油圧ショベルの動作認識のための実画像とシミュレーションを用いた学習法

桂 知弘 [*1], 小松 廉 [*1], 永谷 圭司 [*1], 千葉 拓史 [*2],
茶山 和博 [*2], 山下 淳 [*3], 淺間 一 [*1],

Training Method Using Real Images and Simulation Data for Activity Recognition of Excavators

Tomohiro KATSURA[*1], Ren KOMATSU[*1], Keiji NAGATANI[*1],
Takumi CHIBA[*2], Kazuhiro CHAYAMA[*2],
Atsushi YAMASHITA[*3] and Hajime ASAMA[*1]

[*1]Department of Precision Engineering, School of Engineering, the University of Tokyo
Hongo 7-3-1, Bunkyo, Tokyo 113-8656, JAPAN
[*2]Fujita Corporation, 4-25-2 Sendagaya, Shibuya-ku, Tokyo 151-8570, Japan
[*3]Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, the University of Tokyo
Kashiwanoha 5-1-5, Kashiwashi, Chiba 277-8563, JAPAN

In this study, we proposed a training method using real images and simulation data for activity recognition of excavators. In the proposed method, the activity recognition of excavators is divided into a position detection model, a skeleton detection model, and an activity recognition model. The first two models are trained from real image data, while the activity recognition model is trained from simulation data. We confirmed that the proposed method can achieve the same or better accuracy than the conventional method that uses real video data, even though the proposed method does not use any real video data in the training process. We also obtained suggestions on how to efficiently collect data from simulations.

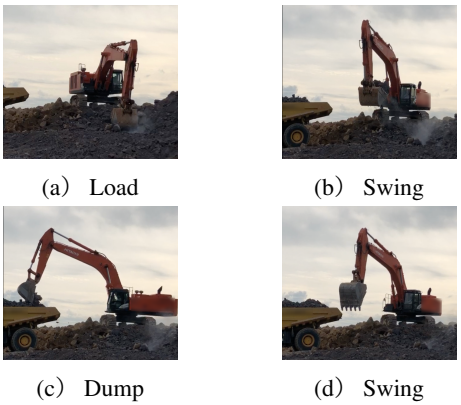
Key Words : Activity recognition, Excavator, Deep learning

1. 緒　　言

油圧ショベルは Fig. 1 のように「Load（掘削）」「Swing（旋回）」「Dump（放土）」「Swing（旋回）」を繰り返す．この各動作時間の記録は，油圧ショベルの生産性を評価する上で重要な指標になる[(1)]．

本研究では，アノテーション付きの実環境の画像を実画像データ，アノテーション付きの実環境の動画を実動画データと呼称する．油圧ショベルの動作認識に関する研究は，機械学習を用いてカメラ映像から動作認識を行う研究[(2)]などが一般的である．しかし，十分な動作認識精度を得るための油圧ショベルの学習用実動画データが不足している．

油圧ショベルの大規模な実画像データセットは公開されており[(3)(4)]，撮影方向のバラエティや，データのバラエティは豊富である．一方，実動画データセットは，Roberts が公開した数百点の実動画データセットのみであり[(5)]，撮影方向のバラエティや，掘削動作自体のバラエティの両面で大きく不足している．十分な量の実動画データセットを作成するには，金銭的・時間的コストが高いため，実動画データ以外からも学習が可能な手法が望ましい．実動画データを使わずに，シミュレーションから学習する研究も行われているが，



(a) Load　(b) Swing　(c) Dump　(d) Swing

Fig. 1: Excavating cycle

[*1] 東京大学大学院工学系研究科（〒 113-8656 東京都文京区本郷 7-3-1）katsura@robot.t.u-tokyo.ac.jp
[*2] 株式会社フジタ（〒 151-8570 東京都渋谷区千駄ヶ谷 4-25-2 修養団 SYD ビル）takumi.chiba@fujita.co.jp
[*3] 東京大学大学院新領域創成科学研究科（〒 277-8563 千葉県柏市柏の葉 5-1-5）yamashita@robot.t.u-tokyo.ac.jp

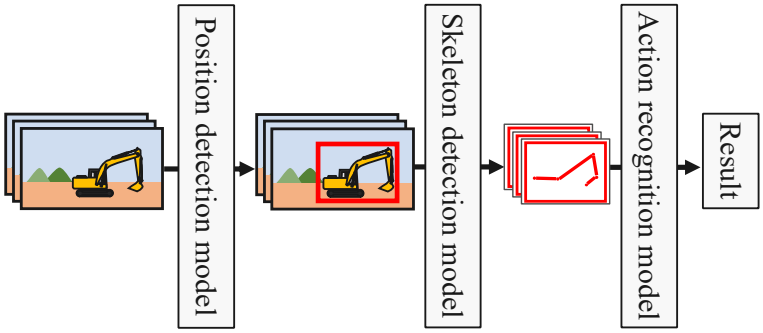


Fig. 2: Overview of the proposed method

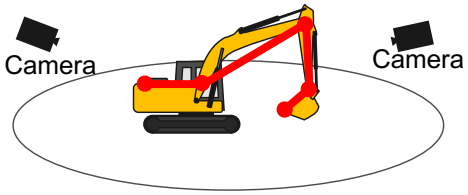


Fig. 3: Data acquisition

日々変化する作業現場に頑健な手法ではない[6].

以上より，本研究の目的を油圧ショベルの動作認識のための実画像とシミュレーションを用いた学習法の提案とする．実動画データ以外からも学習が可能であり，かつ日々変化する作業現場に頑健な手法の提案を目指す．特に，作業現場固有のデータを使用せず，シミュレーションから得たデータと実画像から学習する手法の提案を第 1 の目標とし，シミュレーションからのデータの効率的な集め方の提示を第 2 の目標とする．

第 2 章では，提案手法について述べる．第 3 章では，提案手法の有効性を確認するための実験について述べる．第 4 章では，本研究の結論を述べる．

2. 提　案　手　法

2·1　学習モデルの概要　提案手法では，Fig. 2 に示すように，動作認識の過程を油圧ショベルの位置検出モデル・骨格検出モデル・動作認識モデルの 3 段階に分ける．実動画データと比べ，油圧ショベルの実画像データは撮影の方向のバラエティ，掘削動作自体のバラエティの両面で豊富である．そこで，位置検出モデルや骨格検出モデルは実画像データから学習を行う．実動画データは撮影の方向のバラエティ，掘削動作自体のバラエティの両面で非常に少ない．そこで，動作認識モデルは，シミュレーション等から得られた時系列骨格推移データをもとに学習する．

2·2　動作認識モデルの学習データ取得方法　動作認識モデルの学習データは，シミュレーション等から得る．シミュレーションの概要を Fig. 3 に示す．シミュレーション上で油圧ショベルを動かし，骨格の時系列 3 次元座標を得る．次に，動作認識を行うカメラが設置されうる範囲内の任意の位置で，カメラで撮影したように，多くの 2 次元平面上に骨格の時系列 3 次元座標を同時に投影する．これにより，多様な方向から見た時系列骨格推移データを効率的に得られる．より具体的に説明する．学習データを得るために，油圧ショベルの動作を計 M_{train} 回シミュレーション上で再現する．得られた骨格の時系列 3 次元座標を N_{view} 個の 2 次元平面に投影する．これにより，データを N_{train} 個に拡張できる．同様に，検証データを得るために，計 M_{val} 回だけ再現した動作を N_{view} 個の 2 次元平面に投影し，データを N_{val} 個に拡張する．ただし，$N_{train} = M_{train}N_{view}$，$N_{val} = M_{val}N_{view}$ の関係がある．また，データ拡張のために回転やノイズの追加をした．

3. 検　証　実　験

3·1　実験概要　本実験の第 1 の目的は，シミュレーションから得た時系列骨格推移データを利用して，動作認識モデルの学習が可能か確認することである．第 2 の目的は，学習に最適なシミュレーションデータの効率的な集め方を提示することである．そこで，提案手法を用いて「Load」「Swing」「Dump」を認識し，従来手法の動作認識精度と比較した．提案手法は位置検出モデル・骨格検出モデル・動作認識モデルからなる．それぞれの学習モデルを説明する．その後，従来手法や実験結果について述べ，考察する．

3·2　位置検出モデル　油圧ショベルの位置検出に，Mask R-CNN[7]を用いた．位置検出モデルの学習データとして Moving Objects in Construction Site（MOCS）[8]を，検証データとして The Alberta Construction Image Dataset（ACID）[4]を利用した．データ拡張のために左右反転・回転・トリミングを用いた．

3·3　骨格検出モデル　Fig. 4 に示すように，バケット先端，バケット回転軸・アーム回転軸・ブーム

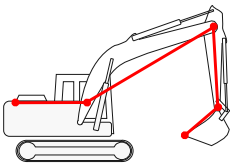


Fig. 4: Excavator skeleton

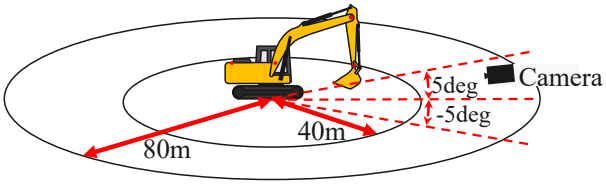


Fig. 5: Data augmentation

の付け根・油圧ショベル本体後方の 5 点のキーポイントを直線で結び，油圧ショベルの骨格とした．骨格検出モデルは HRNet[9]を用いた．骨格検出モデルの学習には，Luo らが公開しているデータセット[10]のうちの 1,000 枚を学習データ，残りの 281 枚を検証データとして用いた．また，データ拡張のために，左右反転と回転を行った．

3·4 動作認識モデル 動作認識モデルの学習データを集めるために，シミュレーションで各動作中の油圧ショベルの各キーポイントの 3 次元座標を取得した．Fig. 5 に示すように，中心の油圧ショベルから水平距離が 40m から 80m の範囲で，かつ仰角が-5 度から 5 度の範囲に動作認識を行うカメラがあると想定した．

学習データを得るために，シミュレータ上で「Load」「Swing」「Dump」を計 M_{train} 回だけ再現し，N_{view} 個の 2 次元平面に投影し，データを計 N_{train} 個に拡張した．同様に，検証データを得るために，動作を計 M_{val} 回再現し，N_{view} 個の 2 次元平面に投影し，データを計 N_{val} 個に拡張した．提案手法の Condition 1 から 7 における M_{train}，M_{val}，N_{view}，N_{train}，N_{val} の関係を Table 1 に示す．動作認識モデルは Pose-SlowOnly[11]を用いた．学習データにはランダムな回転と，各キーポイントにガウス分布に従うノイズを加えた．

3·5 比較する従来手法 同様に，従来手法は位置検出モデルと骨格検出モデルと動作認識モデルからなる．動作認識モデルの学習時は Roberts らのデータセット[5]のみを用いた．学習データとして 222 クリップ，検証データとして 47 クリップを使用した．

3·6 動作認識精度の判定用テストデータ Roberts らのデータセット[5]で，前節で不使用の 179 クリップをテストデータとして用いた．このテストデータは，Load が 42 クリップ，Swing が 90 クリップ，Dump が 47 クリップだけ含まれていた．

3·7 結果と考察 動作認識精度を r，正しく判定された実動画の総数を $N_{correct}$，判定した実動画の総数を N_{video} とすると，以下の式 (1) が成り立つ．

$$r = \frac{N_{correct}}{N_{video}}. \tag{1}$$

動作認識精度を示した Fig. 6 より，従来手法の動作認識精度は 83%である．一方，提案手法の Condition 1 の認識精度は 85%である．適切に学習を行えば，実動画データを学習に使用する従来手法より高い動作認識精度を提案手法は得るとわかった．

次に，シミュレーションからの効率的な学習データの集め方を考察する．まず，従来手法と，$N_{train} = 4000$，$N_{val} = 400$ で固定されている Condition 1 から Condition 4 を比較する．動作認識精度は Condition 1 が最も高く，Condition 4 が最も低い．Condition 1 以外の提案手法は，少量の掘削動作を多量の 2 次元平面に投影し，時系列骨格データの総量を確保しており，時系列骨格データの元の掘削動作のバラエティが少なく，動作認識精度が低下したと考えられる．

次に，$M_{train} = 800$，$M_{val} = 80$ で固定されている Condition 1，Condition 5，Condition 6，Condition 7 を比較する．Condition 5，Condition 7，Condition 6，Condition 1 の順に動作認識精度が上昇する．Condition 5 のように N_{view} が小さすぎる場合，データの総量が少なくなり，動作認識精度が上がらない．一方，Condition 6 や 7 のように，N_{view} が大きすぎる場合，$N_{train} = 12000$ や $N_{val} = 1200$ と比較し，N_{view} が大きくなり，動作認識精度が上がらない．つまり，学習を適切に進めるには，掘削動作を投影する 2 次元平面のバラエティ（N_{view}）と掘削動作自体のバラエティ（M_{train}，M_{val}）が適切な関係にある必要がある．掘削動作自体のバラエティ（M_{train}，M_{val}）に対し，掘削動作を投影する 2 次元平面のバラエティ（N_{view}）は大きすぎても小さすぎても動作認識精度が低下する要因となることがわかった．

4. 結　　言

本研究では，油圧ショベルの動作認識のための実画像とシミュレーションを用いた学習法の提案をした．以下に成果を示す．

（1）油圧ショベルの動作認識を，位置検出モデル・骨格検出モデル・動作認識モデルに分け，前者 2 つは実画像データから，動作認識モデルはシミュレーションから時系列骨格推移データを得て学習する手法を提案した．提案手法は学習に実動画データを使用しないが，従来手法と同程度以上の動作認識精度を持つことがわかった．

Table 1: Number of data

	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Condition 7
M_{train}	800	160	80	40	800	800	800
M_{val}	80	16	8	4	80	80	80
N_{view}	5	25	50	100	1	10	15
N_{train}	4000	4000	4000	4000	800	8000	12000
N_{val}	400	400	400	400	80	800	1200

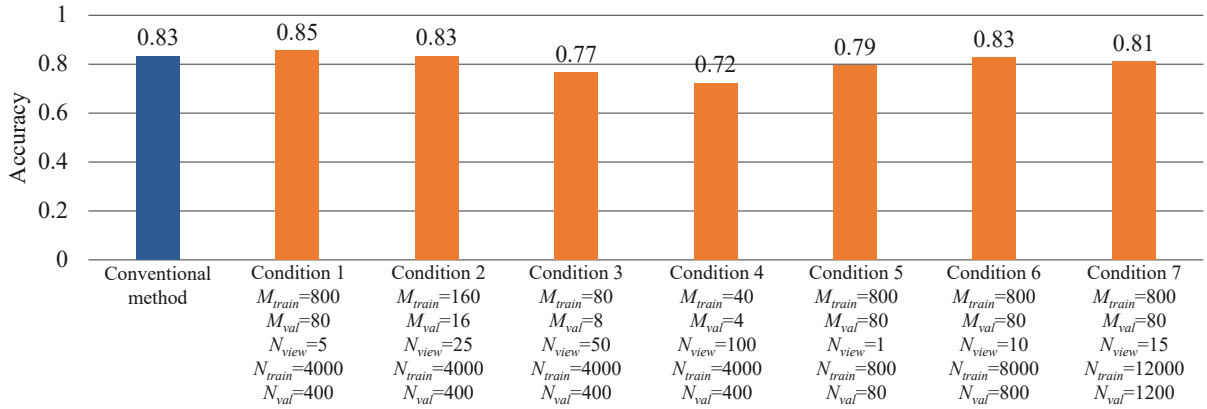


Fig. 6: Accuracy comparison

（2）シミュレーションから時系列骨格推移データを効率的に集める手法を提案した．シミュレーション上で油圧ショベルの骨格の 3 次元座標を得て，複数の 2 次元平面に投影し，効率的に時系列骨格推移データを集められた．これにより，学習データを収集する時間的・金銭的なコストを大きく削減できた．また，掘削動作を投影する 2 次元平面のバラエティ（N_{view}）と掘削動作自体のバラエティ（M_{train}，M_{val}）の適切な関係に示唆を与えた．

参 考 文 献

(1) T. Cheng, J. Teizer, G.C. Migliaccio, and U.C. Gatti, “Automated Task–level Productivity Analysis through Fusion of Real Time Location Sensors and Worker’s Thoracic Posture Data”, Automation in Construction, Vol. 29, pp. 24–39, 2013.

(2) Chen, C., Zhu, Z., and Hammad, A, “Automated excavators activity recognition and productivity analysis from construction site surveillance videos”, Automation in Construction, Vol. 110, pp. 103045, 2020.

(3) Xuehui, A., Li, Z., Zuguang, L., Chengzhi, W., Pengfei, L., and Zhiwei, L, “Dataset and benchmark for detecting moving objects in construction sites”, Automation in Construction, Vol. 122, pp. 103482, 2021.

(4) Xiao, B., and Kang, S. C. “Development of an image data set of construction machines for deep learning object detection”, Journal of Computing in Civil Engineering, Vol. 35, No. 2, pp. 05020005, 2021.

(5) Roberts, D., and Golparvar-Fard, M, “End-to-end vision-based detection, tracking and activity analysis of earth-moving equipment filmed at ground level”, Automation in Construction, Vol. 105, pp. 102811, 2019.

(6) ルイ笠原ユネス, 沈鎮赫, 小松廉, 筑紫彰太, 永谷圭司, 千葉拓史, 山本新吾, 茶山和博, 山下淳, 淺間一, “シミュレータで作成された訓練データのデータ拡張による油圧ショベルの動作認識”, 精密工学会誌, Vol. 88, No. 2, pp. 162-167, 2022.

(7) He, K., Gkioxari, G., Dollár, P., and Girshick, R. “Mask r-cnn”, In Proceedings of the IEEE international Conference on Computer Vision, pp. 2961-2969, 2017.

(8) Xuehui, A., Li, Z., Zuguang, L., Chengzhi, W., Pengfei, L., and Zhiwei, L, “Dataset and benchmark for detecting moving objects in construction sites”, Automation in Construction, Vol. 122, pp. 103482, 2022.

(9) Sun, K., Xiao, B., Liu, D., and Wang, J, “Deep high-resolution representation learning for human pose estimation”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693-5703, 2019.

(10) Luo, H., Wang, M., Wong, P. K. Y., and Cheng, J. C, “Full body pose estimation of construction equipment using computer vision and deep learning techniques”, Automation in Construction, Vol. 110, pp. 103016, 2020.

(11) Duan, H., Zhao, Y., Chen, K., Lin, D., and Dai, B, “Revisiting skeleton-based action recognition”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2969-2978, 2022.