一人称ライフログ映像を用いた非言語行動の抽出による視線推定

久米田 羽月[1,a)]　角 康之[1,b)]　小池 英樹[2,c)]

概要：本研究では，ユーザの振る舞いから重要シーンと注目対象を推定することを目的とし，非言語行動を手がかりにすることで振り返りを容易にする手法を作成することを目指す．本稿では，ユーザの非言語行動に注目し，視線を推定する手法を提案する．胸部の魚眼カメラと頭部の視線計測装置を同時に装着して記録したデータをもとに，MonoEye を用いて姿勢・視線ペアのデータセットを作成した．LSTM を用いたモデルでデータセットを学習させた結果，屋外データを学習させたモデルにおいて，データセットに含まれている，いないに関わらず一部視線を推定できたため，非言語行動から視線を推定できる可能性が示唆された．今後はより多くの状況下でデータを収集して性能向上を図り，興味領域からの重要シーン推定を目指す．

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や注目対象を自動的に推定し，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]．このような視覚的特徴を用いて重要シーンや注目対象を特定する手法は直感的な一方，ユーザの興味を自動的に反映できているとは言い切れない．

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．

そこで本研究では，ユーザの振る舞いから重要シーンと注目対象を推定することを目的とし，非言語行動を手がかりにすることで振り返りを容易にする手法を作成することを目指す．本稿では，ユーザの姿勢に注目し，視線を推定する手法を提案する．

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

[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)] koike@c.titech.ac.jp

はゆっくりと再生され，残りの部分は早送りされる．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，探したいシーンによって使い分けることができる．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効率的に発見でき，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．

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，Hwang らによる MonoEye を用いて姿勢推定を行うことで，非言語行動から視線推定を行う．

3. 非言語行動からの視線推定手法の提案

姿勢から視線推定を行うことについて，Krafka[12] らは，頭部方向を含む顔画像から視線推定が可能であることを示した．また Land ら [13] は，身体運動を伴うアクションを行う際，視線が先行することを明らかにした．これらの研究より，身体運動を伴うシーンでは，非言語行動から視線の推定が可能であると考えた．

本研究では，深層学習を用いて姿勢データから眼球の向きを推定することで視線を特定する手法を提案する．

3.1 データセット

システムを作成するために，一人称視点映像と視線データが対になったデータセットが必要である．そこで実験参加者を募り，屋内外でのデータ収集を行った．屋内でのデータ収集は公立はこだて未来大学の構内で行い，屋外でのデータ収集は近隣の公園および街中で行った．実験参加者は，著者の所属する研究室の学生 7 名であり，一人称視点映像の撮影のために魚眼レンズの付いたウェアラブルカメラと視線計測装置を身に着けた．

収集したデータは次の 3 つに大別される．表 1 はデータセットについての情報である．

3.1.1 理想的データ

1 つ目は，屋内で収集した理想的データ（以下，champ6）である．理想的データは，身体運動と視線の関係が分かりやすい恣意的なデータである．具体的には，周囲を見る際に視線を先行させ，次に頭や上体を動かす，場合によって指さし行為を同時に行うなどの動作をした様子が含まれている．このデータはシステムの作成にあたり，身体運動から視線が推定できることを確認するために収集した．また，学習データに含まれないデータでの検証のため，6 番を除く 5 つのデータからなるサブセット（以下，champ5）も作成した．

3.1.2 屋内データ

2 つ目は，屋内で収集したデータである．このデータには，屋内を散歩しているものに加えて，大学内の購買店で商品を探して購入するといったものがある．後者は商品の検討段階で周囲を見渡す，手を伸ばすといった状況を想定して記録を行った．

3.1.3 屋外データ

3 つ目は，屋外で収集したデータである．この屋外データには，公園や観光地を散歩している様子を記録したデータが含まれている．また，屋外データに加えて屋内データのうち 2 つのデータを含めたデータセット（以下，outdoor6）を作成した．

3.2 データの前処理

収集したデータは，胸部カメラによって記録された一人称視点映像と，視線計測装置によって記録された一人称視点映像，および眼球の向きを含む視線情報である．非言語行動の手がかりとなる姿勢データから眼球の向きを推定

表 1: データセットの概要

データセット	データ番号	参加者	収録時間	説明
champ6	1	A	00:01:21	大学構内の 1 階を歩く.
	2	A	00:02:00	大学構内の 1 階を歩く.
	3	A	00:01:17	大学構内の 1 階から 2 階への階段を歩く.
	4	A	00:01:36	大学構内の 1 階を歩く.
	5	A	00:01:09	大学構内の 1 階を歩く.
	6	A	00:02:42	大学構内の 1 階から 2 階への階段を歩く.
outdoor6	1	E	00:33:52	大学構内を散歩する．H と同時に収集.
	2	B	00:33:51	大学構内を散歩する．G と同時に収集.
	3	G	00:54:05	道南四季の森公園を散歩する．D と同時に収集.
	4	C	00:54:04	道南四季の森公園を散歩する．C と同時に収集.
	5	F	01:36:34	大沼国定公園を散歩する.
	6	D	01:55:53	函館の西部地区を散歩する.
購買店データ	1	H	00:12:26	大学構内の購買店で一番おいしそうなものを 1〜2 個購入する.

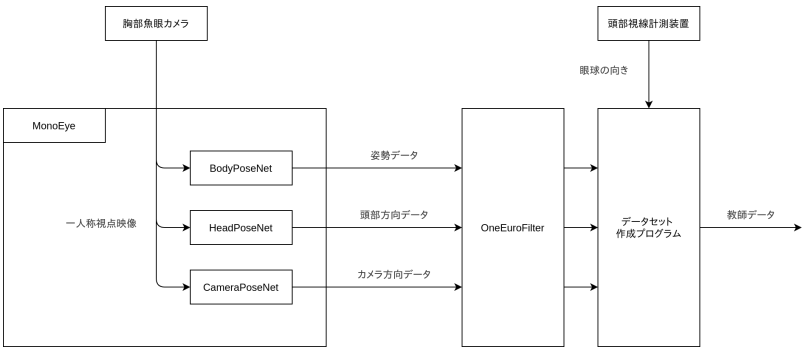


図 1: データの前処理の概略

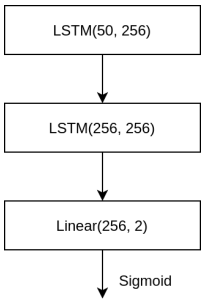


図 2: 深層学習ネットワークの概略

するシステムを作成するため，収集したデータの前処理を行った．図 1 は，データの前処理の流れを記載したものである．

3.2.1 姿勢データの取得

教師データの入力として用いられる姿勢は, 収集した胸部カメラの一人称視点映像から MonoEye を用いて取得した．姿勢データには，BodyPoseNet から得られる 3 次元の姿勢，HeadPoseNet から得られる頭部方向，CameraPoseNet から得られるカメラの向きが含まれる．一人称視点映像は 1 フレームごとに MonoEye へ入力されて姿勢データへと変換される．また，ノイズの低減のため，OneEuroFilter[*1] を用いてデータの平滑化を行った．

3.3 システムの実装

姿勢から眼球の向きを推定するため，LSTM を用いたネットワークを作成した．図 2 はネットワークの概略を示したものである．

3.4 学習

動作確認のため，champ6，champ5，outdoor6 のデータセットをそれぞれ用いて 500epoch の学習を行った（以下，学習させたモデルとデータセットと同様に champ6，champ5，outdoor6 とする）．学習時，損失関数には MSE を用いた．図 3 と図 4，図 5 と図 6 は，champ6 と outdoor6 を学習させた際の学習曲線である．

図 3, 4, 5, 6 の結果に基づき, champ6, champ5, outdoor6 は 250epoch 時点のモデル（以下，250epoch 時点のモデルをデータセットと同様に champ5，champ6，outdoor6）を採用した．

4. 動作確認

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．図 7 は，動作確認のため眼球角度の正解値，眼球方向の推定値，推定値の標準偏差を可視化したものである．ただし，視線の位置は上下左右-90 度から 90 度までの範囲を映像の高さと幅に線形に当てはめたものである．

4.1 結果

図 8 は，champ6 に含まれる 6 番目のデータにおける指さし行為が行われたシーンをそれぞれのモデルで推定させたものである．結果，屋内での理想的データである場合，学習モデルに含まれているデータでは，上体をひねるなど

[*1] https://jaantollander.com/post/noise-filtering-using-one-euro-filter/

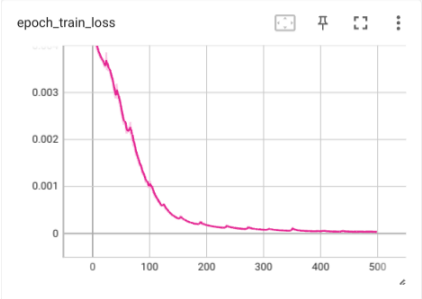


図 3: champ6 を用いたモデルの学習曲線
(trainning loss)

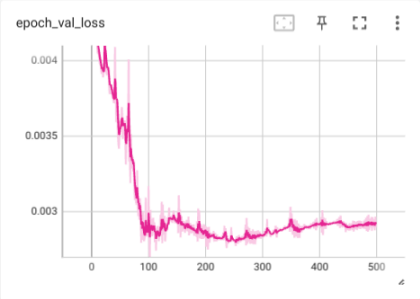


図 4: champ6 を用いたモデルの学習曲線
(validation loss)

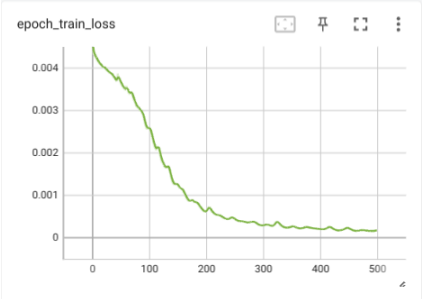


図 5: outdoor6 を用いたモデルの学習曲線
(trainning loss)

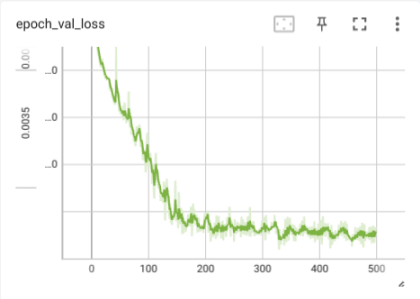


図 6: outdoor6 を用いたモデルの学習曲線
(validation loss)

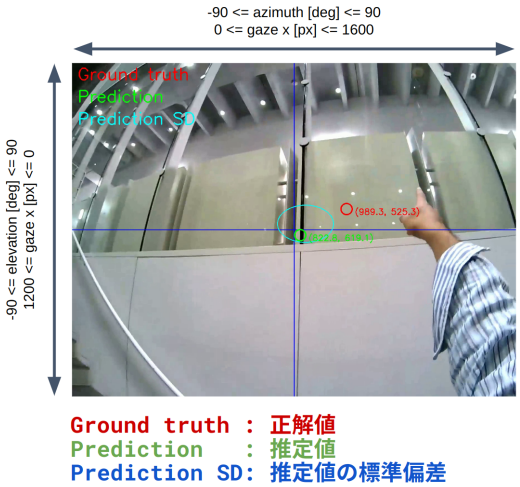


図 7: 可視化プログラムを用いて眼球方向を可視化した画像

の動きを伴って横を見た際には，視線の推定が行えていることが分かった．学習モデルに含まれていない場合では，含まれている場合よりも誤差が大きいものの，視線の推定が行えている可能性があることが分かった．また，屋外データで学習させたモデルでは，指さし行為が行われているシーンでの視線の推定がうまく行われなかった．

図 9 は，outdoor6 に含まれる 3 番目のデータにおける振り向きが行われたシーンをそれぞれのモデルで推定させたものである．結果，屋外のデータである場合，学習モデルに含まれているデータでは，視線の推定が行えていることが分かった．また，理想的データで学習させたモデルでは，振り向きが行われているシーンでの視線の推定がうまく行われなかった．

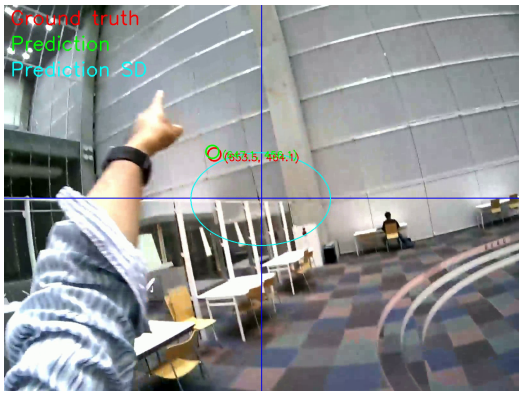
図 10 は，購買データに含まれる振り向きが行われたシーンをそれぞれのモデルで推定させたものである．結果，理想的データで学習させたモデルでは，振り向きが行われているシーンでの視線の推定がうまく行われなかった．屋外データで学習させたモデルでは，振り向きが行われているシーンでの視線の推定が部分的に行われていることが分かった．

5. 考察と今後の展望

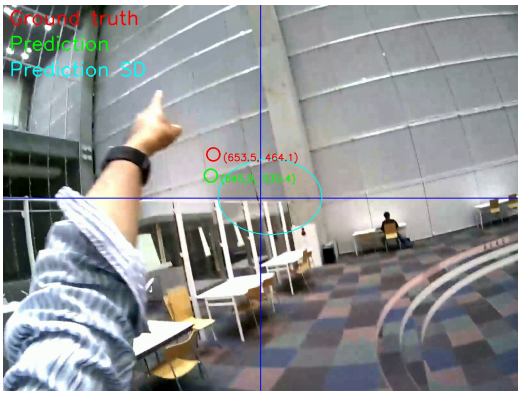
結果より，理想的データセットにおいては，データセットに含まれているデータは推定できていると考えられる．そのため，ネットワークが特徴を学習することが可能であることが分かった．しかし，データセットに含まれていないデータでは全体的な推定の精度が下がっているように見られた．ただし，データセットに含まれているデータの場合でも，部分的に推定できていない箇所があった．逆にデータセットに含まれていないデータの方がうまく推定できている箇所もあったため，データセットの数によって汎化性能に影響が出ていると考える．

屋外データセットで学習させたモデルでは，指さし行為中の視線を推定できていなかった．これは，屋外データセットに指さし行為があまり含まれていなかったことが理由として考えられる．また，理想的データセットでは，自然な状況では行われない行為が含まれており，自然な状況化でのデータから学習させたモデルではうまく推定できなかったことも考えられる．

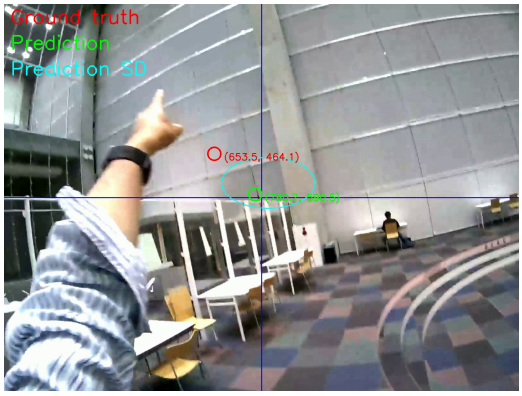
屋外のデータセットにおいては，データセットに含まれ



(a) champ6 モデルによる推定

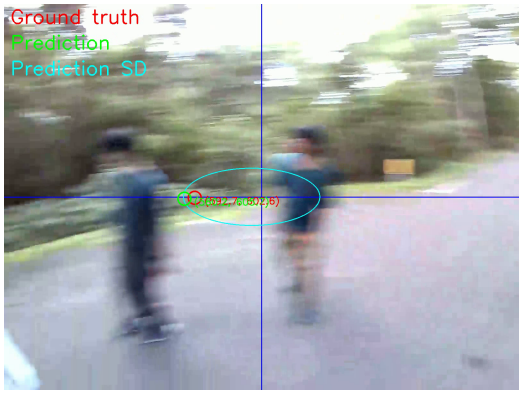


(b) champ5 モデルによる推定

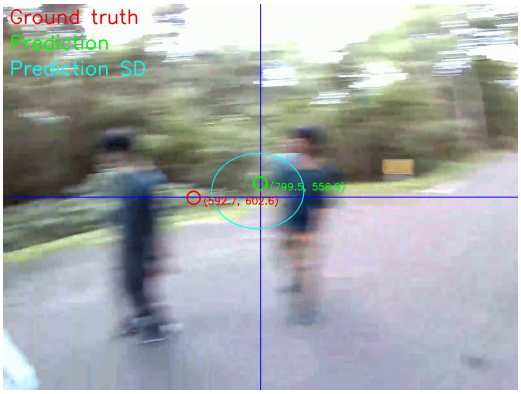


(c) outdoor6 モデルによる推定

図 8: champ6 (6 番) の指さし行為シーン



(a) outdoor6 モデルによる推定



(b) champ6 モデルによる推定

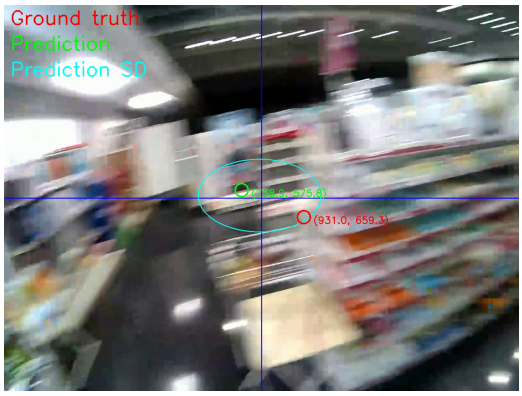
図 9: outdoor6 (3 番) の振り向きシーン

ているデータは推定できていると考えられる．しかし，データセットに含まれない場合，視線の推定ができなかった．

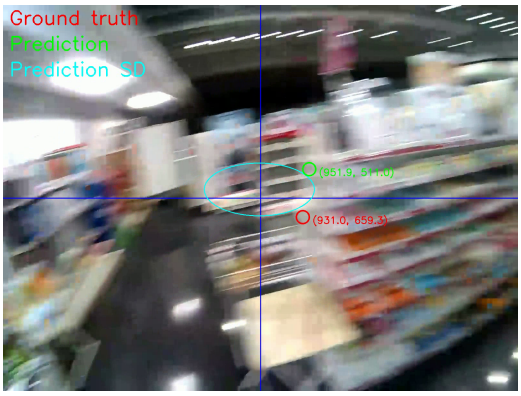
いずれのデータセットにも含まれない購買店データでの結果から，champ6 よりも outdoor6 の方がより正解に近い推定が行えていることが分かる．この理由として，屋外データセットの方がサンプル数が多かったため，より一般的な特徴を学習できた可能性が考えられる．

6. おわりに

本稿では，ユーザの非言語行動から視線を推定することで注目箇所を見つけ出し，実世界の注目対象を発見する手法を提案した．その結果，提案モデルが姿勢から視線を推定できる可能性が示唆された．また，屋外データで学習させたモデルでは，振り向きなどの姿勢の動きを伴う場合，データに含まれていないシーンでもある程度視線を推定することができることが分かった．これは，長時間のデータ



(a) champ6 モデルによる推定



(b) champ6 モデルによる推定

図 10: 購買店データの振り向きシーン

から一般的な特徴を学習できたためと考える．ただし，理想的データに含まれる指さし行為のような恣意的なシーンでは，うまく推定することができなかった．より様々な状況下のデータを学習させることで，より精度を向上させることができると考える．

今後はより多くの状況下でデータを収集することで性能向上を図り，興味領域からの重要シーン推定を目指す．

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, 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.

[12] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Michael F. Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, Vol. 41, No. 25, pp. 3559–3565, 2001.