6ZG-08

バイクの組立作業の習熟をサポートする工程識別システムの構築と機械的フィードバックの効果

中村 光伴 †1　山本 泰生 †1　西村 雅史 †1,2　白澤 怜樹 †3　中野 貴行 †3　青木 崇浩 †1,3
†1 静岡大学情報学部　†2 愛知産業大学造形学部　†3 ヤマハ発動機株式会社生産技術部

1. 研究背景と目的

現代ではロボットや人工知能による自動化が進展しているが，工場では依然として手作業が必要な工程が多く残っている．作業者の技能にはばらつきがあるため，習熟のための教育作業が必要となる．しかし教育にかけられるリソースには限界があり，現場での自己学習が主となるケースもある．本研究では習熟のサポートを目的に，作業を収録して実施工程の正確性や作業時間を作業者にフィードバックするための工程識別モデルおよび Web システムを構築し，その有効性を検証する．提案システムは FIELDS (*Feedback Integrated Expert Level Description System*) と名付けた．

2. 提案システム

2.1 機能

FIELDS が実現する機能は主に 3 つある．1 つ目は作業の収録・開始および収録データの保存，2 つ目は収録の工程識別，3 つ目は識別結果に基いたフィードバックである．フィードバックの観点は，工程抜けがないか，工程の順序が正しいか，時間がかかりすぎている工程がないか，の 3 つである．

図 1 に FIELDS の GUI を示す．画面中央には収録した動画を，画面左のサイドバーには識別した工程列を表示する．動画の右上には熟練者の平均時間との比較を表示する．サイドバーの各セグメントと動画は相互に連携している．セグメントをクリックすると再生位置はその開始時刻へ移動する．動画の再生バーをクリックした時あるいは動画の再生中は，サイドバーは対応するセグメントへ自動でスクロールしてハイライトする．

図 1: FIELDS の GUI

2.2 データフロー

2.2.1 収集データ

収録するデータは帽子のツバに取り付けたアクションカメラの動画である．作業者にはこの帽子を被って作業してもらう．動画は 30 fps で 1 フレームあたり 854 × 480 ピクセルの RGB 画像の系列として表せられる．

2.2.2 前処理

動画はフレームごとに BLIP-2 [2] に入力し，画像埋め込みベクトルの系列に変換する．BLIP-2 は image-to-text といった Vision language の分野で用いられる画像エンコーダである．1 フレームあたりの埋め込みベクトルは 256 次元であり，フレーム数を T とすると動画の埋め込みベクトルの形状は $\mathbb{R}^{T\times 256}$ と表せられる．

2.2.3 工程識別

動画の埋め込みは時間的行動分節モデルである MS-TCN [1] に入力する．MS-TCN はフレーム毎に工程ラベルを出力する．これをランレングス符号化することで (工程番号，工程開始時刻, 工程終了時刻) のタプルの系列を得る．

2.3 アーキテクチャ

図 2 に FIELDS のシステム構成を示す．GUI は Web ブラウザで実現される．API サーバは Web ブラウザがシステムのロジックと対話するための窓口として機能する．Worker サーバは時間のかかる処理や GPU が必要な処理を実行する．API サーバは収録終了の HTTP リクエストを受け取ると，収録をストップするとともにジョブキューに収録の情報を投入する．投入されたジョブは Worker サーバで処理される．カメラの動画は RTMP によりシステム内へストリーミングされる．

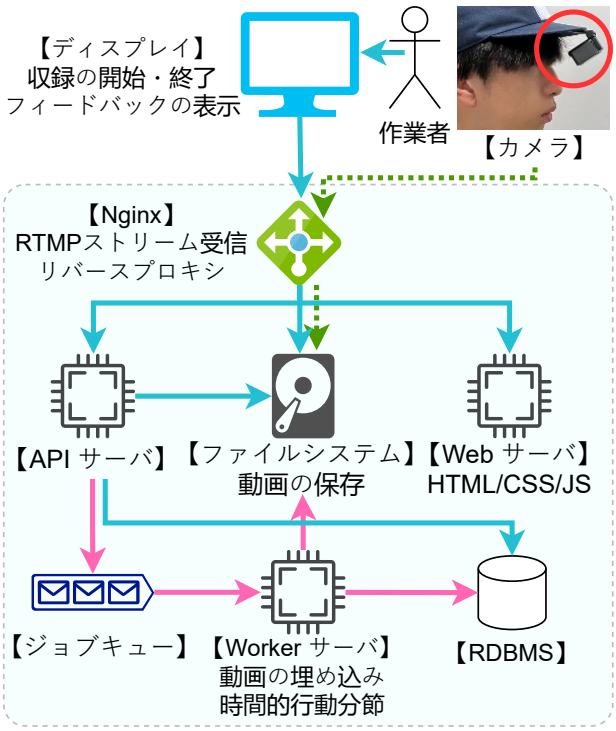

図 2: FIELDS のシステム構成

Development of an action recognition and feedback system for human training in motorcycle assembly works

Kosuke Nakamura†1, Yoshitaka Yamamoto†1, Masafumi Nishimura†1,2, Reiki Shirasawa†3, Takayuki Nakano†3, Takahiro Aoki†1,3
†1Shizuoka University, †2Aichi Sangyo University, †3Yamaha Motor Co., Ltd.

3. システムの有効性の評価

本セクションでは，FIELDS の有効性や，機械的フィードバックが人に与える影響を調べる．

3.1 実験設定

全 35 工程からなるバイクの外装取り付け作業を対象とした．本作業で扱う機材や工程はヤマハ発動機株式会社で実際に扱っているものである．

被験者として 7 名の学生を我々のキャンパス内から集めた．3 名を (A) FIELDS 使用，4 名を FIELDS 不使用の群に分けた．なお，被験者はみなバイクの組立の経験がない．実験は 1 人ずつ行った．被験者 1 人ごとに，(1) 15 分間の作業説明，(2) 30 分間の作業練習，(3) 45 分間の収録 5 回を通しで実施した．(A) 群と (B) 群とで統一した条件は次のとおりである：

- 練習時間の間は，被験者は熟練者の動画を観たり工程のマニュアル書を読むことができる．
- 被験者に対して人によるフィードバックはしない．これは作業者の自己学習を再現するためである．
- 練習後の 5 回の収録では，熟練者の動画やマニュアルを読むことはできない．

(A) 群の被験者には練習時間中に FIELDS を使ってもらった．つまり (B) 群は全くフィードバックが無いのに対し，(A) 群は機械的フィードバックが受けられる．

3.2 評価指標

我々は各被験者の 5 回の収録を目視で確認し，①工程抜け回数，②被験者の工程系列と正しい工程系列との編集距離 (Edit distance) ③作業全体にかかった時間 (開始合図から終了スイッチに触れるまで) の 3 つの指標で両群を比較した．また，定性的な指標として， FIELDS 使用者からの感想や，システムが誤認識した際の使用者の反応を収集した．

3.3 実験結果

図 3 に各被験者の工程抜け回数を示す．図 4 に各被験者の工程の編集距離を示す．2 つの図において，横軸は収録の回を，縦軸はそれぞれ欠落していた工程の数と編集距離を示す．

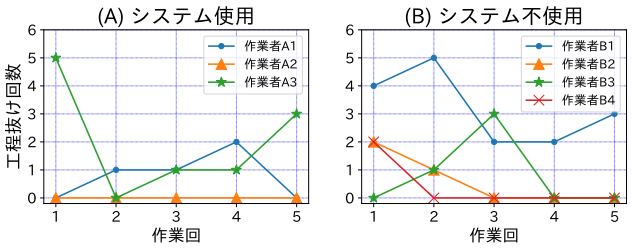

図 3: 各被験者の工程抜け回数

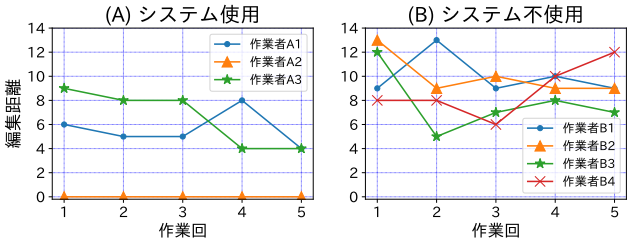

図 4: 各被験者の工程系列と正しい工程系列との編集距離

図 5 に作業全体にかかった時間の推移を示す．この作業時間には工程抜けを考慮したペナルティ時間も加算されている．ペナルティ時間は，とある工程 p の熟練者の平均時間を $t(p)$，欠落した工程の集合を M とすると，$\sum_{p \in M} \alpha \cdot t(p)$ で計算される．ただし，α はペナルティ係数であり本稿では 3 とする．

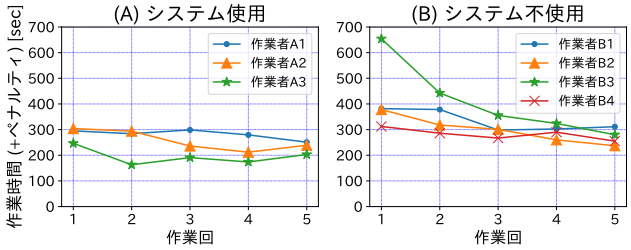

図 5: 各被験者の作業全体にかかった時間 (ペナルティ時間込み)

3.4 システム使用者の反応

(A) 群の被験者がシステムを使用したときの感想として，「工程と動画が連動しているので文章だけ・動画だけより覚えやすい」「時間がかかっている工程が分かるので時間短縮しやすい」といった声が挙げられた．

システムが誤認識したときの被験者の反応を以下に示す．

- 被験者は誤認識された自身の動きに問題があるのだと理解した．例えば，被験者の「ボルト締め付け」の工程をシステムが「カバータンクサイド取り付け」として誤認識したことがある．被験者は自身の動画とともにこのフィードバックを受けて，ボルト締め付けの際にカバーに手を置きながら作業していたことに気付き，以降は不必要に手を置かないよう注意するようになった．
- 被験者が工具を迷って潤滑剤のブラシを持ちかけたとき，システムは「潤滑剤塗布」と誤認識した．被験者はこれを受けて，自身が意外ともたついていたのだと認識した．

3.5 考察

工程抜けに関して，図 3 を見ても (A) 群と (B) 群とで明確な違いは表れなかった．(A) 群でも工程抜けが見られる理由としては，システムの誤認識と，練習時間の短さの 2 つが考えられる．実際，練習時間 30 分ではシステムを 1 回か 2 回しか使うことができなかったので，仮にシステムからのフィードバックがあっても作業に慣れなかった可能性が考えられる．この仮説は工程抜けの回数が上下していることからも裏付けられる．

編集距離については差が見られた． 図 4 をみると，(A) 群は 5 回目では 4 以下に収束しているのに対し，(B) 群は殆どが 7 以上となっている．これは FIELDS からのフィードバックが効果があったことを示唆している．

作業時間に関しても，(A) 群の方が 1 回目から短い時間となっており，システムが作業習熟に役立った可能性が考えられる．

システム使用者の感想からは肯定的な意見が得られた．また，システムからの誤ったフィードバックは，人にとっては動作を改善する良いヒントになりうることがわかった．

4. まとめと今後の展望

本研究では，ウエアラブルなカメラ 1 つで行動分節して機械的フィードバックを与える FIELDS を構築し，その有用性を実証評価した．今後はより効果的なフィードバックの見せ方や，加速度・角速度データを用いた細かい動きに関するフィードバックを模索したい．

参考文献

[1] Y. Abu Farha and J. Gall. MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. In CVPR, 2019.

[2] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In ICML, 2023.