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の提案

梶村 恵矢 * 西村 太一 * 羽路 悠斗 * 山本 航輝 * 崔 泰毓 * 亀甲 博貴 † 森 信介 †

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．

1 はじめに

任意の作業は連続した複数の手順で構成されている．作業には機械による自動化がなされているものもあれば，人手で行われるものもある．科学実験や工作，料理などは人間が行う作業の最たる例である．人間が作業を行う場合，手順書を定めそれを遵守して作業を遂行することで，作業の再現性が担保される．しかし，手順書は作業者の習熟具合によっては時に不明瞭なものとなり，作業の誤りにつながり再現できない結果に陥ってしまう．

特に科学実験について言えば，7 割を超える科学者が他の研究者の再実験に失敗したことがあり，5 割を超える科学者が自身の行った実験の再実験に失敗したことがあると Baker によって報告されている [1]．さらに生化学の分野においては，実に 8 割を超える研究者が他者が行った実験の再現に失敗したことがあるという．このような状況において実験再現性の向上を目指す試みは科学の発展のために必要不可欠である．

作業者は手順書から文字情報を受け取り実行に移すが，作業映像を見ることで受け取る情報量が増え作業に対する解像度が高まるため，映像を見ながら作業することは作業の再現性向上に有効に働くと考えられる．人手で行われる作業の中にはその内容を説明する動画がインターネット上に公開されているものもある．例えば，クラシル[1] は料理をしている場面を定点カメラで撮影した動画をユーザに公開している．定点カメラによる映像は対象の状態を確認する意味で有用である一方で，映像中の作業者の意図や視線情報を把握することが難しい．作業内容を知ることでなく，作業を忠実に再現することを意図とした場合，それらの情報があることは作業者にとって大きなメリットとなる．それらの情報を得るために一人称視点の映像は非常に有用であると考えられる．

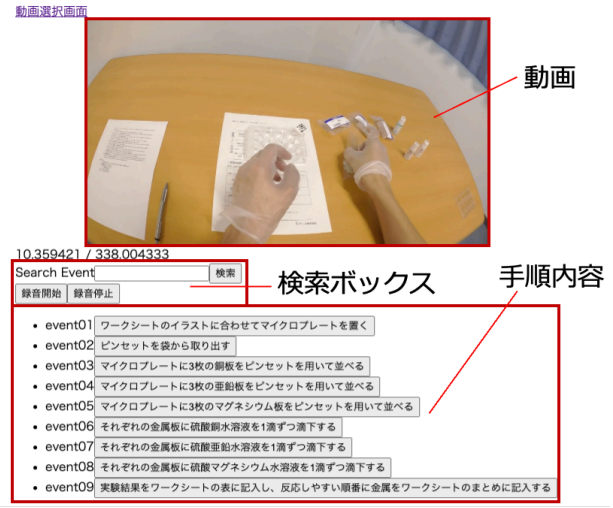

図 1. システム動作画面

以上のことから，本稿では科学実験に限らず，広く人手で行う作業の再現性向上を目的とし，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を提案する．

2 関連研究

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．世界 9 カ国の参加者の日常活動が記録された Ego4D はその中で最大規模のデータセットであり，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 [2]．BioVL2 は生化学実験に特化した数少ないデータセットであり，生化学分野の実験再現性向上のためのいくつかの手法を提案している [7]．HoloAssist はインタラクティブな AI アシスタントの開発を目的としたデータセットであり，視線情報や作業中の手の形など 7 つのモダリティの情報を含んでいる [6]．

* 京都大学大学院情報学研究科
† 京都大学学術情報メディアセンター

[1] https://www.kurashiru.com/

図 2. システムの処理の流れ

スマートグラスを用いたハンズフリーな実験支援システムは Scholl らによって提案されている [5]. このシステムでは，Google Glass 上に手順内容を記述した手続書が表示される．化学実験においてこの手続書の事をプロトコルと言う．またシステムでは，音声認識を用いて表示されるプロトコルを切り替えることが可能である．彼らが提案したシステムと我々が提案するシステムとの差分は作業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では実験映像を作業者に提示する．1 章で述べたように文字情報よりも映像情報のほうが情報量が多いため作業者に実験内容をより詳細に伝えられることが期待される．

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，手順箇所を検索するための検索ボックス，手順内容がそれぞれ含まれている．(図 1) 作業者がこのシステム内で検索ボックスにキーワードや手順内容を入力するか，音声入力によってそれらを伝えることで動画中の対象手順箇所が再生される．これにより，作業者は不安のある手順や文字情報だけでは詳細がわからない手順の映像情報を得ることができ，作業の再現性が高まることが期待される．

3.2 実装

本システムは入力機能や動画再生機能などユーザが操作する画面であるフロントエンドと受け取った入力と各手順とを比較する処理を行うバックエンドからなる．本システムで用いるデータは動画，各手順内容，(動画の最初を 0 秒として) 各手順の開始秒数が必要である．

フロントエンドでは検索ボックスから直接得たテキスト，あるいは録音ボタンで収録した音声がユーザの入力としてバックエンドに送られる．また，バックエンドでの処理の出力に応じて対応する動画中の箇所を再生する．

バックエンドではフロントエンドから送られてきた入力を受け取る．受け取った入力が録音された音声の場合，Whisper [4] を用いた音声認識によってそれをテキスト化する．得られたテキストを形態素解析し，ベクトル化したものを $\boldsymbol{t}$ とする．また検索対象となる，n 個の手順を含むプロトコルを同様にしてベクトル化したものを $\boldsymbol{P} = [\boldsymbol{p_1}, \boldsymbol{p_2}, ..., \boldsymbol{p_n}]$ とする．その後，$\boldsymbol{t}$ と各手順 $\boldsymbol{p_i}$ とのコサイン類似度を計算し，類似度が最も高い手順 $\hat{\boldsymbol{p}}$ を検索結果とする．形態素解析，コサイン類似度の計算には spaCy [3] を用いた．フロントエンドに返す出力は検索結果の手順が動画中で行われる時間（秒）である．

研究室内で撮影したデータを用いて上述した処理の流れを図 2 に示す．この例ではイオン化傾向を調べる実験を行っている．まず，最下段で入力としてテキストあるいは音声を受け取る．次に中段で入力された言語情報と各手順のコサイン類似度を計算する．そして，この例で最も類似度が高いのは手順 2 であるため，手順 2 が始まる箇所から動画が再生される．

4 おわりに

本稿では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した．このシステムは一人称視点の作業映像とプロトコルからなるシステムで，ユーザはプロトコルに基づいたイベント検索を行うことで，手順を一人称映像として確認できる．これにより，作業の再現性が高まることを期待する．

残された課題として作業者の両手が塞がっている状態だとシステムを利用しづらいことがある．システムをハンズフリーで操作可能にすることでこの課題は解決されると考えられる．その実装例として AR ヘッドセットで利用可能な形でシステム実装が挙げられる．また，実際にシステムを用いた実験を行い，提案するシステムの有用性を検証することも課題として残っている．これについては，研究室内で収集したデータや BioVL2 などの既存のデータセットを用いた実験を行っていく．

参考文献

[1] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 2016.

[2] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. 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] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

[5] P. M. Scholl, M. Wille, and K. 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] X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F. V. Frujeri, et al. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023.

[7] 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, 29(4):1106–1137, 2022.