一人称視点映像を用いた マルチモーダル作業支援システム

梶村恵矢 [1]　西村太一 [2]　羽路悠斗 [1]　山本航輝 [1]　崔泰毓 [1]
亀甲博貴 [3]　森信介 [3]
[1] 京都大学大学院 情報学研究科 [2] LINE ヤフー株式会社
[3] 京都大学 学術情報メディアセンター
[1]{kajimura.keiya.48x,haneji.yuto.58c}@st.kyoto-u.ac.jp
[1]{yamamoto.koki.76n,cui.taiyu.33c}@st.kyoto-u.ac.jp　[2]tainishi@lycorp.co.jp
[3]{kameko,forest}@i.kyoto-u.ac.jp

概要

実験や料理など，作業者が手順書に従って作業を行う状況において，不安のある手順や不明瞭な手順を映像として確認できることは作業の再現性向上に有効に働くと考えられる．また，映像中の作業者の視線情報や手元の細かな動作を確認できる点において，実際の作業者が確認する映像として一人称視点動画を用いることはメリットがある．本研究では広く人手で行う作業の再現性向上を目的とし，テキストと音声による入力機能を持った一人称視点動画を用いたマルチモーダル作業支援システムを提案する．また，実験では作業負荷や作業完成度にシステムが与える影響を調べ，システムが負荷の軽減や完成度の向上に寄与するかを考察する．

1　はじめに

作業は連続した複数の手順で構成され，それらを自然言語で記述したものを手順書と呼ぶ．人間が行う作業には科学実験や工作，組み立て，料理などが挙げられる．これらの人手作業において，作業の再現性は記述された手順書に従って各手順を正確に実行することによって担保される．しかし，手順書を定めた場合でも再現性が保たれないことがある．Baker によると 5 割を超える科学者が自身の行った実験の再現に失敗したことがあり，7 割を超える科学者が他者が設定した実験の再現に失敗したことがあると報告されている [1].

人手作業において手順書を定めているにもかかわらず，作業の再現に失敗してしまう原因の 1 つとして，誤った動作を行うことや物体を取り違えることなどのヒューマンエラーが考えられる．手順書は作業に習熟している者にとっては作業内容を明瞭に示したものとなる．一方で，作業に十分に習熟していない者にとっては，見慣れない動作や物体がその中に出てくると，時に作業内容が不明瞭になってしまい，結果としてヒューマンエラーを引き起こしてしまう可能性がある．

手順内容を文字情報だけでなく，動作や物体の視覚的情報を含む映像で確認することができれば，不慣れな作業に対しても作業者は作業内容に対する理解度を高めることができ，作業中の誤りを減らすことが可能になると考えられる．また，映像として一人称視点の作業映像を扱うメリットとして手元の細かな作業や視線情報などを確認することが可能になることが挙げられる．三人称視点による作業映像は作業の全体の状況を客観的に捉えることが可能である．対して，一人称視点の作業映像では映像中の作業者が行っている，より具体的（あるいは局所的）な動作について上述したような情報を通じて知ることが可能であり，さらに，それらの情報から映像中の作業者の意図を推測することが可能となる．これらのことから，作業の誤りを減らすことを目的とすれば，各動作，各手順の詳細が分かりやすい一人称視点の映像が三人称視点の映像と比べて有効であると考えられる．

以上を踏まえて，本稿では人手作業の誤りを減らし，再現性を高めることを目的としたマルチモーダル作業支援システムを提案する．また，2 種類の作業を 6 名の参加者にシステム有無で条件を変えた実験によってシステムが作業の実行時間，負荷，完成度にどのような影響を与えるかを検証する．

2　関連研究

2.1　一人称視点映像のデータセット

近年，様々な研究グループによって一人称視点映像のデータセットが公開されている．Grauman らによって公開されている Ego4D はその中でも最大規模のデータセットである [2]．このデータセットには世界 9 カ国の参加者の日常活動とそれらに付随するアノテーションが収録されており，一人称視点の視覚的認知課題の究明に大きく寄与している．

西村らによって公開されている BioVL2 は生化学実験に特化した数少ない一人称視点映像データセットである [3]．このデータセットは生化学実験を一人称視点で撮影した動画と実験の手順書が紐付けられている．また，映像から手順書を生成する課題にも取り組んでいる．

2.2　作業支援システム

Chang らは How to Video に対してイベント検索を行うコマンドと時間的に映像を操作するコマンドが実装された作業支援システムを提案している．また，実験により，How to Video を参照して実行するタスクをより低いストレスで実行できることを示している [4]．彼らの研究と我々の研究の差分は達成すべき目的にある．彼らがよりストレスフリーなシステムを目的とする一方で，我々の目的はシステムを介して作業の誤りを減らし，その再現性を高めることである．Scholl らはスマートグラス上に手順内容がテキストとして表示され，音声認識により手順内容を切り替えることができるハンズフリーな実験支援システムを提案している [5]．このシステムを用いた実験の参加者全員がスマートグラス上に表示される情報だけで化学実験を成功させることができたことを報告している．彼らの提案するシステムと本稿で提案するシステムの最も大きな差異は利用者に提示する情報のモダリティである．本稿で提案するシステムはテキストだけでなく，実際の作業映像を提示する．

3　提案手法

3.1　システム概要

システムは Web アプリケーションとして実装した．ページ内には作業動画，手順内容，検索ボックスが含まれている (図 1)．システム利用者は検索ボックスにキーワードを入力するか，音声によってそれを伝えることでイベント検索を行う．また，各手順内容に対応するスキップボタンを用いることでも該当イベントの参照が可能である．本稿の実験で扱う作業を用いて上記の処理の流れを図 2 に例示する．まず，最下段で入力を受け取る．次に，中断で入力されたテキストと各手順のコサイン類似度を計算する．最後に最も類似度が高い手順内容が動画で再生される．この例では手順 2 が該当するので，手順 2 が始まる箇所から動画が再生される．

人手作業では作業者の両手が塞がっていて手を使ったシステムの操作が難しい場合が多々考えられる．提案システムは一人称視点の作業映像に対して，イベント検索を行うことで手順内容を映像で確認できるものであり，両手が塞がっている状況においても音声入力によってハンズフリーで操作できるものである．

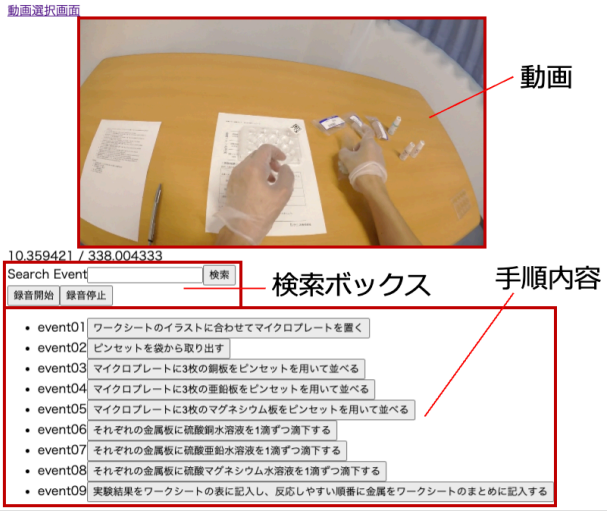

図 1　利用者が操作するシステムの画面

図 2　システムの処理の流れ

3.2 キーワードによるイベント検索

システムは音声あるいはテキストを入力として受け取る．入力が音声の場合は Whisper [6] を用いた音声認識によってテキスト化する．こうして受け取った入力テキストと各手順内容のそれぞれに対して各単語ベクトルの平均値を計算し，それらのコサイン類似度を比較することでキーワードによるイベント検索が実現される．コサイン類似度の計算には spaCy [7] を用いた．入力されたテキストは GiNZA [8] を用いて形態素解析され，トークン化される．その後，各トークンには学習済みの単語ベクトルが付与される．このようにして付与された単語ベクトルの平均値を $\boldsymbol{t}$ とし，検索対象となる n 個の手順内容を同様に処理したものをそれぞれ $\boldsymbol{p_1}$，$\boldsymbol{p_2}$，…，$\boldsymbol{p_n}$ とする．検索結果となる手順 $\hat{\boldsymbol{p}}$ は以下の式で決定される．

$$\hat{\boldsymbol{p}} = \underset{\boldsymbol{p_i}}{\operatorname{argmax}} \frac{\langle \boldsymbol{t},\ \boldsymbol{p_i} \rangle}{\|\boldsymbol{t}\| \|\boldsymbol{p_i}\|}$$

ここで $\langle \boldsymbol{t},\ \boldsymbol{p_i} \rangle$ は $\boldsymbol{t}$ と $\boldsymbol{p_i}$ の内積であり，$\|\boldsymbol{t}\|$ は $\boldsymbol{t}$ のノルムである．

4 実験

提案システムが作業の実行時間，負荷を減少させることができるか，完成度を向上させることができるかを測定するために，被験者実験を行った．実験では研究室内で作成したデータセットを用いた (付録 A)．このデータセットに含まれる 5 種類の作業のうち比較的手順数が多く複雑さの高い以下の 2 種の作業を用いた，

作業 1 :金属のイオン化傾向を比較する化学実験
作業 2 :簡易的な電子回路を組む作業

作業 1 は中学生向けの化学実験キットを用いた作業であり，作業 2 は対象年齢 6 歳以上の知育玩具を用いた作業である．参加者はこの 2 種類の作業を実行した．

4.1 実験参加者

研究室内の学生を対象として実験協力者を募集し，20 代男性 6 名が実験に参加した．協力者を募集する上で，実験で行う作業内容を事前に知らないことを条件とし，その他の条件は設けなかった．参加者を作業 1，2 に対して次のような各 3 名ずつのグループに無作為に分けて実験を行った．この振り分けにより，参加者は作業 1，2 のいずれかで必ず提案システムを操作した．

A :作業 1 でシステムあり，作業 2 でシステムなし
B :作業 2 でシステムあり，作業 1 でシステムなし

4.2 実験手順

それぞれの作業に対して，まず，参加者に対して作業概要の説明と手順書中に現れる物体名の確認を行った．この時被験者がシステムを用いながら作業を行う場合は，システムの機能と基本的な操作方法についても説明を行った．次に，参加者に対して作業を実行する “速さよりも正確性を意識して作業を行う” ように教示を与え，参加者は事前に準備された手順書に従って各作業を遂行した．作業はこちらから指示したタイミングで開始し，参加者が全ての作業手順の完了を申告した時点で終了と判断した．

4.3 評価項目

評価項目は実行時間，NASA-TLX [9]，複数項目に対する 5 段階評価，自由回答によるフィードバックである．実行時間は作業開始時点から作業終了時点までの経過秒数を計測する．

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．計測項目には知的・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 つである．作業者はこれらの 6 つの項目に対して点数をつけ，${}_6C_2$ 通りで各項目同士を比較する．各項目の点数と項目同士の比較によって決定した重みを用いた平均（WWL，付録 B）とを用いて作業負荷を計測する．各項目の点数は高ければ高いほど作業負荷が大きいことを示す．5 段階評価によるフィードバックでは作業難易度，作業完成度を評価した．グループ A については，これに加えてシステムの利便性，システムの操作性も評価した．自由記述によるフィードバックによって上述した評価項目で取りきれないフィードバックを取った．

5 結果と考察

表 1 に各参加者の実行時間を示す．この結果から作業 1 はシステムを用いた場合は実行時間が長くなっていることがわかる．作業 2 についてはシステムの有無による実行時間の変化は見られない．作業 1 についてはシステムを利用する時間が実行時間に

表 1　各作業の実行時間.
A：作業 1 でシステムあり，B：作業 2 でシステムあり

参加者 No.	グループ	作業 1 実行時間	作業 2 実行時間
P1	A	8:06	3:18
P2	A	5:04	4:13
P3	A	12:21	6:45
P4	B	6:23	3:15
P5	B	5:44	4:46
P6	B	5:49	3:42

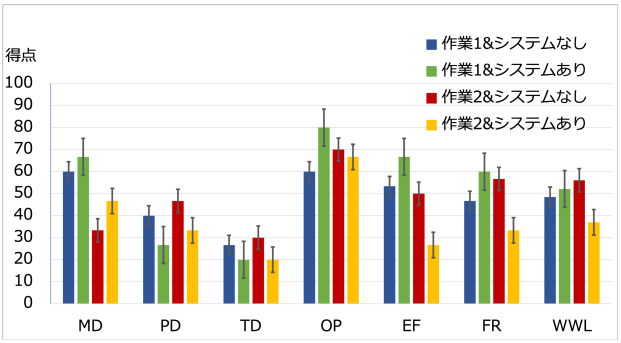

図 3　各条件における NASA-TLX の平均値

影響した結果，実行時間が長くなっていることがわかる． 図 3 に各条件における参加者の NASA-TLX の平均値を示す．各作業について，システム有・無で結果を比較すると作業 1 ではシステムありの場合に 6 項目中 4 項目の得点がシステムなしの場合と比べ高くなっていることがわかる．WWL はシステムありの場合におよそ 5 点高くなっていることがわかる．作業 2 ではシステムありの場合に 6 項目中 5 項目の得点がシステムなしの場合と比べ低くなっていることがわかり，WWL はシステムありの場合におよそ 20 点低くなっていることがわかる．この結果から，作業 2 についてシステムによって作業負荷が軽減しているといえる．

各作業に対するアンケート結果を図 4，5 に示す．これらの結果から作業 1 についてはシステムを用いた場合の方が完成度の評価がまとまっており，用いていない場合は評価がばらけていることがわかる．また，難易度についてはシステムを用いる場合の方が高くなっていることがわかる．作業 2 についてはシステムを用いた場合の方が完成度の評価が高く，難易度が低いことがわかる．システムに対する評価は作業 1 の場合，操作性・利便性ともに平均でおよそ 3.5 であり，作業 2 の場合，操作性・利便性ともに平均で 3 ということがわかる．以上の結果からシステムを用いる場合，自己評価としての作業完成度が高くなると考えられる．また，作業の種類によってシステムの有無で作業難易度が変化することが考えられる．自由回答によるフィードバックで得た回答（付録 C）を一部抜粋して示す．“動画のサイズが若干小さいと感じた”，“テキスト情報と完成イラストとの対応が直感的にわかりやすかった”（P1）

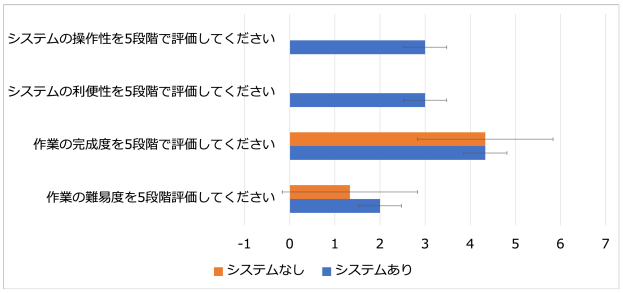

図 4　作業 1 に関するアンケート結果

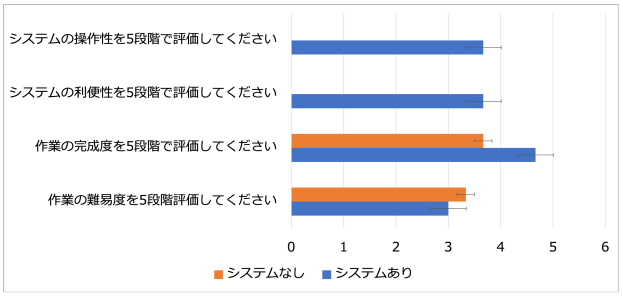

図 5　作業 2 に関するアンケート結果

“音声入力は使えるタイミングがよくわからなかった”，“スキップボタンは反応が早く使いやすかった。”（P5）これらのフィードバックから，システムの利用は一長一短であることがわかる．

6　おわりに

本稿では人手作業における誤りを減らし，作業の再現性を高めることを目的としたマルチモーダルな作業支援システムを提案し，システムが作業の実行時間，負荷，完成度にどのように影響するかを調べるため，2 種類の作業を行う実験によりシステムを評価した．実験の結果から，作業 1 に関して，実行時間，作業負荷は共に増加し，作業 2 に関して，実行時間に変化はなく，作業負荷は減少した．どちらの作業もシステムの利用によって作業完成度に対する自己評価は高まったことがわかった．これらの結果から作業種類によってはシステムの利用が誤りを減らすことができる可能性があると考えられる．

今後の課題として，実験のサンプル数を増やすこと，異なる作業に対して同様の実験を行うこと，システムの UI/UX の向上などが挙げられ，これらの解決に務めていきたい．

参考文献

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. **Nature**, Vol. 533, No. 7604, 2016.

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 18995–19012, 2022.

[3] 西村太一, 迫田航次郎, 牛久敦, 橋本敦史, 奥田奈津子, 小野富三人, 亀甲博貴, 森信介. Biovl2 データセット: 生化学分野における一人称視点の実験映像への言語アノテーション. 自然言語処理, Vol. 29, No. 4, pp. 1106–1137, 2022.

[4] Minsuk Chang, Mina Huh, and Juho Kim. Rubyslippers: Supporting content-based voice navigation for how-to videos. In **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[5] Philipp M Scholl, Matthias Wille, and Kristof Van Laerhoven. Wearables in the wet lab: a laboratory system for capturing and guiding experiments. In **Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing**, pp. 589–599, 2015.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In **International Conference on Machine Learning**, pp. 28492–28518. PMLR, 2023.

[7] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[8]

[9] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In **Advances in psychology**, Vol. 52, pp. 139–183. Elsevier, 1988.

A　実験に用いたデータセット

実験に用いたデータセットは研究室内で作成したものである．このデータセットは以下に示す 5 種類の作業についての各 10 本の映像とそれらと手順内容を紐づけたアノテーションからなる．映像の撮影には Panasonic HX-A500 一人称視点カメラを用いた．作業の選定にあたり，作業の領域，誤りの種類，映像の長さ，物体，動作の多様性を基準とした．表 2 に各作業の内容，手順数，手順あたりの単語数を示す．

表 2　各作業内容，手順数，手順あたりの単語数

作業番号	内容	手順数	手順あたりの単語数
1	金属のイオン化傾向を比較する化学実験	9	12.7
2	簡易的な電子回路を組む作業	8	7.6
3	積み木で指定された形を作る作業	7	18.6
4	蛍光ペンを用いた光の混色を調べる作業	8	17.0
5	ダンボール工作	14	9.6

B　NASA-TLX における WWL の算出方法

まず，NASA-TLX の 6 つの評価項目に対して ${}_6C_2$ 通りの比較を行う．比較の中で各項目が選ばれた回数を重み w_i（0〜5）とし, 各項目の素点を v_i とすると WWL は以下の式で算出され，1〜100 の値を取る．w_i の総和は比較回数と同じ 15 である．

$$WWL = \frac{\sum_{i=1}^{6} w_i \times v_i}{\sum_{i=1}^{6} w_i}$$

この値も各項目の素点と同様、WWL が高ければ高いほど作業者にとって作業負荷が高いことを示す．

C　自由回答によるフィードバック

“動画のサイズが若干小さいと感じた”，“テキスト情報と完成イラストとの対応が直感的にわかりやすかった”（P1）

（作業 1 に関して）“反応するかどうかがわからなかった”，“「スイッチ S2 をモーターと電池ボックスを直列に繋ぐ」について、1) 文法的に理解し難かった、2) 完成図の把握に時間がかかった。”（P2）

“反応の有無を確かめるのが微妙に大変だった。”（P3）

“ワークシートのマイクロプレートを置く位置の上に書かれている「銅、亜鉛、マグネシウム」が、マイクロプレートの後ろに隠れて見えづらかった。”，“並列に繋ぐの意味が分からなかった。または先に S2 とランプを繋げると、正解のようにならないと思います。”（P4）

“適切な場所に滴下するのが難しかった。”，“音声入力は使えるタイミングがよくわからなかった。手を使う作業中なのでテキスト入力はしにくい。スキップボタンは反応が早く使いやすかった。”（P5）