物体追跡のための物体検出しきい値の深層学習による最適推定

三浦　一真†　橘川 拓実†　Sarthak Pathak†　梅田 和昇†
†中央大学
E-mail: miura@sensor.mech.chuo-u.ac.jp

1.　序論

近年，工場内においてカメラから作業工程を把握することによりスケジュール管理等を円滑に行おうとする試みが進められている．この作業工程の把握は，物体検出と物体追跡を組み合わせることにより実現することができる．作業工程は様々であるが，本研究では，人の手で完成までに数日を要する大きな製品を組み立てる作業工程を対象とする．この時，いくつかの問題点が存在する．一般的な作業工程の把握手法として，作業員の骨格点情報を利用した動作認識によって作業工程を把握する手法が存在する[1]．これらの手法は手元の動きに注目したものが多く，作業対象が大きい場合には利用が難しい．加えて，製品の完成までに数日を要する場合には，完成までの組み立て手順に作業員による個人差が増加する．さらに，作業員が製品を組み立てる一連の動作の時系列データが大きくなることから処理が複雑化する．そのため，動作認識を行うことが難しいという問題点が挙げられる．作業員の動作認識から作業工程を把握することが難しい場合，製品の画像から工程進捗の把握を行うことが考えられる．製品の画像から工程進捗を把握する場合，動画に対して物体検出を行い，製品の画像領域を検出し，検出した製品の画像領域から工程進捗の把握を行うことが考えられる．最初に製品の画像を切り出す必要があるため，正確な物体検出が求められる．代表的な物体検出手法として，Faster R-CNN[2] や YOLO(You Only Look Once)[3]が挙げられる．これらの手法では，精度を向上させるために，利用時に予測された確率や Non-Maximum Suppression の実行時に使用する IoU(Intersection over Union)にしきい値を設ける作業が必ず行われる．このしきい値は，学習時の結果や物体検出の実行結果を見ながら手動で一定の値に設定される．しかし，正確に物体検出を行うための適切なしきい値を手動で設定することは難しい．さらに動画には，明るさや背景の変化やぼけ等が生じるため，フレームごとに適切なしきい値も変化すると考えられる．よって，フレームごとに適切なしきい値を設定して物体検出を行うことが，誤検出を減らし物体の画像から作業工程を把握するために重要であると考えられる．また，作業工程の把握には，製品の場所と組み立て進捗の変化を把握することが重要なため，物体追跡が有効である．物体追跡の結果は物体検出の結果に影響を受けるため，物体検出を正確に行うことが重要である．すなわち，製品を正しく管理するに当たり，検出漏れや誤検出を減少させた，正確な物体検出を行うことが重要であると言える．

そこで本研究では，物体検出時にフレームごとに最適なしきい値を設定する．最適なしきい値を深層学習によって推定することで，動画のフレームごとに最適なしきい値を設定し，製品を追跡するシステムを提案する．

2.　関連研究

2.1　物体検出手法

本研究では，物体検出に YOLO を用いる．YOLO は画像を CNN(Convolutional Neural Network)に通すことで，リアルタイムの物体検出を行う．YOLO の検出結果は検出物体を長方形で囲んだ BBox(Bounding Box)として得られる．BBox は検出した物体のクラス分類予測確率として，Confidence Score を持つ．一般的に誤検出の箇所は Confidence Score が低いため，しきい値以上の Confidence Score を持った BBox のみを求めることで，誤検出を防ぎ False Positive を減らす．また，同一物体に何度も検出を行わないように，BBox 同士の重なりを IoU として計算する．IoU にしきい値を設定し，一定以上ならば同一物体に BBox を複数出力しているとみなして，Confidence Score の高いものを残して BBox を削除する．一般的には，Confidence Score と IoU のしきい値は動画に対して一様に設定される．設けたしきい値に基づいて，YOLO は実行される．

2.2 物体追跡手法

物体追跡の手法として，前フレームの情報を利用するものが一般的である．SORT（Simple Online Realtime Tracking)[4]では，物体検出を行ったあと，カルマンフィルタとハンガリアン法を利用して ID 付けして物体追跡を行う．しかしながら，前のフレームでの物体検出時に False Positive が含まれていた場合，次のフレームも誤った物体検出を推定してしまう．SORT を発展させたものに，DeepSORT（Simple Online Realtime Tracking with a Deep Association Metric)[5]がある．主な改良点は物体の外見情報をコサイン距離で計算して ID 付けの精度を向上させる点であり，前フレームの誤検出を次フレームでの推定に使ってしまう問題は改善されていない．ByteTrack[6]は前のフレームに含まれる信頼性の低い検出結果を切り捨てずに考慮することで，正確な物体追跡が行われる．いずれの手法も問題点として，正確な物体検出の結果を利用していない場合，その影響を受けることが挙げられる．

3. 提案手法

YOLO を用いて物体検出を行う．検出した対象の数が前のフレームと異なる場合にのみ，前のフレームと同じ検出数になるように検出数を修正する．これにより，YOLO で検出する段階で，適切なしきい値を用いるように設定され，物体追跡を行う際に対象の数が正しく求まることが期待される．

3.1 深層学習による YOLO しきい値学習

YOLO は YOLOv5[7]を使用する．動画を入力としてフレームごとに最適なしきい値を CNN により学習する．

データセットとして，固定視点の動画から一定間隔（実験では 1 分間隔）で画像を切り出す．画像に対して，YOLO の BBox を求めるために必要な Confidence Score と IoU のしきい値をそれぞれ 0.1 から 0.9 までの範囲で 0.1 ずつ値を変えたものを用意する．9×9=81 通りのしきい値の組み合わせで YOLO を実行して，作業者と検出対象物の 2 つに対して物体検出を行う．その後，検出結果を目視で確認し，正しく物体検出を行えているしきい値の組み合わせを選別する．一枚の画像に対して，正しく物体検出を行えているしきい値の組み合わせが複数ある場合は，Confidence Score，IoU，BBox の平均値をそれぞれ求める．画像とそれに対する Confidence Score，IoU，YOLO の検出結果の平均の値をしきい値学習のデータセットとする．

構築したネットワークの概要を Fig. 1 に示す．画像と YOLO の検出結果の平均の値が入力される．これらの値はそれぞれ畳み込み層を経たのち，全結合層へ入力される．最後に Confidence Score と IoU の推定値が出力される．学習時には用意したデータセットの YOLO の BBox 情報を入力し，推定時には YOLO をデフォルトしきい値で実行した結果の BBox 情報を入力し，Confidence Score と IoU を得る．推定時に入力した BBox 情報には，誤検出や検出漏れが含まれていると考えられるため，得られる Confidence Score と IoU は補正して物体検出に使用する．

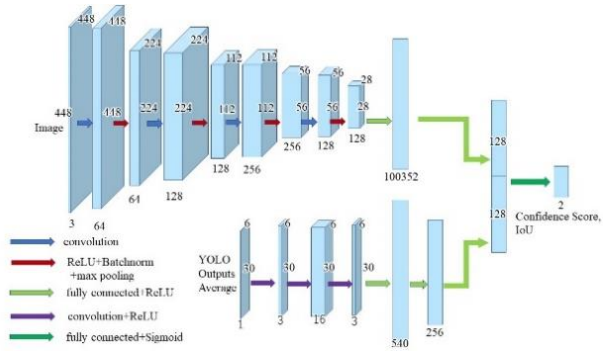


Fig. 1 しきい値学習に用いるモデルの概要

3.2 正解検出数を設けた物体追跡

最初の数フレームに対して，推定したしきい値を用いて物体検出を行う．その間の検出対象物の検出数の平均を算出する．以後のフレームでは，算出した物体検出数を基準として，基準より多く検出された場合は False Positive が生じているとみなして Confidence Score が低い BBox を削除し，基準より少ない場合は False Negative が生じているとみなしてカルマンフィルタを用いて BBox を求める．物体検出数に基準を設けずにカルマンフィルタを用いた場合に比べて，前フレームの誤検出された BBox を次フレームに利用することを防ぐことが期待される．得られた BBox にハンガリアン法を適用し，ID 付けを行う．画像中の検出対象の数が変動するのは，作

業者が検出対象を動かした場合なので，画面端座標に人と検出対象がある場合に，検出数の基準を変更する．

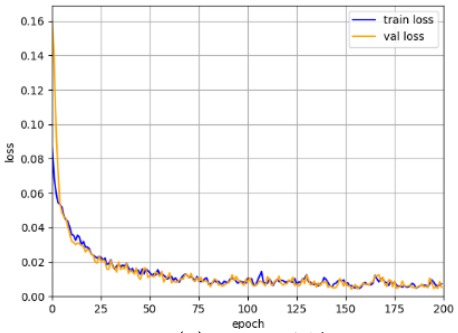
4. 実験

提案手法のうち，YOLO の最適しきい値の学習及び推定に関して実験を行った．

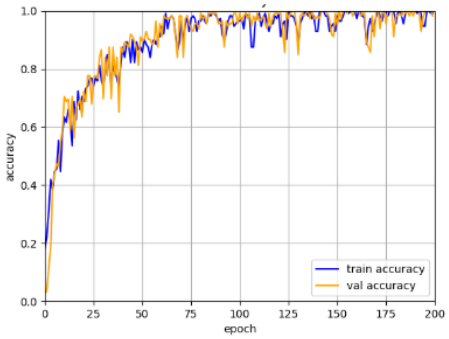
4.1 最適パラメータ推定の学習実験

工場内の動画から 388 枚の画像を用意した．81 通りのしきい値で YOLO を実行したところ，52 枚の画像は False Positive や False Negative を含む結果が出力された．しきい値の組み合わせを変えることで理想的な結果を得られた 336 枚をデータセットとした．215 枚を train に， l54 枚を validation に，67 枚を test に割り振った．学習において epochs = 200, batch size = 4, learning rate = 0.000001 とした．

学習は Loss, Accuracy で評価する．Loss は平均絶対値誤差により評価し，Accuracy は，正解として与えた Confidence Score と IoU の値に対して，差が 0.025 以内であれば正解とした．Accuracy は 80%以上を確認できた．Fig.2 に学習の結果を示す．train loss は右肩下がりを示しており，過学習なく正常に学習が行えたことが確認できた．



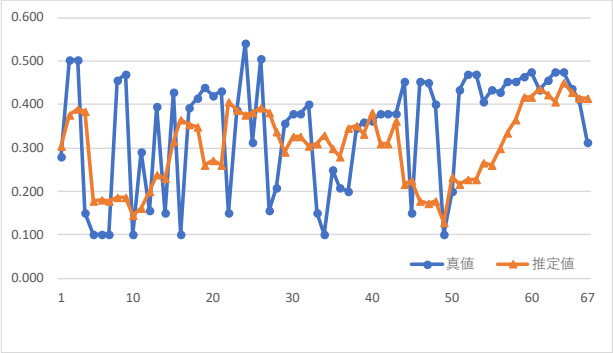
(a) Loss 評価



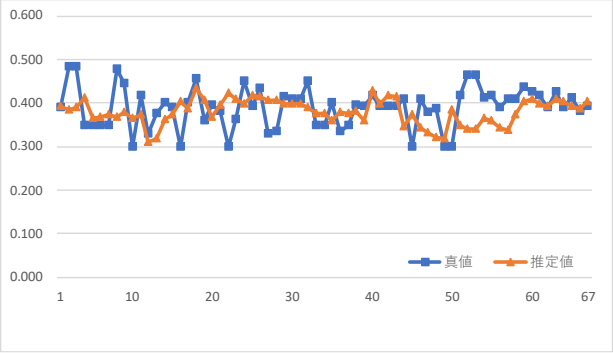
(b)Accuracy 評価
Fig. 2 学習評価

4.2 物体検出の評価実験

推定された最適パラメータを用いた物体検出の評価実験を行った．test に割り振られた 67 枚の工場内画像に対して学習済みの提案モデルを用いて実験した．YOLO は Confidence Score=0.25, IoU=0.45 で実行し，得られた BBox 情報と画像を学習済みの提案モデルに入力することで，Confidence Score, IoU を推定した．Fig.3 に結果を示す．真値と推定値の相関係数は Confidence Score では約 0.29, IoU では約 0.068 となった．IoU では相関は認められなかったが，Confidence Score では弱い相関関係が認められた．相関が小さいのは，正しく物体検出を行えているしきい値の組み合わせが複数ある場合，真値はその平均から定めているためだと考えられる．



(a) Confidence Score 評価



(b) IoU 評価
Fig. 3 パラメータ推定評価

次に推定されたパラメータを利用した物体検出を行った．YOLO のデフォルトしきい値である，Confidence Score=0.25, IoU=0.45 で実行した結果と比較して，正しく物体検出が行われなかった画像は 16 枚から 14 枚に減少し，False Positive と

False Negative はそれぞれ 1 ずつ減少し，合計は 19 個から 17 個に減少した．間違った検出結果の画像に関して Confidence Score のしきい値に注目すると，推定値と真値の差は，推定値の方が平均して約 0.11 大きくなっていることが認められた．そこで，Confidence Score のしきい値をデフォルト値と推定値でそれぞれ 0.11 小さくし，補正したしきい値で物体検出を行った．Table1 に推定パラメータによる物体検出結果を示す．デフォルト値を補正した場合でも，補正前のデフォルト値より検出を正しく行える画像は増加し，推定値を補正した場合でもより検出を正しく行える画像は増加した．しかし，デフォルト値を補正したしきい値と推定値を補正したしきい値を比較すると，推定値を補正したしきい値の方が検出を正しく行えた画像が多いことがわかる．このことから，動画全体に対する設けるべきしきい値が，単にデフォルト値から Confidence Score を 0.11 小さくした値だったため推定値を補正すると検出が正しく行えた画像が増えたというわけではなく，求めた推定値が検出結果の正解数向上に寄与していると考えられる．よって，提案したフレームごとのパラメータ推定手法が有効だということが認められた．

Table 1 推定パラメータによる物体検出結果

	デフォルト	推定値	デフォルト補正	推定値補正
誤検出[個]	2	1	12	6
検出漏れ[個]	17	16	6	7
誤検出,検出漏れ合計[個]	19	17	18	13
エラー画像[枚]	16	14	17	12
正解検出画像[枚]	51	53	50	55

5. 結論と今後の展望

本論文では，しきい値を深層学習によって推定することで，動画のフレームごとにしきい値を動的に設定し，製品を追跡するシステムを提案した．実験によって，検出時に使用するしきい値を画像から適切に推定できることが確認できた．具体的には，67 枚の工場内画像に対して物体検出を行った際に，従来手法と提案手法を比較すると，物体検出が正しく行われなかった画像は 16 枚から 12 枚に減少し，False Positive と False Negative の合計は 19 個から 13 個に減少した．しかしながら，しきい値の調節によって，物体検出に生じる False Positive と False Negative を 0 にすることができない場合も当然存在する．よって，今後は 3.2 節で述べた物体追跡手法を実装し，本手法全体を評価する．

参考文献

[1] 吉川裕，金子真也，浦野雄大，永吉洋登，太田俊広，“映像解析技術を核とした作業認識ソリューション”，*日立評論*，pp. 739-743, 2020.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, *Neural Information Processing Systems (NIPS)*, 2015.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple Online and Realtime Tracking”, *IEEE International Conference on Image Processing (ICIP)*, 2016.

[5] N. Wojke, A. Bewley, and D. Paulus, “Simple Online and Realtime Tracking with a Deep Association Metric”, *IEEE International Conference on Image Processing (ICIP)*, 2017.

[6] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, “ByteTrack: Multi-Object Tracking by Associating Every Detection Box”, *Computer Vision – ECCV 2022*, 2021.

[7] G. Jocher, ultralytics/yolov5, http://github.com/ultralytics/yolov5.