ステレオカメラによる YOLO と 3 次元点群を利用した人物検出

高橋 正裕† Alessandro Moro‡ 池 勇勳† 梅田 和昇†
†中央大学 ‡ライテックス
E-mail: m.takahashi@sensor.mech.chuo-u.ac.jp

1 背景・目的

近年，防犯カメラやマーケティング等への利用を想定し，人物検出や人数カウント等の需要が高まっている．しかし，これらを人手で行う場合，ヒューマンエラーや人件費等の問題が生じてしまう．そのため，これらの分野における自動化が求められている．

人物検出手法の代表例として，画像の輝度勾配を特徴として学習した HOG (Histogram of Oriented Gradients) 特徴量 [1] や，背景差分法を用いた手法 [2] が提案されている．また，近年では深層学習を用いた研究が非常に盛んであり，その中でも物体認識が可能な YOLO (You Only Look Once)[3] は高いリアルタイム性と検出率を誇っている．しかし，これらの手法はカラー画像に対してのみ適用可能であるため，カメラから人物までの距離情報を考慮することができない．そのため，人数カウントにおいて重要であるオクルージョン（画像中における人同士の見かけの重なり合い）への対応が難しい．

そこで本研究では，YOLO の持つオクルージョンへの弱さの解決を目的とする．具体的には，ステレオカメラを用いてカラー画像と 3 次元点群を取得し，YOLO により推定された人物領域内の 3 次元点群を用いて，オクルージョンへの対応を目指す．

2 YOLO の概要

YOLO とは，深層学習をベースとした一般物体検出用アルゴリズムである．本研究で用いる YOLOv3[4] は，主に畳み込みニューラルネットワークを 75 層によって抽出しており，ResNet(Residual Network)[5] と FPN(Feature Pyramid Network) 構造 [6] により検出精度を上昇させている．以上のように，YOLO は 1 つのネットワークによって出力までが完結しているため，Mask-RCNN[7] のような他の深層学習をベースとした手法に比べて，高いリアルタイム性を実現している．しかし，境界ボックスという大局的な領域の出力となってしまうため，Fig. 1 のようにオクルージョン発生時に人物領域が結合する課題が存在し，人ごみにおける人物検出率の低下等を引き起こしていた．

Fig. 1 オクルージョンによる YOLO の失敗例

3 提案手法

オクルージョンが発生すると，画像等の 2 次元情報のみによる判断では，隠れている人物の未検出などが起こりうる．そこで本研究では，ステレオカメラによりカラー画像とともに奥行き情報を取得し，人物の検出を 3 次元で行うことで，オクルージョン発生時も人物の検出を実現する．

提案手法の処理の流れを Fig. 2 に示す．まず，ステレオカメラによりカラー画像と 3 次元点群を取得する．続いて，カラー画像のみを YOLO に入力し，人物候補領域を得る．その後，得られた点群と人物候補領域を用いて点群処理を行い，最後に *k*d-tree によるクラスタリングを用いて人物検出を行う．

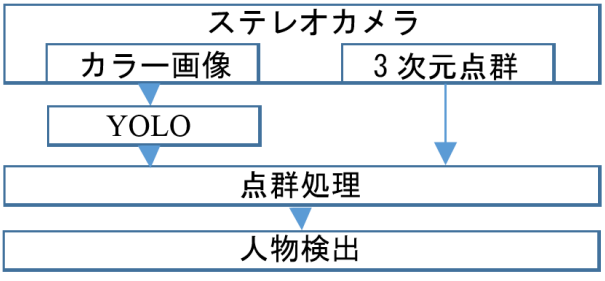


Fig. 2 提案手法のフローチャート

3.1 カラー画像と 3 次元点群の取得

ステレオカメラを用いてカラー画像と 3 次元の点群を得る．その後，カラー画像のみを 2 章で説明した YOLO に入力し，画像中での人物候補領域の境界ボックスを得る．YOLO による検出例を Fig. 3 に示す．

Fig. 3　YOLO による検出例

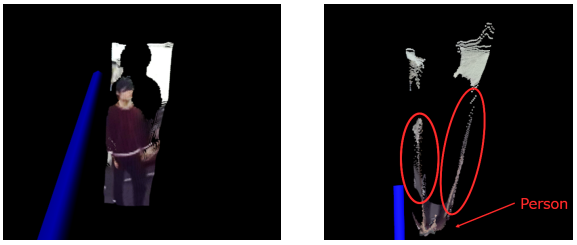
3.2 点群処理

3.1 節で生成した 3 次元点群と境界ボックスを利用し，Fig. 4(a) のように人物候補領域内の 3 次元点群のみを出力する．境界ボックスの位置は画像座標系での値である．3 次元空間における点 (X, Y, Z) と画像中の点 (u, v) の対応は，

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} f_x & 0 & c_u \\ 0 & f_y & c_v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{1}$$

という同次座標によって表現することができる．ここで，カメラの内部パラメータである (f_x, f_y) は焦点距離，(c_u, c_v) は画像中心を表す．

また，Fig. 4(b) に示すように，ステレオカメラで生成した 3 次元点群には多くのはずれ値や誤計測点等が含まれている．そのため，PCL(Point Cloud Library)[8] に実装されている点群分布の統計量に基づく手法を利用し，これらの除去を行う．この時，点群のボクセルにより 3 次元点の総数を減らす処理も同時に行う．Fig.5 に処理後の点群を示す．



(a) 抽出された点群　(b) (a) を上から見た図

Fig. 4　YOLO を用いて抽出した点群

3.3 人物検出

3.2 節で得られた点群に対し，クラスタリングにより人物の点群を取得する．クラスタリングには，kd-tree によるユークリッド距離での最近傍探索を用いる．1 つのクラスに含まれる点の数の範囲を 50〜10000 とし，小さなクラスや大きすぎるクラスを除去することで，人物のみの点群を得る．Fig. 6 にクラスタリング後の点群を示す．

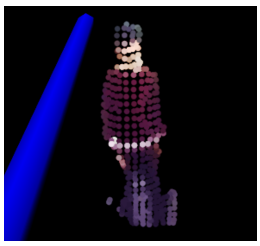


Fig. 5　処理後の点群

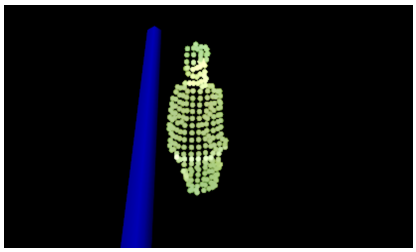


Fig. 6　クラスタリング後の点群

4 人物検出実験

本手法の有効性を検証するために，歩行者 3 人を対象に人物検出実験を行った．ステレオカメラの前をランダムに移動してもらい，その際に取得したデータの中からオクルージョンが発生した 1018 フレームをオフラインで取得し，正確な人物検出が可能かを検証した．ステレオカメラには，ZED[9] を用いた．また，オンラインで取得した全てのフレームと計測時間から，平均の処理速度を算出した．以上の 2 点について，YOLO のみの結果と提案手法の結果とで比較を行った．

Fig. 7 に YOLO による人物の検出結果を示す．紫色の矩形は人物領域の推定結果を表す．また，同様のシーンにおける提案手法の検出結果を Fig. 8 に示す．提案手法ではラベルに対して色分けを行っている．これらを比較すると，YOLO の結果では 3 人のうち手前と中央にいる人物の領域が結合してしまい検出できていないが，提案手法による結果では検出できていることがわかる．

実験結果を Table 1 に示す．オクルージョンが発生した環境における人物検出という困難な対象であることから誤検出が多いものの，提案手法により YOLO による検出結果が改善されていることがわかる．

誤検出の原因としては，ステレオカメラにより取得された点群に多くの誤差が含まれていることが挙げら

れる．Fig. 8 における人物領域の足元に注目すると，床にあたる点群が正確な位置と異なる場所に存在していることが分かる．このように，ステレオカメラはテクスチャの弱い床や壁等の距離を誤計測しやすい性質があるため，はずれ値除去が十分に行われず，クラスタリングの際の領域分割が不十分であったことが考えられる．対策としては，より距離計測精度の優れたセンサに変更することや，色情報等の異なる指標によるノイズ除去が考えられる．

また，処理速度に関しては，提案手法が YOLO のみの手法と比べて 55%程度低下してしまった．この処理速度ではリアルタイム性には問題ないと言えるが，より正確な人物検出のためには処理速度を改善する必要がある．

処理速度が低下した原因としては，2 次元処理から 3 次元処理に拡張したことが挙げられる．特に 3 次元点群に対するはずれ値の除去やクラスタリング等の処理は多くの計算量を必要とするため，その分処理速度が低下したと考えられる．対策としては，GPU による並列処理を導入することが挙げられる．並列処理により，点群処理の速度を大幅に上げることができるので，処理速度の改善が見込まれる．また，点群のノイズ対策も処理速度の改善に繋がる．点群のノイズ除去によりはずれ値が減ることで，クラスタリングの対象となる点群の総数が減り，処理速度の改善が期待できる．

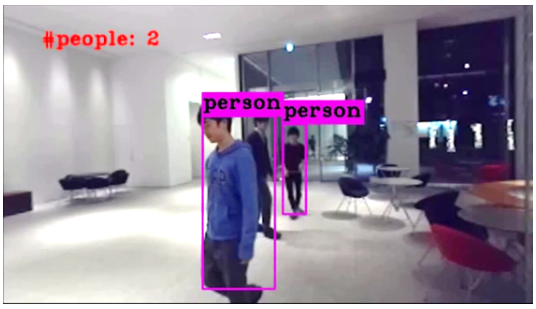


Fig. 7　YOLO による出力結果

Fig. 8　提案手法による出力結果

Table 1　YOLO と提案手法の比較結果

	検出フレーム		平均処理速度 [fps]
	正検出	誤検出	
提案手法	184	834	14.02
YOLO	117	901	30.04

5　結論

ステレオカメラから取得した 3 次元点群及び YOLO を利用することで，オクルージョンに強い人物検出手法を提案し，実験により有効性を検証した．

今後の展望としては，3 次元点群に含まれるノイズの除去手法，ならびに YOLO の検出結果と 3 次元点群を入力とし人物領域のみの点群を抽出するニューラルネットワークを用いた手法の構築を目指す．

参考文献

[1] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR), pp. 886-893, 2005.

[2] 梅田 和昇，寺林 賢司，橋本 優希，中西 達也，入江 耕太 :“ 差分ステレオ-運動領域に注目したステレオ視-の提案, ” 精密工学会誌， Vol.76，No.1，2010.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.

[4] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv:1804.02767.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[6] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2117-2125, 2017.

[7] K. He, G. Gkioxari, P. Doll, and R. Girshick, “Mask R-CNN,” in Proc. of the IEEE Conf. International Conference on Computer Vision (ICCV), pp. 2980-2988, 2017.

[8] Point Cloud Library, http://pointclouds.org/, 2019. 4. 21.

[9] STEREOLABS, https://www.stereolabs.com/zed/, 2019.4.21.