一人称ライフログ映像の追体験による重要シーンラベリング手法の提案

久米田 羽月[1,a)]　角 康之[1,b)]

概要：本研究の目的は，一人称ライフログ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ライフログ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，映像中の重要シーンをラベリングすることに焦点を当てる．重要シーンのラベリングにおいて，VR ヘッドセットを用いた映像の追体験を行うことによって映像中の重要シーンのラベリング作業を効率的に行う方法を提案する．

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現れたとすると，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，振り返りを容易にする手法を作成することである．非言語行動と重要シーンの関係を分析するためには，映像中のどの時間，どの部分が重要に感じられたかのデータを十分に収集する必要がある．そこで本稿では図 1 のように，実世界ライフログ重要シーンのラベリングにおいて，VR ヘッドセットを用いた映像の追体験を行うことによって映像中の重要シーンのラベリング作業を効率的に行う方法を提案する．

本稿では以下，2 章では関連研究について述べる．3 章では提案するラベリング手法について述べ，4 章では実装について述べる．5 章ではシステムの動作確認と結果，考察について述べる．6 章，7 章では，本稿のまとめと今後の展望について述べる．

[1] 公立はこだて未来大学
[a)] u-kumeta@sumilab.org
[b)] sumi@acm.org

図 1 一人称ライフログ映像の追体験による重要シーンラベリング手法

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もつ 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．

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，および分析を行っている．魚眼レンズによって撮影された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. ライフログ映像の追体験による重要シーンのアノテーション手法

これまで我々は，ユーザの非言語行動と重要シーンの関係を分析するため，一人称ライフログ映像のどのシーンが当事者にとって重要であるかを調査してきた．その方法として，当事者に映像を見返してもらい，重要に感じたシーンのアノテーションを行ってもらった．アノテーションを行う際は参加者 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．しかしこの手法にはいくつかの問題点が考えられる．

1 つは，アノテーションを行う際にインタビューを行うことである．インタビューを行うのは，質問によって当事者に記憶の想起を促すためであった．しかし，このインタビューではインタビュアによって質問内容にブレが出てしまい，属人性が生じてしまう可能性がある．また，アノテーション作業では再生と一時停止を繰り返し行うため，映像よりも長い時間が必要になる．ユーザの非言語行動と実世界の重要なシーンとの関係を分析するにあたっては，できる限り多くのラベル付きデータを分析する必要があると考えられるため，ラベリングに多くの時間を費やすことは望ましくない．1 つの解決アプローチとして，アノテータの視線追跡とボタン操作による時間方向のアノテーションを組み合わせることで，映像中のどの時間，どの部分に注目していたかを自動的に行う方法が考えられる．

もう 1 つの問題は，魚眼レンズの映像が一般的な映像と比べて歪んでいることである．一般的な広角レンズを用いた映像では少しの歪みがあっても視聴に問題ない場合が多いが，本研究で使用している魚眼レンズは画角が 280 度あるため，映像の中心近くであっても歪みが大きい．これまでは当事者のみに映像を見返してもらっていたが，今後，当事者以外に映像をアノテーションしてもらう場合を考えると，実験時の主観を思い出せる当事者とは違い映像の歪みによる見づらさがアノテーションに影響する可能性がある．

そこで，魚眼レンズの映像を球に投影し，VR ヘッドセットを用いて視聴することで，歪みを抑えたうえで当事者に近い視点で映像を視聴することができるのではないかと考えた．球にテクスチャとして投影した魚眼レンズの映像を中心から視聴する際，ヘッドセットの向きからテクスチャの座標を割り出すことで擬似的な視線を求めることができる．また，映像の視聴中にコントローラーのボタン操作によってリアルタイムにアノテーションを行うことで，作業の時間短縮が期待できる．

4. システムの実装

魚眼レンズを VR ヘッドセットを用いて視聴するため，Unity[*1]を用いて視聴用のソフトウェアを開発した．このソフトウェアでは，魚眼レンズの映像を球に投影し，球の中心からプラネタリウムのように映像を視聴することができる．球に投影する映像は，ffmpeg[*2]を用いて事前に正距円筒図法へ変換したものを用いる．また，ヘッドセットの向きを測定することで擬似的に視線を計測することができるようになっており，コントローラーのボタンを押すことによって重要シーンのアノテーションを行うことができるようになっている．これらのデータは，動画の視聴後に CSV データとして書き出しが行われる．図 2 はこのソフトウェアの動作の概略を表したものである．

また，書き出された疑似視線データを魚眼レンズの映像上にオーバーレイ表示することによって可視化を行うソフトウェアも作成した．図 3 は視聴用ソフトウェアと視線データの可視化ソフトウェアを含めた概略を表したものである．

5. 動作確認と結果

前章で提案したシステムを用い，実際の一人称ライフログ映像を対象として動作確認を行った．動作確認に用いた映像の収集は次のような手順で行われた．

- 1. 参加者 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．

動作確認に用いた映像は，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て設定したものである．参加者は公立はこだて未来大学の学部 4 年生 2 名である．参加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．

また，動作確認を行ったのは筆者および公立はこだて未来大学の大学院生 1 名，同大学の学部生 1 名である．

動作確認の結果，ライフログ映像を見返すことは可能であったが，次の問題が挙げられた．

- 1. 視野が狭いこと．
- 2. カメラが胸部に取り付けられているため，視点の高さが当事者の視点よりも低いこと．
- 3. 当事者が移動中の映像を視聴すると 3D 酔いが起こりやすいこと．

[*1] https://unity.com/
[*2] https://ffmpeg.org/

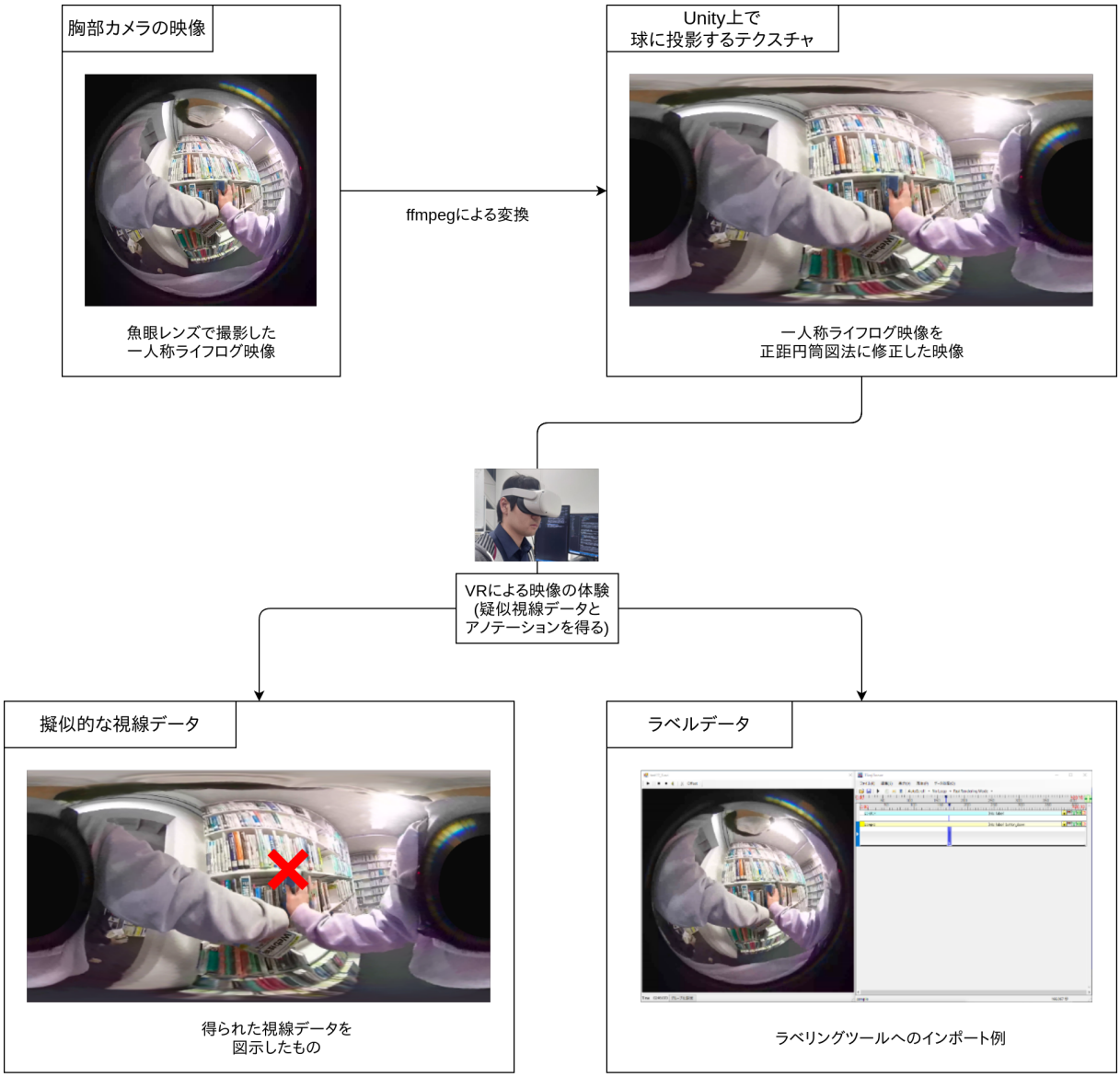

図 2　アノテーションソフトウェアの動作の概略

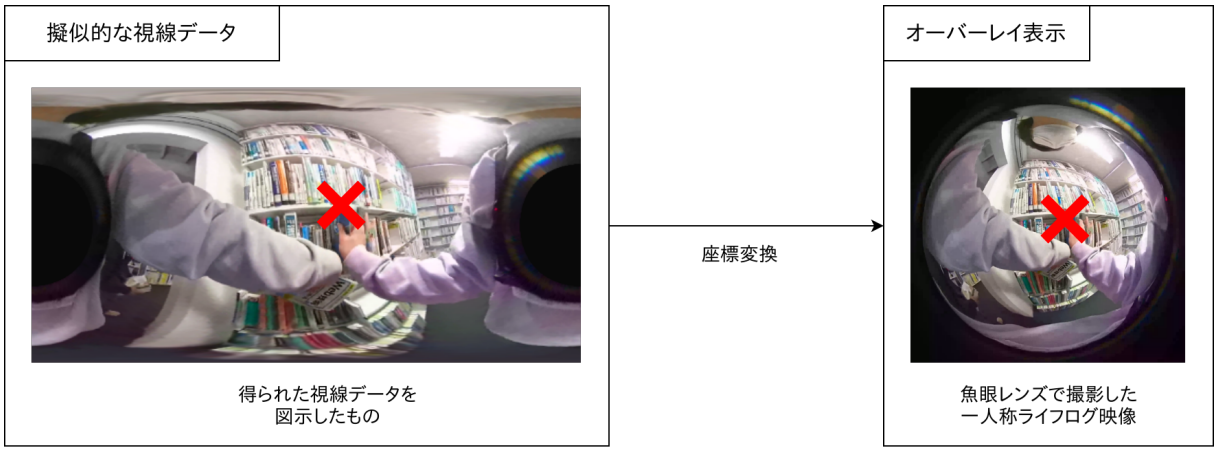

図 3　視線データの可視化ソフトウェアの概略

6. 考察と今後の展望

今回の動作確認によって，当事者が移動中の映像を VR で視聴すると，酔いやすいという問題点が判明した．この問題については，MonoEye で得られるカメラの傾きを考慮することによって低減することができるのではないかと考える．視野が狭いという問題に対しては，VR 空間上のカメラの位置を中心から少し後ろにずらすことで視野を広げることができると考える．その際，周りを見回すと端が歪んでしまう問題が予想できるため，カメラの操作を工夫する必要があると予想する．

また，カメラの位置が低いために視点の高さが当事者の視点よりも低いことについては，今回の用いた手法では解決できない．今後は上記の問題も考慮したうえで，魚眼レンズの映像そのまま視聴してアノテーションを行う場合と比較し，VR による追体験を用いたラベリング作業が有効であるかを確認する必要がある．

7. おわりに

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．本稿では目的の実現のために，映像中の重要シーンをラベリングすることに焦点を当て，VR ヘッドセットを用いた映像の追体験を行うことによって映像中の重要シーンのラベリング作業を効率的に行う方法を提案した．動作確認においては，視野の狭さと 3D 酔いが起こりやすいという問題が判明した．これらの問題は改善を行い，ラベリングを効率的に行う方法を引き続き検討する．今後は，提案手法を用いたラベリングを行い，非言語行動と重要シーンの関係を分析することで，ユーザの何気ない行動から重要シーンの推定を行う手法を検討していく．

謝辞　本研究は科研費（22H03634）の研究助成を受けた．

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, nov 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, 2013.

[3] Yasuyuki Sumi, Tameyuki Etani, Sidney Fels, Nicolas Simonet, Kaoru Kobayashi, and Kenji Mase. *C-MAP: Building a Context-Aware Mobile Assistant for Exhibition Tours*, pp. 137–154. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[4] 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.

[5] M. Blum, A. Pentland, and G. Troster. Insense: Interest-based life logging. *IEEE MultiMedia*, Vol. 13, No. 4, pp. 40–48, 2006.

[6] 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.

[7] 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.

[8] 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.

[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, 2011.

[10] Kai Toyama and Yasuyuki Sumi. Quick browsing of shared experience videos based on conversational field detection. In Kazuya Murao, Ren Ohmura, Sozo Inoue, and Yusuke Gotoh, editors, *Mobile Computing, Applications, and Services*, pp. 40–55, Cham, 2018. Springer International Publishing.

[11] 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.