様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ６ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2020～2023
課題番号：２０Ｈ０４２０５

研究課題名（和文）自己視点・他者視点・固定視点映像の統合解析による人物行動センシング

研究課題名（英文）Human behavior sensing by integrated analysis of first, second, and third-person point of view videos

研究代表者
佐藤　洋一（Sato, Yoichi）

東京大学・生産技術研究所・教授

研究者番号：７０３０２６２７

交付決定額（研究期間全体）：（直接経費）　　14,730,000円

研究成果の概要（和文）：本研究課題では、自己視点・他者視点・固定視点の異なる視点からの映像を統合解析による人物行動センシングに向けて、研究項目１．自己教師有り学習による複数視点映像の総合解析に適した特量表現学習、研究項目２．異なる視点の映像間でのアラインメント、研究項目３．自己視点映像と固定視点映像の統合による人物行動センシング、研究項目４．自己視点映像と他者視点映像の統合によるインタラクション解析、の４項目について研究に取り組み成果を得た。

研究成果の学術的意義や社会的意義
映像に基づく人物行動理解技術は、様々なアプリケーションドメインにおいて書くことが出来ない重要な基盤技術であり、これまでもコンピュータビジョンの分野において活発に研究が進められてきた。これまでは、固定視点映像、自己視点映像、他者視点映像という異なる視点映像が別々に扱われていたのに対し、本研究課題では、他に先駆けて固定視点・自己視点・他者視点という異なる視点映像の統合解析に取り組み、映像に基づく人物行動理解の高度化に貢献する成果を得ることができた。

研究成果の概要（英文）：In this project, we conducted research on the sensing of human behavior by integrated analysis of videos from different viewpoints (self-viewpoint, other viewpoint, and fixed viewpoint) in four aspects: 1. Feature learning for comprehensive analysis of multiple viewpoint videos by self-supervised learning, 2. Spatio-temporal alignment of videos from different viewpoints, 3. Human behavior understanding by integration of self-viewpoint and fixed viewpoint videos, and 4. Interaction analysis by integration of self-viewpoint and other viewpoint videos.

研究分野：computer vision

キーワード：一人称視点映像解析　人物行動理解

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

防犯カメラなどの固定カメラから得られる映像を用いた人物行動センシングは、コンピュータビジョン分野の主要研究テーマの一つとして 1990 年代から活発に研究が進められ、近年では深層学習へのパラダイムシフトによる大幅な性能向上もあり、防犯、犯罪捜査、マーケティング、自動運転など様々な分野で欠かすことのできない基盤技術となっている。

しかしながら、防犯カメラ等から得られる固定視点映像による人物行動センシングには、行動粒度と観測継続性という 2 つの根本的な問題が存在する。固定カメラの観測エリア内での人の検出と追跡、立ち止まりや物を手に取るなどの基本動作の認識、顔や歩容などによる人物認証などの自動解析が実現している一方、人が何に注意を向け、どのように物を取り扱い、誰とどのようなやり取りをどのように行っているのかという、詳細な行動のセンシングはいまだ難しい（行動粒度の問題）。これは、一般的に固定カメラは広い範囲を俯瞰する形で設置されるため、画像解像度と観察方向の制限で個々の人の細かな動作を観察することが出来ないということに起因している。さらに、個々の固定カメラの観測エリアは限られるため、ある人物の行動を長時間継続的にセンシングすることも難しい（観測継続性の問題）。

一方、2010 年代頃からアクションカメラや眼鏡型カメラのようなウェアラブルカメラの映像を対象とした人物行動センシングが注目を集めるようになってきた。撮影エリアや方向が限定され、人物を比較的遠方から捉える固定視点映像とは異なり、装着者自身の視点から撮影された映像（自己視点映像）や相手の視点から撮影された映像（他者視点映像）では、対象人物の行動の様子（手元の作業ややり取りをしている相手の様子）を詳細にかつ継続的に観察可能であり、固定視点映像の場合のような行動粒度と観測継続性の問題が生じない。しかしながら、ウェアラブルカメラで得られる自己視点映像と他者視点映像では、限られた視野でカメラ装着者の近くを撮影するため、自身の全身や周辺状況を直接観測できない（視野範囲の問題）。すなわち、ウェアラブルカメラ映像からの人物行動センシングには視野範囲の制限という根本的な問題が存在する。

これは、互いのウェアラブルカメラから得られる自己視点映像と他者視点映像を用いて人同士のインタラクションにおける行動をセンシングする場合においても同様となる。さらに、視野範囲の限界を別にしても、自己視点映像と他者視点映像のどちらか一方だけではインタラクションにおける行動を正しくセンシングできないという問題が生じる。身振り手振りなどのジェスチャ、表情、アイコンタクトなどは人同士のインタラクションの中で自然に生じるものであり、相手との関係を無視してはその本質的な意味を捉えることが出来ない。例えば、ある表出されたジェスチャはそれ単体で解釈されるべきものではなく、相手から自身へのどのような働きかけを受けて、どのようなタイミングでジェスチャが表出されたのかを考えることで初めてその本質的な意味を捉えることが可能となる。従来研究ではこの点がほとんど考慮されておらず、研究代表者らの研究例も含めて、表出されるジェスチャや表情のカテゴリ間の共起関係の考慮といった表層的なレベルにとどまっていた。

２．研究の目的

本研究課題では、映像にもとづく人物行動センシングについて、固定視点・自己視点・他者視点という異なる視点映像の統合的解析を他に先駆けて実現することにより、

- 固定視点映像とウェアラブルカメラ映像（自己視点映像と他者視点映像）からの人物行動センシングには、それぞれ行動粒度の限界と観測継続性の限界、視野範囲の限界という問題が存在する。
- 自己視点映像と他者視点映像を用いたインタラクション中の人物行動センシングにおいても視野範囲の限界の問題が存在する。
- 人同士のインタラクションにおいて、ジェスチャや表情などは単体としてではなく、相手とのやり取りの中でその本質的な意味を捉えることが出来るが、従来の人物行動センシングではこの点が考慮されていない。

という課題の解決を目指した。

３．研究の方法

自己視点・他者視点・固定視点の異なる視点からの映像を統合解析による人物行動センシングを実現するためには、まずそれぞれの視点映像について人物行動センシングに適した特徴表現を獲得する必要がある。さらに、異なる視点の映像同士が時間的・空間的にアラインメント（すなわち、2 つの映像の時間同期がとれており、かつカメラの幾何的な位置関係が分かっている）されていることも重要となる。その上で、自己視点映像と固定視点映像からの人物行動センシングと、自己視点映像と他者視点映像からのインタラクション解析を実現することができるようになる。これを踏まえて、本研究課題では、研究項目１．自己教師有り学習による複数視点映像

の総合解析に適した特量表現学習、研究項目２．異なる視点の映像間でのアラインメント、研究項目３．自己視点映像と固定視点映像の統合による人物行動センシング、研究項目４．自己視点映像と他者視点映像の統合によるインタラクション解析、の４項目について研究に取り組んだ。

４．研究成果

本研究課題で得られた研究成果のうち主なものについて述べる。まず、研究項目１．自己教師有り学習による複数視点映像の総合解析に適した特徴表現学習について、非局所的な関連性を捉える Non-local Neural Network に対照学習を組み合わせた Cross-View Non-Local Network を考案し、ウェアラブルカメラで撮影された一人称視点映像と外部固定視点カメラで撮影された三人称視点映像のペアから、異なる視点間の相補的関係を捉えた特徴量の学習を実現した（図１）。

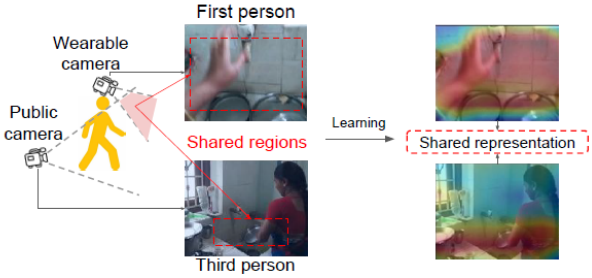

図１　一人称視点映像と三人視点映像のペアからの特徴学習

また、この手法で得られた特徴量が、研究項目２の異なる視点の映像間でのアラインメントに関して、時間的アラインメントに有効に機能することを確認することができた[Zhu+, MIRU 2021]。

研究項目３の自己視点映像と固定視点映像の統合による人物行動センシングについて、Stacked Temporal Attention と呼ばれる動作認識手法を新たに考案し、既存手法を越える高い認識精度を達成した[Yang+, BMVC 2021]。ここでは、映像からの動作認識タスクを考える場合、対象動作とは直接関係しない外乱となるフレームが認識精度の低下を招くという課題に着目し、映像全体のグローバルな特徴を考慮した時間アテンション機構を用いることで、外乱となるフレームが存在する場合においても対象動作を精度良く認識することを可能としている。

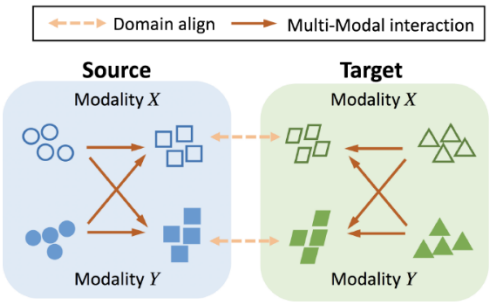

図２異なるモダリティ特徴間のインタラクションによる教師無しドメイン適応

また、少量データからの動作認識モデルの学習にも取り組み、Transformer をベースとした Compound Prototype Matching という手法を提案した。この手法では、少数の学習サンプルから動作クラスの全体特徴と局所特徴を抽出し、クエリサンプルとのマッチングをとることで、複数の動作認識タスクベンチマークデータセットを用いた評価実験において SOTA 手法を越える性能を達成している[Huang+, ECCV 2022]。さらに、一人称視点映像と三人称視点映像などのように大きく撮影条件が異なる場合、片方のドメインのデータで学習された動作認識モデルが他方のドメインでは大幅に性能が低下してしまうというドメインギャップの課題に対して、RGB 画像、オプティカルフロー、音などの異なるモダリティ特徴間のインタラクション（図２）に基づく動作認識モデルの教師無しドメイン適応手法を開発した。一人称視点映像と三人称視点映像の両方のベンチマークデータセットを用いた評価実験により、提案手法が他の教師無しドメイン適応手法を越える性能を持つことが示されている[Yang+, CVPR 2022]。

動作分割タスクに関して、Graph Neural Network を用いて前後の動作との関係に基づき動作識別・動作区間の推定をリファインする手法を提案し、動作クラス推定と動作区間推定ともに、既存の動作分割手法の精度を大幅に越える性能を実現することができた[Huang+, CVPR 2020]（図３）。

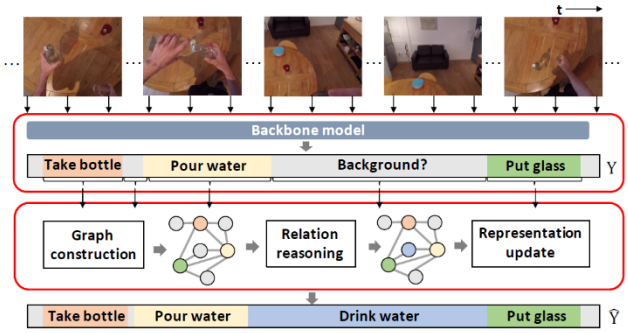

図３　前後動作間の関係に基づく動作分割の高精度化

さらに、映像に基づく詳細な人物行動理解では、手による物体操作の理解が重要となることを踏まえ、Hand-Object Interaction の解析に関する研究に取り組んだ。その結果、手と操作物体のコンタクト状態推定[Yagi+, BMVC 2021]、手物体操作における Fine-grained なアフォーダンスのモデリング[Yu+, WACV 2023]、手領域抽出と手姿勢推定[Ohkawa+, ECCV 2022][Ohkawa+, IJCV 2023]などで主要な成果を得た。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Takehiko Ohkawa, Ryosuke Furuta and Yoichi Sato	４．巻 131
２．論文標題 Efficient Annotation and Learning for 3D Hand Pose Estimation: A Survey	５．発行年 2023年
３．雑誌名 International Journal of Computer Vision	６．最初と最後の頁 3193-3206
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11263-023-01856-0	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kaipeng Zhang and Yoichi Sato	４．巻 26
２．論文標題 Semantic Image Segmentation by Dynamic Discriminative Prototypes	５．発行年 2023年
３．雑誌名 IEEE Transactions on Multimedia	６．最初と最後の頁 737-749
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TMM.2023.3270637	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato	４．巻 32
２．論文標題 Spatio-Temporal Perturbations for Video Attribution	５．発行年 2022年
３．雑誌名 IEEE Transactions on Circuits and Systems for Video Technology	６．最初と最後の頁 2043-2056
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TCSVT.2021.3081761	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計22件（うち招待講演　1件／うち国際学会　15件）

１．発表者名 八木 拓真，大橋 実咲，黄 逸飛，古田 諒佑，足達 俊吾，光山 統泰，佐藤 洋一
２．発表標題 FineBio：密な階層アノテーションを付与したバイオ実験映像データセット
３．学会等名 日本分子生物学会年会
４．発表年 2023年

１．発表者名
八木 拓真，大橋 実咲，黄 逸飛，古田 諒佑，足達 俊吾，光山 統泰，佐藤 洋一

２．発表標題
FineBio: A Fine-Grained Video Dataset of Biological Experiments with Hierarchical Annotation

３．学会等名
日本バイオインフォマティクス学会年会

４．発表年
2023年

１．発表者名
佐藤禎哉、高木基宏、古田諒佑、菅野裕介、佐藤洋一

２．発表標題
一人称視点映像を対象としたfew-shotアクティビティ認識

３．学会等名
画像の認識・理解シンポジウム

４．発表年
2023年

１．発表者名
舘野将寿、八木拓真、古田諒佑、佐藤洋一

２．発表標題
大規模言語モデルを用いた学習カテゴリの自動決定による映像からのオープン語彙物体状態認識

３．学会等名
画像の認識・理解シンポジウム

４．発表年
2023年

１．発表者名
Yuan Yin, Yifei Huang, Ryosuke Furuta, and Yoichi Sato

２．発表標題
Proposal-based Temporal Action Localization with Point-level Supervision

３．学会等名
British Machine Vision Conference（国際学会）

４．発表年
2023年

１．発表者名
Takuma Yagi, Misaki Ohashi, Yifei Huang, Ryosuke Furuta, Shungo Adachi, Toutai Mitsuyama, and Yoichi Sato

２．発表標題
FineBio: A Fine-Grained Video Dataset of Biological Experiments with Hierarchical Annotations

３．学会等名
Joint International 3rd Ego4D and 11th EPIC Workshop（国際学会）

４．発表年
2023年

１．発表者名
Lijin Yang, Quang Kong, Hsuan-Kung Yang, Wadim Kehl, Yoichi Sato, and Norimasa Kobori

２．発表標題
DeCo : Decomposition and Reconstruction for Compositional Temporal Grounding via Coarse-to-Fine Contrastive Ranking

３．学会等名
IEEE/CVF Conference on Computer Vision and Pattern Recognition（国際学会）

４．発表年
2023年

１．発表者名
Yifei Huang, Lijin Yang, and Yoichi Sato

２．発表標題
Supervised Temporal Sentence Grounding with Uncertainty-Guided Self-training

３．学会等名
IEEE/CVF Conference on Computer Vision and Pattern Recognition（国際学会）

４．発表年
2023年

１．発表者名
Yifei Huang, Lijin Yang, and Yoichi Sato

２．発表標題
Compound Prototype Matching for Few-shot Action Recognition

３．学会等名
European Conference on Computer Vision (ECCV 2022)（国際学会）

４．発表年
2022年

１．発表者名
Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato

２．発表標題
Interact before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Action Recognition

３．学会等名
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022) (国際学会)

４．発表年
2022年

１．発表者名
Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Gotsu, and Yoichi Sato

２．発表標題
Fine-grained Affordance Annotation for Egocentric Hand-Object Interaction Videos

３．学会等名
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023) (国際学会)

４．発表年
2022年

１．発表者名
Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Ryosuke Furuta, Kris Kitani, and Yoichi Sato

２．発表標題
Domain Adaptive Hand Keypoint and Pixel Localization in the Wild

３．学会等名
European Conference on Computer Vision (ECCV 2022) (国際学会)

４．発表年
2022年

１．発表者名
Zhenqiang Li, Ling Gu, Weimin Wang, Ryosuke Nakamura, and Yoichi Sato

２．発表標題
Surgical Skill Assessment via Video Semantic Aggregation

３．学会等名
International Conference on Medical Computing and Computer Assisted Invention (MICCAI 2022) (国際学会)

４．発表年
2022年

１．発表者名
Takuma Yagi, Md Tasnimul Hasan, and Yoichi Sato

２．発表標題
Hand-Object Contact Prediction via Motion-Based Pseudo-Labeling and Guided Progressive Label Correction

３．学会等名
British Machine Vision Conference (BMVC 2021)（国際学会）

４．発表年
2021年

１．発表者名
Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato

２．発表標題
Stacked Temporal Attention: Improving First-person Action Recognition by Emphasizing Discriminative Clips

３．学会等名
British Machine Vision Conference (BMVC 2021)（国際学会）

４．発表年
2021年

１．発表者名
Kaipeng Zhang, Zhenqiang Li, Zhifeng Li, Wei Liu, and Yoichi Sato

２．発表標題
Neural Routing by Memory

３．学会等名
The 35th Conference on Neural Information Processing Systems (NeurIPS 2021)（国際学会）

４．発表年
2021年

１．発表者名
八木拓真、Md. Tasnimul Hasan、佐藤洋一

２．発表標題
誘導付き逐次ラベル訂正に基づく映像からの手・物体接触判定

３．学会等名
画像の認識・理解シンポジウム（MIRU 2021）

４．発表年
2021年

１．発表者名
Zhehao Zhu, Yusuke Sugano and Yoichi Sato

２．発表標題
Cross-view Non-local Neural Networks for Joint Representation Learning between First and Third Person Videos

３．学会等名
画像の認識・理解シンポジウム（MIRU 2021）

４．発表年
2021年

１．発表者名
Zhehao Zhu, Yusuke Sugano, and Yoichi Sato

２．発表標題
Cross-View Non-Local Neural Networks for Joint Representation Learning Between First and Third Person Videos

３．学会等名
電子情報通信学会パターン認識・メディア理解研究会

４．発表年
2021年

１．発表者名
Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato

２．発表標題
Toward Visually Explaining Video Classification Networks by Pertubation-based Method

３．学会等名
IEEE Winter Conference on Applications of Computer Vision (WACV 2021)（国際学会）

４．発表年
2021年

１．発表者名
Yifei Huang, Yusuke Sugano, and Yoichi Sato

２．発表標題
Improving Action Segmentation via Graph Based Temporal Reasoning

３．学会等名
IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020)（国際学会）

４．発表年
2020年

１．発表者名 Yoichi Sato
２．発表標題 Analyzing Human Activity and Attention from First-Person Perspectives
３．学会等名 International Workshop on Egocentric Perception, Interaction and Computing（招待講演）（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	菅野　裕介 (Sugano Yusuke) (10593585)	東京大学・生産技術研究所・准教授 (12601)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関