一人称ライフログ映像からの興味領域の切り出し

久米田 羽月[1,a)]　角 康之[1,b)]　Hwang Dong-Hyun[2,c)]　小池 英樹[2,d)]

概要：本研究は，胸に装着したカメラの映像から，装着者の興味領域の切り出しを実現することが目的である．人が日常の気づきに対して行う，指さし行為や頭部運動のような無意識の非言語行動とパターンに着目し，ユーザの興味との関係性について究明を行う．この関係性を明らかにすることで，興味を持ったものについて意識的に記録を行うことなく，ユーザの興味が表出したシーンを判別し，それを切り出すことができる．実世界の対象を切り出すことで，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．本研究の独自性は，広角一人称映像を利用することで，カメラ 1 台で記録が完結し，非言語行動をヒントにして，ユーザの反応に基づいたシーンの推定が可能であるという点である．

1. はじめに

本研究は，ユーザ自身が写り込んだ一人称視点のライフログ映像を用い，ユーザの非言語行動を手がかりにし，実世界の重要シーンを意識せずとも記録することで，ユーザ自身が日常で興味を持ったシーンを苦労なしに切り出し，可視化することが目的である．

ライフログによって，日常の振り返りが可能となり，そのことによる生活の質向上が望まれる．しかしライフログを映像として残した場合，その長さが問題となる．1 日分のライフログ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．標準的な動画再生ソフトウェアであれば，シーク機能を利用して動画の場面を探すことができるが，長時間の映像から特定のシーンを探し出すのは困難である．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．

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，画像に写ったものを手がかりとして解決を試みた研究がある．Ghosh らの研究 [1] は，一人称視点の映像に写った人物や物体の重要度を測り，それをもとに映像を要約するというものである．

中村 [2] はデジタルカメラで撮影したライフログ写真の情報を利用することで，手間なく管理，検索できるシステムを提案している．これは撮影日時や撮影場所，顔認識の結果を使うことで，振り返りを容易にできるというものである．

別のアプローチとして，カメラをつけているユーザ自身の何気ないふるまいから興味を推定し，タグ付けする方法も考えることができる．

そこで，本研究では，**図 1** のように，魚眼レンズによる広角映像を撮影可能なカメラを身に着けるだけで，実世界の重要シーンの切り出しを実現する．図 1 は，会話中に話題となったものを指さすことで，その対象の写ったシーンが重要なものとして自動的に切り出される様子である．ここで使用するカメラは，一般的なアクションカメラに 270 度の超広角魚眼レンズを装着したものであり，身につけたユーザ自身も写り込むようになっている．

本研究では，人が日常の気づきに対して行う，指さし行為や頭部運動のような無意識の非言語行動とパターンに着目し，ユーザの興味との関係性について究明を行う．この関係性を明らかにすることで，興味を持ったものについて意識的に記録を行うことなく，ユーザの興味が表出したシーンを判別し，興味領域の切り出しを行うことができる．本研究の独自性は，広角一人称映像を利用することで，記録はカメラ 1 台で完結し，非言語行動をヒントにして，ユーザの反応に基づいたシーンの推定が可能であるということである．中村 [2] の研究では，意識的に撮った写真からユーザが興味を示しそうなシーンを探索することができた．それに対して本研究では，本人でも忘れてしまうような些細な出来事についても記録，抽出できる部分が面白い点である．

[1] 公立はこだて未来大学
[2] 東京工業大学
[a)] u-kumeta@sumilab.org
[b)] sumi@acm.org
[c)] hwang.d.ab@m.titech.ac.jp
[d)] koike@c.titech.ac.jp

図 1　一人称ライフログ映像による興味領域の切り出し

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 [3] がある．この研究では，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，ユーザの興味の対象を推定している．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．

一方，Kayukawa ら [4] 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ている．そこで，映像中の物体に注目し，ユーザが任意の物体が写り込んだシーンを重要視することができるようになっている．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．本研究では広角一人称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．

一人称視点の映像から，社会的な交流の種類を検出する研究がある．Fathi ら [5] は，一人称視点の映像に写り込む他者の顔の向きから，顔の位置と向きを 3 次元上に置いて解析することで，他者がどの位置におり，誰が誰に，またはどこに注目しているのかを，約 7 割の精度で推定できたとしていた．これは社会的な場での興味対象を，一人称視点の映像から分析するという意味で，本研究と関連している．

その他の興味深いアプローチとして Bolaños ら [6] は，画像の特徴をクラスタリングし，代表的なシーンを抽出することで，映像を要約していた．これは候補となるシーンの中から，似たシーンを排除することができるため，好ましい結果である．

これらの問題を解決するために，Hwang ら [7] 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を提案した．本研究では，Hwang らの研究 [7] を前提として，ライフログ映像の収集，および分析を行う．

3. 非言語行動を活用した興味領域の切り出しの提案

本研究では，Hwang ら [7] による MonoEye システムを用いて，広角一人称映像を分析し，フレームごとの姿勢推定を行う．本研究で提案するシステムの概略を**図 2** に示す．得られた姿勢の時系列データから，ユーザが興味を示したと考えられるシーンを判別し，興味の対象を推定する．具体的には，ユーザがある対象の前に長時間とどまる，「二度見」を行うなどの行動があれば，その対象はユーザを引き付ける要素を持っており，ユーザはその対象に興味があるという仮説が立てられる．このような行動パターンをもとに推定した興味の対象と，ユーザが実際に興味を持った対象を比較し，結果が一致していれば，システムがある程度の精度で興味を推定できると考えられ，興味領域の切り出しを実現することができる．

ユーザの姿勢データから行動パターンを抽出する方法については現在調査中であるが，顔や体の向きなどから興味の対象を絞り込むことができると考える．Higuchi ら [3] は，移動，手の動き，他の人物の写り込みなどを参考にしているため，これらの手法を組み合わせることで，より効果的に分析を行うことができると考える．

4. システムの構築

4.1 一人称視点の魚眼映像による姿勢推定

本稿で提案するシステムでは，MonoEye システム [7] を使用する．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

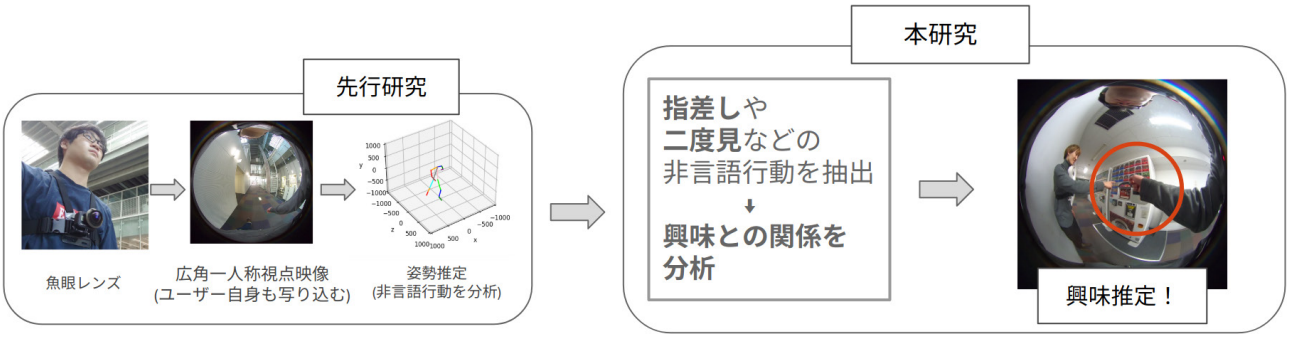

図 2 提案システムの概略

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，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．

4.2 興味推定の方法論

本研究において，ユーザにとって重要なシーンを推定するために重要視する要素は，ユーザの指さし行為と頭部方向である．

角ら [8] 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．そのため，本研究では指さし行為を 1 つの指標とする．角らの研究 [8] 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ないためである．本研究において，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であると仮説を立てる．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．

図 3 は，一人称ライフログ映像と，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，上段が右手，下段が左手についてのグラフである．この図では，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．

ただし，角らの研究 [8] では指さし行為は行為者のみの ego-centric な行為ではなく，会話のパートナーが存在することで成り立つ行為であるとしており，多人数の視線が指さし行為に同期した場合のみ指さし行為が行われたと認定している．本研究においても，会話の中で行われる指さし行為を想定しているため，今後は多人数の視線にも注目する必要があると考える．また，ユーザの頭部方向について，指さし行為以外にも「二度見」のような非言語行動に注目することによって，その瞬間が注目すべきシーンであることも分かると考える．

以上より，指さし行為や，「二度見」のような非言語行動を行ったタイミングがユーザに取って重要なシーンであると考える．さらに，これらの非言語行動に関係する指さし方向と頭部方向を分析することで，映像のどこにユーザの興味対象があるのかということも推測できると期待する．

4.3 システムの評価方法

現在システムは開発中であるため，具体的な評価方法は検討段階である．しかし，システムの目的はユーザの興味領域を切り出すことであるため，システムの評価を行う場合は，実験参加者を募り，アンケートでの調査を行う必要があると考える．

評価観点は，ユーザからみてシステムがどの程度正確に重要なシーンを抽出できているかが重要であると考える．そのため，実験参加者に実際にカメラを身に着けてもらい，一定の時間自由に行動してもらう．その後，実験参加者に

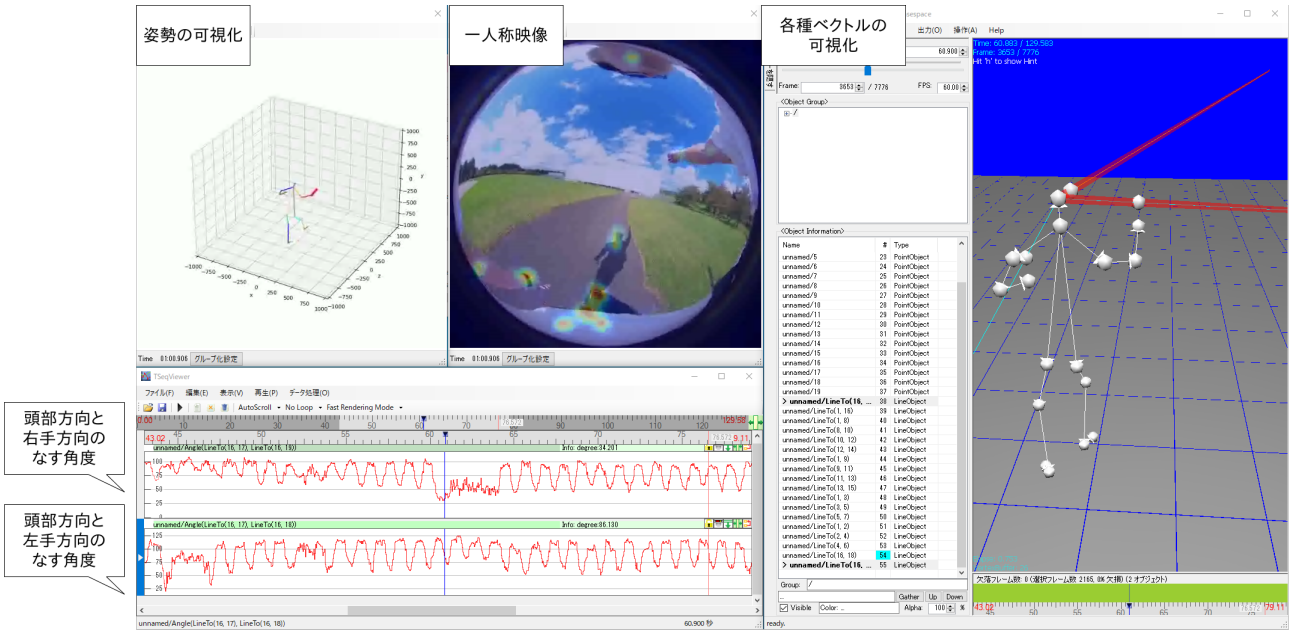

図 3　指さし行為の抽出

撮影した映像を視聴してもらい，興味を持ったシーンをラベル付してもらう．それをもとに，システムの挙げたシーンがどの程度，実験参加者のものをカバーできているかといった評価が 1 つの指標になると考える．

また，Higuchi らの研究 [3] では，映像の中から特定のシーンを探すというタスクの平均走査速度を評価している．これは，長時間のライフログを振り返る際の時間的コストが高いという問題を解決することができているかを測ることができるため，本研究でも参考にできると考える．

5.　日常を想定した興味推定の動作確認

前章で提案した方法を用い，日常的な行動を想定した一人称映像を対象として動作確認を行った．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である．

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．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．

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．このように，前後の時刻より両ベクトルのなす角度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．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．

6.　おわりに

本研究では，魚眼レンズによる広角映像を撮影可能なカメラを身に着けるだけで実世界の重要シーンの切り出しを実現することを目的として，ユーザの非言語行動を利用してユーザにとって重要なシーンを推定する方法を提案した．

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 システム [7] を用いた．

ユーザの興味が現れている重要シーンを推定するにあたって，頭部の方向を示すベクトルと頭部から手首の少し上に伸ばしたベクトルのなす角度が，他のシーンよりも小さくなっている場合に，指さしであるとの仮説を立てた．その結果，指さしが行われているシーンでは，両ベクトルのなす角度が他のシーンよりも小さくなっている傾向があり，両ベクトルのなす角度を特定の閾値を使ってフィルタリングすることで，指さしをある程度判定できることが分かった．ただし，頭を掻くような行為をした場合，指さし行為をした場合よりも両ベクトルのなす角度が小さくなる場合があった．このことから，単に角度が小さい場合を抽出するだけでは，指さし行為とは断定できないと考えられる．今後はより正確に指さし行為を判別できるような方法を考案する必要がある．

また，本稿で指さし判定の失敗例として挙げた頭を掻くような行為も，一種の非言語行動と考えれば，会話の判定や考えを巡らせているシーンの抽出に役立つ可能性がある．複数人での会話における分析や，「二度見」のような他

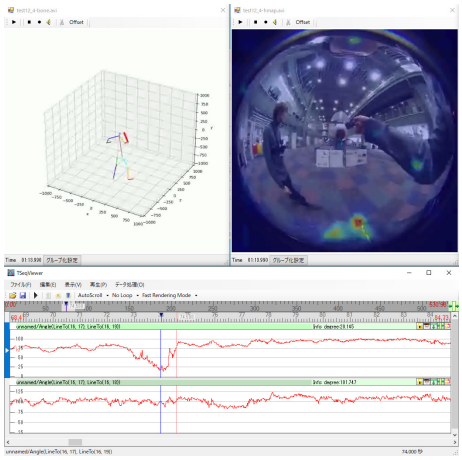

図 4 ロボットを指さすシーン

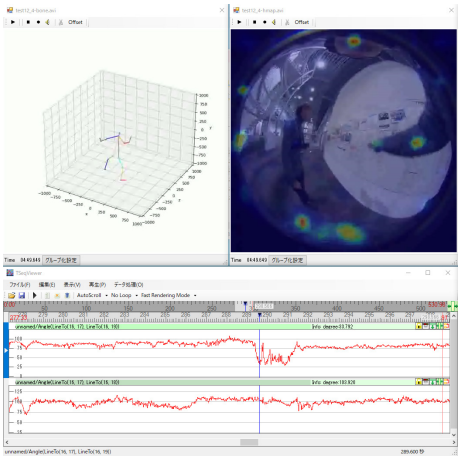

図 5 壁の広告を指さすシーン

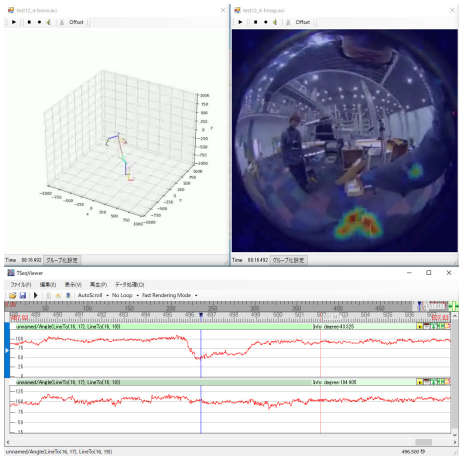

図 6 床のダンボールを指さすシーン

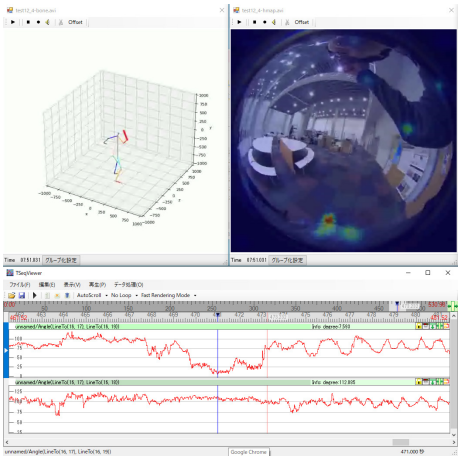

図 7 ユーザが頭を掻くシーン

の非言語行動については検討できておらず，今後の課題となっている．しかし，指さしのような非言語行動の判定を改善するとともに，より多くの非言語行動を扱うことができれば，今後の発展が見込める．

現在は指さしが行われているシーンの候補を抽出する方法を検討したに過ぎないが，非言語行動を手がかりにした興味領域の切り出しの第一歩になったと考える．

参考文献

[1] Joydeep Ghosh. Discovering important people and objects for egocentric video summarization. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pp. 1346–1353, USA, 2012. IEEE Computer Society.

[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, 2013.

[3] Keita Higuchi, Ryo Yonetani, and Yoichi Sato. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 6536–6546, New York, NY, USA, 2017. Association for Computing Machinery.

[4] Seita Kayukawa, Keita Higuchi, Ryo Yonetani, Masanori Nakamura, Yoichi Sato, and Shigeo Morishima. Dynamic Object Scanning: Object-Based Elastic Timeline for Quickly Browsing First-Person Videos. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pp. 1–6, New York, NY, USA, 2018. Association for Computing Machinery.

[5] Alircza Fathi, Jessica K. Hodgins, and James M. Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1226–1233, 2012.

[6] Marc Bolaños, Ricard Mestre, Estefanía Talavera, Xavier Giró-i Nieto, and Petia Radeva. Visual summary of egocentric photostreams by representative keyframes. In *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–6, 2015.

[7] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, pp. 98–111, New York, NY, USA, 2020. Association for Computing Machinery.

[8] 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, 2011.