建築生産のデジタル化に向けたAI活用
－単眼カメラを用いた作業者検出と追跡ならびに作業者動作認識の検証－

石岡　宏晃
（技術研究所）

AI for Disitization of Construction Management
－Single Camera Worker Detection, Tracking and Action Recognition in Construction Site－

Hiroaki ISHIOKA

建築生産のデジタル化に向けたAI活用の事例として、現場管理者が作業を振り返ることが可能となるデータ取得手法を構築した。我々は日本の建築現場内の映像から機械学習用データセットを構築した。次にそれを用いて建築現場用にカスタマイズされた作業者検出と追跡アルゴリズムならびに作業者動作認識アルゴリズムを構築した。性能評価により、作業者検出は87.9%の精度を示し、動作認識は平均精度で60.2%を示した。最後に作業者検出からの出力を作業要素に変換するデータ取得手法を設計し、実装例を示すことで、建築現場内のデータ取得へのAI活用の有用性が示された。

This paper shows an example of AI for digitization of construction management. We built a method to acquire digital data of the site work for the site manager's review. Firstly, we created a new dataset for machine learning from Japanese construction sites' movies. Secondly, construction domain specific algorithms of worker detection, tracking, and action recognition were developed. In the evaluation, the worker detection showed 87.9% AP, and the action recognition showed 60.2% mean accuracy. Finally, by developing a data converts method from output of worker detection into work elements, the usefulness of using AI for data acquisition was clarified.

1. 緒言

日本国内の建設産業の労働者数の減少[1)]を受け、建築生産における生産性向上が強く求められている。筆者は生産性向上実現に向けて重要な点として、新規入職者からベテランまで、現場内で作業するすべての職人が、自身の技能を発揮するうえで受ける種々の制約を最小化することが重要であると捉えている。そのために必要なのは2点あり、一つは作業の前提となる工程計画と作業計画に、当該作業空間である建築現場の実状を適切に反映すること、もう一つは作業実施中に発生した事象が作業に与える影響を適切に配分すること、である。いずれも作業の状況を最適化計算に乗せるために事象をモデル化する必要があり、そのためには作業の状況をデジタルデータとして取得するところから実現しなければならない。

建築現場内の作業のデータ取得は、全世界の建設業界で共通の課題であり、多くの研究者が建築現場内の作業のデータ取得に取り組んできた。Tarakら[2)]は文献レビューによりこれらの研究を3つの分類：拡張情報技術を用いた手法(原文：Enhanced IT technologies)地理情報技術を用いた手法(原文：Geospatial technologies)そして、画像処理技術を用いた手法(原文：Imaging technologies)として取りまとめた。Aminら[3)]もまたBIMとコンピュータビジョン(以下、CV)に焦点を当てた文献レビューによりそれらの研究の進展について記述した。これら全世界的な取り組みがあるものの、欧米諸国と日本の建築現場には違いがある[4)]点に注意が必要である。特に画像処理技術を用いた手法の適用に当たっては、映像に映り込む作業者の特徴を考慮する必要がある。

日本の建築生産においては、総合建設会社に所属する現場管理者が現場に常駐し、工事の進捗を管理する。現場管理者は、同じ作業日に複数の専門工事業者が同フロアで工事を実施できるように作業場所を細かに割り振ることで工期を短縮することが一般的である。そのため、建築現場内の映像には同日に複数工種が映り込むことが一般的である。日本では現場管理者が高い安全意識を持った現場管理ルールを適用している現場が多く、丈

夫で保護性に優れた長袖の作業着の着用が一般的である。その結果、一般の商品としてのカラーバリエーションが豊富である。また、職人の所属企業がその企業独自のユニフォームを用意している場合もある。特記すべきは、日本以外の国々の建設現場で着用が一般的である鮮明な色をした安全ベストは、日本の建築現場ではその着用は一般的ではない点である。さらに、日本ではヘルメットと安全帯の着用ルールは作業者に浸透しており、安全帯に装着する工具袋等のアタッチメントも多くの商品が販売されている。以上のように日本独自の服装の特徴が存在する。日本独自の機械学習用データセットを構築する必要がある。

本稿では、日本の建築現場において有効なデータ取得手法の構築に向けて、はじめに独自の学習用データセットを構築したうえで、世界的に進展する画像処理技術を用いた手法から、作業者検出と追跡のアルゴリズムならびに動作認識アルゴリズムの適用を図る。加えて、それらのアルゴリズムの出力データをもとに日々の工事の進捗を把握するために価値のある作業要素データに変換する手法を示す。

2. 文献調査

2.1. 建築現場のデータセット

建築現場を対象とした研究において、日本以外の他国においても、独自のデータセットを構築している例がある。Jun ら[5]は動作認識のための 11 動作の短い動画群を集めたデータセットを作った。Kaijian ら[6]はデータアノテーションをアウトソーシングにより実施する際の発注条件を変えてデータ品質を比較した。Mohammad ら[7]はパワーショベル・ホイールローダー・トラックの画像データセットを作成し、各種の画像処理技術による検出精度を比較した。近年においては、Mingzhu ら[8]は建設重機と作業者の画像データセットを作成している。いずれの場合も、独自の建設現場で画像が取得されており、日本の建設現場ではない。依然として、日本の建設現場の画像データセットは独自に構築される必要がある。

2.2. 物体検出と追跡

1)物体検出：ここ 10 年で物体検出技術には大きな進展があった。深層学習の時代以前では、エンジニアの手作業による特徴量表現に関する研究によって文献は占められている。例えば、HoG 特徴量[9]、SIFT(Scale-Invariant Feature Transform)[10]、DPM(Deformable Part Model)[11]などの様々な特徴量表現方法が、歩行者検出を目的としてカスタマイズされた形で示された。これらの手法が物体検出を著しく発展させる一方で、一つの特徴量表現で汎用的に多種の物体の特徴を表すことは困難であったため、これらの手法は手作業による調整に多くの労力を研究者に強いることとなった。

ここ数年で深層学習手法がよく知られるようになったことで、普遍的な特徴量を、大量に収集されたデータセットから学習できる、多くの現代的な物体検出手法が示され、特徴量を手作業で定義する必要性を不要のものとした。種々のアプローチの中で、RPN(Region Proposal Network)という領域の特徴を扱うネットワークを加えた物体検出手法[12-17]が知られている。これは、全体の処理時間の短縮を目的として、画像全体をカバーする異なる寸法とサイズのアンカーを定義して前景と背景との位置を学習しながら、併せて物体の分類のネットワークの学習を行うことで、物体の存在する領域を早期に検出対象とする手法である。

本稿では、高精度の一般物体認識器としてよく知られている Faster-R CNN[14]を使用し、日本の建築現場の建築作業者という特定の場面において性能を発揮するように適応させる。

2)複数物体追跡：物体種別の分類と画像上の物体の位置の検出の後、複数物体追跡は、フレームごとに検出された物体を動画を通して関連させ、物体の移動を出力することを目的とする。そのため、最近の複数物体追跡手法は、物体検出結果を用いた追跡(tracking-by-detection)という処理手順で用いられることが多い。具体的には、すべてのフレームの検出結果が与えられると、追跡器は近い場所で検出されたオブジェクト同士に同じ ID を割り当てる。例えば、Bewley ら[18]は SORT と呼ばれる手法を提案し、動きの情報と、逐次的な速さのカルマンフィルタを用いた動きのモデルを検出による追跡手法を示した。動きの特徴に加えて外観の特徴を考慮するために、Deep-SORT[19]や MOTS[20]といった深層学習ベースの手法が、物体の再分類の一つの方向性として物体の外観の特徴を学習するために組み込まれた。これら深層学習技術を用いた手法は、十分なデータセットが提供される場合において、カルマンフィルタベースの手法よりも良い性能を発揮したものの、リアルタイムの速度で映像を処理する能力には欠けている。本稿では、カルマンフィルタベースの追跡手法を採用し、好ましい処理速度を得るとともに、日本の建築作業者の外観の特徴の差異の不足(似

通ったユニフォームとヘルメットとを着用する場合)に対応する。

2.3. 動作認識

物体の位置や姿勢あるいは軌跡を学習するのとは異なり、動作認識は物体の動きと周囲の環境との相互作用に与えられた名称が有する内的なロジックをデータから学習することを目的とする。CV分野における深層の畳み込みニューラルネットワーク(以下、CNN)の発展とともに、多くのCNNベースの手法は伝統的な動作認識手法[21),22)]の性能と比較して顕著に高い性能を示した。我々は広範にわたるビデオ動作認識手法を2つのジャンル、すなわち2D手法と3D手法、にカテゴライズする。

2D手法は、単眼2D画像におけるCNNの発展を用いて、映像の各フレームにこれらCNNによる分類を適用し、結果を集計して時系列を推定する[23)]。動作内の一時的な細かな動きを考慮するために、外観のモデルとしての特徴と、動的な変化とを分けて取り扱い、冒頭あるいは最後に統合するTwo-stream[24,25)]が提案された。これらの手法の中でSimonyanら[26)]は初めに、オプティカルフロー(連続するデジタル画像間の動きをベクトルで表したもの)を入力として取得する一時的な流れを組み込んだTwo-stream ConvNetアーキテクチャを提案した。Wangら[27)]はTemporal Segment Networksを提案した。これはTwo-stream構造と重みづけられた平均とを融合することで疎な一時的なサンプリング戦略をとる。そのほかの手法、例えばCRFやLSTM[28),29)]は、性能向上への異なるアプローチを示した。

3D手法は、映像から直接3DのCNNによって時空間特徴量を探すものである[30-33)]。それらの中で、C3D[30)]は最初に時空間特徴量を学習するための映像データにおける3Dカーネル(データを高次元の特徴空間に写像するための関数)を提案した手法であり、それは長期間の一時的情報を捉えるために作成された。Carreiraら[31)]は、C3Dを改良したi3Dを提案した。これはImageNetの事前学習2Dカーネルを3Dに発展させ、Two-stream構造のオプティカルフロー情報を利用した。彼らは新たな大規模な動作認識データセットKineticsを公開し、他のベンチマークデータセットとの性能比較を示している。STCNet[32)]はSTCブロックを3D_ResNetに挿入することで、空間的特徴と時間的特徴との間の多チャンネルでの相関を捉えた。Slowfast[33)]は、slow-pathにより空間の意味を捉え、fast-pathにより動作の時系列を高解像度で捉えた。CNN以外では、グラフ問題として動作認識をモデル化しグラフニューラルネットワークによりこの問題を解く手法がある[34)]。本稿では、安定性と複雑なシナリオにおける動作認識への強さを理由として、Two-stream 3D-ConvNet[31)]をベースとする。我々はこの手法を日本の建築現場のシナリオ、すなわち高頻度でオクルージョン(他の物体の後ろに検出対象が一部あるいは全部が隠れること)が発生するとともに複雑な動作を有する場面、において最善の性能が得られるように適応させる。

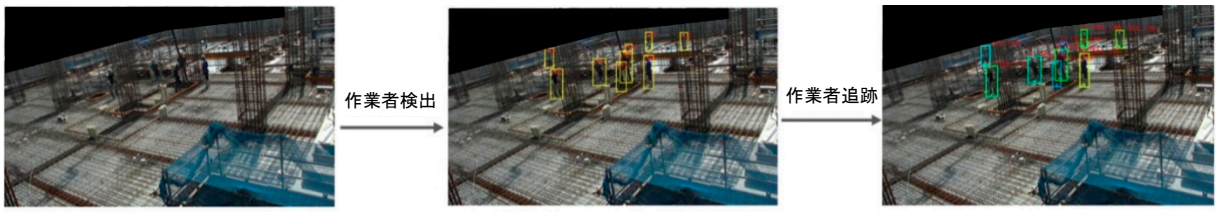
2.4. データ利用

データ利用方法は種々のデータ取得技術とセットで示された。例えば、Eiriniら[35)]は4Dの軌跡データを基に、作業者が作業場所で停止した時間を得て、作業時間として用いたが、移動時間の作業を把握する方法は示していない。Yeら[36)]は作業空間の使用性を把握するために、BLEビーコンにより取得された作業者たちの位置情報を基に時空間ヒートマップを示したが、センサ情報のみでは、ヒートマップで示される位置の意味を理解する困難さを有している。本稿では、現場管理者が作業を振り返る場面で利用できるようデータ変換方法を設計する。

3. 研究手法

3.1. データセット構築

映像に映り込んだ作業者の個人情報保護への配慮として、測定対象とする現場の選定ののちに許可の交渉、映像撮影、データアノテーション(映像に正答とする情報を付与する作業)の手順でデータセットを構築する。アノテーションにおいては、作業者検出の正答としてのバウンディングボックス(作業者を囲む矩形領域、以下Bbox)、作業者追跡の正答としての作業者ID、そして作業者動作認識の正答としての動作種別、の3つを、映像のすべてのフレームに付与する。建築作業と建築現場内の映像に不慣れなアノテータが容易に判別できるように一般的な用語で定義する。本稿では、はじめに多くの作業において要素動作として含まれるであろう基本動作として、歩行(Walk)、しゃがみこむ(Crouch)、立ち上がる(Stand-up)、運ぶ(Carry)、置く(Place)、持ち上げる(Pick-up)、の6つの動作種別を定義した。判断ルールを**表－1**に示す。アノテータには、図表等を併記したマニュアルにより提示する。なお、アノテーションの結果、定義した動作に該当しない“その他



図－1　作業者検出と追跡の処理手順の概要

(other)”が非常に多いことが明らかになり、**表－2** に示す 2 つの静止動作、すなわち、立っている(Standing)としゃがんでいる(Crouching)の判断基準を追加した。

3.2.　作業者検出と追跡

我々の作業者検出と追跡の処理手順を**図－1** に示す。映像から 1 枚の画像が与えられ、初めに作業者検出器が処理し当該画像中の Bbox のリストを出力する。この段階ではすべての Bbox は黄色で描画されている通り、まだそれぞれの作業者の ID が付与されていない。つぎに、すべての作業者検出の結果が作業者追跡器に与えられ、時系列的情報によってそれぞれの作業者の ID が付与される。追跡結果には、それぞれの枠が個別にカラーリングされている通り、作業者の ID を知ることができる。

関連研究で指摘した通り、我々の使用した Faster-RCNN[14]は一般物体認識のために設計されたネットワークである。前景である作業者と背景とを区別するという我々のタスクに適応させるためには、我々は Faster-RCNN の最後の数層の次元と属性を変えて、COCO データセット [37]で定義された 100 近いクラス群から、作業者か否かという 2 クラスのみを出力するかたちに変える必要がある。また、精度よい作業者検出には、より我々のデータセットにおける作業者のサイズとアスペクト比に一致するようにアンカーのサイズとアスペクト比を変える必要がある。最適な性能を得るために、はじめに ImageNet データセット [38]や COCO データセットを学習させて一般物体の特徴量を得る。そのうえで、我々独自のデータセットによるファインチューニングをおこなう。この事前学習が検出性能を向上させた結果については、事前学習なしで直接我々のデータセットを学習させたモデルとの比較実験結果として後述する。

作業者検出器をカスタマイズしたうえで、我々は検出された作業者を経時的に追跡する。具体的には、我々はカルマンフィルタベースの追跡手法である SORT[18]を用いる。これにより、追跡に関する学習

表－1　アノテータの業務定義

[人物の矩形領域(バウンディングボックス)]
・説明：人物を最小サイズで取り囲んだ閉じた四角
・開始条件：肩より上の全て(顔や頭を含む)が見えている、あるいは全身の半分が見えている。
・終了条件：開始条件が維持されなくなる。
・注意：
　・人物の見えている部分をすべて囲むこと。物体に隠れた部分は囲まない。
　・人物がほかの物体に働きかけている場合、その物体を把持している手などの部分は囲うが、把持した物体のすべてを囲わなくてよい。
　・矩形領域は人物を囲う最小となるようにする。多少のずれは許容されるが 10 フレームごとにもっともらしい矩形領域となるように調整すること。

[動作]
・すべての動作種別への注意
　・6 つの動作種別がある。すべて人物ひとりによる動作である。歩行(Walk)、しゃがみこむ(Crouch)、立ち上がる(Stand-up)は、他の物体への働きかけのない基礎動作である。持ち上げる(Pick-up)、置く(Place)、運ぶ(Carry)は、他の物体への働きかけを有する対話的動作である。
　・一人の人物は 6 つの動作種別のうちのひとつに属するか、いずれにも属さない。
　・6 つの動作種別ですべての動作をカバーしていない。人物は“その他(other)”という動作に属することができ、その場合にはラベルを付与する必要はない。
　・定義が満足されたときにのみラベルを付与すればよい。例えば、ある男性がしゃがみ込む動作を全身の 90%が隠れた状態で開始する場合がある。我々はこの場合にラベルを付与する必要はない。例えば、ある男性が持ち上げる動作をする際にその手と持ち上げる物体が完全に隠れているならば、我々はこの動作にラベルを付与する必要はない。

[動作の定義]
・歩行(Walk)
　・説明：人物が何ももの を持たずに歩いている動作。
　・開始条件：人物が動き始め、かつ一歩目の足が床から離れた。
　・終了条件：人物が歩くのをやめた。
　・注意：歩行が明らかであり、少なくとも 2 歩歩いている場合とする。人物がわずかな動きや 1 歩で位置を変える動作は歩行とはみなさない。
・しゃがみこむ(Crouch)
　・説明：人物がしゃがんでいない状態から、体を曲げる、座る、しゃがむ、膝をつく、などをする動作。
　・開始条件：人物の膝、腰、あるいは背中が曲がり始める。
　・終了条件：人物の膝、腰、あるいは背中が動きを止めるか、完全に曲がり終える。
　・注意：この動作が明らかであること。例えば人物が少量の曲げで見下ろすなどの場合はしゃがみこむ動作とはみなさない。
・持ち上げる(Pick-up)
　・説明：人物が物体に接触し、持ち上げ、重力に逆らって物体を把持する動作。
　・開始条件：人物が持ち上げはじめ、物体が重力に反して支持され始める。
　・終了条件：人物が物体の把持を終え、物体が安定した姿勢に至る。
・運ぶ(Carry)
　・説明：人物が幅広い/長い/大きい/重い物体とともに歩行あるいは移動する動作。
　・開始条件：人物が物体とともに歩行あるいは移動を開始する。
　・終了条件：人物が移動をやめる、あるいは物体の重量の支持をやめる。
　・注意：物体は幅広い/長い/大きい/重い物体とする。小さなカバンやリュックサックによる運搬はこの動作とはみなさない。
・置く(Place)
　・説明：人物が物体への接触を無くし、下ろし、物体の重量の支持をやめる動作。
　・開始条件：人物が物体を下ろしはじめ、あるいは置き始める。
　・終了条件：人物が物体との接触を無くす、あるいは物体の重量の支持をやめている、あるいは体の動きが止まる。
・立ち上がる(Stand-up)
　・説明：人物がしゃがみあるいは立っている状態ではない状態から立ち上がる。
　・開始条件：人物の膝、腰、あるいは背中が曲がった状態から回復し始める。
　・終了条件：人物の膝、腰、ある　いは背中が曲がった状態から完全に回復し終える。

表－2　静止動作 2 種の定義

・立っている(Standing)
　上体を起こした状態でひとところに立っている。足は直径 1m の範囲内に少なくとも 2 秒間とどまっている。上体は回転していてよいが、動いていない。
・しゃがんでいる(Crouching)
　足あるいは胴体が合計で少なくとも 135 度曲がっており、その状態を少なくとも 2 秒間保持している。腕は動いていてよいが、胴体は静止している。

の必要性を除外し、かつリアルタイムでのモニタリングへの活用も可能な処理速度とすることができる。画像上の追跡を超えて、我々はホモグラフィ変換技術を用いた2.5Dでの追跡手法も開発する。結果として、我々は平面図上での作業者の軌跡を可視化することができ、現場管理者が作業者の動きを理解しやすくできる。最終的な出力である作業者の時系列の連続した画像という出力結果は、後述する作業者動作認識器のインプットとして使用することで、それぞれの作業者の動作という情報を取得しうる。

3.3. 作業者動作認識

映像からの作業者検出と追跡に引き続き、我々はTwo-stream 3D-ConvNetアーキテクチャによる動作認識に取り組む。これはRGB特徴量を学習するCNN(以下、画像ストリーム)とオプティカルフロー特徴量を学習するCNN(以下、フローストリーム)とによるi3Dを踏襲する。前節で示した通り、作業者検出器の出力から、同じ作業者の連続したBboxを集計したムービークリップを作成し、RGB特徴量の学習に使用することができるが、本稿においてはアノテーションデータを使用する。フローストリームへの入力データはOpenCVに既存の機能を用いて生成する。移動前後のBbox内の画像同士を用いてオプティカルフローを計算すると背景の動きを大きく捉えてしまうため、オプティカルフローは映像全体で計算したのちに、対象のBbox周囲の値を取得する。画像ストリームはRGB画像群を入力とし、フローストリームはオプティカルフロー画像群を入力とする。それぞれの映像群は連続した60フレームから取得される。二つの特徴量マップを取得したのちに、それらの平均をとる。最終的に、9つの動作種別の該当確率(例えばWalk80.0%, Carry15.0%, Standing5.0%, 他すべて0.0%、のように計100%とする値)が推定結果として出力される。

学習データを増やすために、水平フリップとランダム移動によるデータ拡張をおこなう。同じ映像サンプルには同じ拡張手法が適用される。

3.4. データ利用

現場管理者が作業を振り返る場面で利用できるように、データ変換により得られる作業要素を設計する。現場管理者が日々の進捗が想定通りかどうかを把握するためには、現場内の作業者の行動について５Ｗ１Ｈの情報が必要となる。本稿では**表－3**に示す通り、“いつその作業がなされたか”を動画ファイルの持つ時刻データ、“誰がその作業をおこなったか”を作業者追跡が付与する作業者IDデータ、“どこでその作業者は作業をおこなったか”を作業者検出結果の2D座標と固定カメラの姿勢から計算する3次元座標、“その作業はどのようなものか”を時刻と位置からリスト検索により得る作業名、そして“作業者はどのようにその作業を進めたか”を作業者動作認識の結果、として設計する。“なぜその作業をおこなったか”については計画に従った建物の完成という自明な理由である場合と、計画外の複雑な事情による理由である場合とがあり、その推定は今後の課題とし、本稿ではデータを確認する現場管理者の主観に委ねるものとする。なお、固定カメラの姿勢から3次元座標を計算するためには、床面等の基準平面を見下ろす位置へのカメラ固定が条件となる。

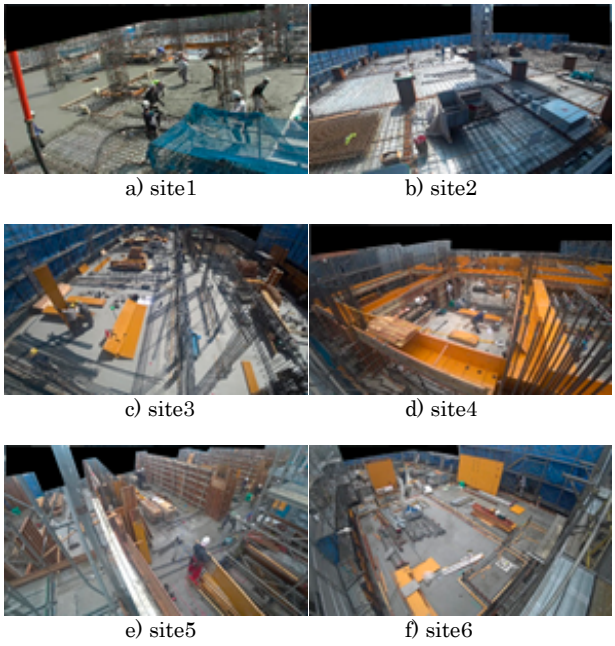
表－3 作業要素の取得方法

When	“いつその作業がなされたか” ・動画ファイルの持つ時刻データ
Who	“誰がその作業をおこなったか” ・作業者追跡が付与する作業者IDデータ
Where	“どこでその作業者は作業をおこなったか” ・作業者検出結果の2D座標と固定カメラの姿勢から計算する3次元座標
What	“その作業はどのようなものか” ・時刻と位置からリスト検索により得る作業名
Why	“なぜその作業が実施されたのか” (・現場管理者の主観による判断)
How	“作業者はどのようにその作業を進めたか” ・作業者動作認識の結果

4. 結果と考察

4.1. データセット構築

1)現場の選定：当社が管理する建築現場から、躯体工事中の6つの現場(site1～site6)を選定した。
2)許可の交渉：現場管理責任者ならびに当該現場が構築する建物の発注者の担当者から映像撮影の許可を得た。現場内の作業者には映像を撮影する事実と撮影の目的とを書面により掲示した。今回の撮影により取得した映像に対して、作業者の映像と作業者個人を特定する情報である名前とは紐づけない。
3)映像の撮影と映像の選出：撮影は2つのカメラ(Sony-FDRX3000, JVC-GZRY980)を用いておこない、建物の外周に構築された仮設足場に、金属製の治具を用いて固定した。多種の作業が撮影されるように、複数の設置場所を移動させながら撮影した。撮影後に映像を確認し、選出と編集により、17の場面の52の映像ファイルを作成した。6つの現場の代表的な映像から画像を**図－2**に示す。**図－2**に示す通り、site1は晴れの日のコンクリート打込み作業、site2は晴れの日の床配筋作業、site3は晴れの日の型枠運搬作業、site4とsite5は曇りの日の型枠関連



a) site1　b) site2　c) site3　d) site4　e) site5　f) site6

図－2　撮影対象現場のサンプル画像

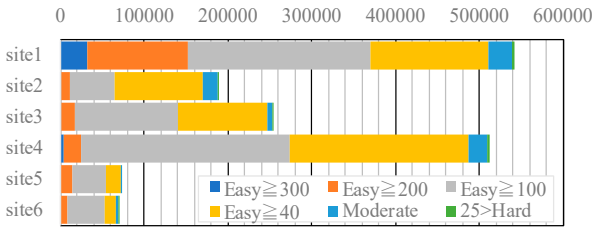
表－4　現場ごとの映像の概要

	天候	映像数	フレーム数	FPS	長さ
site1	晴れ	9	50535	29.97	28’06”
site2	晴れ	5	30591	29.97	17’01”
site3	晴れ	7	61695	29.97	34’19”
site4	曇り	9	64637	29.97	35’57”
site5	曇り	4	24825	29.97	13’48”
site6	雨	18	38565	29.97	21’27”
計	-	52	270848	-	150’37”

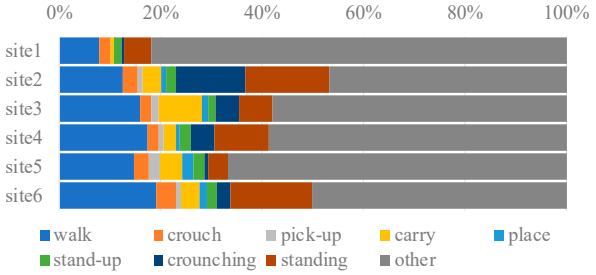
作業、site6は雨の日の型枠設置作業と仮設足場構築作業であった。**表－4**は現場ごとの映像の長さを示している。フレーム数は合計で27万フレームであり、合計で2.5時間分の映像であった。

4)作業者の位置と属性のアノテーション：アノテーションにより付与されたBbox数の集計結果を**図－3**に示す。site1とsite4は映像中に10名以上の作業者が映り込んでいたため、Bbox数が多くなっていることがわかる。図中の色分けは、Bboxの縦のサイズ(ピクセル数)で塗り分けている。ほとんどの作業者の画像サイズは40ピクセル以上の“Easy”のカテゴリに含まれることがわかる。また、動作種別の内訳を**図－4**に示す。“その他”カテゴリが半数以上を占めており、6つの動作種別もバランスが悪いことがわかる。

将来的なデータセット構築においてはデータ数のバランスのために、映像の長さを映像に映り込んだ人数を考慮して選出することが重要であり、動作種別の定義においては動作タグの付与数がアンバランスにならないよう配慮した定義あるいは撮影シーンの選定をすることが重要だといえる。



図－3　現場ごとのデータ数とサイズの内訳



図－4　現場ごとの動作種別の内訳

4.2.　作業者検出の評価

我々の作業者検出器の性能を、標準的な測定基準である、IoU(Intersection of union)が0.5以上であることをしきい値とした平均検出精度(AP: Average Precision)を用いて評価する[注1)]。評価に当たっては、正答データ(GT：Ground Truth)に含まれる極小の作業者画像により、検出精度が過小評価されることが懸念されるため、3段階の難易度レベルを縦のピクセル数40, 25, 1をしきい値として設けた。例えば、Easyレベルでの評価においては、ピクセル数が40に満たないGTは除外して、よりカメラに近く検出が容易な作業者のみを用いて評価を行う。我々のデータセットで学習した我々の検出器を、COCOデータセットで事前学習したFaster RCNN[14]とmask RCNN[39]と比較した。

評価においては2種類のデータ分割方法を用いた。本稿では全現場評価と交差現場評価と呼称する。全現場評価では、すべての映像の頭から70%を学習に用い、次の15%を検証に、後ろの15%を性能評価に用いた。全現場検証の結果を**表－5**に示す。**表－5**から明らかなとおり、カスタマイズなしの一般物体検出器では建設現場の作業者検出ではよい成績を示さず、我々の検出器よりも低い検出精度となった(Easyレベルで Mask RCNNが33.5%に対して、我々の検出器は66.8%)。また、我々の検出器に対して一般物体データセットCOCOを事前学習させ、我々のデータセットでファインチューニングすることにより、検出精度は向上(Easyレベルで66.8%か

ら87.9%に向上)した。このことは、多くの一般物体を含んだCOCOデータセットによる事前学習は、建設現場映像から作成したデータセットにおける作業者検出性能の向上を助ける働きがあることを示している。交差現場評価では、異なる現場間で特徴量を学習できるのかを評価している。これは、まず一つの現場(本稿においてはsite1の9つの映像)をテストデータとして選択し、残りの現場の映像を学習と検証に用いるというものである。この方法により、テストデータと同等の映像が学習時に使用されることがないため、検出の難易度は全現場評価の場合よりも高くなる。結果を**表－6**に示す。**表－6**からわかる通り、我々の検出器は学習していない建設現場の映像に対しても使用に耐える検出精度(Easyレベルで77.5%)を示していることがわかる。このことは、将来的に建設現場管理者が新しい現場で収集した映像に対しても我々の検出器を使用できる可能性を示している。また、全現場評価と同様に、我々の検出器は一般物体検出器Faster RCNNよりも良い性能(Easyレベルで50.7%に対して77.5%)を示した。

表－5　全現場評価における性能

Method / AP	Easy (>=40)	Moderate (>=25)	Hard (>=1)
Mask RCNN (COCO)	31.7	31.3	31.3
Faster RCNN (COCO)	33.5	32.1	31.9
Our Detector (Ours)	66.8	66.5	66.4
Our Detector (COCO+Ours)	**87.9**	**87.4**	**86.2**

表－6　交差現場評価における性能

Method / AP	Easy (>=40)	Moderate (>=25)	Hard (>=1)
Faster RCNN (COCO)	50.7	48.7	48.3
Our Detector (COCO+Ours)	**77.5**	**72.8**	**72.7**

4.3.　作業者動作認識

作業者検出と同様に、我々のデータセットを用いて作業者動作認識器の性能評価を行った。具体的には、はじめにKineticsデータセットで事前学習されたモデルを我々のデータセットでファインチューニングした。データセットはランダムに80%を学習セットとし、20%を評価セットとした。動作種別ごとの再現率(Recall)を評価対象とし、動作認識器全体の精度はすべての動作種別(6つの動作＋2つの静止動作＋その他、計9種)の平均をもって検出精度とした。作業者の動作の周辺環境を考慮するため、GTのBboxを周囲に10%拡張することで、背景をより多く含めた映像を使用した。学習率は0.0005とし、重み減衰率0.0001とした。バッチサイズは4とした。動作種別ごとの例を**図－5**に示す。実際のデータは60フレームからなるビデオ形式であり、**図－5**は各ビデオクリップが取得されている作業者周囲を正方形に切り出した例であることに注意されたい。

図－5　動作種別9種の例

性能評価においては、該当確率一位の動作種別を回答として採用した。評価の結果を**表－7**に示す。**表－7**に明らかなとおり、我々の動作認識器は一般的なST-GCN[34]やi3D[31]よりも良い性能を示した(平均精度33.1%と54.0%に対して60.2%)。ただしST-GCNの性能はデータセットのラベリングにStandingとCrouchingのない構成の際に評価された値である。

平均精度で最も良い性能を示した我々の動作認識器においてもその性能は60.2%であり、今回の計9種の動作を明確に認識できているとは言えない。Otherの動作種別においては、我々の認識器は最も悪い性能を示しており、これは我々が再現率を評価に用いたためである。未定義のOtherの再現率が高いということは、一般的にそのほかのクラスにおける高い誤検知(False-Positive)の結果によるものである。また、Placeクラスの性能が極めて低いことに関しては、我々のデータセット内の動作種別の

表－7　既往の動作認識器との性能比較

Method	Mean Class Accuracy	Walk	Crouch	Stand-Up	Pick-Up	Carry	Place	Standing	Crouching	Other
ST-GCN	33.1	34.7	49.2	49.5	14.7	8.1	**18.7**	-	-	**57.2**
i3D	54.0	**80.7**	37.2	85.9	38.2	**79.6**	5.6	73.8	30.3	55.0
Our Method	**60.2**	75.5	**84.1**	**90.5**	**59.0**	61.1	6.1	**82.2**	**34.3**	50.0

表－8　片方の特徴量のみを考慮した場合との比較

Method	Mean Class Accuracy	Walk	Crouch	Stand-Up	Pick-Up	Carry	Place	Standing	Crouching	Other
RGB	51.6	44.0	82.0	86.0	56.0	**77.0**	0.0	71.0	16.0	28.0
Flow	56.2	68.0	82.0	83.0	49.0	40.0	**18.0**	77.0	**39.0**	48.0
RGB + Flow	**60.2**	**75.5**	**84.1**	**90.5**	**59.0**	61.1	6.1	**82.2**	34.3	**50.0**

データ量ならびに複雑性の不均一が原因である。我々のデータセットからは、現実の建設現場における Place の発生頻度の低さ(例えば Walk や Standing と比較して)を読み取ることができる。加えて Place の動作は 1 秒にも満たない短い時間で実施されていた。これらの事実が Place の検出の困難さに影響していると考えられる。

画像ストリームとフローストリームとのそれぞれがどの動作認識に寄与しているかを明らかにするために、除去実験(Ablation study)を実施した。結果を**表－8**に示す。はじめの 2 行はそれぞれ画像ストリームのみ(RGB)、あるいはフローストリームのみ(Flow)を用いて最終的な統合をしない場合の結果である。これにより、平均精度においても動作種別ごとの再現率においても、片方のネットワークのみを用いた場合の方が性能を落としている場合がほとんどであることがわかる。しかし片方のストリームのみの結果を比較することにより、Pick-up や Carry といった他の物体の操作を含む動作においては、画像の特徴を用いる方がよい性能を示すことがわかる。一方で、Walk や Place といった動的特徴の強い動作は、動きの特徴の学習により認識性能が高められていることがわかる。

4.4.　データ利用

前節に示した通り、作業者動作認識器の検出性能には不安が多いことから、本稿におけるデータ変換技術の実装は“How：作業者はどのようにその作業を進めたか”の要素を除いておこなった。我々の手法を用いて変換されたデータの例を**図－6** に示す。**図－6** で作業者を囲う Bbox は作業者検出器の出力をそのまま描画した枠であり、作業者 ID(図中の id001)は追跡器からの出力値をそのまま描画している。Bbox 内の斜めの直線は、計算により求められた実空間上の床からの垂直線であり、求めた XYZ 座標(図中の X:11.3m,Y:16.5m,Z:0.0m)と作業者の身長(図中の h:1870mm)も併せて描画している。なおこれらの数値は、Bbox の中心ピクセルが、床から作業者の身長の半分の高さにあると仮定して身長 2500 ㎜から 10mm ずつ減じながら反復計算し、Bbox 内に垂直線が収まった際の値を採用している。Bbox が作業者周囲にバッファを有しているため、実際の作業者よりも大きな身長として出力されている。垂直線の下端から伸びている太い線は当該作業者の移動の軌跡である。同一の作業者 ID を有したデータを、数フレーム前まで検索して位置を描画することで表現した。作業名(図中の Work:Board)については、変換システムが別途有する“工区×XYZ 座標対応表”(CSV 形式のテキストデータ)と“工区×作業対応表”(CSV 形式のテキストデータ)とを用いて、日付と XYZ 座標により検索して取得している。目視により、対応表から正しく検索されていることが確認された。

実装した変換計算プログラムにより、いつ・どこで・誰が・何をした、という作業要素がデータとして取得できることが確認された。データの集計により、工区ごとの作業人数、単位時間ごとの人数の推移をグラフとして取得可能であることも確認されており、これによって、現場管理者はグラフから全体を振り返り、気になる点を見つけた場合には、別途保存した映像に立ち返り詳細を振り返ることが可能となったといえる。

a) 変換前　　b) 変換後

図－6　作業者検出と追跡結果からのデータ変換例

5. 結言

建築生産のデジタル化に向けた AI 活用の事例として、現場管理者が作業要素を振り返り可能となるデータ取得手法を構築した。我々は日本の建築現場内の映像からデータセットを構築し、建築現場に適

用させた作業者検出と追跡アルゴリズムならびに作業者動作認識アルゴリズムを提案した。性能評価により、我々の作業者検出器は全現場評価において87.9%、交差現場評価において77.5%の検出精度を示し、我々の作業者動作認識は平均精度で60.2%を示した。最後に作業者検出結果を用いたデータ変換手法を設計し実装例が示された。提案手法が実務において満足した性能を有しているか否かは今後の検証により確認する必要がある。

本稿の制約として、この計測手法は単眼カメラを用いるため、オクルージョンには対応できない。これは、複数カメラを使用して死角をなくすことで対応できる。実務での利便性を考慮すれば、複数カメラ間でのデータ連携の仕組みも必要となる。

現場管理者が施工計画に必要とする情報をデジタルデータで取得し分析する技術を構築することにより、より高度で生産性の高い建築生産が実現するものと考える。当該技術構築に向けて画像処理AIの活用可能性を引き続き検討する。

謝辞

本稿は米国カーネギーメロン大学との共同研究を通して得た成果をまとめたものである。同校のKris Kitani准教授、Xinshuo Weng, Yunze Manにここに感謝いたします。また現場計測にご協力いただいたご関係者の皆様にもここに感謝いたします。

＜注釈＞

注1) 追跡器の性能については検出器の未検出・誤検出の影響を含み、相互の分割が困難である点もあり、本稿では割愛する。

＜参考文献＞

1) 建設業ハンドブック 2019, (一社)日本建設業連合会, https://www.nikkenren.com/publication/pdf/handbook/2019/2019_04.pdf
2) Tarek O. and Moncef N.: Data Acquisition Technologies for Construction Progress Tracking, Automation in Construction 70, pp.143-155, 2016
3) Amin A. et al.: Sensor-based Safety Management, Automation in Construction 113, 103128, 2020
4) 前田邦夫: 現代アメリカ建設学, 開発問題研究所, 1987.02
5) Jun Y. et al: Automatic Recognition of Construction Worker Activities Using Dense Traj ectories, 2015 Proc. of the 32nd International Symposium on Automation and Robotics in Construction (ISARC), pp.1-7, 2015
6) Kaijian L. and Mani G.: Crowdsourcing Video-based Workface Assessment for Construction Activity Analysis, 2015 Proc. of the 32nd International Symposium on Automation and Robotics in Construction (ISARC), pp.1-8, 2015
7) Mohammad S. et al: Evaluating the Performance of Convolutional Neural Network for Classifying Equipment on Construction Sites, 2017 Proc. of the 34rd International Symposium on Automation and Robotics in Construction (ISARC), pp.509-516, 2017
8) Mingzhu W. et al: Predicting Safety Hazards Among Construction Workers and Equipment Using Computer Vision and Deep Learning Techniques, 2019 Proceedings of the 36th International Symposium on Automation and Robotics in Construction (ISARC), pp.399-406, 2019
9) N. Dalal and B. Triggs: Histograms of Oriented Gradients for Human Detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp.886-893 vol. 1, 2005
10) D. Lowe: Distinctive Image Features from Scale-invariant Keypoints. International Journal of Computer Vision (IJCV), Vol.60, pp.91-110, 2004
11) P. Felzenszwalb, et al.: A Discriminatively Trained, Multiscaled, Deformable Part Model, 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1-8, 2008
12) R. Girshick, et al.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.580-587, 2014
13) R. Girshick: Fast R-CNN, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) pp.1440-1448, 2015
14) S. Ren, K. He, R. Girshick, and J. Sun.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Advances in Neural Information Processing Systems 28 (NIPS 2015), pp.91-99, 2015
15) T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie: Feature Pyramid Networks for Object Detection, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936-944, 2017
16) Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T. et al. Selective Search for Object Recognition. Int J Comput Vis Vol.104, pp.154–171, 2013
17) J. Dai, Y. Li, K. He, and J. Sun.: R-FCN: Object Detection via Region-based Fully Convolutional Networks, Proceedings of the 30th International Conference on

Neural Information Processing Systems (NIPS 2016), pp.379-387, 2016

18) A. Bewley, G. Zongyuan, F. Ramos, and B. Upcroft: Simple Online and Realtime Tracking, IEEE International Conference on Image Processing (ICIP), pp.3464-3468, 2016

19) N. Wojke, A. Bewley, D. Paulus: Simple Online and Realtime Tracking with a Deep Association Metric. IEEE International Conference on Image Processing (ICIP), pp. 3645-3649, 2017

20) P. Voigtlaender et al.: MOTS: Multi-Object Tracking and Segmentation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.7934-7943, 2019

21) H. Wang, A. Kläser, C. Schmid and C. Liu: Action Recognition by Dense Trajectories, 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3169-3176, 2011

22) H. Wang and C. Schmid: Action Recognition with Improved Trajectories, 2013 IEEE International Conference on Computer Vision (CVPR), pp.3551-3558, 2013

23) A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei: Large-Scale Video Classification with Convolutional Neural Networks, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1725-1732, 2014

24) Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman: Convolutional Two-stream Network Fusion for Video Action Recognition, Conference on Computer Vision and Pattern Recognition (CVPR), pp.1933-1941, 2016

25) Chen, Yunpeng, et al.: A^ 2-nets: Double Attention Networks, Advances in Neural Information Processing Systems (NeurIPS 2018), pp.352-361, 2018

26) Simonyan, Karen, and Andrew Zisserman.: Two-stream Convolutional Networks for Aaction Recognition in Videos, Advances in Neural Information Processing Systems (NIPS 2014), pp.568-576, 2014

27) Wang L. et al.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, European Conference on Computer Vision (ECCV2016), vol 9912, pp.20-36, 2016

28) G. A. Sigurdsson, S. Divvala, A. Farhadi and A. Gupta: Asynchronous Temporal Fields for Action Recognition, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.5650-5659, 2017

29) J. Donahue et al.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2625-2634, 2015

30) Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri: Learning Spatiotemporal Features With 3D Convolutional Networks, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.4489-4497, 2015

31) Joao Carreira, Andrew Zisserman: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.6299-6308, 2017

32) Diba A. et al.: Spatio-temporal Channel Correlation Networks for Action Classification. European Conference on Computer Vision (ECCV2018), pp.299-315, 2018

33) C. Feichtenhofer, H. Fan, J. Malik and K. He: SlowFast Networks for Video Recognition, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp.6201-6210, 2019

34) Sijie Yan, Yuanjun Xiong, Dahua Lin: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, AAAI Conference on Artificial Intelligence, 2018

35) Eirini K. and Ioannis B.: Trajectory-Based Worker Task Productivity Monitoring, 2018 Proc. of the 35th International Symposium on Automation and Robotics in Construction (ISARC), pp.1145-1151, 2018

36) Ye H. et al: A Visualization System for Improving Managerial Capacity of Construction Site, 2017 Proceedings of the 34rd International Symposium on Automation and Robotics in Construction (ISARC), pp.388-395, 2017

37) T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. Zitnick: Microsoft COCO: Common objects in context, European Conference on Computer Vision (ECCV), pp.740-755, 2014

38) J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei.: ImageNet: A Large-Scale Hierarchical Image Database. 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.248-255, 2009

39) K. He, G. Gkioxari, P. Doll´ar, and R. Girshick: Mask RCNN. 2017 IEEE International Conference of Computer Vision (ICCV), pp.2980-2988, 2017