

先端国際共同研究推進事業
2023 年度採択
「Top 研究者/Top チームのための ASPIRE」
AI・情報分野

2023 年度～2024 年度
年次報告書（公開版）

研究課題名 人間中心のビジョン・メディア技術に関する国際共同研究ネットワークの構築

日本側研究代表者 佐藤 洋一 東京大学 生産技術研究所 教授

相手側研究代表者 ・Kris Kitani, Associate Research Professor,
The Robotics Institute, Carnegie Mellon University
・Dima Damen, Professor, Department of Computer
Science, University of Bristol
・Cees Snoek, Professor, Informatics Institute,
University of Amsterdam

研究期間 2024 年 2 月 1 日～ 2029 年 3 月 31 日

1. 研究成果の概要

① 研究構想にかかる成果

本課題では、コンピュータビジョン分野とマルチメディア分野で日本のトップ研究チームと米国、英国、オランダのトップ研究チームが緊密に連携し、人間中心のビジョン・メディア技術として、映像にもとづく人物行動理解技術の意味的深化と、映像を通して人・社会を見る技術の高度化に関する研究を進めている。

映像からの人物理解の意味的深化

研究項目 1-1「大規模言語モデルを活用した複雑な行動の理解」について、2024 年度は大規模言語モデルを用いたオープン語彙物体状態認識に関する研究に取り組み成果を得た[Tateno+, LPVL'24]。この研究成果をもとに、2024 年度後半から日本側研究チームの博士学生 1 名が相手国側研究チームである英国ブリストル大学の Prof. Dima Damen の研究グループに長期滞在し、共同研究の取り組みを開始している。また、研究項目 1-2「アフォーダンスに基づく詳細な行動の理解」について、人が手で物体を扱いながら行う行動（Hand Object Interaction, HOI）の理解のための要素技術として、HOI に関わる対象物体領域セグメンテーション[Ouyang+, ECCV'24]、ならびに映像からの 3 次元手姿勢推定[Liu+, CVPR'24]について研究に取り組み成果を得た。前者はトップ国際会議 ECCV'24 のオーラル発表論文（採択率 2.3%）に選ばれている。さらに、映像に基づく人物行動理解に関する研究として、一人称視点映像と複数モーションセンサの動きデータを用いたマルチモーダル動作認識[Zhang+, ECCV'24]、シーン内の対象物体を探す視覚探索タスク時にスキャンパス予測と探索対象推定[Nishiyasu+, Gaze WS'24](Best Poster Award 受賞)について成果を得た。また、研究項目 1-3「人物行動の表現・生成基盤モデルに基づく行動理解」について、視線推定の基盤となる特徴表現を大規模事前学習により獲得する手法に取り組んだ。

映像を通して人・社会を見る技術の高度化

研究項目 2-1「マルチメディア解析・検索・マイニングの高精度化・大規模化・高速化」について、2024 年度は、以下のような取り組みを行った。まず、追加の学習データを一度しか参照することができないオンライン継続学習に対して、知識蒸留を用いた Momentum Knowledge Distillation 法を提案し、画像認識精度を向上させることに成功した[Michel, ICML'24]。本成果はフランスの Univ Gustave Eiffel との共同研究成果である。機械学習における学習が不安定になる Double Descent と呼ばれる問題に対する理論的な解析を、KTH・牧惇人教授と共同研究を開始し、Neural Tangent Kernel 法を使った有望な結果を得た。他にも、AI が社会に受容されるための基礎技術について日本チームで検討した。

さらに、映像検索の高精度化について集中的に検討した。複数の視覚・言語事前学習モデルの利用による検索、LLM による複雑な検索クエリの単純な複数のクエリへの分解、Video LLM を用いたリランキング等の技術の組み合わせにより、世界的にも競争力の高い映像検索エンジンを構築した。その成果として、TREC 映像検索(Ad-hoc Video Search)タスクにおいて一位、また国際会議 MMM2025 併催イベントの Video Browser Showdown (ビデオブラウザ対決)においてエキスパート部門と総合部門で一位を獲得した。この他、広く使われている視覚・言語事前学習モデルである CLIP について、実は類似検索ではあまり性能が出ないことを指摘し、その解決方法を提案した。

研究項目 2-2「マルチメディアアーカイブの解析による人・社会のふるまい解析」について、放送映像アーカイブの解析について検討を進め、放送映像中の政治家について検出・追跡し、その共起関係等の出現状況を解析する総合的なフレームワークを提案した。また、放送映像アーカイブ中に出現する人物の男女比を測定し、その年単位の経時変化を可視化するとともに、男女識別機の学習データのバイアスによる性能劣化の問題の解決に取り組んだ。

また、SNS のアカウントの人気の変化の予測・分析の取り組みを予定していたが、昨今 SNS 各サービスがデータのダウンロードを厳しく規制し、AI の学習に必要なデータの取得がままならない状況となっていることから、当初予定した研究を一時的に中断し、代わりとして SNS 上の人の感情や行動の認識に焦点を当て、ホテルの評判サイトを用いた具体的な批評内容の可視化を実現した。また、観光に特化した LLM を実現し、観光地認識、仮想レビュー生成、

ルート推薦などを実現した。

② 国際ネットワーク構築・拡大に関する成果

人間中心のビジョン・メディア技術を軸として、コンピュータビジョン分野とマルチメディアの分野の壁を越えた国際共同研究ネットワークの構築を目指し、2024 年度は国際ワークショップ等のイベントの企画・実施と海外研究者を講師として招いた講演会の開催を進めた。

国際ワークショップの実施

国際ワークショップの実施について、人間中心のビジョン・メディア技術に関する国際ワークショップを東京で実施したのに加えて、若手研究者を含めた本研究課題のメンバーによる国際会議・国際ワークショップの企画・実施を積極的に進めた。具体的な実施内容と得られた成果は以下の通り。

- International Workshop on Human-Centered Vision and Media Technologies : ASPIRE チーム主催で、2024 年 4 月 12 日、東京大学生産技術研究所において開催された。相手側チーム University of Amsterdam の Assoc. Prof. Yuki M. Asano と Assist. Prof. Pascal Mettes, さらにオランダ Delft University of Technology の Assist. Prof. Xucong Zhang による招待講演と、日本側研究チームによる研究ポスター発表を行い、その後開催された懇親会と合わせ、最新研究に関する情報交換と、博士学生などの若手を含む人的ネットワーク強化で大きな成果を得ることができた。
- 日本側研究チームのメンバーが General Chair, Program Chair, Program Coordinator, Workshop Organizer などの中心的役割を果たし、合計 5 件の国際会議・ワークショップ (ICMR 2024, ETRA 2024, EgoVis 2024, HANDS 2024, MMM 2025) を開催した。これにより、人間中心のビジョン・メディア技術に関わるマルチメディアコンテンツ検索とモデリング、一人称視点映像解析、人の動作における手のセンシングと理解の分野において、国際的な人的ネットワークの強化で成果を得た。

海外研究者による講演会の開催

さらに、海外研究者による講演会を 14 回実施し、コンピュータビジョン、マルチメディア、HCI の分野における最新の研究成果に関する知識を得るとともに、人間中心のビジョン・メディア技術に関する意見交換を行った。

- 講演会講師 : Marc Pollefeys (Professor, ETH/Microsoft), Zhu Li (Professor, University of Missouri), David Lindell (Assist. Professor, University of Toronto), 他 11 名

③ 国際頭脳循環の促進に資する若手研究者の人材育成に関する成果

本研究課題では、国際頭脳循環の促進に資する若手研究者の人材育成にかかる取り組みとして、日本側研究チームの大学院学生や若手研究者を海外研究機関に長期派遣することで、コンピュータビジョン分野とマルチメディアの分野の将来を担う若手研究者の研究能力とコミュニケーション能力の向上を図るとともに、研究者としてのキャリアに長期的に資する国際的な人的ネットワークの構築を目指す。さらに、相手国側研究者の日本滞在を受け入れることによって、国際研究ネットワークの強化とともに、日本側研究チームの若手研究者や大学院学生のコミュニケーション能力等の向上を目指す。また、大学院生が国際会議に参加する機会を設けることにより、将来を担う研究者としての視野を広げるとともに、人的ネットワーク構築に資することを目指した。

若手研究者の海外長期派遣

本年度は、日本側研究チームから博士学生 5 名と修士学生 1 名を海外研究機関に長期派遣した。

- D1 学生 (東京大学生産技術研究所 佐藤洋一研)、滞在先 : 英国 University of Bristol・Prof. Dima Damen、2024/9/5～2025/8/31、大規模言語モデルを用いたオープン語彙物体状態認識に関する研究
- D3 学生 (東京大学生産技術研究所 佐藤洋一研)、ドイツ University of Stuttgart・Prof. Andreas Bulling、2024/4/3～2024/11/1、人の行動によるタスクと視線の関係理解に関する研究
- D2 学生 (東京大学生産技術研究所 菅野研)、オランダ Delft University of Technology・Assit. Prof. Xucong Zhang、2024/7/23～2024/12/11、視線推定基盤モデルの事前訓練に関する研究

- D3 学生（東京大学生産技術研究所 佐藤洋一研）、米国 Meta Reality Labs・Dr. Takaaki Shiratori、2024/6/1～2024/11/30、セルフコンタクトを含む条件下での身体・手 3 次元姿勢・形状推定に関する研究
- D3 学生（国立情報学研究所 佐藤真一研）、オランダ University of Amsterdam・Prof. Cees Snoek、2024/9/6～2025/2/20、映像解析における映像中の動作の詳細分類並びに時間解像度向上に関する研究
- M1 学生（東京大学大学院情報理工学系研究科 山崎研）、スウェーデン KTH Royal Institute of Technology・Prof. Atsuto Maki、2024/8/1～2025/1/31、機械学習の学習時に見られる Double Descent の理論的解析

海外研究者の滞在受け入れ

本年度は、相手側研究チームのメンバーに加え、関連分野の海外研究者、合計 9 名を短期・長期で受け入れ、人間中心のビジョン・メディア技術に関する共同研究の実施や情報交換を行った。

- Vincent Oria (Professor, New Jersey Institute of Technology, アメリカ)、2024/3/14～2024/3/23、国立情報学研究所 佐藤真一研、画像深層学習モデルの差分インデクシングに関する研究
- Antoine Laurent (Professor, ルマン大学、フランス)、2024/2/18～2024/3/17、国立情報学研究所 佐藤真一研、映像アーカイブ中の人物のジェンダー解析
- Pascal Metters (Assist. Professor, University of Amsterdam, オランダ)、2024/4/2～2024/4/12、国立情報学研究所 佐藤真一研、Hyperbolic Embedding の応用に関する検討
- Yuki Asano (Assoc. Professor, University of Amsterdam, オランダ)、2024/4/11～2024/4/17、国立情報学研究所 佐藤真一研、画像・映像事前学習モデルに対する外部知識の組み込み
- Xucong Zhang (Assist. Professor, TU Delft, オランダ)、2024/4/6～2024/4/14、東京大学生産技術研究所 菅野研、アピアランスベース視線推定・人体姿勢推定に関する講演・研究議論を実施
- Tillmann Ohm (PhD student, Tallinn University, エストニア)、2024/4/1～2024/4/14、東京大学生産技術研究所 菅野研、コンピュータサイエンスとアートとの融合に関する検討
- Man To Tang (MS student, CMU, アメリカ)、2024/6/1～2024/8/31、東京大学生産技術研究所 佐藤洋一研、相手側研究者の Assoc. Prof. Kris Kitani (CMU) との共同研究として、Personalized 3D Avatar と VR ダンストレーニング支援システムに関する研究を実施
- Christian Beker-Asano (Professor, Stuttgart Media University, ドイツ)、2024/5/1～2024/7/31、東京大学生産技術研究所 菅野研、視線推定技術のアンドロイド応用に関する研究議論を実施
- Frank Y. Shih (Professor, New Jersey Institute of Technology, アメリカ)、2024/9/11～2024/9/24、国立情報学研究所 佐藤真一研、画像キャプションと画像生成に関する研究

国際会議への参加

2024 年度は、大学院生を中心とした日本側研究チームのメンバーにより、33 回の国際会議参加があった。国際会議としては、CVPR 2024、ECCV 2024、ACMMM 2024、NeurIPS 2024、ICML 2024、SIGGRAPH 2024 など、コンピュータビジョン、マルチメディア、機械学習の各分野のトップ国際会議が中心となっている。

若手研究者による国際会議・ワークショップ運営等に関する顕著な実績

日本側研究チームのメンバーである菅野裕介（東京大学生産技術研究所、准教授）が、アイトラッキング分野のトップ会議の一つである ACM Symposium on Eye Tracking Research and Applications (ETRA 2024) の General Chair を務めた。これに加えて、コンピュータビジョン分野のトップ会議や主要会議である CVPR 2025、ICCV 2025、BMVC 2024 の Area Chair を務めている。また、Yifei Huang（東京大学生産技術研究所 佐藤（洋）研、特任研究員）が、活躍する若手研究者としてコンピュータビジョン分野のトップ国際会議 CVPR 2025 の Area Chair に選出された。さらに、D3 学生（東京大学生産技術研究所 佐藤（洋）研）が、大学院学生ながら ECCV 2024 併催の Observing and Understanding Hands in Action Workshop (HANDS 2024) の

Workshop Organizer を務め、同ワークショップの企画・運営において主要な役割を果たした。

2. 研究実施体制

研究テーマ	中心となる研究者氏名	所属機関・部署・役職名
研究テーマ1	佐藤洋一 菅野裕介 Dima Damen Kris Kitani	東京大学・生産技術研究所・教授 東京大学・生産技術研究所・准教授 University of Bristol, Department of Computer Science, Professor Carnegie Mellon University, The Robotics Institute, Associate Research Professor
研究テーマ2	佐藤真一 山崎俊彦 Cees Snoek	国立情報学研究所・コンテンツ科学研究系・教授 東京大学・大学院情報理工学系研究科・教授 University of Amsterdam, Informatics Institute, Professor
研究テーマ3	佐藤洋一 佐藤真一 山崎俊彦 菅野裕介 Dima Damen Kris Kitani Cees Snoek	東京大学・生産技術研究所・教授 国立情報学研究所・コンテンツ科学研究系・教授 東京大学・大学院情報理工学系研究科・教授 東京大学・生産技術研究所・准教授 University of Bristol, Department of Computer Science, Professor Carnegie Mellon University, The Robotics Institute, Associate Research Professor University of Amsterdam, Informatics Institute, Professor

3. 代表的な業績（原著論文、プレスリリース、表彰など）

- Liangyang Ouyang, Ruicong Liu, Yifei Huang, Ryosuke Furuta, and Yoichi Sato, "ActionVOS: Action as Prompts for Video Object Segmentation", in Proc. European Conference on Computer Vision (ECCV 2024), pp. 1-20, October 2024. CV 分野のトップ国際会議 ECCV 2024 におけるオーラル発表に採択（採択率 2.3%）
- Yoichi Sato, IAPR (International Association for Pattern Recognition) Fellow, December 2024.
- Maorong Wang, Nicolas Michel, Ling Xiao, Toshihiko Yamasaki, "Improving Plasticity in Online Continual Learning via Collaborative Learning," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), pp. 23460-23469, 2024.
- Bao Tran Gia, Tuong Bui Cong Khanh, Tam Le Thi Thanh, Thuyen Tran Doan, Khiem Le, Tien Do, Tien-Dung Mai, Thanh Duc Ngo, Duy-Dinh Le, and Shin'ichi Satoh, NII-UIT at VBS2025: Multimodal Video Retrieval with LLM Integration and Dynamic Temporal Search, Video Browser Showdown, MMM 2025, Best Overall VBS System, Best Expert VBS System
- Nicolas Michel, Maorong Wang, Ling Xiao, Toshihiko Yamasaki, Rethinking Momentum Knowledge Distillation in Online Continual Learning, in Proc. Proc. International Conference on Machine Learning (ICML 2024), PMLR 235:35607-35622, 2024.
- TREC 映像検索(Ad-hoc Video Search)タスクにおいて一位、また国際会議 MMM2025 併催イベントの Video Browser Showdown (ビデオブラウザ対決)においてエキスパート部門と総合部門で一位を獲得した。