝転んだりしながら映像を見るのとでは，身体的インタラクションが大きく異なる．スマートフォンやタブレット PC は，カメラ，マイク，加速度センサなど，多くのセンサを持ち，利用者の行動や反応をつぶさに観測することができる．
そこで本研究では，タブレット PC による映像閲覧者の反応を上記のセンサ類で計測し，映像に対する集中度や興味の変化を自動的に測量することを目的とする．本論文では，閲覧者の顔の向きとその変化，思わずあげる発声，タブレット PC の把持動作の変化などに注目し，閲覧者の関心度を計る可能性について検討する．

Estimating video viewer's interests by tablet PC

Sadao Komaki† and Yasuyuki Sumi†

According to the spread of smart phones and tablet PCs, our behavior of video viewing is drastically changing. We can view video contents anywhere and any short time, and then, we tend to give an attention to them only if they catch our attention. Also, view viewing is becoming social behaviors, i.e., feeding back comments, recommending to friends. Another noteworthy change is bodily interactions between viewing devices and viewers. In the case of old-fashioned style, we view the videos on TV monitors located away. But, in the case of smart phones and tablet PCs, we view the videos on them grasped by hands during walking and doing any other things. In this paper, we measure viewers' reaction during viewing videos such as face moving, voice, and grasping status with camera, microphone, and acceleration sensors built in tablet PCs. We developed a logger of the sensed data and explored the possibility of automatically estimating change of viewers' concentration and interests along with provided video.

1. はじめに

近年，YouTube，ニコニコ動画に代表されるような，映像クリエーターが制作・公開した映像コンテンツを手軽に視聴できるユーザ投稿型映像閲覧サービスが普及しつつある．映像コンテンツの閲覧者（以下，閲覧者と呼ぶ）は，映像コンテンツから受けた感銘をコメントの記述やお気に入りへの登録といった明示的な行為を通して明文化を行っている．映像クリエーターは自身の作品に残されるこれらの明文化されたリアクションを拾い読みし，今後の作品づくりへのモチベーション維持に繋げていると考えられる．

そこには，映像クリエーターが映像を投稿し，閲覧者は映像作品から受けた感銘を明文化して映像クリエーターに伝達する，映像クリエーターはそのリアクションを糧に，より良質な作品作りを志す，といった循環が起きていると考えられる．

しかし，閲覧者の総数に対しコメントの記述などを行う閲覧者は少数である．その原因として，閲覧者は映像閲覧という本来の目的である行為に重ねて意見・感想を明文化するという行為に負担を感じている可能性が考えられる．

近年，ユーザの発話やしぐさを計測し，文脈と照らし合わせることで人の行為や会話の意味を認識・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的に用いることでユーザの発話やしぐさの計測を行うマルチモーダルセンシングが注目されている．

筆者らは，映像閲覧者のふるまいをマルチモーダルにセンシングした結果から，自動的に閲覧者の興味の移ろいを推定し，映像クリエーターに閲覧者のリアクションを伝達することが出来ないかと考えた．その為にはまず，閲覧者が閲覧に利用しているデバイスのセンサを興味推定に利用できるか否かの検証が必要となる．本研究では検証に用いるデバイスとして，多様なセンサを持ち，映像閲覧のためのデバイスとしても利用されているタブレット PC に着目した．

本論文では，タブレット PC の持つ三軸加速度センサ，フロントカメラ，マイクロ

† 公立はこだて未来大学
Future University Hakodate

ⓒ2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.22
Vol.2012-UBI-33 No.22
2012/3/13

フォンを利用して，映像閲覧者のマルチモーダルデータの収集を試みる． また，収集したマルチモーダルなデータを用いて閲覧者の興味を推定する可能性について検討する．

2. 関連研究

本章ではマルチモーダルセンシング，映像閲覧者の興味推定の関連研究について述べる．

(1) マルチモーダルセンシング

我々人間は会話において，視線，ジェスチャ，うなずき，あいづちといった非言語行動によって様々な意図を表現する．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，人間同士のコミュニケーションの様子を記録した音声・映像，行動データを用いて人と人，人と物のインタラクション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収集されたデータは発言内容を構造化したデータと，発言場面の映像や音声のデータから構成されており，それらのデータをリンクさせることで，ログの閲覧者は会話の場における発言内容と共にその場における発言者の様子や表情などを振り返ることができる．

インタラクション分析の典型的手法としては，カメラやマイクを用いて人と人，人と物のインタラクションをマルチメディアログとして記録し，手作業でアノテーションを施すことによって分析を行うものである．また，昨今ではマルチモーダルなデータから特徴を抽出し，半自動でインタラクションの分析を行う試みも行われている．

角らは、IMADE (Interaction Measurement, Analysis, and Design Environment) ルームと呼ばれる環境を構築した．この環境は，人同士のインタラクションに関する様々な種類のマルチモーダルデータ，具体的には，映像，音声，移動，視線，生体反応といったデータを統合的に計測するために設計された[2]．

IMADE を用いた研究では，会話の微視的な分析（例えば，発話交替や視線の時間構造分析など）だけではなく，巨視的な分析（つまり，会話グループの生成・分解や移動などのダイナミクスの分析）も目的としている．

IMADE ルームにはビデオカメラ，マイク，モーショントラッキングシステムが配置されている．また，ユーザは視線計測の為のアイトラッキングモジュールや，モーショントラッキングの為の反射材を身につけなくてはならない．

本研究では映像閲覧行為に日常的に用いられるタブレット PC を利用して人と物（映像コンテンツ）のインタラクションの分析を行うことを目的としている．このことから，特別な設備や装備を必要とする角らの研究と，普段利用しているデバイスをセンシングに用いる本研究とに差異があると言える．

(2) 映像閲覧者の興味推定

映像コンテンツに対する閲覧者の反応を追うことで閲覧者の興味の移ろいを測る試みは，映像閲覧サービスや映像を利用した広告などに応用されている．

YouTube では, YouTube Analytics というサービスを通じて投稿した動画の視聴率の推移を調べることが出来る[3]． YouTube で動画を閲覧するユーザは，動画を見続ける必要がなく，自身の興味の移ろいに従って閲覧行為を能動的にコントロールすることができる．閲覧者は興味を惹かれ続ければ視聴を持続し，興味を失えばその動画の視聴を止めることができる．YouTube Analytics では，閲覧者が再生を止めたタイムコードを記録し，視聴率を算出している．このサービスは閲覧者の興味を推定する方法の１つとして挙げることが出来る．

人は興味を惹かれるものに対して無意識に視線を向けたり，体の向きを変えたりする．映像閲覧行為中に閲覧者が興味を映像に対して向けているならば，目線や顔の向きはディスプレイに対して向けられていると考えられる．この考えを広告配信システムにおける，映像広告に対する興味推定に応用する研究例として南竹らの研究が挙げられる[4]．南竹らは閲覧者の顔向きが閲覧者の興味・関心と一致していることを明らかにした．また，映像に対しての関心の有無を推定する手段として閲覧者の顔向きを計測することの優位性を明らかにした．

映像閲覧者の集中度を自動的に計測する研究は，主に視線計測や映像の視覚的特徴との関係で議論されてきた（例えば，[5]）．

本研究は，固定カメラによる詳細な視線計測やコンテンツ映像の事前の分析に頼らず，タブレット PC に内蔵されたセンサで閲覧者の反応を近似的に理解することを目指す．また，タブレット PC での映像閲覧は，閲覧者の身体的自由度が高く，かつ，閲覧者とデバイスの身体的接触性が高いことに注目し，閲覧者が思わずあげる声や体の動き（あるいは静止）に着目することが，本研究の特徴である．

3. 収集するマルチモーダルデータ

本章では本研究で収集するマルチモーダルデータと，収集の目的について述べる．

3.1 閲覧者の顔向きを捉えた映像

２章（２）で述べたように，南竹らによって映像に対しての関心の有無を推定する手段として閲覧者の顔向きを計測することの優位性が明らかになっている．そこで顔向きによる興味推定の可能性を検証するため，閲覧者の顔向きを捉えた映像を記録す

ⓒ2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.22
Vol.2012-UBI-33 No.22
2012/3/13

ることとした．

3.2 閲覧者の発話を記録した音声

映像閲覧を行なう際，閲覧者が発話する場合がある．コンテンツの内容に対して驚きや発見があった際に感嘆の声を挙げるなどがそれに該当する．また，『これ見て！見て！』と他者に声をかけて誘い，一緒に映像を閲覧するといった行為も日常的に目にすることができる．このように閲覧者の発話のタイミングやその内容が映像に対する興味推定の基準となり得ると考えた．そこで，閲覧者の発話を録音することとした．

3.3 閲覧者の行為に伴う筐体の傾き

テレビのような据え置き型のデバイスで映像視聴を行う際，ユーザはデバイスを運びながら視聴することがない．しかし，タブレット PC は持ち運ぶことを前提として設計されており，その利用方法においても「ながら操作」が出来るようになっている．手にとって映像を閲覧している状況と，テーブルに置きながら視聴している状況，移動しながら視聴している状況などの状況の違いによって閲覧者の興味がどのように変化するのかを計測することは有意義であると考えた．そこで，閲覧者の閲覧中の行為に伴う筐体の傾きを記録することとした．

4. マルチモーダルデータを記録するためのロガー

本章では 3 章で述べたマルチモーダルデータを記録するために製作したタブレット PC 上で稼働するロガーについて説明する．

4.1 ロガーの概要

制作したロガー（以下，ロガー）と閲覧者の関係を示したシステム図を（図．1）に示す．ロガーは映像コンテンツを再生しながらタブレット PC のフロントカメラ，モノラルマイクロフォン，三軸加速度センサを通じて閲覧者のふるまいをセンシングする．それぞれのセンサは閲覧者の顔向きを捉えた映像，閲覧者の発話，閲覧者の閲覧中の行為に伴う筐体の傾きを取得する．また，再生されている映像のタイムコードを記録する．

ロガーは映像コンテンツの再生装置としても機能する．閲覧者はロガー上で再生・一時停止，シーケンスバーを利用した巻き戻し・早送り，音量調節という基本的な操作を行うことができる．

ロガーの開発プラットフォームは XCode 4.2 を利用した．アプリケーションを実行する環境は，閲覧者の顔向きを捉えた映像を録画する必要性からフロントカメラを持つ Apple 社の iPad2 を選択した．

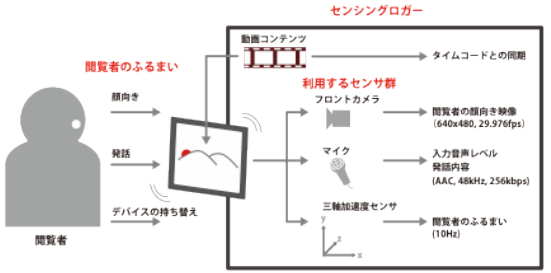

図 1 ロガーと閲覧者の関係を示すシステム図

4.2 ロガーで利用するセンサ

映像閲覧時のマルチモーダルデータを収集するためのセンサはすべて iPad2 に標準装備されているものを利用した．

閲覧者の顔向きを捉えた映像を録画するためのカメラにはフロントカメラを利用した．ロガーはこのカメラで捉えた映像を解像度 640*480，29.976fps，3.5Mbit/s のデータレートで記録する．なお，映像を視聴する際に一般的に用いられる，フロントカメラを左側にして筐体の長辺を横向きにする Landscape Left Style にした際，映像視野角は垂直方向に 45 度，水平方向に 60 度となる．このカメラで録画することで単独で閲覧する場合のみならず，複数人で閲覧した場合の姿勢も記録できることが期待できる．

映像閲覧者の発話を録音するためのマイクロフォンにはモノラルマイクロフォンを利用した．このマイクロフォンには映像閲覧者の発話以外に，スピーカーから発せられた映像コンテンツの音声も入力される．マイクロフォンで取得した音声は Advanced Audio Coding によって 48kHz,16bit,256kbps のデータレートで記録される．

閲覧者の閲覧中の行為に伴う筐体の傾きを取得するセンサには三軸加速度センサを利用した．加速度の変化と映像コンテンツのタイムコードは 10Hz 間隔で同期的に記録され，CSV 形式でロガーのデータ領域に保存される（図．2）．

ⓒ2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.22
Vol.2012-UBI-33 No.22
2012/3/13

[time]	[X]	[Y]	[Z]
0:00:00	0.013458	-0.42009	-0.891144
0:00:00	0.004425	-0.409302	-0.915344
0:00:00	0.003571	-0.426682	-0.870026
0:00:00	-0.002151	-0.494934	-0.845978

図 2 コンテンツのタイムコードと同期した三軸加速度のログ

5. ロガーで記録できた映像閲覧者のマルチモーダルデータ

本章では，製作したロガーを用いた予備検討のためのデータ収集，及びその結果得られたデータについて述べる．

5.1 データ収集の概要

ロガーを通じて閲覧者のマルチモーダルデータを記録することができるかを検証するためデータ収集を行った．データ収集の対象者は，公立はこだて未来大学の学部生3名，教員1名とその家族2名とした．データ収集で用いた iPad2 には後述する 2 種類の映像コンテンツを転送し，ロガーを起動後に自動再生されるように設定した．普段の生活と同様の環境で映像を閲覧できるよう，対象者に対しては閲覧場所や時間などは指定しなかった．対象者には筐体を手で支えて閲覧してもらった（図．3)．

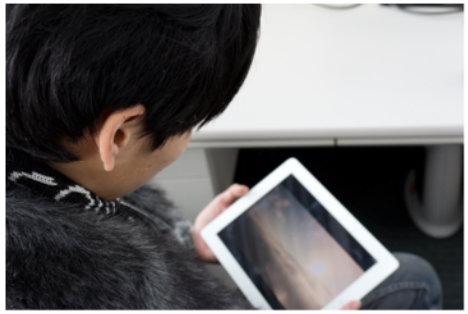

図 3 データ収集の様子

5.2 データ収集に用いた映像

映像閲覧に用いた動画データは映像 A：海外のコマーシャル 30 作品を一繋ぎにした約 40 分の映像，映像 B：筆者が作成した函館市の風景を被写体にした 4 分弱の微速度撮影映像作品の二種類である．

コマーシャル映像は閲覧者に商品やサービスの印象を強く残すという役割があるため，見せ方が面白おかしかったり過激だったりするものが多い．これは閲覧者のリアクションを収集する上で有効なコンテンツだと考えた．

微速度撮影映像は圧縮された時間の流れを体感できる映像コンテンツである．題材を函館市の風景とすることで，はこだて未来大学の学生．教員にとって見慣れた風景を普段と異なった見方で観ることができるため，新鮮さを感じることができる．その結果，対象者のリアクションを引き出すことが出来ると考えた．

5.3 データ収集で得られたフロントカメラの映像

データ収集によって得られた映像を目視で観察した結果，大きく分けて１．筐体を持って視聴をしているシーン，２．「ながら」視聴しているシーン，３．閲覧者が増減するシーン，４．閲覧者が発話しているシーンの４つのシーンに分けることが出来た(図．4)．

図 4 録画映像から大別できた 4 つのシーン

5.3.1 筐体を持って視聴しているシーン

このシーンでは閲覧者が筐体を持ち続けながらディスプレイを継続的にのぞき込んでいる様子が録画されていた．

ⓒ2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.22
Vol.2012-UBI-33 No.22
2012/3/13

5.3.2 「ながら」視聴しているシーン

このシーンでは閲覧者が筐体を持って視聴しつつ，並行して他の行動をしている様子が録画されていた．映像からは食事，歩行，ゴミ捨てなどが並行して行われていた行動として確認できた．

5.3.3 閲覧者が増減するシーン

このシーンでは閲覧者が画面から居なくなったり，閲覧者が増えたりする様子が録画されている．閲覧者が一人，二人で閲覧するシーンでは閲覧者の顔の全体像が確認できた．しかし，３人以上で閲覧しているシーンでは顔が見きれて確認できない閲覧者がいた．

5.3.4 閲覧者が発話しているシーン

このシーンでは閲覧者が口を開けて何かしらの発話をしている様子が録画されていた．複数人で閲覧していた場合は，一時的に発話者が目線をディスプレイから他の閲覧者の方へ向けることもあった．

5.4 データ収集で得られた音声データ

データ収集によって得られた音声（図．5）を再生して内容を確認すると，コンテンツが再生されているだけのシーン，閲覧者の生活音が録音されているシーン，閲覧者の発話が収録されているシーンに分けることが出来た．（図．6）

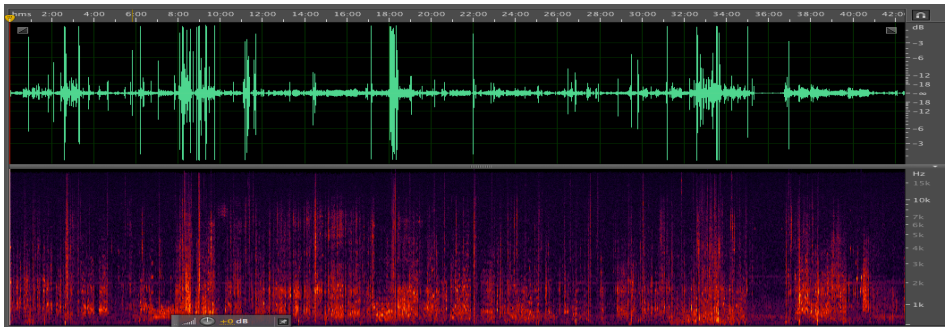

図 5 データ収集によって得られた音声データ

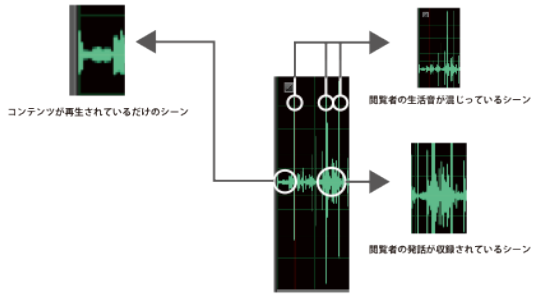

図 6 音声から判断した 3 つのシーン

5.4.1 コンテンツが再生されているだけのシーン

このシーンでは閲覧者が物音を立てたり発話したりせず，再生しているコンテンツから発せられた音声しか収録されていない．

5.4.2 閲覧者の生活音が録音されているシーン

このシーンでは閲覧者がドアを開閉した音や椅子に腰掛けた時に椅子が軋む音などの生活音が混ざっている．この時に発せられる音は極めて大きく，可視化された波形データでもその特徴が極めて大きく現れている．

5.4.3 閲覧者の発話が収録されているシーン

このシーンでは閲覧者の発話が収録されている．確認できた発話には，コンテンツに対して思わず発せられた驚きの声や笑い声，「あ，この会社の CM か」といった映像コンテンツの内容に関する発話，閲覧者が複数人いる状況下での「あの CM もう一回見ようよ」といった会話，閲覧中にかかってきた電話への応答といったコンテンツに関係のない発話などがあった．

5.5 閲覧者の閲覧中の行為に伴う筐体の傾きデータ

データ収集から得られた加速度の変化をグラフにすることで可視化を試みた．また，グラフを映像編集ソフト上で閲覧者の映像・音声と照らし合わせ，閲覧者のどのような行為と筐体の傾きの変化が同期しているのか検証を行った（図.12）．

その結果，筐体を置いている時に現れる変化，筐体を持って移動している際に現れる変化，複数人で閲覧している際に現れる変化を推測できた．また，閲覧者がコンテンツの内容で驚きを示したときに現れた変化を推測できた．

ⓒ2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.22
Vol.2012-UBI-33 No.22
2012/3/13

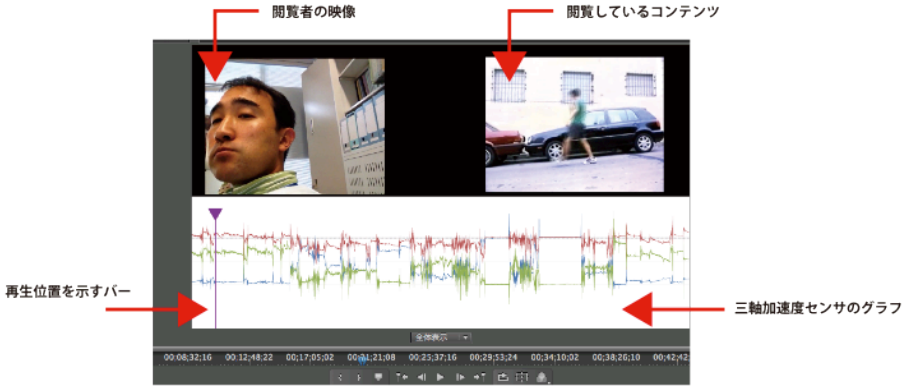

図 7 映像編集ソフトを利用した三軸加速度とシーンの照らし合わせ

5.5.1 筐体を置いている時に現れる変化

筐体をテーブルなどに置いて放置している時に現れた加速度の変化を（図.10）に示す．この時，加速度の変化は著しく低下しほとんど変化がみられなくなる．

図 8 筐体を置いている時に現れる加速度の変化

5.5.2 筐体を持って移動している際に現れる変化

筐体を持って移動している時に現れた加速度の変化を（図．11）に示す．この時，加速度の変化の速度が早まり，X 軸，Y 軸，Z 軸のすべての軸が一定のリズムで変動する．

図 9 筐体を持って移動している際に現れた加速度の変化

5.5.3 複数人で閲覧している際に現れる変化

筐体を持って複数人で閲覧している時に現れた加速度の変化を（図．16）に示す．この時，加速度の変化は長時間に渡って穏やかに推移している．この時，映像からは一人が筐体を支え，他の閲覧者が覗き込めるように保持するような姿勢をとっていることが確認できた．

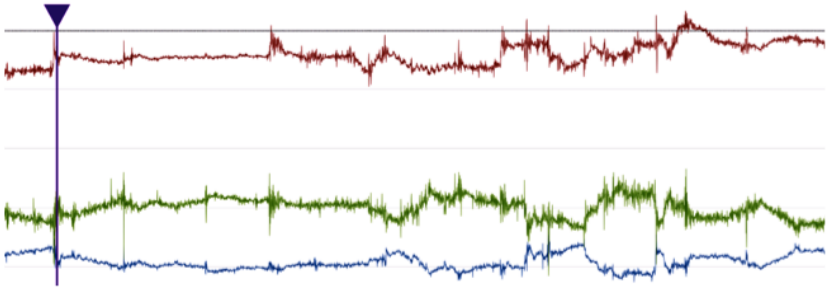

図 10 複数人で閲覧している際に現れた加速度の変化

6. 考察

本章ではデータ収集によって得られたフロントカメラの映像，音声，閲覧中の行為に伴う筐体の傾きデータに対する考察と，タブレット PC で収集したマルチモーダルデータを用いて閲覧者の興味を推定する可能性について考察する．

6.1 フロントカメラの映像についての考察

フロントカメラで撮影した映像データから，映像閲覧中の閲覧者の顔向きを目視で

ⓒ2012 Information Processing Society of Japan

Vol.2012-MBL-61 No.22
Vol.2012-UBI-33 No.22
2012/3/13

判断出来ることが分かった．また，録画された映像から，筐体を持って視聴をしているシーン，「ながら」視聴しているシーン，閲覧者が増減するシーン，閲覧者が発話しているシーンといった閲覧者の状況を推察できることが分かった．

しかし，閲覧者の顔向きがディスプレイに対して垂直より右向きにずれていることが確認できた．これはフロントカメラの配置がディスプレイの左寄りであるためと考えられる．また，３名以上の閲覧者が同時に映像閲覧を行った場合，計測可能な人物と計測不可能になる人物がいることが確認できた．これはフロントカメラの画角が影響していると考えられる．

6.2 音声データについての考察

モノラルマイクで録音した音声データから，コンテンツが再生されているだけのシーン，閲覧者の生活音が混ざっているシーン，閲覧者の発話が収録されているシーンに分けることが出来た.また,閲覧者の発話が収録されているシーンの音声をもとに，閲覧者の関心の対象を測ることが出来た．

6.3 閲覧中の行為に伴う筐体の傾きデータについての考察

三軸加速度センサの変化量を示すグラフだけでは閲覧者の状況を推察することが出来なかった．これはタブレット PC を手に持ちながらの閲覧行為中は絶えず筐体が傾いていることで，加速度の変化の特徴が掴みづらいことが原因として考えられる．

しかし，筐体の傾きの変化のデータと映像閲覧中の閲覧者映像・音声を同期して確認することで，筐体を置いている時に現れる加速度の変化，移動中に現れる加速度の変化，複数人で閲覧する時に現れる加速度の変化を推測することができた．

6.4 タブレット PC で収集したマルチモーダルデータを用いて閲覧者の興味を推定する可能性についての考察

6.1 に述べたように，フロントカメラを用いて映像閲覧中の閲覧者の様子を録画することで，閲覧者の顔向きや状況を推察できることが分かった．２章（２）で述べたように，閲覧者の顔向きを撮影することで興味推定できる可能性は高まる．

6.2 で述べたように，タブレット PC のマイクを用いて録音して音声から，映像閲覧中の発話の内容を知ることができた．また，その内容から閲覧者の興味の内容を確認することができた．

6.3 で述べたように，三軸加速度センサを用いて映像閲覧者の行為に伴う筐体の傾きを計測することができた．また，閲覧者が驚いた際に現れた特徴的な加速度の跳ねを観測することが出来た．

対象者の一人は，一人で映像 A を閲覧した際に思わず笑ってしまったシーンを，他者にそのシーンを見せるために再度再生するといった行為を行なっていたことが分かった．このことから，対象者がその時再生していたシーンに何かしらの興味を示していたと推測できた.これは,その時間軸に再生されていたコンテンツを閲覧者の映像，音声，再生時間から多角的に分析することで初めて明らかになった．

以上のことからタブレット PC で収集したマルチモーダルデータを用いることによって，閲覧者の興味を推定でき得ると考察する．

7. おわりに

本論文では，タブレット PC の持つ三軸加速度センサ，フロントカメラ，マイクロフォンを利用して，映像閲覧者のマルチモーダルデータの収集を行った．また，タブレット PC で収集したマルチモーダルデータを用いることによって，閲覧者の興味を推定でき得るという結論を導いた．

今後はより多くの対象者からデータを収集することで，各データの特徴をより明確にしたいと考えている．また，録画された閲覧者の映像を視線計測ソフトで解析し，データの正確性を検証するといった試みを行う．

参考文献

1) 角康之:マルチモーダルデータに基づいた多人数インタラクションの構造理解，情報処理学会研究報告. HCI，ヒューマンコンピュータインタラクション研究会報告 2011-HCI-145(7), 1, 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) Google："YouTube Analytics". https://www.youtube.com/analytics
4) 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効果測定,表示支援ツール，情報処理学会インタラクション 2010， 2010.
5) Ryo Yonetani, Hiroaki Kawashima, Takatsugu Hirayama, and Takashi Matsuyama, "Gaze probing: Event-based estimation of objects being focused on", The 20th International Conference on Pattern Recognition (ICPR 2010), Aug. 2010

ⓒ2012 Information Processing Society of Japan