2V-05

# 慣性情報と音情報を用いた作業行動の自動分節化

武井久実† 中村圭佑† 塩野由紀‡ 中野貴行‡
青木崇浩‡ 山本泰生† 西村雅史†
静岡大学大学院総合科学技術研究科† ヤマハ発動機株式会社生産技術部‡

## 1. はじめに

生産現場における作業行動の分析・改善は日々行われているものの，人手による分析には多くの時間と労力を必要とするため，自動化が強く望まれている．

行動認識の手法として，大きく分けてセンサベースの手法と画像ベースの手法がある[1]．前者は対象者や対象物に IMU（Inertial Measurement Unit）等のセンサを取り付け，行動に関する情報を収集する．一方，後者はカメラによって動画データを収集することで行動認識を実現する．IMU 等のセンサの装着は対象者の負担となる場合もあるが，画像に比べてデータサイズが圧倒的に小さいことやプライバシー保護の観点では有利とされ，日常生活行動の認識では広く利用されている．ただし，得られる情報は限定的であるため，これまで工場等での作業行動認識で利用されることは少なかった．

本研究では IMU で得られる慣性情報に加え，音情報の有効性について確認するとともに，画像ベースの認識との性能比較を行う．

## 2. 提案手法

### 2.1. 作業行動の文節化

ここでは，各種センサ情報を用いてフレーム単位で作業行動を認識し，工程全体の作業を自動分節化することを目的とする．これによって作業行動の時間や順序に関連した分析が容易になると期待される．

### 2.2. センサ

作業者の行動を検出するためのセンサとして，IMUを３箇所（右手，左手，額），マイクを１箇所（右手）に装着する．IMUは加速度センサとジャイロセンサを持っており，３軸加速と３軸角速度を計測可能である．一方，マイクからは作業音を集音する．

なお，加速度，角速度データはサンプリング周波数100Hzで収集し，標準化を行ったのち，３軸分を連結したものを特徴量とする．一方，音についてはサンプリング周波数16KHz，フレーム周期10msec（=100Hz）で収集し，12次元のMFCCを求め，これを特徴量とした．

### 2.3. 行動認識モデル（MS-TCN）

ここでは動画の分節化において高い性能が確認されている MS-TCN[2]（Multi-Stage Temporal Convolutional Network）を各種センサ情報に対して使用することで行動認識を行う．TCN は時系列情報を処理可能な CNN であり，一定の間隔を空けて情報を畳み込むことで広い範囲の時間情報を考慮した識別が可能となっている．MS-TCN はさらにそれを多段化することで分節化性能を高めたモデルである．実際我々の予備実験においても，一般的な多層 LSTM と比較して，大幅な性能改善が得られることを確認している．

### 2.4. 各センサに対する MS-TCN の統合

多種のセンサを併用する場合，それぞれの特徴量ごとに MS-TCN を構築し，図１に示すように，対数尤度の加算平均を求めて行動識別を行う（Late Fusion）．なお，IMU から得られる特徴量と音声特徴量はフレーム周期 10msec で同期している．

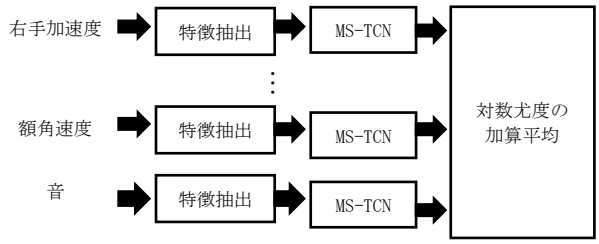


図１ 慣性情報と音情報を用いた MS-TCN モデル

## 3. 評価実験

### 3.1. 実験データ

実際の工場での新人教育用の組み立て作業工程をベースとし，図２に示すような，ガイドピン取付，ボルト締付，Ｏリング組付といった，12 種類の行動からなる部品組み立て作業を設計した．

この作業を２０代男女 4 名が，それぞれ 50 回実施したものを，先に記した 3 個の IMU（ATR-

---

Automatic Segmentation of Assembly Operations Using Inertial and Sound Information
Kumi Takei†, Keisuke Nakamura†, Yuki Shiono‡, Takayuki Nakano‡, Takahiro Aoki‡, Yoshitaka Yamamoto†, Masafumi Nishimura†
†Graduate School of Integrated Science and Technology, Shizuoka University
‡Manufacturing Technology Center, Yamaha Motor Co., Ltd.

Promotion, AMWS020），1 個のマイク（Audio Technica, AT9904）で収集した．一連の作業には平均約 100 秒，各行動には平均約 1〜30 秒を要していた．このデータを学習及び評価に使用する．

また，比較のため，作業行動の動画データも３視座（前，右，左）から同時に収集した．

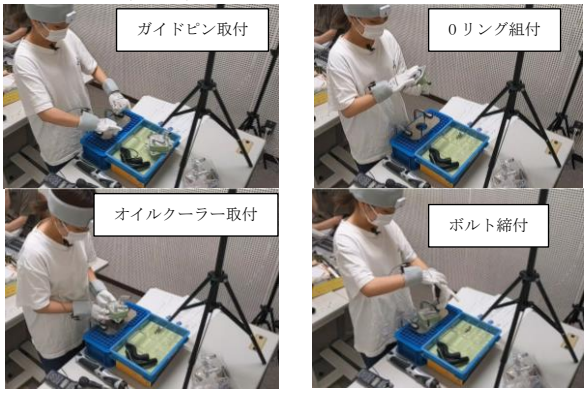


**図 2 本実験で認識を行う動作例**

### 3.2. センサベースの評価実験結果

3 名分のデータを学習用に，残り 1 名分を評価用とし，交差検証で認識性能評価を行った．表１にフレーム単位での評価結果を示す．

表１ 特徴量毎の認識性能(F 値)

| 使用した特徴量 | F1-score |
|---|---|
| 音のみ | 0.89 |
| 加速度（右手のみ） | 0.75 |
| 角速度（右手のみ） | 0.81 |
| 音+（加速度+角速度：右手のみ） | 0.90 |
| 音+（加速度+角速度：３箇所） | 0.91 |

全ての情報（３箇所の IMU+音）を用いた結果，認識性能としてはF 値で0.91 が得られた．IMU については右手だけの情報を使った場合の性能も示したが，手先を使う作業動作であるため，加速度よりも角速度の有効性が確認された．一方で複数の IMU を併用したことによる改善幅は小さかったことが分かる．

また，予想に反して慣性情報よりも音情報が作業行動認識において非常に有効であることも示唆された．ただ，今回は防音室内でのデータ収録であったため，実際の工場環境での有効性については更なる検討が必要だと考えている．

誤認識の多くは時間的に継続時間の短い動作，特に異なる種類の部品をケースから所定の位置に移動させるといった，他の行動と部分的に非常に類似した動作部分で起きており，センサのみを用いる手法の限界も感じられる．

### 3.3. 画像ベースの手法との性能比較

センサベースの作業行動認識性能を，これまで数多く検討されて来た画像ベースの手法と比較した．先に示したように，慣性情報及び音情報の収集と同時に収録した動画データを使用する．ここでは被験者の正面，及び左右の 3 視座から撮影した動画をそれぞれ OpenPose によって処理し，骨格情報を抽出する．さらに，山本ら[3]の手法に基づいて特徴抽出を行ったのち，MS-TCN を用いて学習，評価した．

センサ情報（音，加速度，角速度）及び画像情報を用いた場合のフレーム単位の F 値を表 2 に示す．結果として，３視座の情報を全て使った場合には及ばないものの，今回提案したセンサ情報による作業行動認識の性能は画像情報を用いる場合に近いことが分かる．今回の比較実験では画像に対して骨格情報抽出や特徴点抽出といった処理を加えているため，元の画像情報を十分に活用できているわけではないが，データ量や処理時間を考慮すると，作業行動認識のための情報源として音を中心としたセンサ情報が有望であることは間違いない．

表 2 センサーベースと画像ベースの性能(F 値)

| 音+（加速度+角速度：右手） | 画像情報（正面 1 視座） | 画像情報（3 視座） |
|---|---|---|
| 0.90 | 0.89 | 0.93 |

## 4. おわりに

本研究では，慣性情報と音情報を用いることが作業行動認識において大変有効であることを確認した．その中でも音情報は非常に有効であることが分かった．また装着位置の影響は少なく，スマートウォッチ等のデバイスでも，作業行動認識のための十分な情報収集が可能であることが示唆された．

## 参考文献

[1] E. Garcia-Ceja et al., “Multi-view stacking for activity recognition with sound and accelerometer data”. Inf. Fusion. 40, pp.45-56, 2018.

[2] Y. A. Farha et al., “MS-TCN: Multi-stage temporal convolutional network for action segmentation.” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[3] 山本泰生ほか, “姿勢推定を用いた組立作業の工程分解”, 画像応用技術専門委員会報告 Vol.36 No.4, pp.8-15, 2021.