

人物検知を用いたマーチングプレイヤーの位置推定による 練習補助システムの提案

竹中亜瞳^{†1} 福原義久^{†1†2}

マーチングとはプレイヤーが歩きながら楽器の演奏やダンスを行う総合芸術である。50 人以下の小編成から 90 名以上の大編成まで多様な規模が存在する。マーチングには演奏技術だけでなく、演奏と動きの調和やフォーメーションの綺麗さ、動きの技術が必要とされ、全てのプレイヤーには予め定められた位置まで正確に移動することが求められる。本研究では定点カメラで撮影した映像から人物検知を用いて各プレイヤーのフロア上の位置を正確に読み取る方法を提案する。これにより個々のプレイヤーへの的確な指示が行えるようになり、練習の効率化が期待される。

Practice assistance system using person detection to estimate the position of marching players

AMI TAKENAKA^{†1} YOSHIHISA FUKUHARA^{†1†2}

Marching is an integrated art form in which players play instruments and dance while walking, and can range in size from small groups of 50 or fewer to large groups of 90 or more. In marching, the accuracy of the stride is one of the most important factors for evaluation, and all players are required to maintain an accurate position. In this study, we propose a method to accurately read each player's position on the floor using person detection from video captured by a fixed-point camera. This will enable us to give precise instructions to each player, and is expected to improve the efficiency of practice.

1. はじめに

吹奏楽は、管楽器を主体として演奏される音楽の総称であり、座奏やマーチングなどの演奏形態が存在する。マーチングとは演奏しながら、その演奏曲の曲想にあった動きを加える総合芸術である。演奏しながら行進をするストリート・マーチや、曲に合わせて様々なフォーメーションを展開するフィールド・ショー、広い体育館などのフロアでドリルやフォーメーションを展開するフロア・ドリル、ステージ上で行うステージ・ドリルと 4 種類の発表形態があり、それによって 50 人以下の小編成や 90 名以上の大編成などがある。ドリルとは演奏しながら動きによって曲を表現することで、繰り返し練習することからドリルと呼ばれている。

本研究では日本のマーチングバンドの編成規模からして、最も主流な形態であるフロア・ドリルを基に進めた。(以降フロア・ドリルの形態をマーチングと表す)。

マーチングは演技の構成が示された指示書(以降これをコンテと表す)に従い、フロア上に記された等間隔のポイントを目印にフォーメーションを組む。コンテは方眼に各プレイヤーの 1 セット毎の座標と拍数、動作方法が記されている。そのためフロアを真上から見ると方眼上に多様な図形を描いているように見える(図 1)。マーチングの評価は演奏技術だけでなく、演奏と動きの調和やフォーメーションの綺麗さ、動きの技術が関係する。そのため、練

習では初めは音楽の演奏と合わせずに複数のフォーメーションを連続でつなげ、ある程度形になるまで動きの反復練習を行う。

その際、指導者は正面からプレイヤー間の間隔や全体のバランスを見て、乱れている箇所を判断し、指摘する。一方、フロア上のプレイヤーは隣との間隔や自分が所属している列を見て位置のずれを修正する。

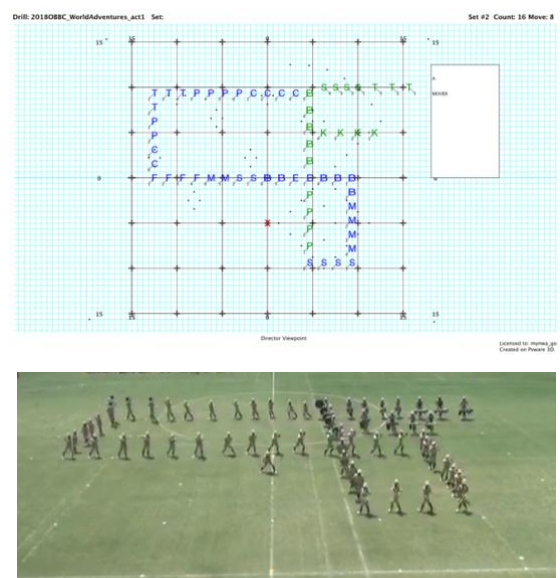


図 1 コンテ(上)と実際の映像(下)。

^{†1} 武蔵野大学 データサイエンス学部
Faculty of Data Science, Musashino University
^{†2} 武蔵野大学 アジア AI 研究所
Asia AI Institute, Musashino University

プレイヤーはコンテの指示にしたがって正確に移動することが求められるが、位置の確認と把握には、歩幅に関する正確な距離感覚が必要である。長年の経験者でもこの感覚を習得している人は少ない。

反復練習において、この位置の確認に要する時間の割合は少なくなく、効率よく各プレイヤーの位置の確認と修正がおこなえるかは練習の質に大きく影響する。

そこで、本研究では、定点カメラで撮影した映像から人物検出を用いて各プレイヤーのフロア上の位置を正確に読み取ることでコンテ上の座標との比較を行う手法を提案する。これにより位置を確認する時間の短縮や個々のプレイヤーへの的確な指示が行えることで練習の効率化が期待される。

2. 人物検出手法の比較

本研究では、マーチングのプレイヤーの位置を映像中から推定するために、人物検出手法を用いる。人物検出とは物体検出の一部で、画像や動画の中で人物が存在する位置や大きさを機械学習や深層学習を用いて検出するものである。学習データから人物の特徴を抽出し、それを満たす物体を画像や動画から見つけることで人物検出ができる。人物検出を含む物体検出では R-CNN, SSD, Faster R-CNN など様々なアルゴリズムが提案されている。

マーチングはフォーメーションによってプレイヤーの重なりが生じるため演技動画に映るプレイヤーの一人々々を区別しがたい。そのため全プレイヤーを検出することは困難であることが予想される。我々はマーチング動画に適する人物検出の方法を定めるために以下の手法を試行した。

- HOG 特徴量に SVM を組み合わせた手法
- Haar-like 特徴量分類器
- Mask R-CNN(MMdetection)

各技術について以下に述べる。

2.1 HOG 特徴量+SVM

HOG(Histograms of Oriented Gradients)[1]とは、局所領域における輝度（色，明るさ）の勾配方向をヒストグラム化したものである。そのヒストグラムを特徴量としたものが、HOG 特徴量である。局所領域を複数のブロックに分割し、各ブロックの勾配をヒストグラム化することで、物体の形状変化に強い特徴量を得ることができる。

SVM(Support Vector Machine)[2]は教師あり学習を用いるパターン認識モデルの一つである。少ない教師データで高い汎用性を持つことを特徴としている。特に分類のタスクで用いられる。HOG で人物の輝度の変化をベクトル化した特徴を抽出し、その特徴量を SVM で学習することで画像の中の人物と人物ではない部分を分類し、人物を検出する。

この手法は形状変化にロバストであるとされるため、さ

まざまな姿勢をとる人物の検出に有効ではないかと考えた。なお、実験には OpenCV を使用した。

2.2 Haar-like 特徴分類器

Haar-like 特徴分類器[3]とは画像の明暗差により物体の特徴を抽出する手法である。画像の一部をさまざまな大きさで抜き出し、それぞれで局所的な明暗さを算出する。算出したいくつもの特徴を組み合わせることで、物体を判別する。なお、実験には OpenCV を使用した。

2.3 Mask R-CNN

Mask R-CNN[4]は物体検出やセグメンテーションにおける優れた手法の一つである。物体検出で結果として得られた検出領域を対象にピクセル毎にクセグメンテーションを行い、より細かい検出領域を得ることができる。これによって、マーチングのフォーメーションで重なって見える人物でも正確に検出できると考え採用した。

実験には MMdetection を使用した。MMdetection は香港中文大学(CUHK)が公開しているフレームワークであり、COCO や CityScapes などの一般的なデータセットを含む 360 を超える学習済みモデルを利用可能である[5]。

2.4 比較実験結果

HOG 特徴量+SVM と Haar-like 特徴分類器, Mask R-CNN を比較してどの手法がマーチングの映像に適しているか調べた。図 2, 3, 4 は 3 つの手法の出力画像である。

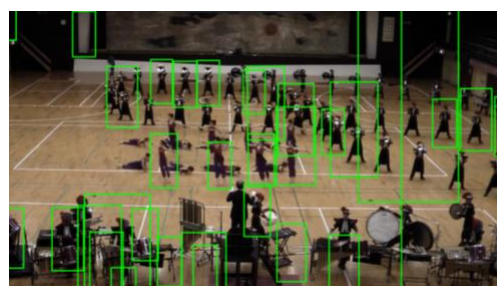


図 2 HOG 特徴量+SVM.

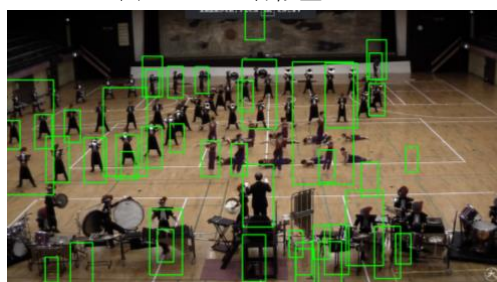


図 3 Haar-like 特徴分類器.

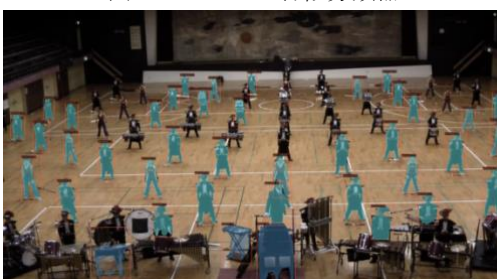


図 4 Mask R-CNN.

HOG 特徴量+SVM と Haar-like 特徴分類器では人物検出はできているが、人物が密集する箇所や楽器と人物が重なっている箇所は検出領域が重なり、何を人物と検出しているかが一見判断しにくい。比較して Mask R-CNN ではセグメンテーションにより、人物の境界が明確である。また、画像手前の打楽器が密集している箇所でも人物を検出でき、境界の精度も高い。しかし、これらの手法に共通して言えることは、図5のように打楽器を持っている人物は認識されていないことと画像の中央の明るい箇所は検出されるが、左奥や右奥の暗い箇所は人物検出がうまくいっていないということである。また、Mask R-CNN は物体検知モデルを使用したため画像手前の楽器や台も検出してしまっている。

以上のことから、すべての条件において瑕疵のない手法はないものの、人物の境界が明確な点と楽器に重なる人物も検出していること、他手法と比較して精度が高いことから本研究では Mask R-CNN を使用することとする。

3. システム構成

図6に提案システムの構成を示す。提案システムは映像から人物を検出し、座標の抽出を行うシステム、フロアを真上から見た時の方眼上の座標に変換するシステム、正しい位置と実際のプレイヤーの立ち位置とのずれや、閾値を超えるずれを持つプレイヤーの可視化を行うシステムの3つから構成される。

3.1 ユーザ入力

ユーザは以下の2つを入力する必要がある。

- 位置を確認したい演技シーンの動画または撮影した画像



図5 Mask R-CNN の拡大画像。

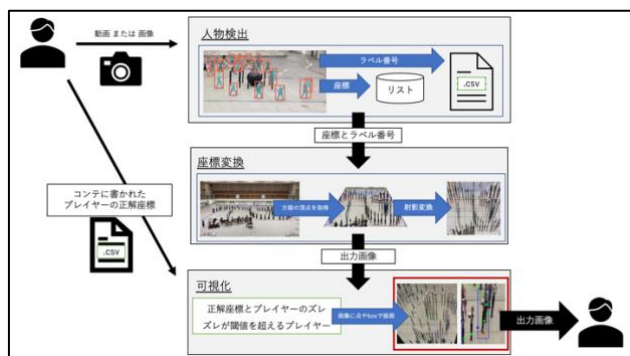


図6 システム構成図。

- 位置を確認したい演技シーンのコンテ情報 (プレイヤーの正しい座標)

プレイヤーの正しい座標は、csv ファイルに1プレイヤーを1行としてx座標が1列目、y座標は2列目に格納する。

3.2 人物検出と座標の抽出

人物検出では、MMDetection に公開されている Mask R-CNN を使った物体検知モデルを使用した。位置を確認したい演技シーンの動画または画像を受け取り、Mask R-CNN を用いて人物を検出し、取得した座標とラベル番号を保存する。取得した座標は、画像内で検出した全ての人物の座標であり、ラベル番号は、フロア上の方眼内にいるプレイヤーのみを抽出し、外側にいる指揮者やその他の人物と区別するために使用した。これは位置を確認したいシーン毎に行う。

3.2.1 入力データの処理

入力された演技のデータが動画である場合は、20フレーム毎の画像を利用する。コンテでは一つのフォーメーションを分割し、複数の局面で座標を指定している場合があるので、これに対応できるように動画の場合は1秒に1回より小さい単位でフレームを取得する。取得した画像の中から位置の確認をしたいシーンを探し、Mask R-CNN への入力情報とする。

3.2.2 人物の検出

Mask R-CNN による人物検出を行う。この際、入力した画像の解像度やフォーメーションによって異なるが、大多数のプレイヤーが検出できるように正答率の閾値を低めに設定する。図7は検出した人物の周りを検出領域で囲み、ピクセル単位でセグメンテーションを行った結果である。

プレイヤーがコンテで指示された座標に立つ場合、足を閉じていれば両足の土踏まずを、足を広げている場合は指導者から右足か左足か体の中心か指定された箇所がコンテの座標の位置になる。今回はどの場面も共通して体の中心を立ち位置とする。検出領域の底辺のy座標をプレイヤーの足元のy座標、左辺のx座標と右辺のx座標の midpoint のx座標をプレイヤーの足元のx座標として座標を抽出する。



図7 人物検出の結果。

3.3 フロア上の座標をコンテ上の座標へ変換

座標変換では人物検出で取得した画像上の座標をコンテに記されている正しい座標と比較するための変換を行う。三次元空間では深度があるため、カメラで撮影した映像には遠近感が生じる。人物検出で取得した座標での 10 ピクセルの間隔は実際のフロアではカメラから離れるほど間隔が広がっていく。これではコンテの方眼上に記された座標と比較する際にカメラから離れた位置にいるプレイヤーほど誤差が大きくなってしまう。そのため映像の奥行きを考慮して、取得した座標を変換する必要がある。今回は入力画像からフロア上の方眼の 4 つの頂点を取得し、射影変換を用いた。

射影変換とは任意の形の四角形から別の形の四角形に変換する変形のことをいう。類似技術としてアフィン変換があるが、これは平行四辺形枠を別の平行四辺形枠に割り付ける変換である。正方形・矩形・菱形は平行四辺形の特別な形状である。射影変換は対辺が必ずしも平行ではない一般的な四辺形に変換できる。

図 8 は射影変換の一例を示す。左の画像は文字が書かれた紙の画像である。斜めから撮影しているため本来長方形の紙は対辺が平行でない四角形として映っている。赤枠で囲まれている部分を、射影変換を用いて長方形すると紙に対して垂直に撮影した画像になる。

図 9 はあるシーンの射影変換前の画像と変換後の画像を示す。左の画像はフロア上の方眼を赤枠で示したものである。右の画像は左の画像の赤枠のみを抽出し、射影変換で正方形にしたものである。

図 9 の右の画像と図 10 は同一視点のものといえる。右の画像に描画されている赤点は人物検出で取得したプレイヤーの立ち位置である。これがコンテに記されている座標と比較する立ち位置となる。

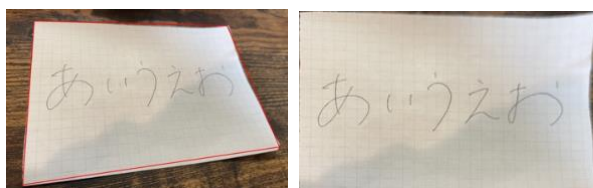


図 8 射影変換の一例。
(左：変換前 右：変換後)

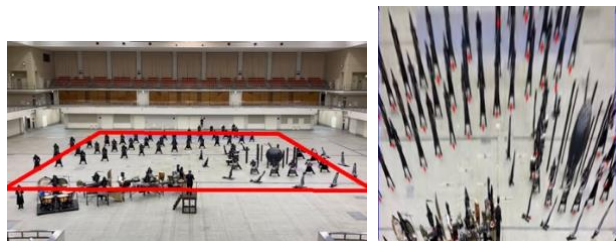


図 9 あるシーンの変換前と変換後。
(左：変換前 右：変換後)

3.4 可視化

最後に射影変換で出力した画像に正しいプレイヤーの立ち位置と、実際のプレイヤーの立ち位置や方眼の線、閾値を超えるずれを持つプレイヤーを明確に表示する。

まず、ユーザが入力したコンテ情報（プレイヤーの正確な座標）に基づいて正しい立ち位置を画像に描画する（図 11 の緑点）。また、射影変換で生成した画像にコンテの表現に合わせて縦横六等分した線を黒色で描画する。また、図 10 の水色の方眼のように細かい線を灰色で描画する。

図 11 にはプレイヤーの正しい座標と方眼の線、実際のプレイヤーの立ち位置を描画した画像を示す。赤点は座標変換の際に描画した実際のプレイヤーの立ち位置である。

次に、正しい立ち位置の座標と人物検出で取得したプレイヤーの座標の誤差を画像から読み取り、csv ファイルに記録する。誤差は x 方向の誤差と y 方向の誤差を足した合計で記した。各プレイヤーの誤差の閾値を 2.0 に設定し、誤差がそれを超える場合は青枠で強調し、そのプレイヤーが映るように画像から抜き出す。

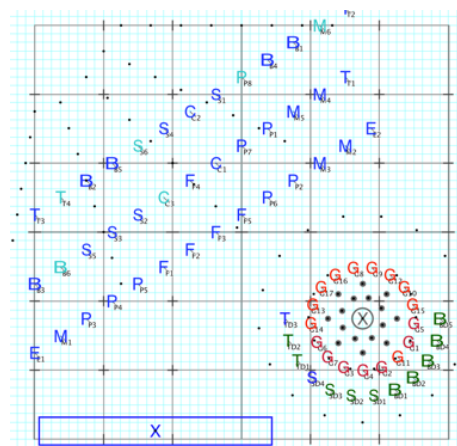


図 10 図 9 と同一シーンのコンテ。

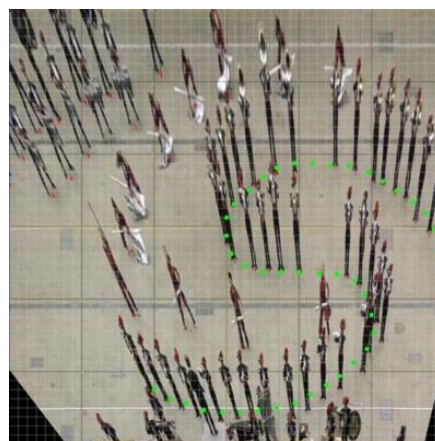


図 11 立ち位置と方眼の線を描画した画像。

図 12 には閾値を超えるプレイヤーを強調表示した画像を示す。青枠は人物検出で取得したプレイヤーの足元の座標から上に 400px, 下に 50px, 左右に 50px に移動した 4 点を頂点とする長方形で人物を囲むようにした。また、閾値を超えるプレイヤーの正しい位置を青点で表示することで他のプレイヤーと区別している。強調表示しているプレイヤーの足元の赤点を拡大し、正確な位置と比較しやすいようにした。図 13 は強調表示した箇所を切り抜いた画像である。図 12 と同様に人物検出で取得したプレイヤーの足元の座標から上に 450px, 下に 100px, 左右に 100px 移動した 4 つの点からなる長方形を抜き出して、新たな画像として保存する。図 12,13 の画像をユーザに出力する。

4. 実験結果

表 1 に示したような 8 種のシーンでシステムの評価を行った。人物の服装、フロアーの色などの背景、カメラの位置、画像の解像度、フォーメーションなどが異なる。表 2 ではそれぞれの解像度とカメラ位置、服装の内容、人物検知の際に設定した閾値を示す。閾値は 0~1 の間である。表 3 にそれぞれの人物検出の結果を示す。

また、図 14 に誤判定となったプレイヤーの人物検出の結果と最終出力画像を示す。図 15 に正確に判定ができたプレイヤーの人物検出の結果と最終出力画像を示す。

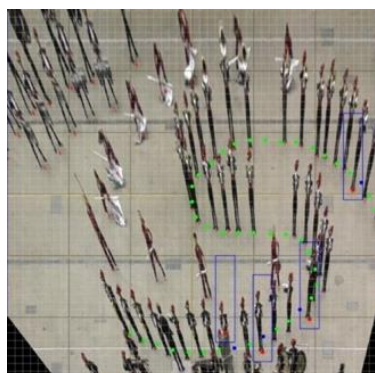


図 12 閾値を超えるプレイヤーの描画画像。



図 13 閾値を超えるプレイヤーの拡大図。

5. 考察

実験結果からシーン 4~8 で人物検出の精度が高いことがわかった (表 2,3)。各シーンの特徴を比較すると、まずシーン 7,8 は解像度が非常に高いことがわかる。このことから、フォーメーションが複雑であっても解像度が高ければ、精度良く人物の検出ができると考えられる。また、シーン 4,5,6 では解像度は低いもののプレイヤーの服装の輝度が高い傾向が見られる。このことから、服装と精度にはなんらかの関係があるのではないかと考えられる。

図 14 の青枠で囲まれたプレイヤーは実際には閾値を超えるわけではない。しかし、手前の楽器に体の下半分が隠れているため人物を判定されていても検出領域で囲む部分が上半身のみになり、正しい立ち位置と大きく誤差があると出力された。このことから隠れた体を推定して足元を特定する必要があることがわかった。

背景や服装の違いでの出力結果の違いはなかったが、フォーメーションの形によって検出がうまくいかないケースがみられた。

表 1 実験で使った 8 種類の画像。

シーン 1	シーン 2	シーン 3	シーン 4
シーン 5	シーン 6	シーン 7	シーン 8

表 2 8 シーンの特徴。

シーン	1	2	3	4	5	6	7	8
解像度	360,640	360,640	360,640	360,640	360,640	360,640	1800,2880	1090,1920
カメラ位置	正面上	正面上	正面上	正面上	正面上	正面上	正面上	右正面
服装	黒色	黒色	黒色	ペーजू	白色/黒色	ペーजू	水色/黒色	黒色
閾値	0.13	0.1	0.14	0.2	0.1	0.1	0.1	0.1

表 3 8 シーンの人物検出結果。

シーン	1	2	3	4	5	6	7	8
プレイヤー数	69	69	69	61	68	61	58	41
検出数	59	60	52	57	64	55	58	38
検出率	0.85	0.87	0.75	0.93	0.94	0.90	1.0	0.93

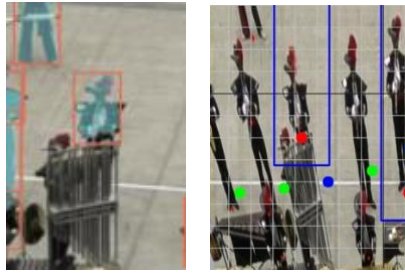


図 14 誤判定がでたプレイヤー。
(左：人物検出の結果 右：最終出力画像)

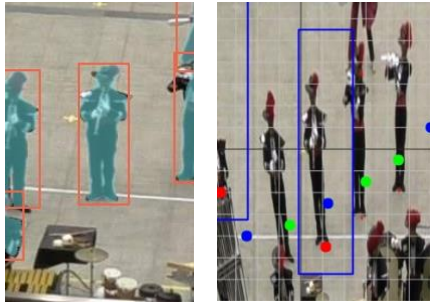


図 15 正確に判定できたプレイヤー。
(左：人物検出の結果 右：最終出力画像)

図 16 では定点カメラは正面にあり，検出されていないプレイヤーは定点カメラに対して垂直方法に位置している．人物が重なっている場合でも Mask R-CNN では検出できるが，このように垂直に一直線で重なると検出が難しいことがわかった．正面に向かって直線で並ぶ動きはマーチングの中で多々あるため，マーチング映像から新たに学習をおこなうなど人物検出モデルの精度向上が望まれる．図 16 では，体の向きによって取得した足元の位置にずれが生じている例を示す．赤点はプレイヤーの立ち位置を示す．マーチングではコンテに書かれた形を正確に表現するために，体のどこをコンテに書かれた座標に置くのかを指定される．しかし，図 16 からわかる通り，左を向いている場合は右足，正面を向いている場合は体の中心，右を向いている場合は左足が立ち位置となる．

これでは，正しい立ち位置と比較するときに正確なずれが算出できない，人物の姿勢を取得し，その姿勢をもとに，プレイヤーの立ち位置を推定するといった改善が必要であることがわかった．

6. まとめ

本研究では，定点カメラから Mask R-CNN を用いて人物検出を行い，コンテ上でプレイヤーがどこにいるか推定し，正しい立ち位置とのずれを可視化する練習補助システムを実現した．このシステムは全プレイヤーが歩幅に関する正確な距離感覚を持つことなく，正しい立ち位置とのずれを確認することができ，ずれが大きいプレイヤーを抽出することで，練習の効率化が期待できる．

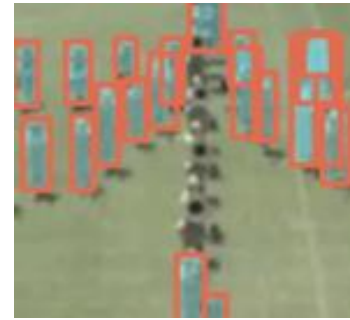


図 16 人物が検出されない例．



図 17 体の向きによって足元の位置にずれが生じる例．

7. 今後の展望

実験結果より，カメラに映らない足元を推定し，より正確に足元を特定することや，マーチングに特化した人物検出モデルの作成，体の向きなどの姿勢を考慮した位置の推定などを，実現していきたい．

謝辞 本研究で使用したマーチングの演技データは福岡大学附属大濠高等学校吹奏楽部に提供して頂いたものである．ここに記して謹んで感謝の意を表する．

参考文献

- 1) Navneet Dalal, Bill Triggs: Histograms of Oriented Gradients for Human Detection (2005)
- 2) 津田宏治: サポートベクターマシンとは何か Overviews of Support Vector Machine (2000)
- 3) 立花智哉, 山下淳, 金子透: Haar-like 特徴を使った顔検出と Mean-shift トラッカによる複数視点人物追跡システム(2007)
- 4) Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick: "Mask R-CNN", (Submitted on 20 Mar 2017 (v1), last revised 24 Jan 2018)
- 5) OpenMMLab: MMDetection
<https://github.com/open-mmlab/mmdetection>
- 6) 小坂谷達夫, 山口修: 顔認識のための射影変換に基づいた三次元正規化法(2005)