全天球カメラにより配信される正距円筒図法動画からのリアルタイム人物検出

林田　和磨†　　横山　昌平††

† 首都大学東京 システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6-6
†† 首都大学東京大学院 システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
E-mail: †hayashida-kazuma@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし　近年，安価な全天球カメラが普及し，VR，AR をはじめ様々なシステムの撮影手段としてこの製品が利用されている．その名の通り一度に全方位撮影できる点などから，防犯カメラなどへの利用も期待されている．撮影した全方位画像は，正距円筒図法という投影法で全景を平面で見ることができるパノラマ動画像に変換され配信，あるいは保存される．しかし，この正距円筒図法の特性として，変換されたパノラマ動画像は，画像の上部下部で幾何学的歪みが発生し，既存の手法では人物検出が難しい．そこで本研究では，YOLOv3 という既存の物体検出手法を用いて，歪んだ物体の画像をそのまま機械学習させ検出する手法を提案する．提案手法の学習に用いるアルゴリズムは既存のままで，歪み画像をデータセットとし転移学習することで検出を可能にするため，高性能のマシンや多くの処理時間を必要とせずに実現できる．最終的にリアルタイム処理での検出結果から，従来の手法と比較する．

キーワード　全天球カメラ，物体検出，学習，YOLOv3，正距円筒図法，歪み

1　はじめに

近年，RICOH の THETA[1]シリーズをはじめ，一度の撮影で全方位を記録できる全天球カメラが幅広いユーザーへ販売されるようになった．Google Street View[2]や不動産物件の内見，また現在では主に VR（Virtual Reality:仮想現実）や AR（Augmented Reality:拡張現実）を用いた視聴を体験するための撮影デバイスとしてこの製品が利用されている．

全天球カメラで撮影された動画像は一般的に正距円筒図法という投影法により全景を平面で見ることができる全天球パノラマ動画像に変換され，保存または配信される．これは撮影した動画像を既存のファイル形式で保存，配信するための処理である．その動画像が全天球イメージとして全天球球面動画像が生成されることにより VR や AR での視聴が可能になっている．アプリケーションなどを用いて動画像中の任意の方向を閲覧できる全天球球面動画像に対し，全天球パノラマ動画像は一度に全景を見ることができる平面画像となっているため，通常の動画像と同じように扱うことができる．

全天球パノラマ動画像はこのような利点に対し，動画像の上部及び下部にある被写体は緯度が高くなるほど，幾何学的歪みが発生し物体の実際の形，大きさとは異なるものに変形してしまうという特性がある．球形の地球を正距円筒図法で変換した世界地図を例にとると，南極大陸は約 13880000km^{2}で 6 大陸の中でも小さい大陸だが，地図上ではどの国，どの大陸よりも大きく見えている．つまり被写体が球面における北極点，南極点に近いほど，変換された後の被写体の歪みは大きくなる．

我々は，全天球カメラを用いて物体検出を行う研究を進めている．物体検出とは，入力された動画像上の物体が何なのか，どこにあるのかをバウンディングボックスを用いて同時に検出することである．物体検出の手法として，近年では深層学習を利用したものが主流となっており， 歩行者検出，顔検出，医療画像処理やサーチエンジンでの画像検索など，様々な分野へ応用されている．ラベル付けされた大量のデータセットを機械学習することで，そのクラスにおける特徴量を自動で抽出し，未知の入力動画像中の物体を検出できる仕組みになっている．画角に制限のない全天球カメラを物体検出に用いることで，実環境下でのリアルタイム認識処理の効率化につながると考えた．

以上より，本研究では全天球パノラマ動画像上の被写体が歪んでしまう領域（高緯度領域）においての物体検出を可能にすることを目標とする．また，物体検出する対象を”人”とし，図 1 のように大きく歪んで投影された全天球パノラマ動画像上での人物検出を可能にする手法を提案する．

既存の物体検出手法の一つである You Only Look Once version 3（以下 YOLOv3）[3]を用いて，高緯度領域に人が映った全天球パノラマ動画像を入力したところ，ほとんど歪まない領域（低緯度領域）に人が映った全天球パノラマ動画像を入力したときに比べ，“人”の認識率が著しく低下することが分かった．したがって，正距円筒図法で変換された全天球パノラマ動画像上の高緯度領域において，被写体が歪んで変形することにより物体検出が困難になることが明らかになった．この問題に対し我々は，高緯度領域で歪んでしまった人の動画像を“人”のクラスとして再学習させることにより解決できるのではないかと考えた．提案する手法は，既存のアルゴリズムをそのまま利用し

1：https://theta360.com/ja/
2：https://www.google.com/intl/ja/streetview/
3：https://pjreddie.com/darknet/yolo/

図 1　正距円筒図法で変換された全天球パノラマ画像の例

たものであるので，特別な処理を加えずに解決したことが研究的な強みである．つまり，従来の動画像に対する学習，検出と同じようなプロセスで，全天球パノラマ動画像上の高緯度付近での人物検出を可能にした．

本論文の構成は以下の通りである．2 章では，関連研究について述べる．3 章では，提案手法について述べる．4 章では，実験の結果と考察について述べる．5 章では，本研究のまとめと今後の課題について述べる．

2　関 連 研 究

本章では関連研究について述べる．本研究のように全天球パノラマ動画像上の歪みを指摘し，高緯度領域での物体検出を可能にするための研究はほかにも行われている．

井上 [1] らは全天球カメラを用いて物体検出，トラッキングを行い，視覚障害者へ向けた支援システムを開発した．全天球カメラで撮影した動画像の投影法は撮影点を中心に 8 面に分けるキューブマップを用いて，各面の画像は通常のカメラで撮影したような歪みの小さい動画像を生成して物体検出を行った．またトラッキングする上では，8 面キューブマップは各面が別々の画像となることが原因でうまく追跡できないため，物体検出した 8 面の画像を元の正距円筒図法の座標で変換し投影することで対応した．しかし，キューブマップで処理する画像が増えたことと，投影法が違う 2 つの処理を統合したシステムであったため，処理速度の遅さが問題となった．

若狭 [2] らは全天球カメラを用いて地域の緑の可視率を測定するシステムを提案した．伝送される正距円筒図法で変換された動画像から，投影法の一種であるランベルト正積円筒図法で変換された動画像を用いて，緑を認識するシステムを実装した．極に近い高緯度領域の緯線間隔をさらに狭め，低緯度領域での物体の面積が正しくなるような動画像を生成した．しかし高緯度領域での歪んだ物体の検出，認識をする上では，この手法は正距円筒図法で変換された高緯度領域の歪みをさらに大きくしてしまうことになるため，歪みの問題は根本的に解決されていない．

Taira [3] らは全天球カメラを用いて，正距円筒図法変換により発生した歪み部分から特徴量を抽出するための手法を提案した． 正距円筒図法で変換された一枚の全天球パノラマ動画像に対し，対象とする被写体が映っている位置が低緯度領域にある時はそのまま特徴量を抽出し，高緯度領域にある時は球面を回転させて注目領域を低緯度領域に移動させることにより，低緯度領域のみでの特徴量抽出を可能にした．また様々な角度からの撮影動画像が必要なＳｆＭ（Structure from Motion）においても，少量の全天球パノラマ動画像から特徴量抽出とカメラ位置推定を行い，有用性を示した．ただし，このシステムを用いた物体検出やトラッキングについての有用性は指摘されていなかった．

また，中澤 [4] らは，ボールに全天球カメラを内蔵した全天球ボールカメラを用いて，動いているボール内で全天球動画を撮影し，その動画像の視点を固定する手法を提案した．Taira らの手法を用いて 1 枚の全天球パノラマ動画像を球面に投影し，カメラ座標系のＸ軸周りに $\pi/6$ ごと回転させた画像を 6 枚生成する．それらの全天球パノラマ画像を低緯度領域のみで特徴量マッチングを行い元の全天球パノラマ動画像に戻すことで，すべての領域から特徴量を抽出し視点固定動画像の出力を可能にした．しかしこの手法は，一枚の全天球パノラマ画像に対し存在領域が変化した 6 枚の全天球パノラマ動画像が生成されてしまうので，計算処理に時間を取られてしまう可能性がある．

これらの研究に対して本研究では，特別な処理，機材，処理時間を必要としない，被写体が歪んで変形してしまった全天球パノラマ動画像の学習のみによって，高緯度領域での物体検出を可能にする手法を提案する．

3　提 案 手 法

本研究では，全天球パノラマ動画像上の歪み部分での物体検

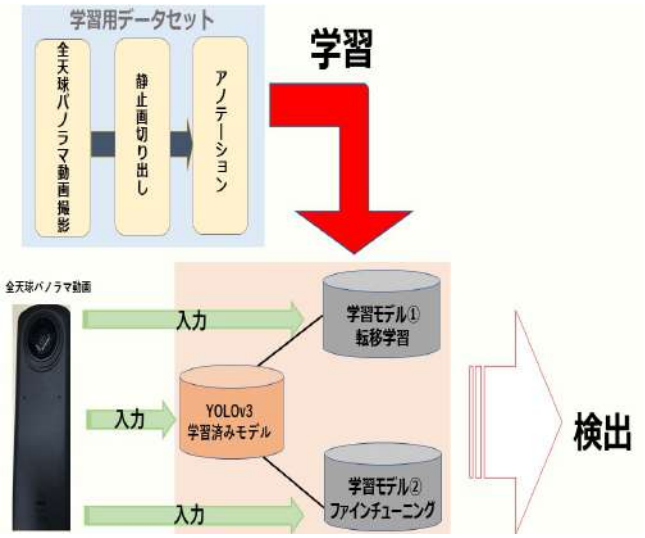

図 2　提案手法の流れ

出を可能にするための手法として，歪んだ物体をそのまま学習させることで検出を可能にするという単純な手法を提案する．さらに，用意するデータセットや学習，検出アルゴリズムなどに対し特別な処理は一切加えていない．つまり，通常の動画像での学習，検出に用いられていた既存のアルゴリズムのまま，全天球パノラマ動画像の学習を行い検出を可能にする．そのため，マシンパワースペックや検出にかかる処理速度は，結果として大きく関与しないと考えられる．

本研究で提案する手法の流れは図 2 の通りである．

全天球カメラとして RICOH THETA Z1 を用いて撮影をし，歪み画像の学習，また入力動画像に対する物体検出において深層学習アルゴリズムベースの物体検出手法である YOLOv3 を用いて実験を行う．

3.1　撮　　影

全天球パノラマ動画像の撮影デバイスとして全天球カメラ RICOH THETA Z1 を用いる．RICOH の THETA シリーズは二枚の魚眼レンズで構成されており，それらのレンズで撮影した動画像はカメラ内で繋ぎ位置補正，ゆがみ補正，ブレンド処理などが施された後，正距円筒図法形式で保存され，PC や RICOH THETA 専用アプリなどを用いて全天球パノラマ動画像，球面動画像の閲覧が可能になっている [5]．正距円筒図法形式での 4K ライブストリーミング機能もあるため，今後リアルタイム認識処理を行う際はこの機能を用いる．本研究ではこのカメラを用いて，正距円筒図法に変換したときに人が歪むような動画をいくつか撮影し，静止画として保存したものをデータセットとする．

図 4 のようにカメラを上に持ち，全天球パノラマ動画像に変換した際に高緯度領域に人が来るように約 30 秒の動画を 5 つ撮影した．撮影した動画の解像度は幅が 3840px，高さが 1920px，フレームレートは 29.97fps となっている．撮影した約 30 秒の動画をフレームで切り出し，学習させるための歪み静止画像として合計 2679 枚用意した．なお，この全天球パノラマ動画撮影では 1 名のみを対象とした．

図 3　RICOH THETA Z1

図 4　撮 影 風 景

また，スマートフォンや一般のデジタルカメラでの動画撮影も行った．3. 4 節でも説明するが，これは低緯度領域でのみ人物検出をするモデルを作成するために行った．歪みがなく視野角が限られている通常の動画を 12 本撮影した．同じようにフレームで切り出し，通常静止画像として合計 9144 枚用意した．なお，これらの動画は 4 名を対象として撮影した．

動画上の対象人物とデータセット静止画像数との内訳を表 1 に示す．

表 1　撮影した画像の内訳

撮影対象者番号	全天球パノラマ画像数	通常静止画像数
No.1	2679	1935
No.2	0	3325
No.3	0	3036
No.4	0	848

3.2　YOLOv3(You Only Look Once version 3)

YOLOv3 は，深層学習アルゴリズムを用いた物体検出手法の一つであり，検出や学習における高速処理や誤認識の少なさに非常に定評があるプログラムである．深層学習ベースの YOLOv3 は，画像全体から特徴量を自動で抽出し学習，検出を行うもので，物体検出手法の中でも画像認識を回帰問題として

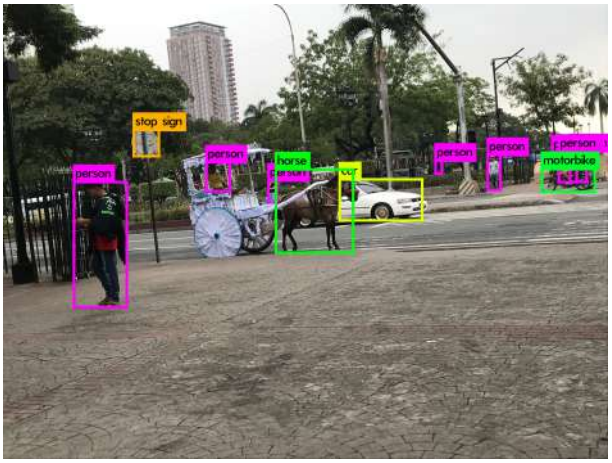

図 5　YOLOv3 での物体検出

図 6　YOLOv3 による全天球パノラマ動画像の物体検出結果

扱っている手法である．ラベル付けしたデータセットをもとに存在位置を学習し，信頼度とその物体が存在する条件付き確率の積から物体の存在領域を確率で抽出する仕組みになっている．

はじめに，YOLO のホームページから事前学習済みモデルをダウンロードした．これは coco dataset[4] を用いて 80 クラスの学習をしたモデルで，“人”は学習済みであるが，1 章で記述したように全天球パノラマ動画像を入力したとき歪み部分はうまく検出できないことを確認した．その結果を図 6 に示す．

独自のデータを学習する際は，その学習したクラスのみの検出となるため，物体検出対象は”人”のみとする．YOLOv3 を用いた学習，検出はすべてコマンドプロンプト上で行い，本研究では効率化を図り GPU を用いて実装する．

3.3　ラベル付け（アノテーション）

撮影した動画を静止画として切り出し，その静止画に対しラベル付けを行う．本研究では動画像の教師データ作成ツールとして，python 上で実装される LabelImg を用いてラベル付けを行う．図 7 のようにマウスで存在領域を指定すると，クラス番号と座標データが一枚の画像に対応したテキストファイルとして保存される．したがって，用意した画像の枚数分，画像に関するテキストデータを用意する必要がある．保存されるテキストファイルの例を表 2 に示す．表 2 における数値は左から，クラス番号，バウンディングボックスの中心 x 座標， バウンディングボックスの中心 y 座標， バウンディングボックスの幅

4 : http://cocodataset.org/#home

図 7　LabelImg を用いた座標作成

w， バウンディングボックスの高さ h となっている．

表 2　LabelImg で保存されるテキストデータの例

クラス番号	x	y	w	h
0	0.436587	0.622449	0.379436	0.662338
0	0.415710	0.610390	0.389875	0.791299

3.4　学　　習

画像と座標データが書かれたテキストファイルの二つを教師データとして深層学習していく．全天球パノラマ動画像の高緯度領域の歪み部分を学習させる検出器は 2 つ用意した．また比較のために，スマートフォンや一般のデジタルカメラを用いて撮影した動画像を学習させた低緯度領域でのみ人を検出する検出器を 1 つ用意した．また，通常の静止画像を変換させ，全天球パノラマ画像上の歪みを再現した画像を学習させる検出器を 1 つ用意した．

本研究では上述の通り検出器を 4 つ用意するのだが，どこで学習を止めるかは，学習状況によって判断する．学習の試行回数と損失を表すグラフをもとに損失の推移が収束したら，十分に学習された状態であるとする．グラフの例を図 8 に表す．

また，後に紹介する転移学習やファインチューニングという言葉の定義は分野によってさまざまであるので，本研究では，学習済みネットワークを固定して新しいクラスに対し行う学習手法を転移学習，既存のクラスに対して一部の重みを再度学習

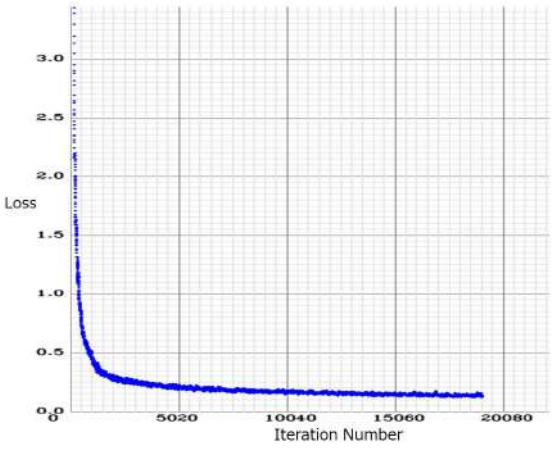

図 8　損失の推移

する手法をファインチューニングとする．

検出器 1: 全天球パノラマ動画像を転移学習

検出器 2: 全天球パノラマ動画像をファインチューニング

検出器 3: 通常の動画像を転移学習

検出器 4: 通常動画像を変形したデータセットをファインチューニング

3.4.1 転移学習を用いた学習

一般に転移学習とは，ある領域で学習したモデルを別の領域での学習に適応させ学習の効率を上げる技術である．YOLOv3 は畳み込みニューラルネットワーク構造になっており，学習の際はその畳み込み層にて特徴量を抽出する仕組みになっている．しかし，何も知識がない機械にたくさんのデータを与えても，それらのデータの特徴を捉えることは難しく，多くの処理時間が必要になる．それに対して，大量の知識，知恵がある学習済みモデルを再利用するこの手法は，動画像上のどのような部分に着目すべきかという特徴抽出に長けており，未知のデータセットを与えたとしても学習の効率が非常に良く検出精度も高い．従来の学習手法よりも学習時間を短縮でき，少ないデータセットで高い精度を出せるという点が最大のメリットである．

転移学習では，事前に学習したモデルを学習済みネットワークとして固定し，追加した層のみを使用して学習できるため，新しいクラスに対する学習の際に有効である．以上の点を踏まえ，全天球パノラマ動画像上の高緯度付近にいる人物を，新しいクラスとして定義し学習を進める．

YOLOv3 は学習済みモデルとして 80 個のクラスが定義されており，その中で人のクラスは”person”という名前で定義されている．転移学習を用いたこのモデルでは，81 個目の新しいクラスとして”person2”クラスを定義し，歪んだ人の動画像を学習させる．2679 枚の全天球パノラマ動画像を転移学習にあてる．

転移学習を用いて学習を行ったモデルを検出器 1 とする．

3.4.2 ファインチューニングを用いた学習

転移学習と同様，事前に学習したモデルを学習済みネットワークとして固定するのだが，出力層のみを学習させる転移学習に対し，ファインチューニングは学習済みモデルの重みを初期値として再学習し微調整する技術であるので，既存のクラスに対する学習の際に有効である．

本手法では，全天球パノラマ動画像上の高緯度付近にいる人物を，既存のクラス”person”として再学習を行う．検出器 1 の学習で用いた全天球パノラマ動画像と同じ 2679 枚を，ファインチューニングの学習にあてる．

ファインチューニングを用いて学習を行ったモデルを検出器 2 とする．なお，検出器 1，2 は学習試行回数がともに等しいものを使用する．

3.4.3 通常の静止画像をデータセットとした学習

全天球パノラマ動画像ではない通常の画像のみで学習させたモデルを作成した．本研究の目的は全天球パノラマ動画像上の高緯度領域における歪んだ人の検出であるので，これは検出器 1，検出器 2 と比較するためのモデルである．

また，この学習にはファインチューニングを用いた．スマートフォンや一般のデジタルカメラで撮影した 9144 枚の通常静止画像を学習にあてる．このモデルを検出器 3 とする．

3.4.4 通常画像を変換したものを学習

全天球パノラマ動画像の高緯度領域での検出を可能にするために，検出器 1，2 では撮影した全天球パノラマ動画像をそのままデータセットとした．しかし現在，全天球パノラマ動画像のオープンデータセットは通常画像に比べ非常に少なく，独自のデータセットのみでは学習効率が悪い．

そこで通常画像を球面の一部にマッピングし，その画像を球面の極に近い位置で正距円筒図法変換することにより，本来の全天球パノラマ動画像の歪み部分を再現した画像を作成した．通常画像は検出器 3 の学習に用いた 9144 枚の画像のうち，正面で上半身が映っている 1952 枚を採用した．そして，図 9 のように 3940 × 1920 サイズの単色の背景の一部に画像を貼り付けたものを球面マッピングした．

球面マッピングには Cube2DM[5] というアプリケーションを利用した．Cube2DM は通常カメラレンズ，魚眼レンズ，ドームマスター (等距離射影)，全球パノラマ (正距円筒図法) など，様々な投影形式で動画像を入出力できるアプリケーションで，仰角や方位角も設定できる．この Cube2DM を用いて通常画像を全球パノラマ形式として入力し，仰角，方位角を調整して全球パノラマ形式として歪みのある全天球パノラマ画像を再現した．変換した画像例は図 10 の通りである．

上記のようにして通常画像から全天球パノラマ動画像を再現したものをデータセットとして，ファインチューニングを用いて学習させた．このモデルを検出器 4 とする．

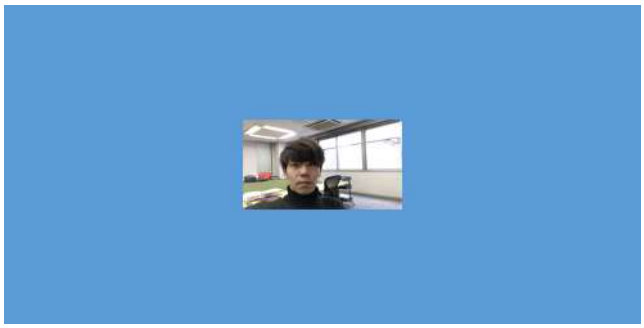

図 9　Cube2DM に入力する画像

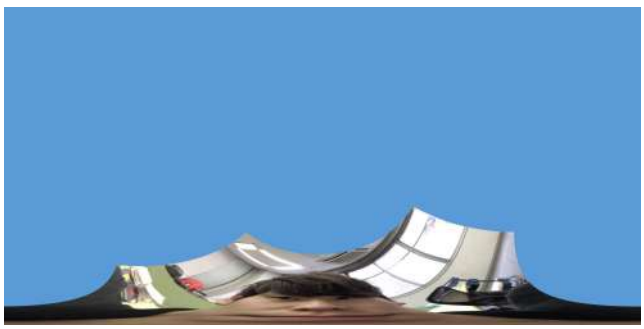

図 10　Cube2DM により変換された画像

5 : http://T.NOMOTO.org/Cube2DM/

4 評価実験

本章では，前章で紹介した 4 つの検出器と YOLOv3 の学習済み検出器の合計 5 つの性能を比較するため，リアルタイムストリーミングでの入力ではなく，全天球パノラマ動画をそれぞれ検出器に入力する方法で実験を行う．RICOH THETA Z1 で映り方が異なる約 15 秒の 3 つの動画を撮影し，正距円筒図法で変換した全天球パノラマ動画を入力し，人物検出を行った．なお，入力する動画の撮影者は，4 つの検出器の学習においてデータセットとした人とは異なる人で撮影している．

入力動画 1: 画像下部 (高緯度領域中央) に人が映っている動画
入力動画 2: 画像下部 (高緯度領域両端) に人が映っている動画
入力動画 3: 画像下部，中央部，下部に人が映っている動画

各領域での認識検出結果をもとに既存の手法と比較し本手法を評価する．なお本研究では，YOLOv3 を用いたフレーム毎の認識検出結果として表示するものは認識精度が 50 %以上のものを採用する．

4.1 入力動画 1 の検出

正距円筒図法で変換した全天球パノラマ動画下部において上半身が歪むように撮影した約 15 秒 (フレーム数 468 枚) の動画を入力する．入力動画の例を図 11 に示す．5 つの検出器に動画を入力したときの認識精度を記録し，5 フレームごとで平均をプロットし作成したグラフを図 11 に示す．また，5 つの検出器と正しく人を認識できているフレーム数の内訳を表 3 に示す．

表 3　全フレーム (468 枚) のうち正しい検出数

YOLOv3	検出器 1	検出器 2	検出器 3	検出器 4
71	467	448	17	458

図 12 からわかるように，検出器 1，検出器 2，検出器 4 がともに高い精度で検出できていることが分かる．このことから，全天球パノラマ動画像上の歪みにより変形した物体検出において，本研究で提案した学習のみによる対策は有効であるといえる．

またデータセットに関するアプローチとして，通常画像を球面マッピングし，全天球パノラマ動画像の歪みを再現した検出器 4 でも十分高い精度で検出できていることから，全天球パノラマ動画像を学習させなくても正距円筒図法変換による高緯度領域での認識率低下は対応できることが分かった．

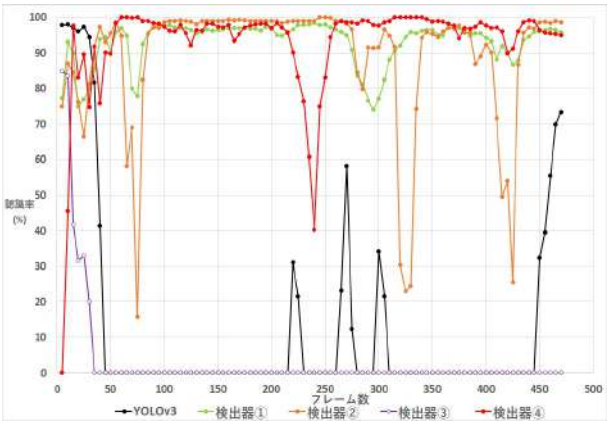

図 12　入力動画 1 の人物検出結果 (x 軸:フレーム数　y 軸:認識率)

YOLOv3 の学習済み検出器による検出もわずかにみられた．体の一部が低緯度領域 (画像中央) に出たときや高緯度領域での歪みが小さいときに検出をしていた．歪みがない通常画像のみを学習させた検出器 3 は，ほとんど検出が見られなかった．

4.2 入力動画 2 の検出

正距円筒図法で変換した全天球パノラマ動画像下部において上半身が歪むように，かつ，1 人が両端の境目で二人に分断するように撮影した約 14 秒 (フレーム数 422 枚) の動画を入力する．また入力動画は，初めからおよそ 300 フレーム目付近までは高緯度領域，それ以降は二つに分かれたまま低緯度領域に映るように撮影した．入力動画の例を図 13 に示す．

5 つの検出器に動画を入力したときの認識精度を記録し，5 フレームごとで平均をプロットし作成したグラフを図 14 に示す．また，入力動画が動画像両端で二つに分かれたときは別々に検出するので，二つの認識率の平均を取り記録した．

表 4　全フレーム (422 枚) のうち正しい検出数

YOLOv3	検出器 1	検出器 2	検出器 3	検出器 4
175	61	266	100	31

転移学習により学習した検出器 1 は，フレームによってはうまく検出ができなかったり，二つのうち片方しか認識ができないフレームが多くみられた．

それに対しファインチューニングにより学習した検出器 2 は，4.1 節に比べ精度は下がったが，二つに分断してもうまく検出ができたといえる．4.1 節，4.2 節での検出器 1 と検出器 2 に

図 11　入力動画 1

図 13　入力動画 2

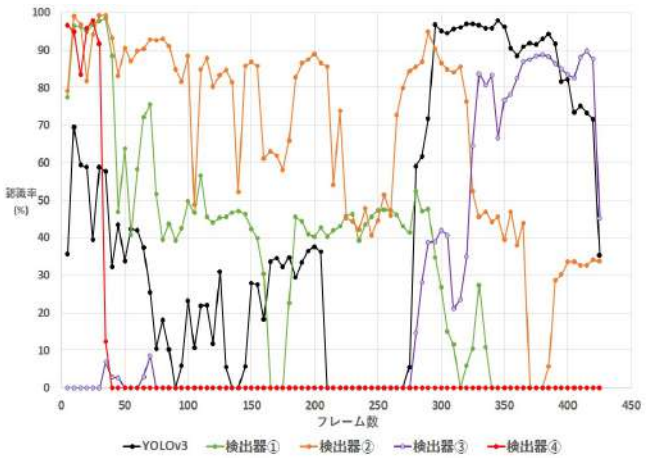

図 14　入力動画 2 の人物検出結果 (x 軸:フレーム数　y 軸:認識率)

よる検出結果を比較すると，検出器 2 に対し行ったファインチューニングによる学習の有効性が確認できる．本来，ニューラルネットワークの畳み込み層は出力側に行くにつれて具体的な特徴を含む構造になっており，転移学習は全体の重みを変更せず，その出力層において学習する．ファインチューニングは同じクラスに対し，ほかの層の重みを初期値から誤差逆伝搬法により更新し，出力層において再学習する．検出器 2 は，畳み込み層において重みを微調整し，”人”として局所的な特徴を上手く再学習できたことで，このような結果が得られたのではないかと考えられる．

通常画像で全天球パノラマ動画像の歪みを再現したデータセットを学習した検出器 4 は，ほとんど検出が見られなかった．検出器 4 に与えたデータセットのほとんどが，画像下部中央付近で歪ませたものだったため，その領域でしか認識ができない検出器になったという可能性がある．また，従来の物体検出手法とは異なり，YOLOv3 は 1 枚の画像の全範囲を学習に利用する．今回，検出器 4 に与えたデータセットの背景はすべて単色だったため，全天球パノラマ動画像で学習するよりも学習効率が悪いということが要因の一つとして挙げられる．

YOLOv3 の学習済み検出器と通常画像を学習させた検出器 3 は，ともに低緯度領域に移動したおよそ 300 フレーム目以降の検出が多くみられた．また，300 フレーム目よりも前の歪み方が小さいフレームにおいては，若干の検出がみられた．ただリアルタイム処理へ適応させ実用化するには難しい結果であると考えられる．

4.3　入力動画 3 の検出

正距円筒図法で変換した約 15 秒 (フレーム数 428 枚) の全天球パノラマ動画を入力する．入力動画は，初めからおよそ 60 フレーム目付近までは動画像下部の高緯度領域で上半身が横向きで歪むように撮影した．60 フレーム目から 240 フレーム目付近までは低緯度領域で，それ以降は動画像上部の高緯度領域で上半身が歪むように撮影した．入力動画の例を図 15，図 16 に示す．

5 つの検出器に動画を入力したときの認識精度を記録し，5

図 15　入力動画 3(画像下部での歪み)

図 16　入力動画 3(画像上部での歪み)

フレームごとで平均をプロットし作成した，YOLOv3 の学習済み検出器，検出器 1，検出器 2 のグラフを図 17 に，YOLOv3 の学習済み検出器，検出器 3，検出器 4 のグラフを図 18 に示す．また検出器によって，誤認識がいくつかみられたので除外し認識率=0 として記録した．

表 5　全フレーム (428 枚) のうち正しい検出数

YOLOv3	検出器 1	検出器 2	検出器 3	検出器 4
364	85	183	198	75

図 17 から，検出器 1，2 は画像下部の高緯度領域 (60 フレーム目付近まで) において，4.1 節よりも検出数が減り，認識率も下がっていると感じられる．また，それ以降のフレームにおいては，検出器 2 のほうが検出器 1 に比べ多く認識できているが，十分良い精度で認識できているとは言えない．高緯度領域 (画

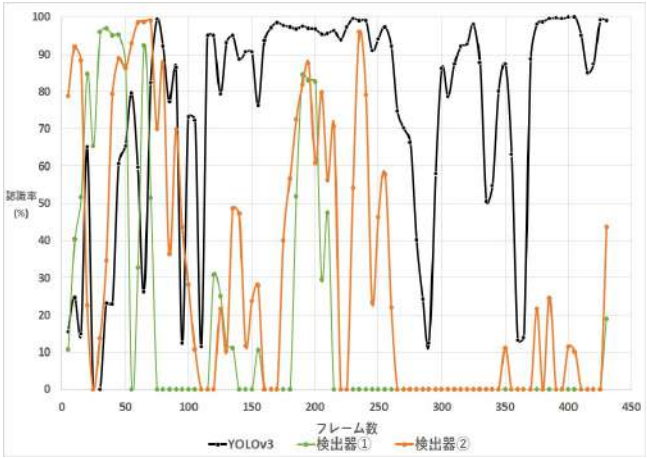

図 17　入力動画 3 の人物検出結果 1 (x 軸:フレーム数　y 軸:認識率)

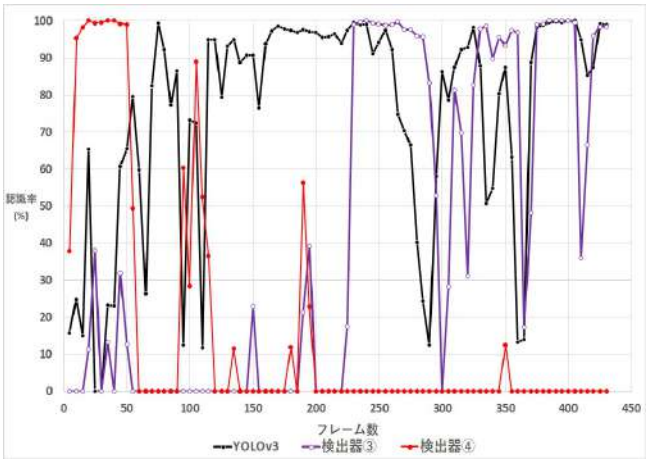

図 18　入力動画 3 の人物検出結果 2 (x 軸:フレーム数　y 軸:認識率)

像下部) の被写体が歪んだデータセットを学習させるだけでは，低緯度領域においての検出や，高緯度領域での歪み方や位置が変化したときなどの検出は上手くいかないことが分かった．

240 フレーム目以降 (画像上部での歪み) において，YOLOv3 の学習済み検出器には多くの検出が見られた．肩から頭にかけては高緯度領域で歪んでいても，体の一部が低緯度領域にあるため，”人”として上手く認識できている．YOLOv3 の既存の検出器は，coco dataset という大量の画像を学習する上で，”person”として認識するための特徴を細かく捉えていることが分かった．被写体が歪んだとしても，目や腕などの特徴的な部分の歪みがそれほど大きくなければ，人を検出しやすくなるのではないかと考えられる．このような YOLOv3 の性能の良さに恩恵を受けたのが，低緯度領域でもいくつか検出がみられた，ファインチューニングにより学習させた検出器 2 ではないかと考えられる．

図 18 から，検出器 3 はおよそ 60 フレーム目から 240 フレーム目付近にかけての検出 (低緯度領域) は，カメラと被写体との距離が近くなるにつれて多くなった．この結果から，カメラとの距離 (被写体の映り方) も検出に大きく影響することが明らかになったので，他の検出器を学習させる時もデータセットの選び方には注意しなければならない．

また検出器 4 では，ほかの検出器ではほとんど見られなかった誤認識が多くみられた．認識しなかった 353 枚のうち 58 枚誤認識をしていて，被写体ではない歪みに対して反応した．4.2 節で述べたように，これも学習に与えたデータセットの不備によるものであると思われる．

これらの実験結果を受け，本研究で提案した検出器に与えるデータセットは被写体との距離，歪み方，動画像における位置など，様々なパターンを考慮して集める必要があることが分かった．

5　まとめ　今後の課題

本論文では，全天球パノラマ動画像上の被写体の歪みが発生する高緯度領域での物体検出を可能にする手法を提案した．高緯度領域で歪んだ物体の動画像をそのままデータセットとして既存のアルゴリズムで学習させ，検出を行った．5 つのモデルで検出された認識結果から，高緯度領域での認識率低下に対し，学習のみによる本手法の有用性が確認できた．また，学習や処理に多くの時間をかける必要がない点も高く評価できる．

そして，歪んだ物体の学習においてはファインチューニングを用いることで，比較的効率の良い検出が行えることも分かった．高緯度領域での認識精度はこれからデータセットを増やし，学習を進めることで解決していく．

また学習用のデータセットは全天球カメラで撮影したものではなく，通常の画像を疑似的に歪ませた画像の学習によっても，学習時と同じような条件下においての検出は多くみられた．

また本論文では扱っていないが，THETA Z1 を用いてリアルタイムで物体検出を行ったところ，どの検出器も評価実験と同じような結果が得られた．

ただ，すべての検出器の学習において，画像上の位置やカメラとの距離，歪み方などを考慮し，学習させるデータセットを選ばなければならないことが結果より確認できた．今回は学習用データセットも検出入力動画もカメラとの距離が限られている，かつ，被写体の向きが正面に映っているものを採用したので，さらなる応用に向け後頭部での歪みや体全体での歪みなど，様々な歪みのパターンを考慮しなければならない．同時に，撮影条件によって発生するモーションブラーなどの対処も考えなければならない．

今後の課題として，実験によって浮上した問題点を考慮し，さらなる精度の向上ために，大量の全天球パノラマ動画像や通常画像を追加データセットとして学習させることがあげられる．coco dataset を学習した YOLOv3 の学習済み検出器は低緯度領域では十分な精度で検出できるため，最終的には coco dataset と本研究で使用した全天球パノラマ動画像のデータセットを合わせて学習させた検出器を作成したいと考えている．

新たなシステムへの応用としては，本研究で提案した手法を用いて監視カメラシステム等への実用化を検討している．

またリアルタイム検出では，現状 USB 接続のみの入力となっているため，無線状態でのリアルタイムストリーミング入力による物体検出を行うシステム構築も検討している．

文　献

[1] 井上慶彦, 岩村雅一, and 黄瀬浩一. 全方位カメラを用いた物体検出とトラッキング—視覚障害者支援システムの実現に向けて—. Technical Report 20, 大阪府立大学, may 2018.

[2] Kazuki Wakasa and Shin'ichi Konomi. Green weaver: Participatory green mapping and networking for fostering sustainable communities. 09 2015.

[3] Hajime Taira, Yuki Inoue, Akihiko Torii, and Masatoshi Okutomi. Robust feature matching for distorted projection by spherical cameras. *IPSJ Transactions on Computer Vision and Applications*, 7:84–88, 2015.

[4] 中澤 正和 and 小池 英樹. 全天球カメラ内蔵ボールにおける視点固定手法. **日本バーチャルリアリティ学会論文誌**, 22(4):485–491, 2017.

[5] 庄原 誠 and 竹中 博一. 全天球撮像デバイス“ricoh theta”の開発. **日本写真学会誌**, 77(3):234–237, 2014.