一人称視点映像からの指さし抽出による重要シーン推定

久米田 羽月[1]　角 康之[1]　小池 英樹[2]

概要：本研究の目的は，一人称視点映像からユーザの何気ない行動を手がかりにすることで，実世界の重要なシーンを発見・可視化することである．実世界の重要シーンを切り出すことで，ユーザ自らが興味を持った部分を効率的に思い出すことに活用できる．本研究では魚眼レンズを用いた一人称視点映像を利用することで，カメラ 1 台で記録が完結し，非言語行動を手がかりにして，ユーザの反応に基づいたシーンの推定を試みている．本稿では目的の実現のため，非言語行動の 1 つである指さし行為に着目し，映像から指さし行為を発見する方法について検討した．次に映像から指さし行動を抽出することで，重要なシーンをどの程度発見できるのかを確認した．その結果，頭部方向と指さし方向から指さし行為が行われたシーンをある程度抽出することができた．また予備実験の結果，指さし行為からいくつかの重要シーンを推定できる可能性が示唆された．

Important Scene Estimation by Finger Pointing Extraction from First-person Video

UZUKI KUMETA[1]　YASUYUKI SUMI[1]　HIDEKI KOIKE[2]

1. はじめに

ライフログを分析することによって，その持ち主がいつどのような行動をとったのかを知ることができる．具体的な例として，食事や睡眠などのスケジュールを記録しておくことで生活習慣を正す [1]，撮影した写真の記録から，その時間にはどこに居たのか思い出すことができる [2] などがある．また，角ら [3] は展示会ツアーにおいて，来場者の位置情報や興味に基づいて案内を行うモバイルアシスタントを構築した．このように，ライフログの利用者である持ち主自身がその生活の実態を振り返ることや，ソフトウェアシステムが生活を手助けするために役立てることができる．

近年ではウェアラブルカメラのような手軽に撮影できる機材が登場し，一人称視点の映像を撮影する機会は増えている．これは長時間撮影できるためライフログとして映像を残すことができるが，1 日分の映像を後から見返したり分析したりする場合，全てを見返すためには 1 日かかるため，振り返りのコストが高くなる．もし映像に含まれる特徴から利用者にとって重要なシーンを自動的に推定し，ハイライトすることができれば，利用者は効率的に振り返りが行えると考える．

映像やフォトストリームからユーザにとって重要なシーンを推定し，振り返りを容易にするという目的を持った研究はいくつも存在する [4][5][6][7]．特に一人称視点映像の振り返りを容易にすることを目的にした研究では，画像処理を用いることで重要なシーンを発見するアプローチを取っているものが多い [6][7][8]．このように，写っているもの自体に注目して重要なシーンを探し出す研究がある一方で，ユーザ自身の何気ない振る舞いに注目して重要なシーンを探すということも考えられる．

角ら [9] によると，複数人で会話をしている場合，指さしは会話の中で参照している対象物を示す行為であり，会話の内容の理解を測るのに役立つとしている．会話の中に現実世界の対象物があらわれたとすると，そのシーンはユーザにとって興味があるか，あるいは重要なシーンであると考えられる．

本研究の最終目的はユーザの非言語行動を手がかりにすることで実世界の重要なシーンを発見・可視化し，振り返

[1] 公立はこだて未来大学
Future University Hakodate
[2] 東京工業大学
Tokyo Institute of Technology

りを容易にする手法を作成することである．そのため本稿では図 1 のように非言語行動の 1 つである指さし行為に着目し，映像から指さし行為を発見する方法について検討する．次に映像から指さし行動を抽出することで，重要なシーンをどの程度発見できるのかを確認する．

本稿では以下，2 章では関連研究について述べる．3 章では前提となる指さし行為を抽出する方法論について述べ，4 章ではその動作確認と結果，考察について述べる．5 章では指さし行為の発見による重要シーンの推定について，予備実験とその結果，考察について述べる．6 章では，本稿のまとめと今後の展望について述べる．

2. 関連研究

一人称視点の映像からユーザが興味のある出来事を見つけることを目的とする関連研究に，Higuchi らの研究 [6] がある．Higuchi らは撮影された一人称視点の映像から，移動，手の動き，他の人物を手掛かりとして，そのシーンを強調して表示できるインタフェースを備えた EgoScanning を提案した．EgoScanning では，重要と判断されたシーンはゆっくりと再生されるようになっており，残りの部分は早送りされる．同時に，重要な部分を引き伸ばし，そうでない部分は縮めて表示する伸縮タイムライン (Elastic Timeline) を提案しており，重要なシーンは赤く強調される．また，ユーザがどの手がかりに注目しているかを入力できるようになっており，探したいシーンによって使い分けることができるようになっている．結果として，提案されたシステムはユーザの興味の対象を有効に見つけられることができ，より細かい手がかりを読み取ることで難しいシナリオにも対応できると結論付けられている．一人称視点の映像を用い，移動や手の動きに着目する点については本研究との共通点である．

一方，Kayukawa ら [7] は，手や人が写っていることの判別だけでは，文脈によってあまり効果的でないことを指摘した．そこで，映像中の物体に注目し，物体検出システムを用いて 80 の物体カテゴリを検出した．これによってユーザは，任意の物体が写り込んだシーンを重要視してシーンを検索することができるようになっている．

また，Higuchi らや Kayukawa らの研究と似た目的を持つ Toyama ら [10] の研究がある．Toyama らは音環境の比較によって，会話の参加者やの位置を分析することができる，コンテキスト・アウェアなアプリケーションを実現することを目的とした．音環境の類似性に注目することによって，会話の参加者やの位置を分析することができるとした．

ここまでは，一人称視点映像を要約する研究について触れたが，映像に写ったもののみを参考にすると，文脈によってはあまり効果的でないという課題があった．本研究では，ユーザ自身の振る舞いに着目した興味領域の推定を行うため，姿勢データを活用することが必要である．

角ら [9] の研究のように，インタラクションに注目した研究にはモーションキャプチャが有効に使われてきたが，IMADE ルームのように大掛かりな設備を必要とする場合もあった．固定カメラを使う手法では大掛かりな設備が不要になるが，事前にカメラを設置しなければならず，撮影できる場所が限られる．頭部につけたカメラを利用する方法では，一人称視点のような映像を用いて姿勢を推定することができるため，撮影できる場所に制限はないが，推定できるのは上半身だけであるなどの制限が存在する．

Hwang ら [11] は，ユーザの胸部に取り付けられた超広角魚眼レンズで撮影した映像を分析し，3 次元での姿勢推定を行うシステムである MonoEye (以下，MonoEye) を提案した．MonoEye は，魚眼レンズを使って撮影された一人称視点映像を利用し，カメラを身に着けたユーザ自身の全身の姿勢を推定することができる．MonoEye はユーザの各関節の位置，頭部の方向，カメラの向きを映像から推定することができ，ポータブルなモーションキャプチャを実現している．図 2 は，MonoEye が行う処理の概略を表したものである．

本研究ではこれまでに述べた問題を解決するために，Hwang らの研究 [11] を前提として，ライフログ映像の収集，および分析を行う．広角一人称視点映像を用いることによって，従来の映像では写り込まなかった会話相手などの情報や，装置を身に着けているユーザの姿勢データを利用できる．これによって，よりユーザの意思に近いシーンの判別ができるため，従来研究よりも，効果的にユーザの興味の対象を推定できると考える．

3. 指さし抽出の方法論

本稿では指さし行為に注目するため，映像からユーザの指さし行為を見つける必要がある．

角らの研究 [9] では，指さし対象を精度良く推定するため，肘から手首に伸ばしたベクトルと目から掌に伸ばしたベクトルの 2 種類の指さしベクトルを定義し，比較している．結果として，後者の目から掌に伸ばしたベクトルのほうがより高精度であると報告していた．そのため，本研究では指さしベクトルとして頭部から手首の少し上に伸ばしたベクトルを使用する．これは指さしを行う際，視界の上で指が注目対象と完全に被るようには指ささないためである．

本研究ではまず，ユーザの頭部方向と指さしベクトルのなす角度が一定の閾値より低くなった場合に指さし行為が行われていると仮説を立てる．これは，あるシーンにユーザが興味を惹かれるものが写っているとき，ユーザはその方向を向いている可能性が高く，その向きと指さしベクトルが近ければ指さしであるという想定に基づいている．

本稿で提案するシステムでは，MonoEye を使用する．

図 1　指さし行為の抽出による重要シーンの推定

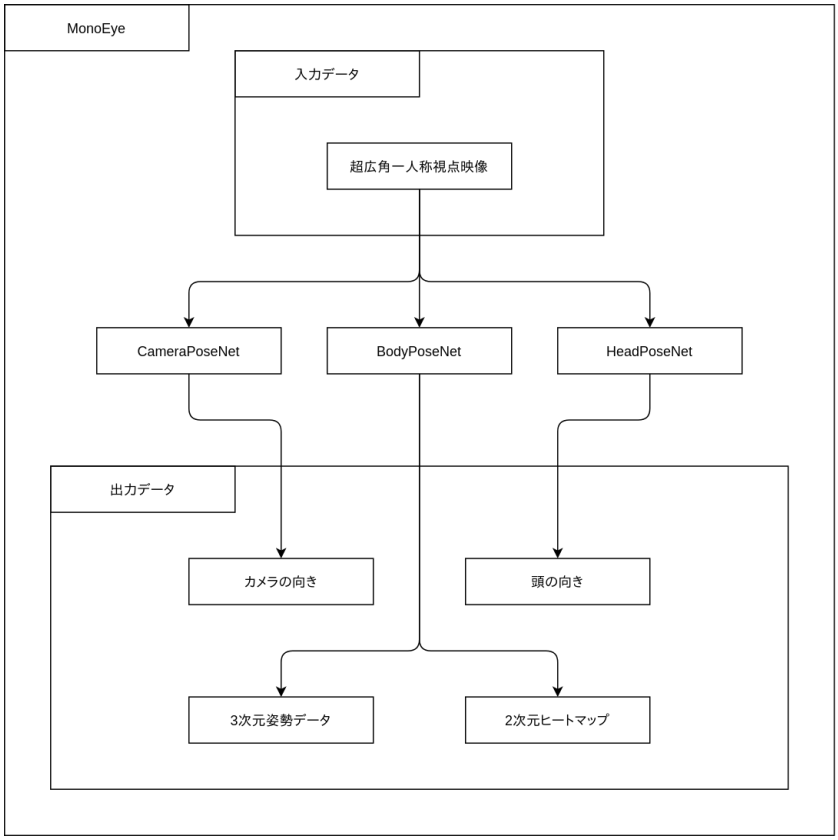

図 2　MonoEye が行う処理の概略

MonoEye では，ユーザの胸部に超広角魚眼レンズを取り付け，その映像を学習済みのネットワークに入力することで，ユーザの姿勢を得ることができる．MonoEye の出力にはユーザの姿勢，ユーザの頭部方向，カメラの向きが含まれており，これらを統合して扱うことで，ユーザがどの時間，どの方向に注目しているかを推定できる．このシステムを用いて動画からフレーム・バイ・フレームでの姿勢推定を行い，ユーザの姿勢からユーザ自身の興味対象，すなわち，ユーザにとって重要なシーンを推定することを目指す．

図 3 は，一人称視点映像とそこから推定された姿勢，頭部方向と指さしベクトルのなす角度を同期させて表示しているものである．図の左下に表示している折れ線グラフは頭部方向と指さしベクトルのなす角度を表しており，上段

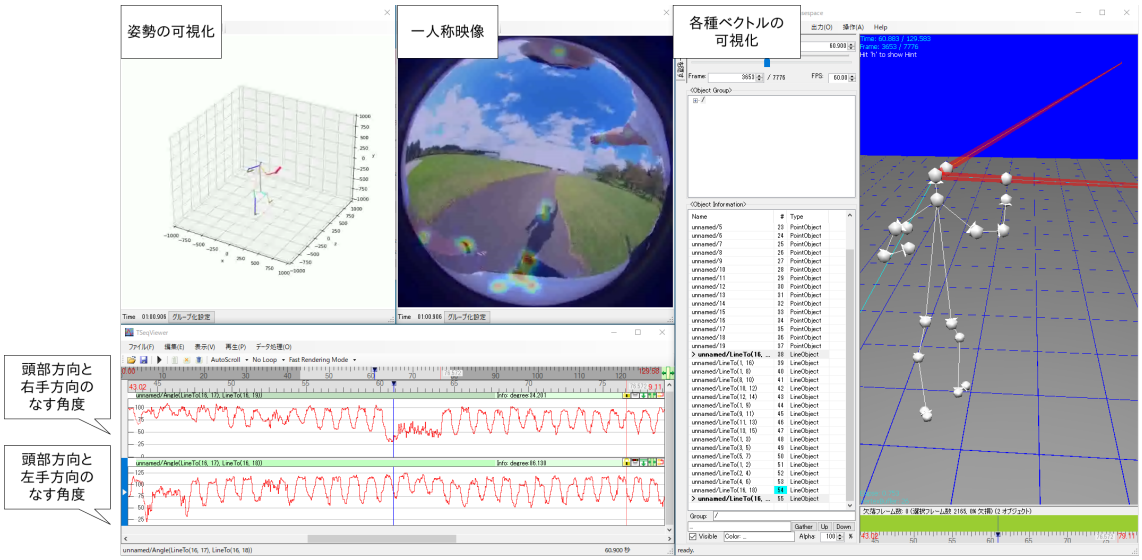

図 3　指さし行為の抽出

が右手，下段が左手についてのグラフである．図 3 では，右手の指さしベクトルと頭部方向のなす角度が他よりも小さくなっているシーンを参照している．図 3 のグラフを参照すると，写っているシーンの前後の時刻では右手の指さしベクトルと頭部方向のなす角度が 45 度から 100 度の範囲であるのに対し，同シーンの直近の時刻では 40 度から 70 度ほどの値が継続している．このように角度の情報に基づいて当該シーンの一人称視点映像を参照すると，実世界においてユーザが指さしに類する行為を取っていることが分かる．よって，この方法によりユーザが指さしを行ったシーンを判定することができると考える．

4.　指さし抽出の動作確認

前章で述べた指さし推定の方法論について検討した．本章では，日常的なシーンを想定した一人称視点映像を用いて，検討したシステムの動作確認を行う．動作確認では指さし行為に注目し，姿勢から指さし行為を判定できるかの検証を行う．対象となる映像は，建物の中を 2 人で歩きながら会話を行っている様子とした．まず，MonoEye SDK[*1] を用いた Python によるプログラムを作成し，映像から姿勢データを取得し，CSV ファイルとして出力した．次に，プログラムから出力されたファイルを iCorpusStudio[12]，および，同梱の 3 次元でのデータの可視化を行うことができる MotionDataUtility，時系列グラフを作成することができる TSeqViewer を用いて視線ベクトル，頭部方向ベクトルを分析した．

分析では，右手，左手の各指さしベクトルについて，頭部方向ベクトルとの角度を時系列のグラフとして出力した．

4.1　結果

頭部方向と指さしベクトルのなす角度が小さくなっているシーンを参照すると，図 4，図 5，図 6 のように，指さしのシーンを見つけることができた．図 4，図 5，図 6 の各グラフにおいて，青い矩形で囲んだ部分が見どころのシーンである．それぞれのグラフは 3 と同様に，上段が頭部方向と右手方向のなす角度，下段が頭部方向と左手方向のなす角度である．

3 つのシーンについて，図 4 は，建物内に設置されたロボット，図 5 は壁に貼り付けられている広告，図 6 は床のダンボールをそれぞれ指さしていることが分かる．ノイズが載っているが，いずれのシーンでも指さしのために腕を動かしており，頭部方向と指さしベクトルのなす角度が徐々に小さくなっている様子が分かる．

一方，図 7 は，ユーザが頭を掻いているシーンである．この時，頭部方向と指さしベクトルのなす角度は，指さしのシーンよりも比較的小さくなっている．

4.2　考察

動作確認では，頭部方向と指さしベクトルのなす角度が小さくなる時間を探すことによって，指さしを行っているシーンを見つけることができた．

図 4，図 5，図 6 を見ると，普通に立っている場合や歩行している場合に，頭部方向と指さしベクトルのなす角度が大きくなることは想定のとおりであった．動作確認では，図 4 では前方のやや離れた物体，図 5 は比較的近くの壁，図 6 では足元を指さしているシーンを例に出した．

[*1] MonoEye SDK は MonoEye を利用するための実装である。この SDK は共同研究という立場で提供されているため、一般には非公開である。

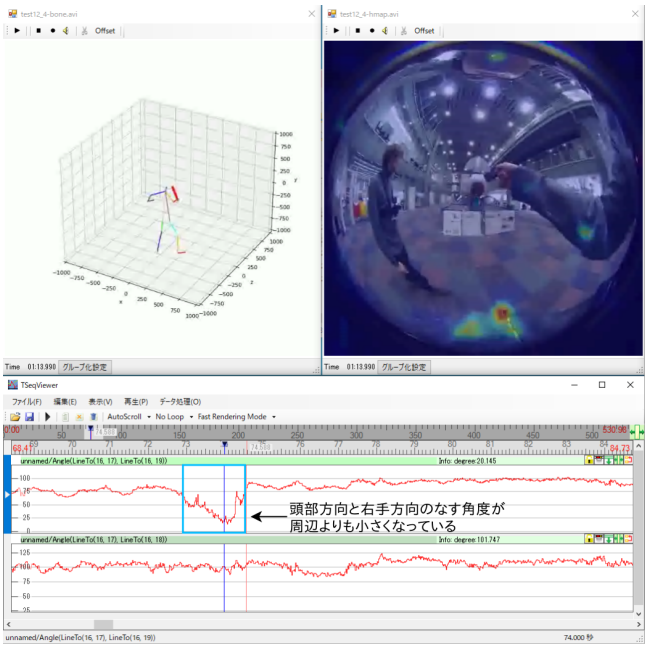

図 4　ロボットを指さすシーン

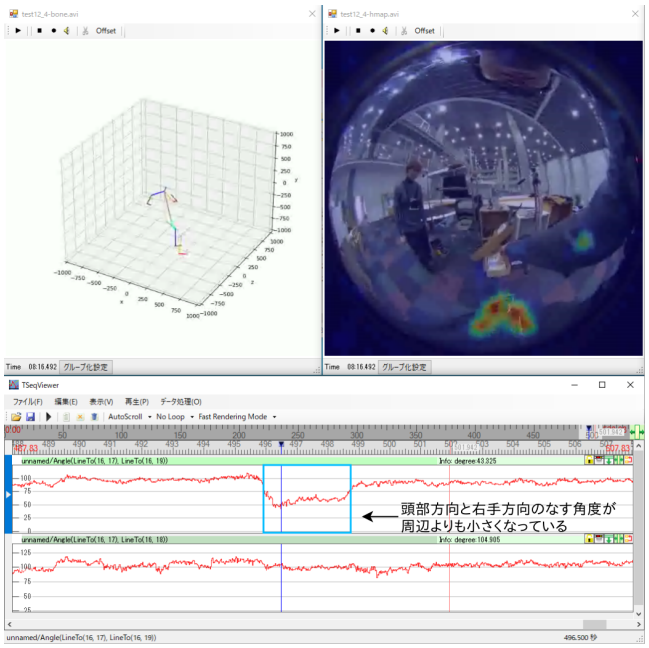

図 6　床のダンボールを指さすシーン

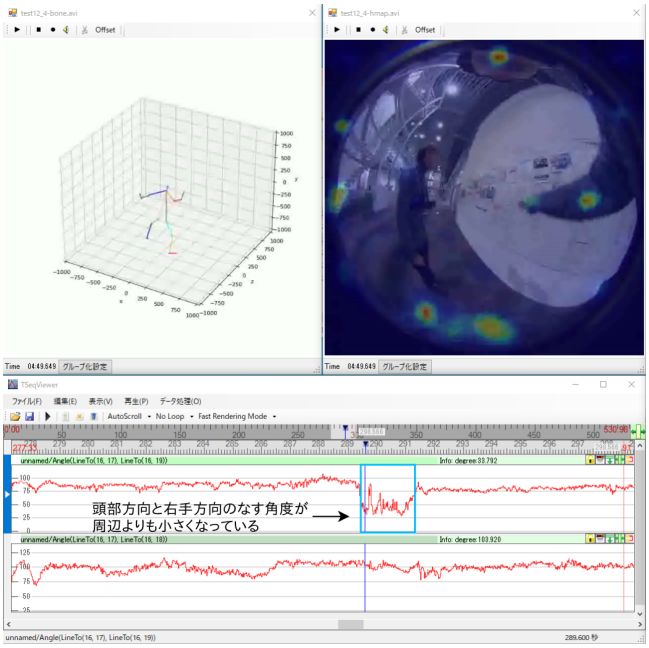

図 5　壁の広告を指さすシーン

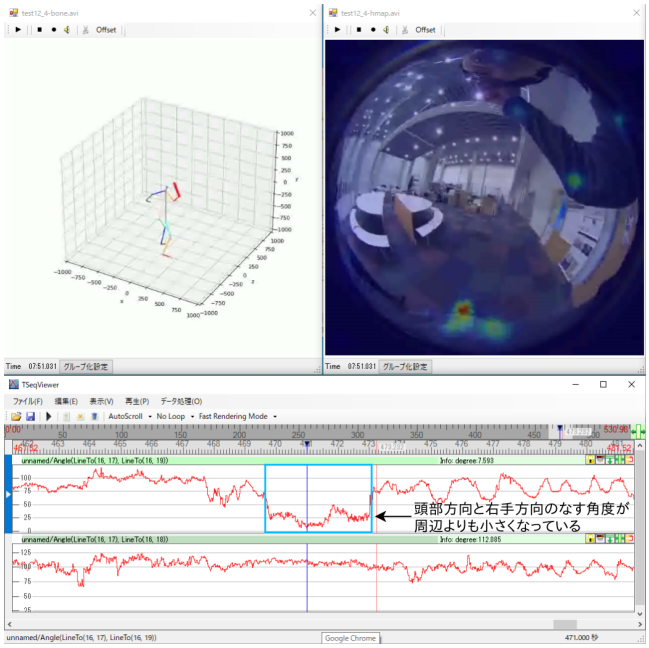

図 7　ユーザが頭を掻くシーン

図 4，図 5，図 6 では，指さしの直前から両ベクトルのなす角度が徐々に小さくなり，指さしの瞬間が周辺のシーンの中ではもっとも小さくなっている．これらの主要なシーンはそれぞれ違った場所の映像から抽出したものであるが，3 シーンとも似た傾向を示して指さしが行われていた．これによって，指さしの判定には両ベクトルのなす角度を参考にできることが分かった．

しかし，前後の時刻より両ベクトルが小さくなっていたとしても，指さしとは限らないケースがあった．例えば，ユーザが頭を掻くなど，自分の頭部付近に手を近づけた場合に両ベクトルのなす角度が小さくなり，指さしを行った場合よりも小さくなることが分かった．これは指さしの場合と似た角度の傾向を示すため，システムが正確に指さしだけを判別するには，閾値の上限と下限の 2 つを考慮する必要があると考える．

さらに，ベクトルのなす角度も映像全体を通して一定ではないため，閾値も一定ではなく，直近の状況を反映して変化させる必要があると考える．

5. 指さし行為の発見による重要シーン推定

前章では，指さし推定がある程度機能することを確認した．本章では，前章で提案した方法が妥当であるかを判断するため，実際に日常的な活動を想定した一人称視点映像を対象として予備実験を行う．

5.1 予備実験の概要

本稿で行う予備実験は，指さし行為を発見することによって，映像の重要なシーンを発見することができるかを確かめることが目的である．例えば，協力者が本を探す過程で話題に出た本を指さすようなシーンが発見できることを期待する．また，指さし行為ではないが本に手を伸ばす，手にとって読んだりするなどの行為も同様に注目シーンとしてされるのも悪くない結果として期待できる．一方で，単に頭を掻いたり，顔を触っているシーンなどが検出されるのは期待しない誤検出である．逆に手の動きが伴わない，視線移動のみで本を探している部分や，立ち止まって会話をしているシーンなどは検出することができないと考える．

この実験に用いる映像の収集は以下のような手順で行われる．

- 1. 協力者 2 名を募集し，2 人 1 組である目的を達成してもらうように依頼する．
- 2. 活動の様子を一人称視点映像として記録する．
- 3. その後，協力者に 1 名ずつ自分の一人称視点映像を視聴してもらう．
- 4. 映像の視聴中，インタビューを行いながら協力者が重要だと思ったシーンをアノテーションしてもらう．

今回は大学内のライブラリで勉強会の本を 2 冊探すという，日常でありうる用事を目的とした．協力者は公立はこだて未来大学の学部 4 年生 2 名である．協力者には一人称視点映像を記録できるカメラを装着してもらい，活動の様子を写した一人称視点映像を収集した．

映像の収集とアノテーションを行った後，指さしベクトルのグラフと比較を行い，指さし行為から映像の重要なシーンを発見することができるかを確認した．なお指さし行為の判定を行うにあたり，前章での結果を踏まえて，指さし行為を判定するための条件に指さしベクトルが一定の閾値内に収まっていることを加えた．今回の実験では，グラフを目視確認したうえで，指さしベクトルと頭部方向の成す角度に上限 45 度と下限 30 度の閾値を設けた．また，アノテーションを行う際は協力者 1 名と実験者が同席し，なぜ映像のような行動を取ったのかなどのインタビューを同時に行った．

5.2 結果

今回は撮影後のアノテーションが完了している 1 名についての結果を取り上げる．

図 8 は，約 30 分の映像のうち，依頼を達成したシーンを含む最後の 10 分のデータである．3，4 段目のグラフはそれぞれ右手と左手について，指さしベクトルと頭部方向の成す角度が 30 度以上 45 度以下であるときに 1，それ以外を 0 としたグラフである．5 段目は実験後のアノテーションによって，当事者が自分にとって重要だったと考えられるシーンにつけた注釈である．なお，注釈の色分けはすべての注釈に一意につけられているもので，分類的な意味はない．データを見ると，指さしが行われているシーンと，当事者が重要であると感じたシーンは必ずしも一致していないことが分かる．例えば，図 9 や図 10 では指さしであると判定したシーンは当事者が注釈をつけた重要シーンの中に時間的に含まれている．図 9 は，当事者が目に入った本を手に取ったシーン，図 10 は当事者が，もう 1 人の協力者が書籍の自動貸出機を上手く扱えていないことが印象に残ったシーンである．しかしながら，図 11 や図 12 では当事者が注釈をつけた重要シーンの中で指さし行為は行われていない．また図 13 のように，指さしが行われていると判定されていても，当事者はそれほど重要なシーンだったと考えていないケースもあった．

5.3 考察

結果は表 1 のように 4 種類に分類できる．以降，それぞれの例がどのケースであるかを踏まえて考察する．

今回取り上げた 1 名の当事者が映像の最後の 10 分につけた注釈について，9 箇所の注釈のうち指さしが行われたと判定されていたのはおよそ 3 箇所であった．図 9 と図 10 は，指さし判定と注釈が重なった 1 のケースである．図 9 は当事者が目に入った本を取ったシーンである．これは指

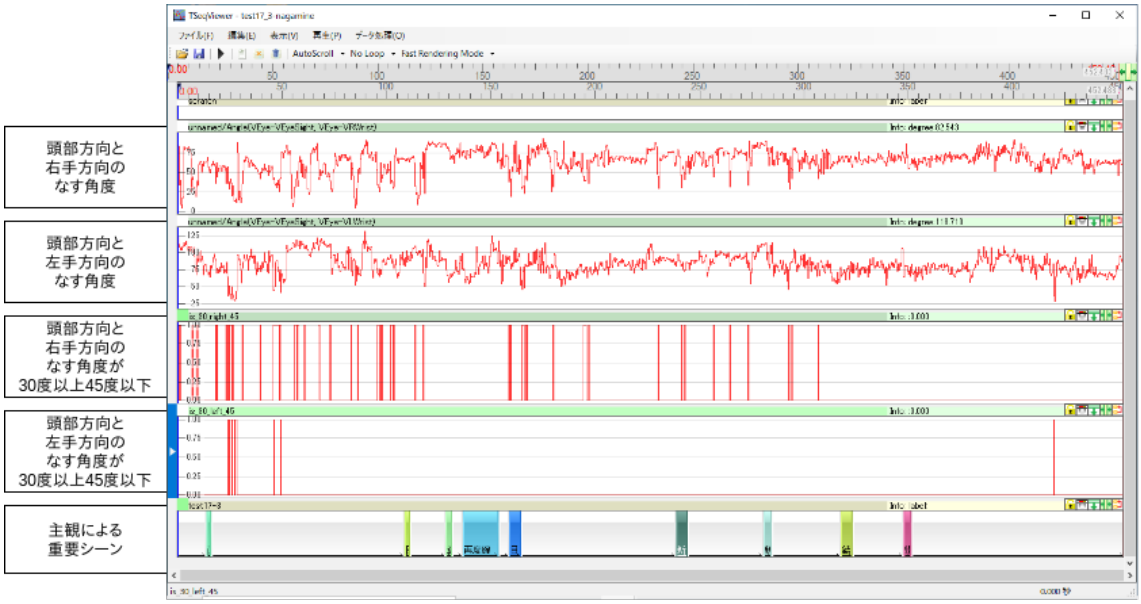

図 8 各グラフとラベルの配置

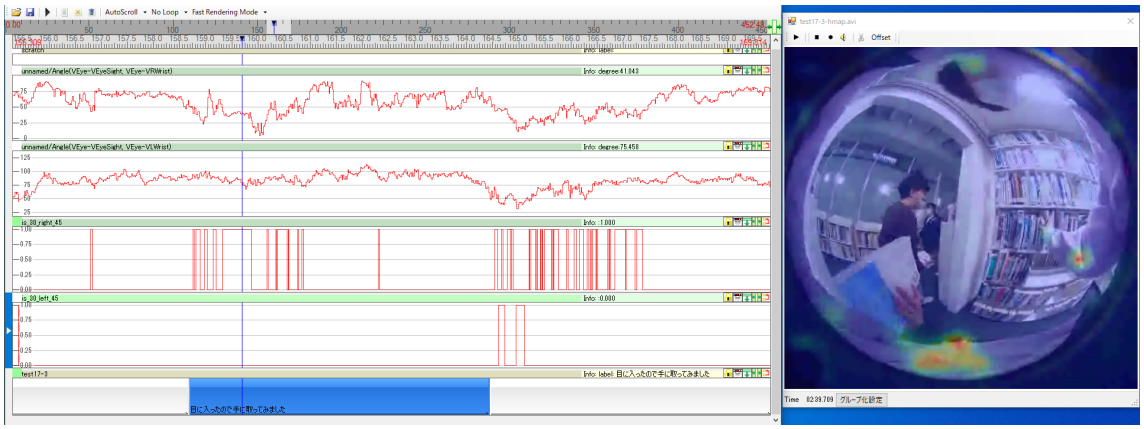

図 9 【正解例】提案手法による指さし検出と当事者による重要シーン注釈が重なる例 1

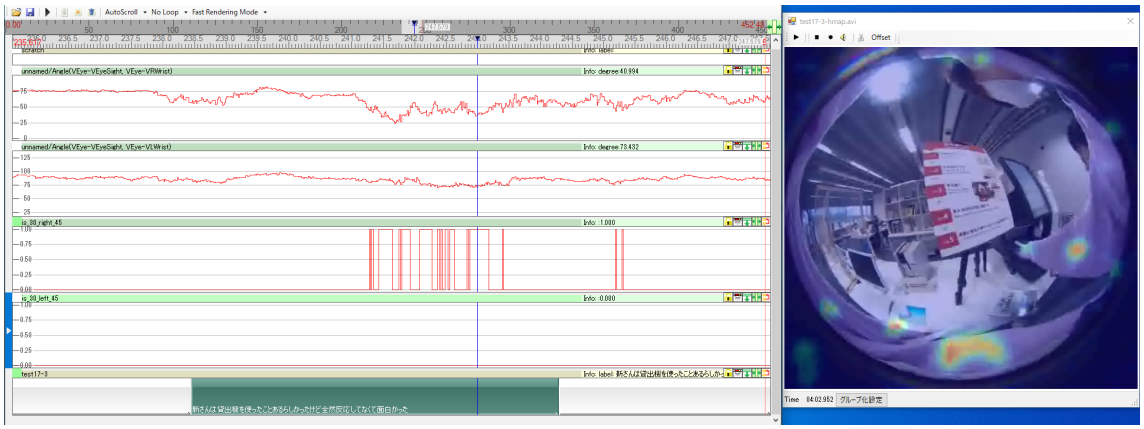

図 10 【正解例】提案手法による指さし検出と当事者による重要シーン注釈が重なる例 2

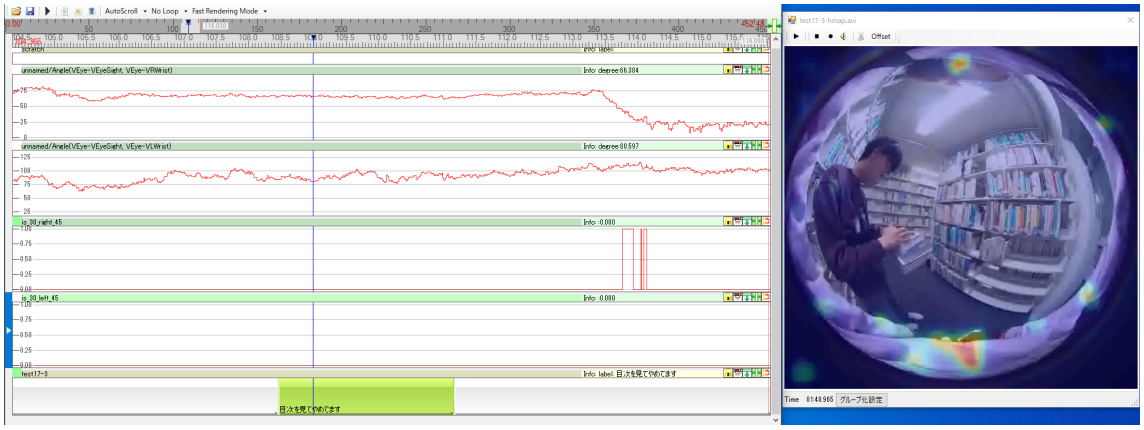

図 11 【検出漏れ】当事者による重要シーン注釈がなされているが提案手法による指さし検出がなされていない例 1

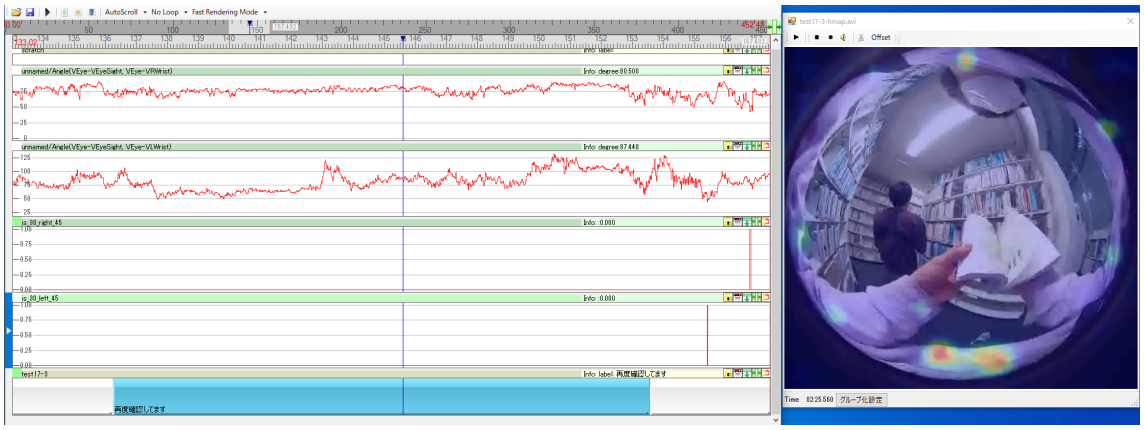

図 12 【検出漏れ】当事者による重要シーン注釈がなされているが提案手法による指さし検出がなされていない例 2

図 13 【誤検出】提案手法によって指さし検出されたものの当事者による重要シーン注釈がなされていない例 1

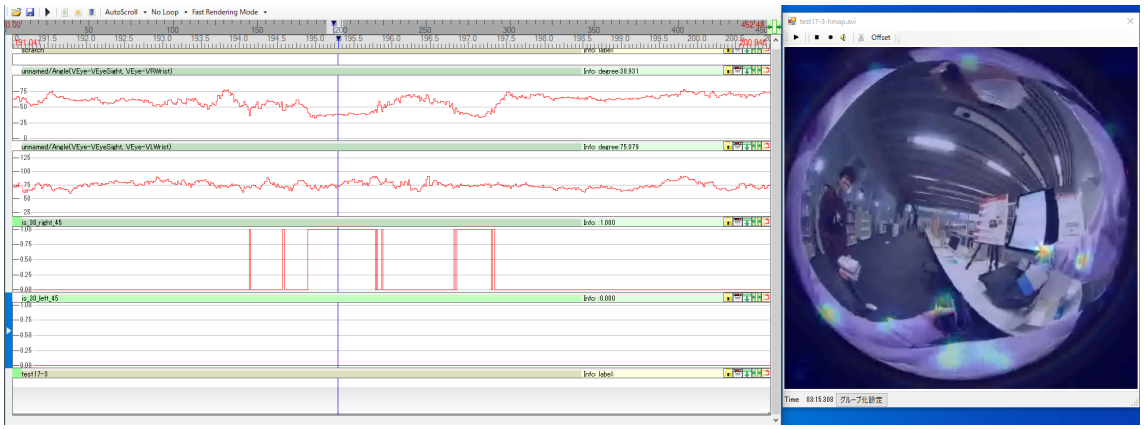

図 14 【誤検出】提案手法によって指さし検出されたものの当事者による重要シーン注釈がなされていない例 2

表 1 判定結果と当事者の注釈についての対応関係

-	指さし行為の判定がある	指さし行為の判定がない
当事者の注釈がある	1. 【正解例】 提案手法による指さし検出と 当事者による重要シーン注釈が重なる例	2. 【検出漏れ】 当事者による重要シーン注釈がなされているが 提案手法による指さし検出がなされていない例
当事者の注釈がない	3. 【誤検出】 提案手法によって指さし検出されたものの 当事者による重要シーン注釈がなされていない例	-

さし行為ではないものの，本に向かって手を伸ばすという行為が抽出できている例である．図 10 は当事者が貸出機を上手く扱えていないもう 1 人の協力者を手助けをするために手を伸ばしたシーンである．これも指さし行為ではないものの，手の動きが伴っているため重要シーンとして抽出できている．

図 11 と図 12 は，指さし判定はないが注釈がある 2 のケースである．図 11 は立ち止まって周りを見渡しているシーンである．ただし当事者の注釈では「目次を見てやめてます」とあるが，映像では当事者が目次を見ていないため，当事者が誤って注釈をつけた可能性がある．しかし，もう 1 人の協力者が本の目次を見ていることを注釈していたとしても，当事者の手に動きがないため，システムが重要シーンとして発見することは難しいと言える．このように立ち止まって身体に動きがない場合には，今回用いたシステムがこのシーンを重要として検出するのは難しい．

図 12 は当事者が選んだ本を再度確認しているシーンである．このシーンでは本をもって読んでいるが，指さし行為とは判定されていないので，重要シーンとして検出していない．図 12 のグラフを見ると，左手の指さしベクトルと頭部方向のなす角度が周りと比較して若干小さくなっているが，閾値の範囲に収まっていないために検出されていないと考えられる．閾値を調整すれば検出できる可能性はあるが，指さしほど手の動きがないため，見つけるのは難しいと考える．

図 13 と図 14 は，指さし判定はあるが注釈がない 3 のケースである．図 13 は当事者が本に手を伸ばしているシーンであり，指さし行為として判定されているが，当事者の注釈はない．また図 14 は当事者が貸出機に手を伸ばしているシーンであり，指さし行為として判定されているが，同様に当事者の注釈はない．これは一見すると擬陽性であるが，本に手を伸ばすシーンや貸出機に手を伸ばすシーンは，本を選ぶ，本を借りるのに必要なシーンであるとも考えられる．そのため，重要なシーンが見つけられないよりも問題ではなく，むしろ意識しなかった重要シーンを抽出できている可能性も考えられる．

結果として，1 のケースのように，期待通りに重要シーンを見つけられた例が存在することが分かった．また，3 のケースのように重要シーンとして当事者に注釈されてはいないものの，明らかな誤検出ではないことから，ユーザが気づいていない重要シーンを抽出できた可能性がある．ただし，3 のケースであっても，頭を掻いたり顔を触ったりするものそれほど重要でない考えられるため，誤検出として注意すべきであると考える．

6. まとめ

本研究の目的は一人称視点映像からユーザの何気ない行動を手がかりにすることで，実世界の重要なシーンを発見・可視化することである．本稿ではその実現のため，非言語行動の 1 つである指さし行為に着目し，映像から指さし行為を発見する方法について検討した．次に映像から指さし行動を抽出することで，重要なシーンをどの程度発見でき

るのかを確認した．

動作確認及び予備実験では，非言語行動の 1 つとして会話の中で参照している対象物を示す行為である指さしに注目した．指さしの判定に必要な視線ベクトルと頭部方向を得るために，一人称ライフログ映像から姿勢を推定可能な MonoEye を用いた．

予備実験の結果，指さし行為に近い動作を抽出し，それを手がかりにすることで，一部の重要シーンを推定できる可能性が示唆された．今回の動作確認では分析対象の人物が 1 名であるため，今回の結果のみをもって，指さし行為からの重要シーンの推定が完全に動作するとは言えない．しかし，注釈のうちいくつかは一致していたことから，注目する非言語行動の 1 つとして指さし行為を扱うことは有用であると考える．指さし行為から重要なシーンをどの程度探すことができるかは，依頼や教示内容を再考したうえで，より多くの協力者を募って確認する必要があると考える．今後は，実験を繰り返して確認すると同時に，他にどのような非言語行動を使うことができるのかを検討していく．

謝辞

本研究の基盤システムである MonoEye システムの主な開発者の Dong-Hyun Hwang 氏には応用システム開発にあたって多大なるご尽力を頂きました．ここに深く感謝します．

参考文献

[1] 竹内俊貴，田村洋人，鳴海拓志，谷川智洋，廣瀬通孝：ライフログとスケジュールに基づいた未来予測提示によるタスク管理手法，情報処理学会論文誌，Vol. 55, No. 11, pp. 2441–2450 (2014).

[2] 中村聡史：LifelogViewer(ライフログビューア)，コンピュータ ソフトウェア，Vol. 30, No. 1, pp. 1_20–1_25 (オンライン)，DOI: 10.11309/jssst.30.1_20 (2013).

[3] Sumi, Y., Etani, T., Fels, S., Simonet, N., Kobayashi, K. and Mase, K.: *C-MAP: Building a Context-Aware Mobile Assistant for Exhibition Tours*, pp. 137–154 (online), DOI: 10.1007/3-540-49247-X_10, Springer Berlin Heidelberg (1998).

[4] Ghosh, J.: Discovering Important People and Objects for Egocentric Video Summarization, *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, USA, IEEE Computer Society, pp. 1346–1353 (2012).

[5] Blum, M., Pentland, A. and Troster, G.: InSense: Interest-Based Life Logging, *IEEE MultiMedia*, Vol. 13, No. 4, pp. 40–48 (online), DOI: 10.1109/MMUL.2006.87 (2006).

[6] Higuchi, K., Yonetani, R. and Sato, Y.: EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, New York, NY, USA, Association for Computing Machinery, pp. 6536–6546 (online), DOI: 10.1145/3025453.3025821 (2017).

[7] Kayukawa, S., Higuchi, K., Yonetani, R., Nakamura, M., Sato, Y. and Morishima, S.: Dynamic Object Scanning: Object-Based Elastic Timeline for Quickly Browsing First-Person Videos, *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, New York, NY, USA, Association for Computing Machinery, pp. 1–6 (online), DOI: 10.1145/3170427.3189085 (2018).

[8] Bolaños, M., Mestre, R., Talavera, E., Giró-i Nieto, X. and Radeva, P.: Visual summary of egocentric photostreams by representative keyframes, *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–6 (online), DOI: 10.1109/ICMEW.2015.7169863 (2015).

[9] 角　康之，矢野正治，西田豊明：マルチモーダルデータに基づいた多人数会話の構造理解，社会言語科学，Vol. 14, No. 1, pp. 82–96 (オンライン)，DOI: 10.19024/jajls.14.1_82 (2011).

[10] Toyama, K. and Sumi, Y.: Quick Browsing of Shared Experience Videos Based on Conversational Field Detection, *Mobile Computing, Applications, and Services* (Murao, K., Ohmura, R., Inoue, S. and Gotoh, Y., eds.), Cham, Springer International Publishing, pp. 40–55 (2018).

[11] Hwang, D.-H., Aso, K., Yuan, Y., Kitani, K. and Koike, H.: MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera, *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, New York, NY, USA, Association for Computing Machinery, pp. 98–111 (online), DOI: 10.1145/3379337.3415856 (2020).

[12] Sumi, Y., Yano, M. and Nishida, T.: Analysis Environment of Conversational Structure with Nonverbal Multimodal Data, *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/1891903.1891958 (2010).