第1回 AI / Analytics カンファレンス
『間違いだらけのAI導入失敗から生まれる目からウロコのAI活用～AIの使い方次第で、DXの妄想スパイラルから抜け出せる～』

～動画像認識AIの課題と実用化のポイント～

2022/02/15

樋口未来
グローバルウォーカーズ株式会社　取締役CTO

Global Walkers

Global Walkers株式会社概要

世界で通用する、ビジネスで「使えるAI」を実現することを目的とし、2016年6月に創業いたしました。AI（機械学習/ディープラーニング）とコンピュータビジョン（画像処理技術）を中心としたテクノロジーのノウハウを保有し、自社技術開発・基礎研究を通じて
3次元姿勢推定、文字認識アプリケーションの継続的な開発、
またAIの学習に極めて重要な教師データ作成に注力しています。
これらAI技術とデータ作成サービスをもとに「使えるAI」を実現いたします。

名称　：Global Walkers株式会社
役員　：森川和正　代表取締役社長
　　樋口未来　取締役CTO
　　小松徹　　取締役
　　天野哲也　監査役
　　六川浩明　監査役
従業員：日本国内体制25名
　　(非常勤・アルバイトを含む)
　　ミャンマー現地法人（従業員50名 ）
住所　：
　　東京都品川区西五反田2丁目25-2飯嶋ビル5階
資本金：7182万5000円
子会社：Global Walkers (Myanmar) Co., Ltd.

取引先/納入先

株式会社ALBERT
いすゞ自動車株式会社
伊藤忠テクノソリューションズ株式会社
カシオ計算機株式会社
京セラ株式会社
シャープ株式会社
株式会社ゼンショーホールディングス
セントラル警備保障株式会社
株式会社大都製作所
大日本印刷株式会社
株式会社電通
株式会社電通国際情報サービス
凸版印刷株式会社
トヨタ自動車株式会社
トヨタ紡織株式会社
日産自動車株式会社
株式会社NTTデータ
株式会社パスコ
株式会社日立製作所
株式会社日立ソリューションズ
フォワードシステム株式会社
富士ソフト株式会社
株式会社フューチャースタンダード
株式会社本田技術研究所
三菱UFJリサーチ&コンサルティング株式会社
国立研究開発法人産業技術総合研究所
国立研究開発法人海洋研究開発機構
一般財団法人 首都高速道路技術センター
一般社団法人 日本国際紛争解決センター
地方独立行政法人東京都立産業技術研究センター
有限責任監査法人トーマツ
国立大学法人東京大学
国立大学法人京都大学
学校法人早稲田大学

順不同

Global Walkers

自己紹介

樋口未来(ひぐち・みらい)
プロフィール

日立製作所 日立研究所に入社後、自動車向けステレオカメラ、監視カメラの研究開発に従事。
2011年から1年間、米国カーネギーメロン大学にて客員研究員としてカメラキャリブレーション技術の研究に携わり、2016年にグローバルウォーカーズ株式会社を創業。
東京大学大学院博士課程 単位取得済み退学。

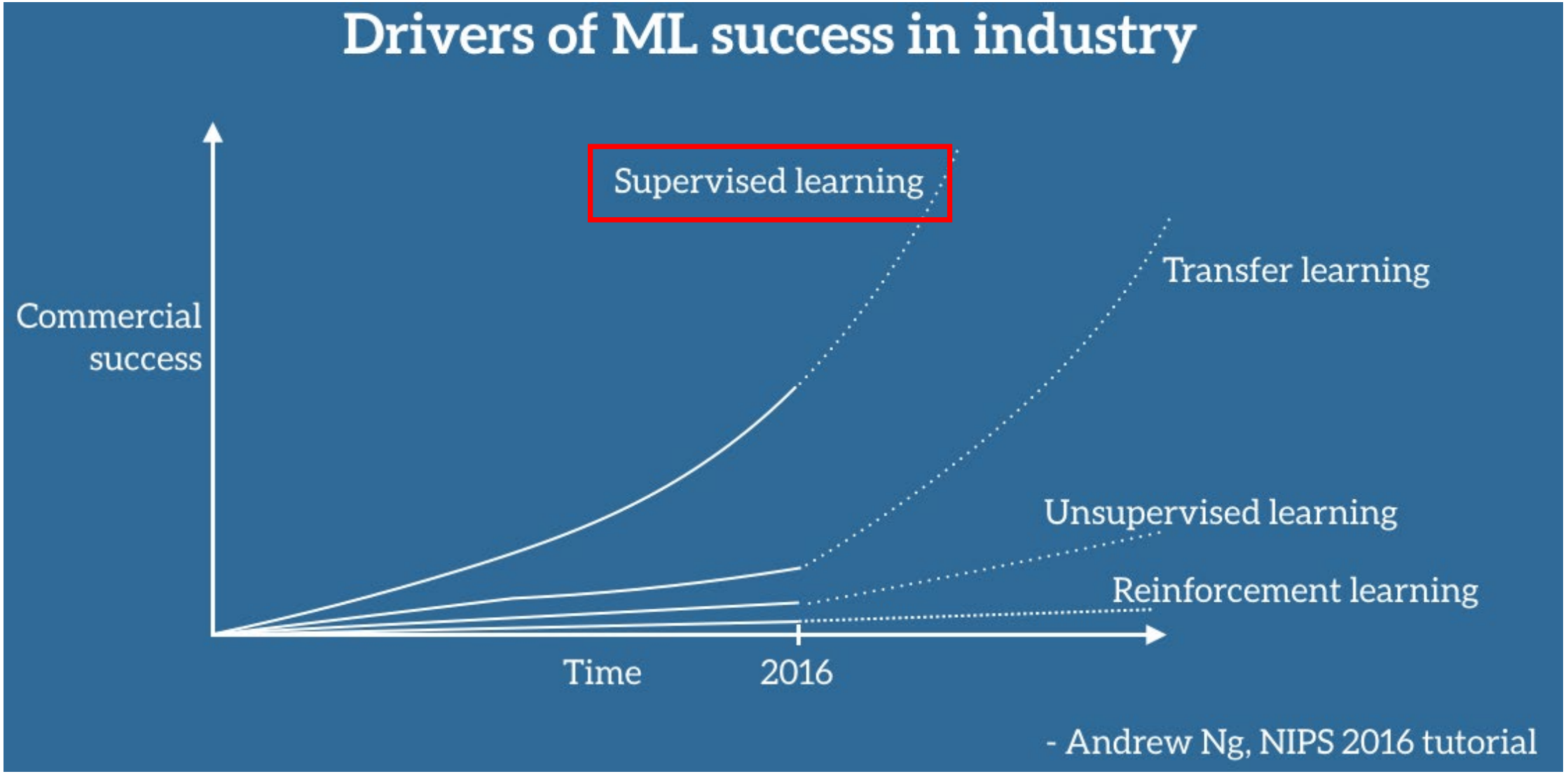
専門：　コンピュータビジョン、機械学習
プログラミング言語：　Python, C, C++, Matlabなど
寄稿：マイナビニュース連載
「機械の目が見たセカイ　-コンピュータビジョンがつくるミライ」
http://news.mynavi.jp/series/cv_future/menu.html
趣味：　バイク、自転車、スノーボード、料理、など
twitter: @at_mirai

コンピュータが目を手に入れることで飛躍的に進化？

ロボットの目（物を見つける、人を見つける、人の表情を認識するなど）
人間が外界から得る情報の8割以上が、視覚から
⇒コンピュータの目を実現できれば、応用範囲も爆発的に広がる

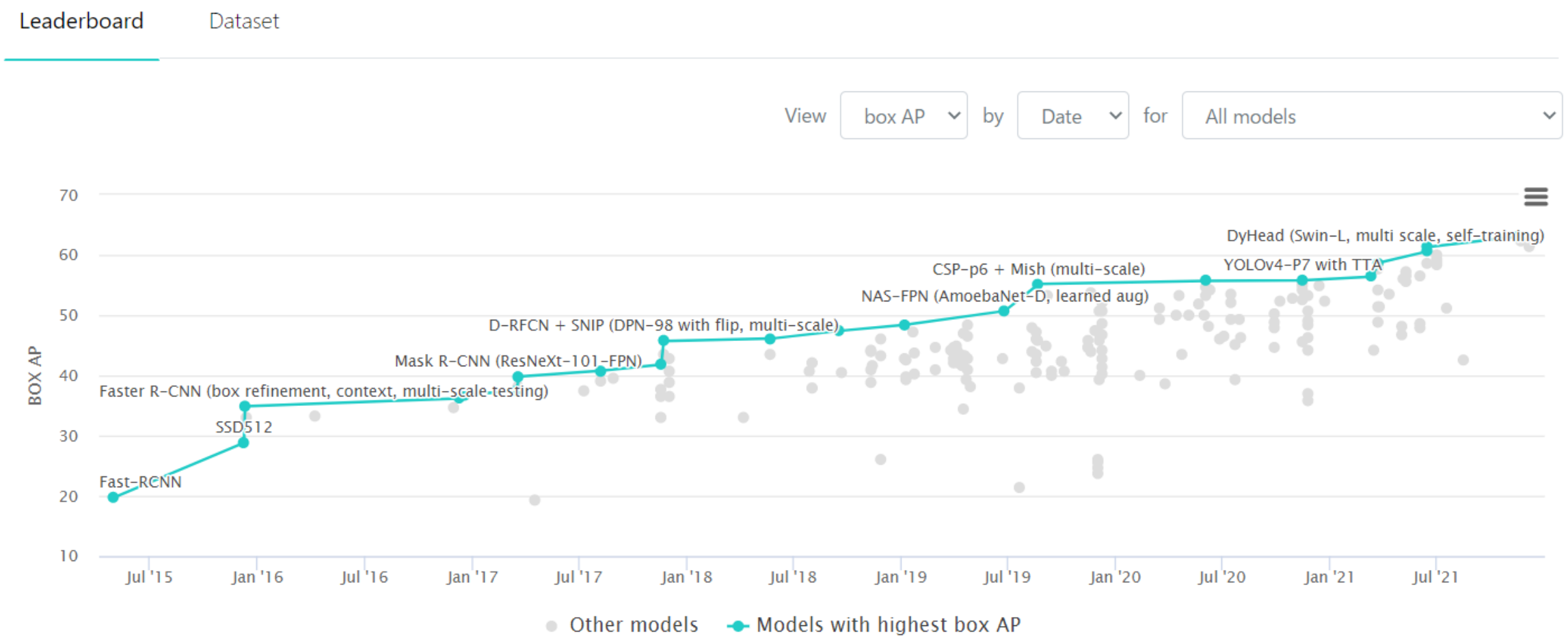
本日のテーマ

機械学習とは？



物体検出・姿勢推定では差別化が難しくなった？

物体検出・姿勢推定のAIは、この5年程で飛躍的に進化



[COCO test-dev Benchmark (Object Detection) | Papers With Code]

物体検出・姿勢推定では差別化が難しくなった？

物体検出・姿勢推定のAIは、この5年程で飛躍的に進化　⇒　差別化が困難

これらの例では、動画像を1フレームずつ処理しているのみであり、
物体の追跡、動作の認識などは行っていない

発表内容

1．動画像認識の課題

2．動画像認識の例
　2．1．トラッキング
　2．2．動作認識

3．実用化のポイント

１．動画像認識の課題

・データサイズ
- Full HDかつRGBだと約600万byte (6MB)/image
  30fpsとすると180MB/sec
- 30fpsの動画100時間は1080万枚の画像からなる

・複数の対象が存在する場合に対応付け（トラッキング）が必要
- 動作認識を行うためにはそれぞれの人をトラッキングする必要がある

・認識対象の曖昧さ
人の検出・顔の検出などと異なり認識対象が曖昧かつ多種多様
例：　動作の開始フレームが曖昧
例：　食事するという動作は複数の細かい動作から構成
　　（箸を手に取る、食材を箸で挟む、食材を口元に運ぶ、・・・）
例：　認識したい動作は様々
　　工場では「ネジ閉め」など、飲食店では「コーヒーを注ぐ」など

発表内容

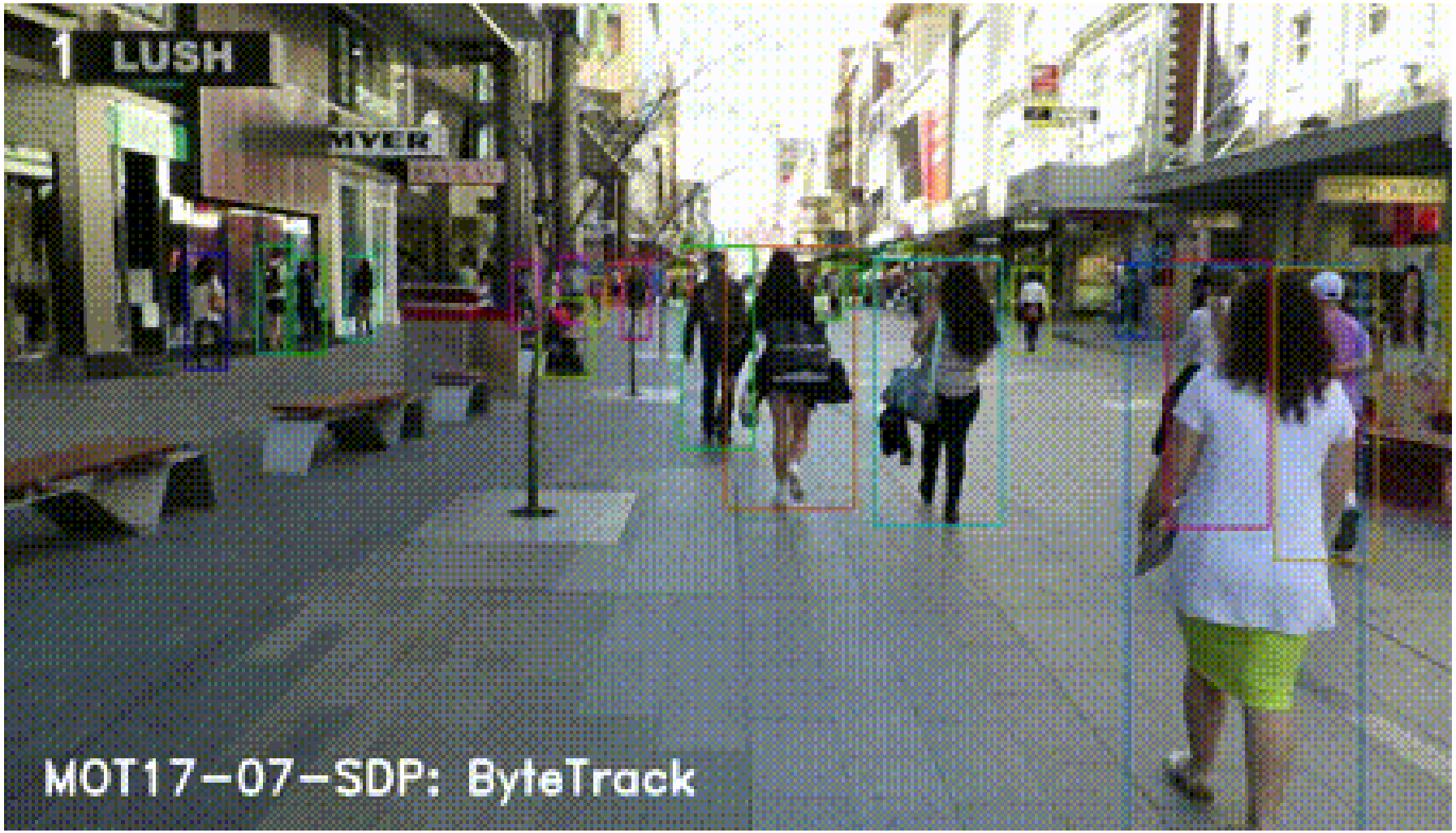
1．動画像認識の課題

2．動画像認識の例
　2．1．トラッキング
　2．2．動作認識

3．実用化のポイント

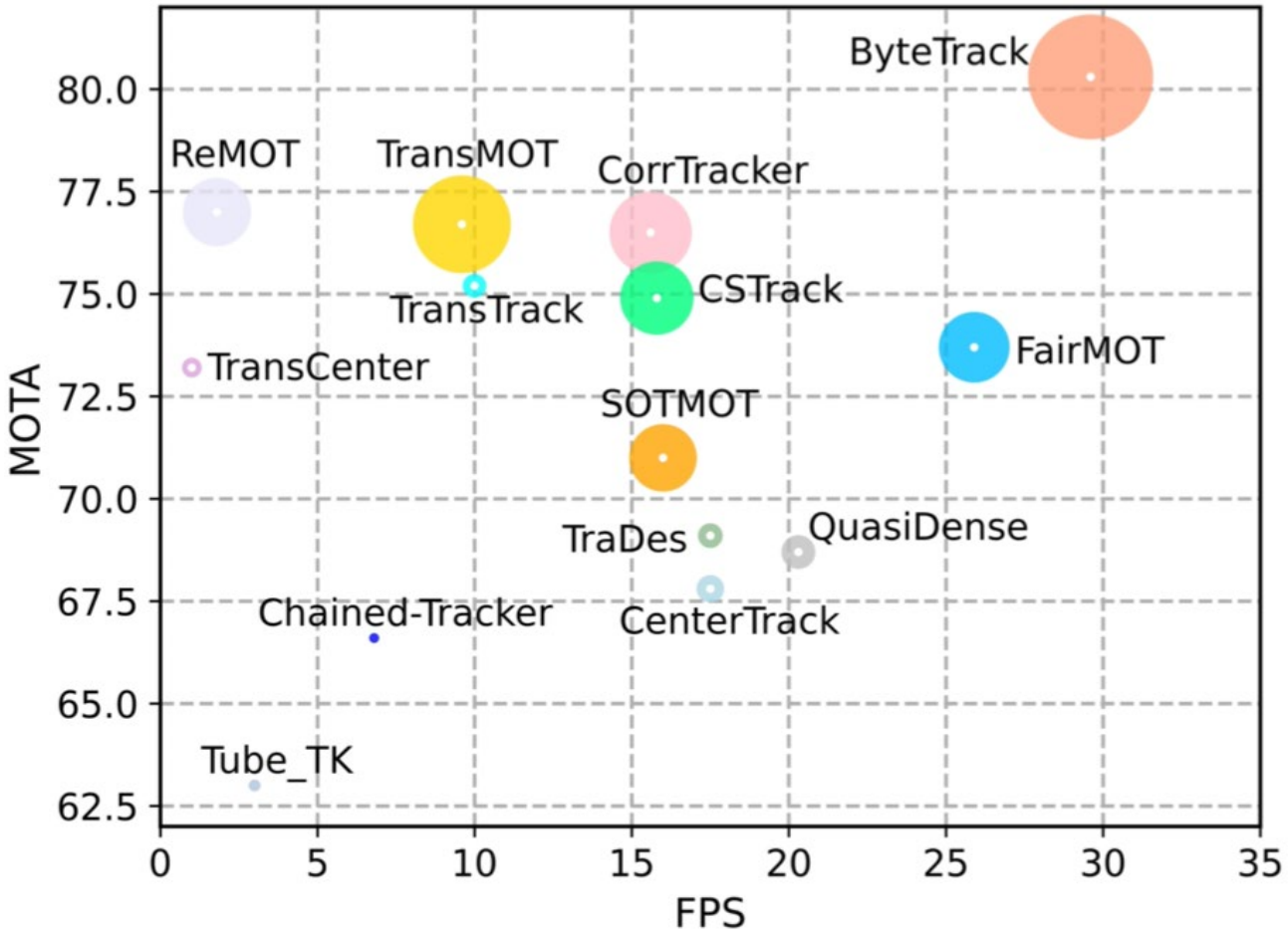
２．１．トラッキングとは

異なるフレームの同じ物体（人）に同一のIDを付与
※下記動画では同じIDを同じ色で可視化



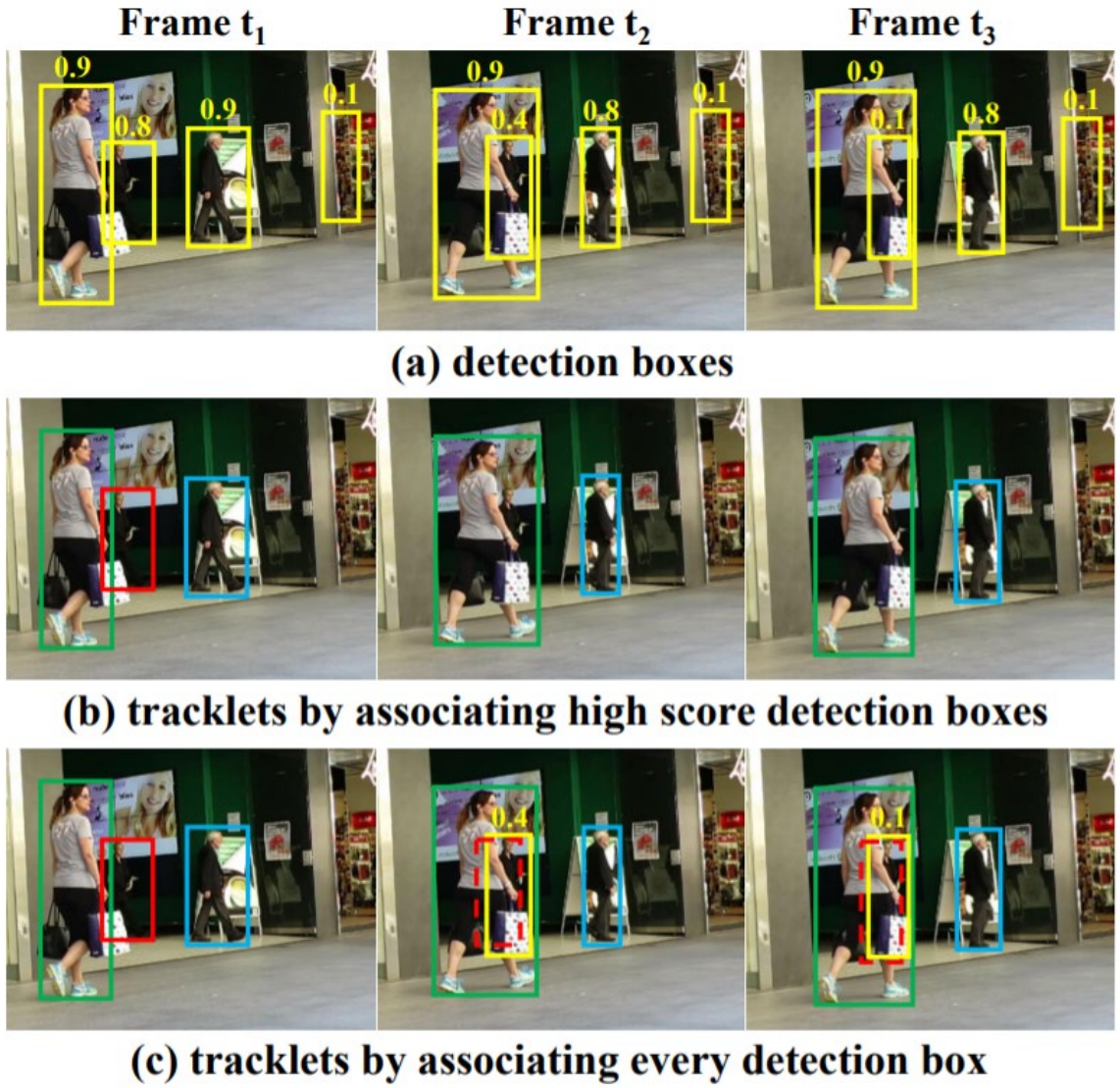
２．１．トラッキングとは

動画像処理では処理速度も重要　⇒　2021年に発表されたByteTrackを紹介



https://arxiv.org/abs/2110.06864

２．１．ByteTrack

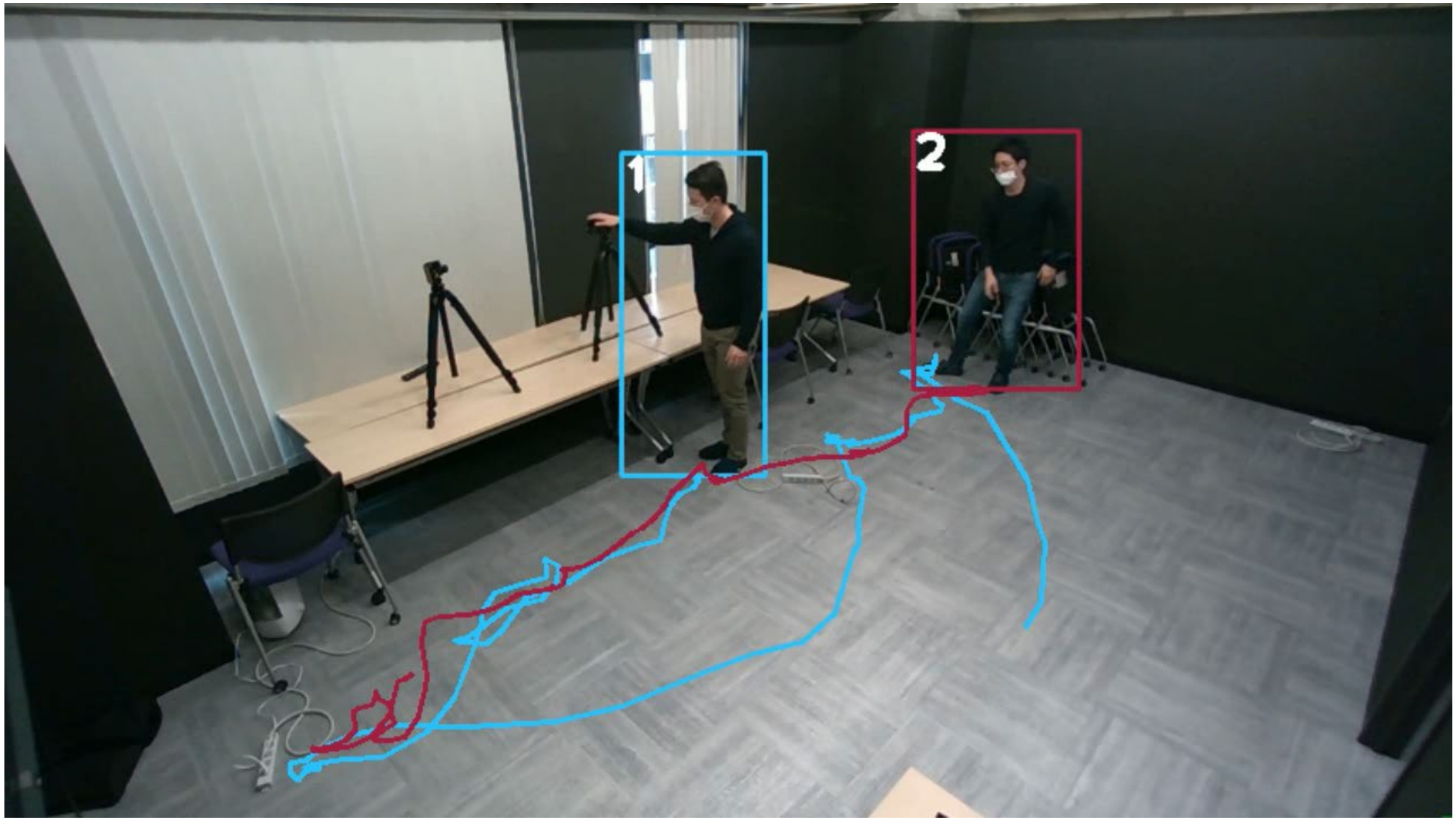


【STEP1】
物体検出モデルにより人を検出

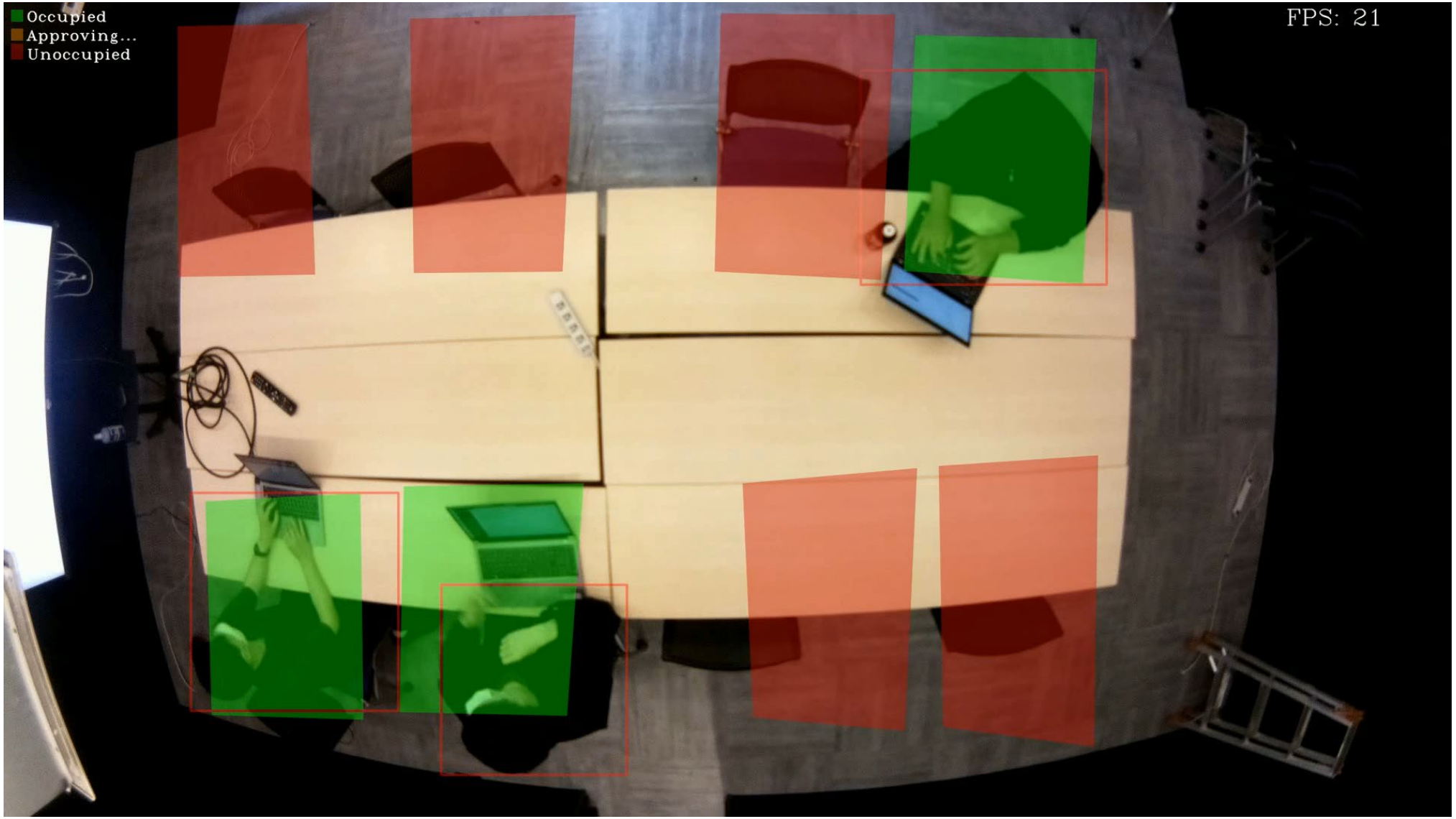
【STEP2】
スコアの高い検出結果のみを
過去のトラッキング結果と対応付け

【STEP3】
スコアの低い検出結果のみを
過去のトラッキング結果と対応付け

２．１．トラッキングによる動線の計測



２．１．トラッキングによる滞留の検出



発表内容

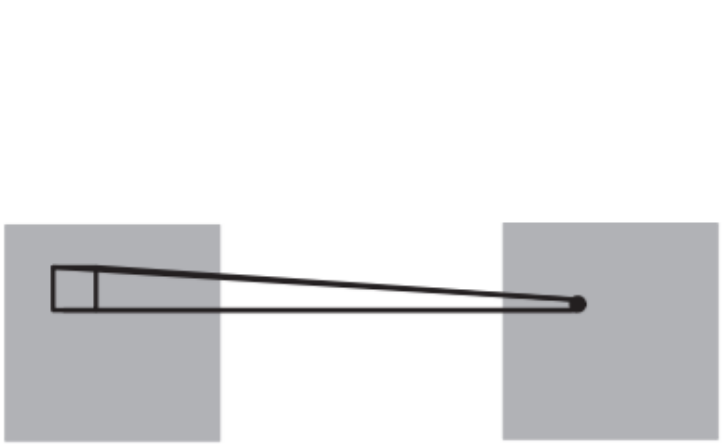
1．動画像認識の課題

2．動画像認識の例
　2．1．トラッキング
　2．2．動作認識

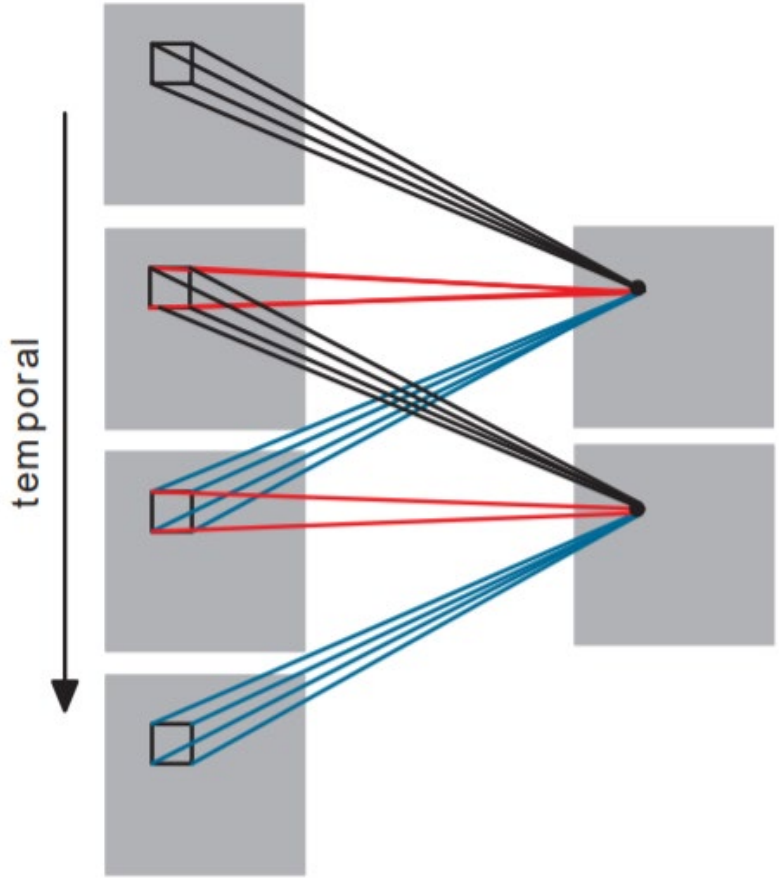
3．実用化のポイント

２．２．3D CNNによる動作認識

3D Convolutional Neural Networksなどが提案されてきた
⇒　課題：　計算コストが高い、認識結果が背景に引っ張られやすい



(a) 2D convolution



(b) 3D convolution

S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. IEEE PAMI, 35(1):221–231, 2013.

２．２．姿勢推定モデル＋時系列モデルによる動作認識

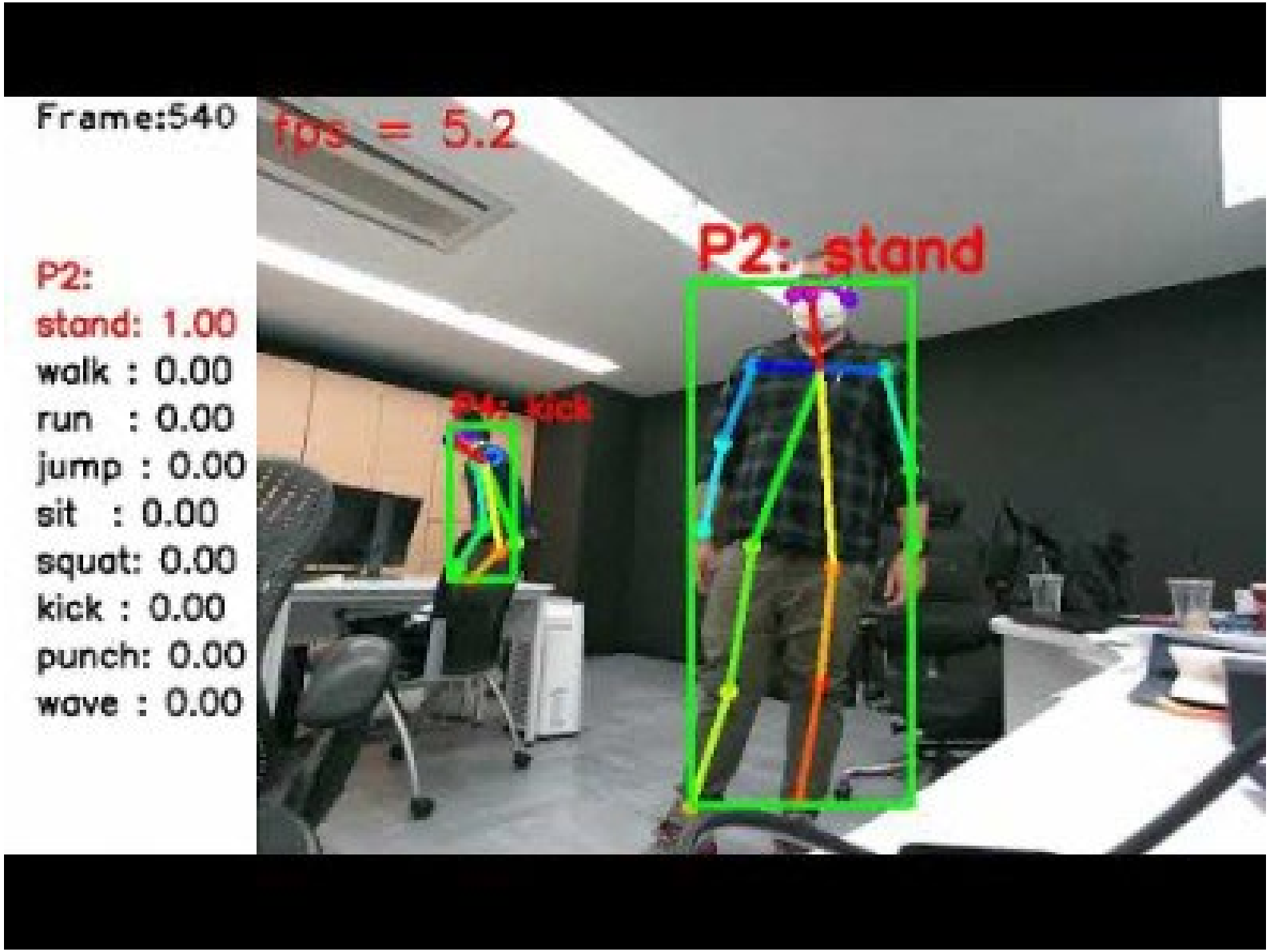
各フレームに対して姿勢推定処理

	Right shoulder	Right Elbow	Right wrist	・・・
時刻t	421, 231	411, 262	423, 292	・・・
時刻t-1	420, 230	408, 260	421, 290	・・・
時刻t-2	418, 231	405, 258	420, 287	・・・
時刻t-3	417, 232	403, 256	416, 285	・・・
時刻t-4	415, 229	405, 259	419, 288	・・・
時刻t-5	413, 231	408, 261	423, 291	・・・
時刻t-6	410, 230	411, 263	424, 293	・・・

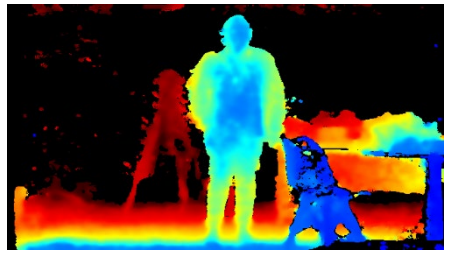
各関節座標の時系列データ

時系列モデル

２．２．動作認識結果の例



２．２．動作認識に3次元姿勢推定を活用することも可能



RGBデータとDepthデータ両方を
用いて2D pose を推定

Depth

RGB

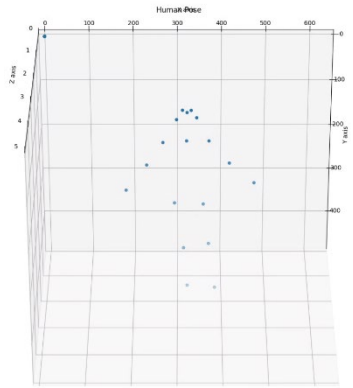
Optimize model

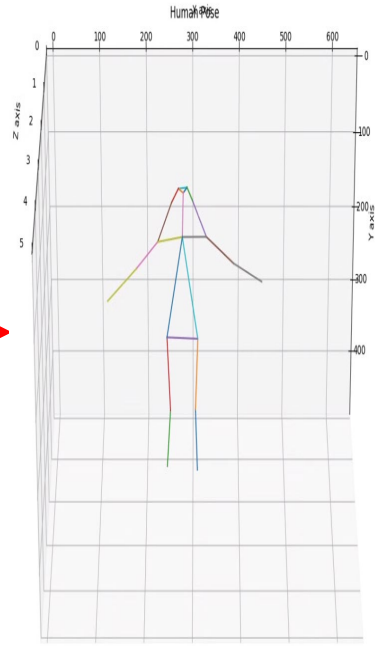
Key points (coordinates x,y)

Depths locator

Key points (x,y,z)

Output Animator





RGB image dataset

Human Pose Model

Tensor-RT optimization

発表内容

1．動画像認識の課題

2．動画像認識の例
　2．1．トラッキング
　2．2．動作認識

3．実用化のポイント

３．１．データセットの準備

■トラッキング向けのデータセット
動画像のデータセットを作成するには膨大な労力が必要
例： 30fps × 100時間 = 1080万枚の画像
さらに同一物体に同一IDの付与

３．１．教師データ不足

■動画像の効率的なアノテーション例

３．１．データセットの準備

■トラッキング向けのデータセット
動画像のデータセットを作成するには膨大な労力が必要
例： 30fps × 100時間 = 1080万枚の画像
　　さらに同一物体に同一IDの付与

■動作認識用のデータセット
動作の開始フレーム・終了フレームをタグ付け
例：　ネジを閉め始めたフレーム番号、ネジを閉め終えたフレーム番号

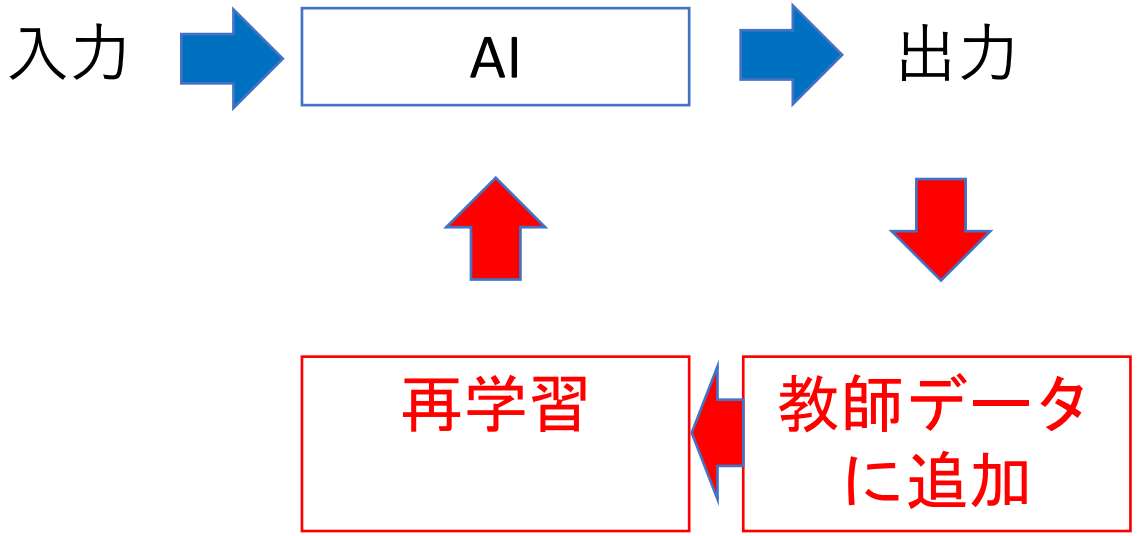
３．１．データセットの準備

【量】
・多い方が良い
　撮影方向の偏りは好ましくない（真正面の実から撮影など）
　背景が同じシーンばかりは好ましくない
　動作のスピードのばらつき、動作の個人差なども考慮する必要がある

【クオリティ】
・データを人手にて付与する際の定義を統一
・認識したい対象に対するデータの付与漏れ、付与間違いが無い

３．２．データ不足に対する対策

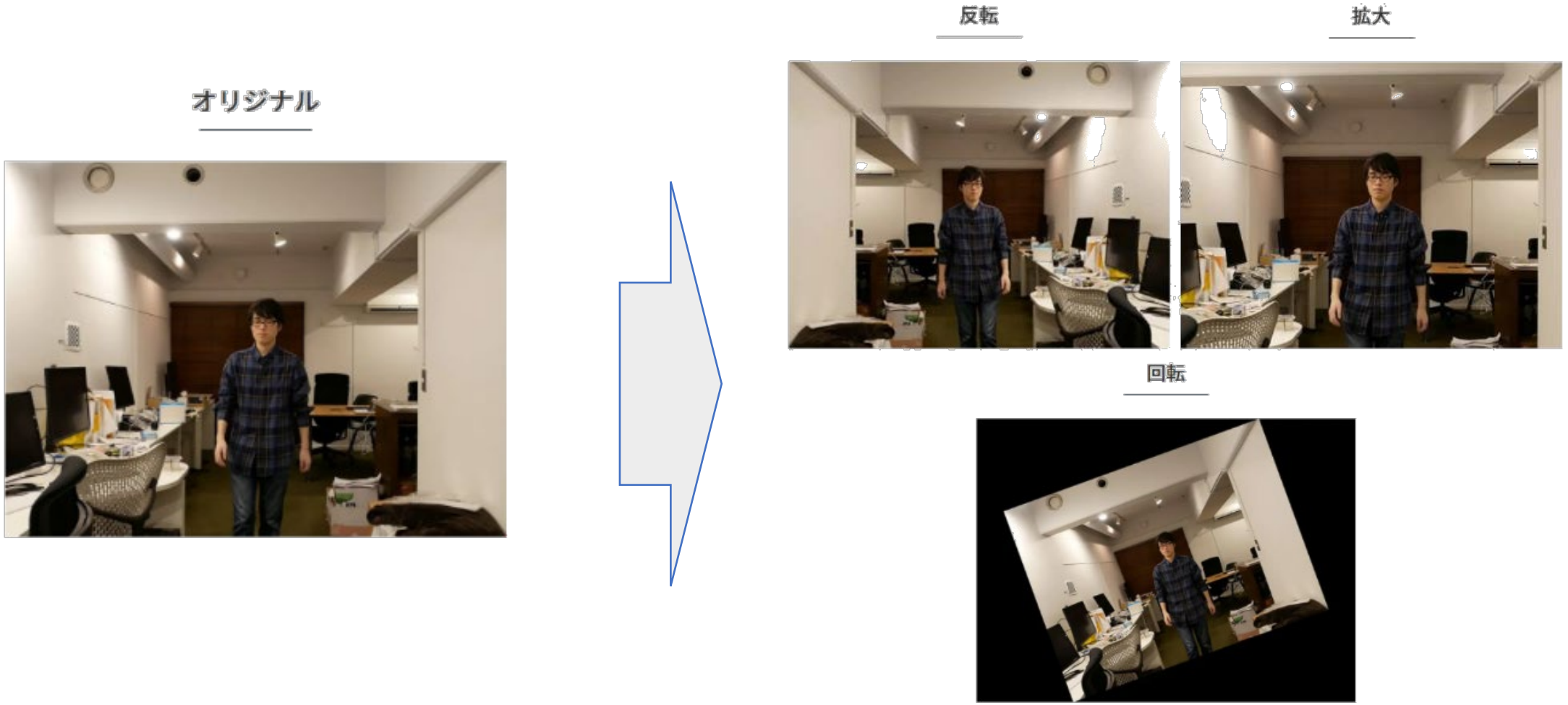
開発着手時に作成したデータセットのみでは精度が不十分であり、
必ず誤認識・未認識が発生



【精度改善に必要なこと】
・試験運用などブラッシュアップの期間を設定
・認識に失敗したシーンを収集する仕組みを想定
・収集したシーンを教師データに追加し、再学習・チューニング

３．３．データ不足に対する対策

データオーグメンテーション（データの拡張）
例：1万枚の教師データを10万枚に拡張



最後に

【動画像認識AIが得意なこと】
・大量の動画を処理する
　人の大まかな動きを計測し続ける
　人の動作時間を計測し続ける
　⇒　統計データを可視化、分析

【動画像認識AIが苦手なこと】
・未知の動作を認識する
　落としたネジを拾うという動作を学習させていなかったとすると誤認識
　の要因となり得る
・撮影方向が異なるシーンで動作を認識する
　カメラを原点とした場合に、各関節の座標値が撮影方向に大きく影響を受ける
・背景が異なるシーンで動作を認識する
　3D CNN等のアプローチを採用する際に注意が必要

最後に

【動作認識AIの開発・実用化の難しさ】

・ビジネスで活用できる動作認識AIを実現するためには深い知識が必要
　姿勢推定、時系列モデル、座標変換など
　※ライブラリが充実してきているためそこそこのものは手軽に作れる
・処理が重い
　GPU、マルチコアの利用、専用ハード（FPGA、ASIC）の設計が必要
・１００％の性能を出すことが非常に難しい
　１００％の精度でなくても成立する運用、設計が重要！
・開発用に大量のデータが必要
・継続的な精度向上が重要