

一人称視点映像と手順書に基づく作業支援システムの構築

亀甲 博貴（京都大学）

概要

手順書によって指示された作業を実施するにあたり、作業書の理解不足などが原因で作業を間違えることが多々ある。手順書の実施誤りは科学実験における再現性にも大きく影響し、誤りを回避あるいは検出する作業支援システムの需要は高い。本研究では、Vision&Language 技術を用いて、一人称視点映像を活用した作業支援システムを構築することを目的とする。構築するシステムは、作業者が着用する AR デバイス上に手順書と正しい作業手順の一人称視点映像を提示し、また音声入力による映像のハンズフリー操作機能を実装することで、より没入感のある作業支援を実現する。支援システムを用いた評価実験により、提案システムが作業の正確性向上や誤り防止に貢献することが確認された。また、支援システムによる誤り検出を行うモデルを構築するため、作業誤りを含む作業実施映像のデータセット構築を行った。加えて、手順書テキストを活用した誤り検出手法を提案し、構築したデータセットによる評価実験により、作業書テキストが誤り検出に有効であることを確認した。

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

- 京都大学 学術情報メディアセンター
- mdx

1.2 課題分野

- データ科学・データ利活用課題分野

1.3 参加研究者の役割分担

■亀甲 博貴 課題代表者。研究の計画と進捗の管理。

■森 信介 課題副代表者。本研究プロジェクトの全体統括。

■梶村 恵矢 AR デバイスを用いた作業支援システムの開発と評価。

■羽路 悠斗 作業映像の収録と整備。映像と手順書の対応関係および作業誤りを検出するモデルの開発と評価。

2 研究の目的と意義

手順書によって提示された作業を実施するにあたり、作業者の理解不足などが原因で作業を間違えることが多々ある。例えば科学実験においては、Nature による 1,500 人以上の研究者を対象とした調査によると、70% 以上の研究者が他の研究者の実験を再現することに失敗した経験があるとしている [1]。手順書に基づく作業は、科学実験や製造業、あるいは日々の生活においても料理などを筆頭に多くの場面で行われており、これらの作業を支援するシステムが構築できれば、その適用範囲は広いと考えられ

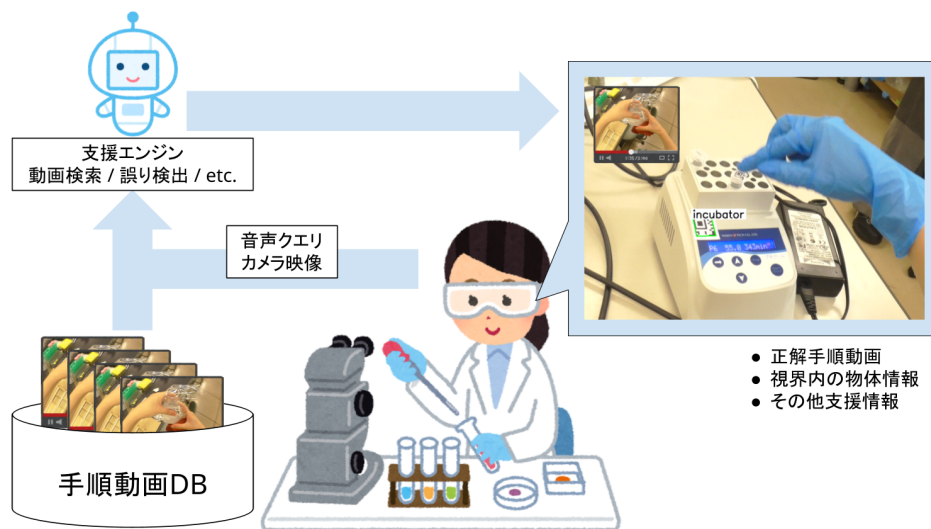


図1 提案システムの全体像。本年度はこのうち映像提示システムと誤り検出手法の構築に取り組んだ。

る。本研究では、作業映像の検索や提示、作業誤りの検出などの機能を持つ作業支援システムを構築することを目的とする。

図1に、本研究で提案する作業システムの概要を示す。提案システムは、手順動画のデータベースに含まれる作業映像と作業手順テキストを作業者に提示することで、作業者がその作業内容を理解しやすくすることが主機能である。また、作業者の一人称視点映像を収録し、これを管理したり映像中から作業誤りを検出したりすることで、作業者支援や作業記録の作成を行う。本年度はこのうち映像提示システムと作業誤り検出手法の構築に取り組んだ。

3 当拠点公募型研究として実施した意義

本研究は科学実験の再現性向上への貢献を念頭に置いたものである。実験の再現性向上は特に自然科学分野の全体において重要なトピックであり、本研究の実施によって期待できる成果は公益に資するものであると考えている。

本研究で提案するシステムは、最終的には

各々の研究者が実施する作業映像を管理し、これらに高性能な機械学習モデルを適用することを想定している。そのようなデータ管理システムを構築することは、情報科学分野を除く多くの研究者にとって困難または高い作業コストがかかることが想定される。そのため、このようなシステムをクラウドサービスとして提供する意義は大きい。一方で、このようなデータのうち特に未発表のデータは、その研究内容が公知となるまでは機微情報として扱う必要があると考えられる。そのようなサービスを信頼性の高い学術機関運用の計算資源上で提供する意義は大きいと考えており、本研究はmdxがそのようなサービスを提供することの可能性を示すものである。

4 前年度までに得られた研究成果の概要

該当なし

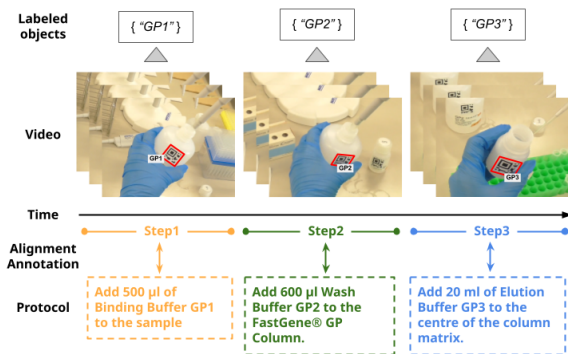


図2 BioVL-QR データの例.

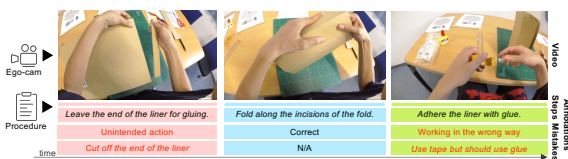


図3 EgoOops データの例.

5 今年度の研究成果の詳細

5.1 一人称視点の作業実施映像データセットの構築

本研究を実施するにあたり、2種類の作業実施映像データセットを構築した。

5.1.1 BioVL-QR データセット

BioVL-QR データセット [2] は、生化学実験において、作業プロトコルに従って実施された作業の一人称視点映像を収録したデータセットである。本データセットの特徴として、作業に用いる器具などに micro QR コードを貼り付けた状態で映像を収録している。特に生化学実験においては、無色透明な液体な似た形状の容器に入った試薬などを扱うことが多く、これを映像情報のみで識別することは困難である。しかしながら用いる試薬の誤りは実験の結果に大きな影響を与えるため、誤り検出システムはこれらの識別を行う必要がある。このような映像を支援システムで扱うために、映像の特徴量とは別に QR コードによる識別情報を活用する

ことで、作業支援システムの精度向上を図る。

データの例を図 2 に示す。構築したデータセットは Web 上で公開している^{*1}。

5.1.2 EgoOops データセット

EgoOops データセット [3] は、5種類の作業手順書を準備し、これに従って実施された作業の一人称視点映像を収録したデータセットである。BioVL-QR データセットと同様に、作業に用いる器具などに QR コードを貼り付けた状態で映像を収録している。本データセットの特徴として、一部の作業において、作業の誤りを指示することで、作業誤りを含む映像を多く含むことが挙げられる。誤り検出や異常値検出といった課題の研究はそのような検出すべきデータが少ないことが研究を困難にしているが、本データセットは擬似的に誤りデータを多く含むデータセットであるため、このような研究を行う上で有用であると考えられる。

データの例を図 3 に示す。構築したデータセットは Web 上で公開している^{*2}。

5.2 作業支援システムの構築

5.2.1 システム概要

本研究では、作業手順テキストとそれに紐付いた一人称視点映像を提示する作業支援システムを構築した。図 4 に、構築した作業支援システムの UI を示す。構築したシステムは Web ベースシステムと AR デバイスベースシステムの2種類である。それぞれ提示する情報は同じであり、手順テキストとそれに従って実施された一人称視点映像が表示される。各手順テキストに対応する映像中の区間は事前に対応付けられており、手順テキストを指定することで映像のスキップや検索機能を提供している。Web ベースシステムは PC のブラウザ上で表示され

^{*1} <https://nishii10mo.github.io/BioVL-QR/>

^{*2} <https://y-haneji.github.io/EgoOops-project-page/>

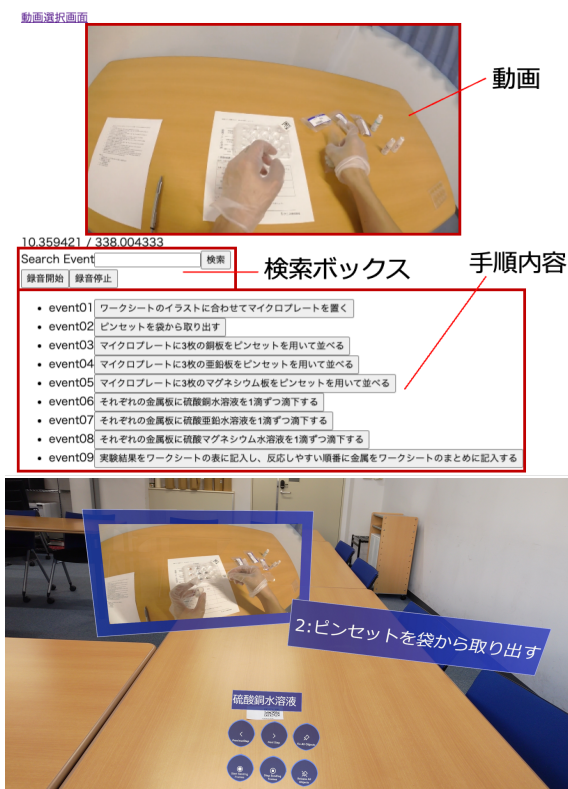


図4 構築した作業支援システムの画面。上: Web ベースシステム。下: AR デバイススペースシステム。

一方で、AR デバイススペースシステムは AR デバイスである HoloLens2 上で動作する。

AR デバイススペースシステムを構築するにあたり、mdx に構築した手順動画データベースと HoloLens2 の間で通信を行うことによって映像を提示するシステムを構築した。これにより、映像および手順テキストは mdx のストレージに保存され、複数の AR デバイスから同時にアクセスできるようになっている。一方で、作業中の映像を mdx に保存する機能は実装できておらず、HoloLens2 のローカルストレージに保存される状態である。

5.2.2 評価実験

構築したシステムの有効性を評価するため、システムの有無による作業への影響を調査し

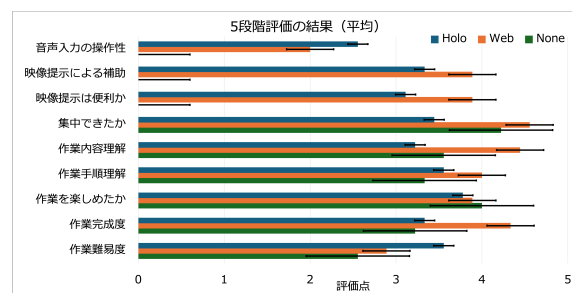


図5 5段階評価。

た。実験は EgoOops データセット中に含まれる作業のうち金属イオン反応を確認する作業・電子回路の工作・段ボールの考察の3種類の作業を対象とし、システムが提示する映像は EgoOops データセットに含まれる映像のうち誤りのないものを用いた。実験には9人の被験者が参加し、それぞれの被験者は3種類の作業を実施した。被験者はある作業についてシステムがない状態で、別の作業について Web ベースシステムを用いて、最後の作業について AR デバイススペースシステムを用いて実施した。

評価として被験者の5段階の主観評価、NASA-TLX [4] による作業負荷の評価、作業誤りの目視計測を行った。主観評価は作業難易度・作業完成度・作業を楽しめたか・作業手順の理解・作業内容の理解・集中して取り組めたか・映像の提示は便利か・映像の提示は作業理解の補助となるか・音声認識の操作性の各項目について被験者が評価を行った。NASA-TLX は作業に対する主観的な作業負荷を計測する指標である。計測項目は知的・知覚的要求 (MD)・身体的要求 (PD)・タイムプレッシャー (TD)・作業成績 (OP)・努力 (EF)・フラストレーション (FR) の6項目である。被験者はこれらの項目に対して100点満点で点数をつける。点数は高いほど作業負荷が大きいことを示す。

これに加え、作業誤りを表1に示すように4種類に分類し、これを目視計測した。

誤り分類	説明
手戻り	他の誤りを回復するために 以前の手順に戻って再度手 順を実行するもの
手順誤り	ある手順以前の全ての手順 を完了する前にその手順を 実行するもの
実行ミス (軽微)	ある手順中の動作に関する ミスで、即時手順を再開可 能なもの
実行ミス (重大)	ある手順中の動作に関する ミスで、作業の回復が不可 能な致命的なもの

表 1 作業誤りの分類.

図 5 に、5 段階評価の結果を示す．結果から、映像提示が作業の理解に効果的であることが示唆される．Web ベースシステムと AR デバイスベースシステムを比較すると、多くの項目で Web ベースシステムの方が高い評価を得ている．本実験の被験者は全員が AR デバイスの利用経験がなく、AR デバイスより PC の方がインタフェースとして慣れているのが要因であると考えられる．

図 6 に、各作業の NASA-TLX のスコアを示す．金属イオンの反応実験において、HoloLens2 を用いた場合の負荷が高い結果となった．この作業はイオン反応による色の変化を観察するものであるが、被験者からの聞き取りによると、HoloLens2 のバイザーが色付きであるため、色の変化が分かりにくいという意見があった．導入する分野によっては、デバイスの選定が重要であることが示唆される．

表 2 に、作業誤りの数を示す．結果から、支援システムは特に各手順の実行ミスを減少させる効果があることが示唆される．

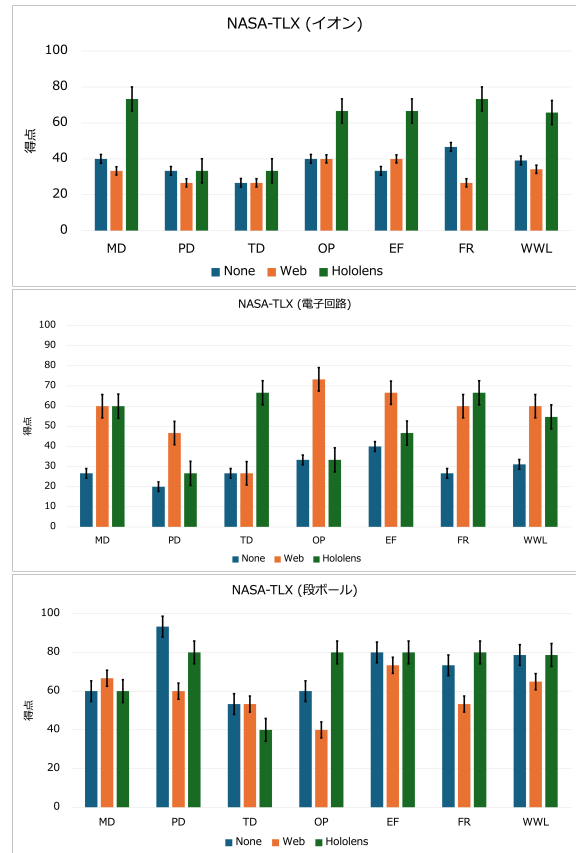


図 6 各作業における NASA-TLX のスコア．上から順番に金属イオン，電子回路，段ボールの作業を示す．

5.3 誤り検出手法の構築

本研究では、手順書テキストと映像が与えられたときに、映像中の作業誤りを検出する課題に取り組む．本研究では、誤り検出を

- 映像中の、各手順書テキストに対応する区間の推定 (アライメント)
- 各区間が正しいか、誤りを含むか、誤りを訂正する作業を行っているかの 3 値分類

の 2 段階に分けて行う．

5.3.1 手法概要

本研究では、アライメント手法として StepFormer++ を提案する．手法の概要図を図 7 に示す．StepFormer++ は自己教師あり学習

誤り種別	金属イオン			電子回路			段ボール		
	AR	Web	None	AR	Web	None	AR	Web	None
手戻り	1	0	2	3	0	5	7	2	1
手順誤り	4	0	3	2	1	9	11	6	4
実行ミス (軽微)	0	2	6	3	4	2	2	2	10
実行ミス (重大)	0	0	0	0	0	0	1	0	1
誤り総数	5	2	11	8	5	16	21	10	16

表 2 誤りの数. AR: AR デバイスペースシステムを用いた場合. Web: Web ベースシステムを用いた場合. None: 支援システムを用いなかった場合.

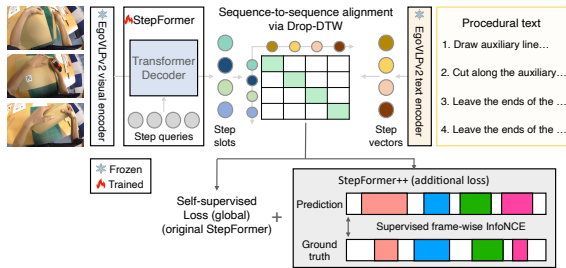


図 7 StepFormer++ の概要.

手法	mAP@tIoU			
	0.1	0.2	0.3	平均
ActionFormer	1.8	0.1	0.1	0.7
提案手法	2.6	2.5	2.5	2.5
GT steps (oracle)	34.7			

表 3 作業誤り検出の結果. “GT steps (oracle)” は, 映像とテキストのアライメントを予測結果ではなく正解を用いて行った場合の上限値.

によって構築される映像とテキストのアライメント手法である StepFormer [5] をベースに, 教師あり学習のための損失を追加したものである. また, 誤りクラス分類は事前学習済みモデルによる特徴量抽出と多層パーセプトロンによる分類器の学習を行う.

表 3 に, 手法の性能を示す. ベースライン手法として, ActionFormer [6] を用いた手法と

比較した. ベースライン手法に比べて提案手法が高い性能を示しており, これは手順書を参照することがより高精度な作業誤り検出に寄与することが示唆される. 一方で, 正しい手順区間を用いた場合の性能と比較すると大きな差があり, これは映像と手順書テキストのアライメント精度が誤り検出精度に大きく影響しており, アライメント精度を向上させることが誤り検出精度の向上に必要であることを示している.

6 今年度の進捗状況と今後の展望

申請時に設定した各課題について, その達成状況および今後の展望を以下に示す.

■映像の収録およびアノテーション BioVL-QR データセットと EgoOops データセットの 2 種類のデータセットを構築し, 整備・公開した.

■映像と手順書の対応関係の自動獲得および作業誤りの検出 映像と手順書のアライメント手法および作業誤り分類手法を提案し, 構築したデータセットに対して評価実験を行った. ベースライン手法に比べると提案手法は高性能であったが, 実用的な精度実現のためにはさらなる精度向上が必要である.

■作業支援システムの構築 一人称映像の提示・検索機能を持つ支援システムを構築した.

誤り検出手法などの機能は実装していないため、実用的な誤り検出手法が構築できたら支援システムへの導入を予定している。

■実サービスへ向けたシステム全体の検証 構築したシステムが作業に与える影響について評価実験を行った。誤り検出など他の機能が実装でき次第、それらの性能を評価する予定である。

参考文献

- [1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, Vol. 533, No. 7604, pp. 452–454, May 2016.
- [2] Tomohiro Nishimoto, Taichi Nishimura, Koki Yamamoto, Keisuke Shirai, Hirotaka Kameko, Yuto Haneji, Tomoya Yoshida, Keiya Kajimura, Taiyu Cui, Chihiro Nishiwaki, Eriko Daikoku, Nat-suko Okuda, Fumihito Ono, and Shinsuke Mori. Biovl-qr: Egocentric biochemical vision-and-language dataset using micro qr codes, 2025.
- [3] Yuto Haneji, Taichi Nishimura, Hirotaka Kameko, Keisuke Shirai, Tomoya Yoshida, Keiya Kajimura, Koki Yamamoto, Taiyu Cui, Tomohiro Nishimoto, and Shinsuke Mori. Egooops: A dataset for mistake action detection from egocentric videos referring to procedural texts, 2025.
- [4] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, Vol. 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988.
- [5] Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G. Derpanis, Richard P. Wildes, and Allan D. Jepson. Stepformer: Self-supervised step discovery and localization in instructional videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18952–18961, 2023.
- [6] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, Vol. 13664 of *LNCS*, pp. 492–510, 2022.