Vol.2021-CG-184 No.15
Vol.2021-DCC-29 No.15
Vol.2021-CVIM-227 No.15
2021/11/5

スポーツ映像における視線集中を利用したボール検出

中谷 千洋[1,a)]　川嶋 宏彰[2,b)]　浮田 宗伯[1,c)]

概要：コンピュータビジョンを用いたスポーツ映像解析において重要な要素の一つであるボール検出の精度向上のために，人群の視線がボールに集中しやすい事実を利用した手法を提案する．バレーボールやバスケットボールなどのスポーツ映像では，ボールは小さく高速に動くために検出が難しい．また，一般的なスポーツ映像では，視線を推定できるほど人物頭部が高解像度撮影されていない．しかし，ボール検出と比べて，人体の一部である頭部の検出や方位推定は容易であり，かつ，目線だけでボールを追うようなケースでなければ頭部方位によって視線を近似できる．さらに，ボールはシーン中に一つしかないのに対してプレーヤーは複数人存在するため，数人の頭部方位推定にミスがあったとしても多数決的にボールの位置を特定しやすい．そこで，この「頭部集合の方位群によって近似した視線集中領域」を「一般的な物体検出によるボール検出の結果」と統合することで，ボール検出精度を向上する手法を提案する．人群の頭部方位を用いない場合に比べ，提案手法によってボール検出の精度を向上できることを実験的に検証した．

1. はじめに

スポーツ動画解析はコンピュータビジョンにおいて重要な話題のひとつであり，戦術解析やシーン検索などの応用がある．スポーツ動画解析のためには，プレーヤーのそれぞれの行動を認識する個人行動認識 [1], [2], [3], [4]，複数人が協調した行動を認識するグループ行動認識 [5], [6], [7], [8]，ボール検出 [9], [10], [11] などが重要な技術である．本研究ではその中でもボール検出に注目する．

バレーボールやバスケットボールなどのスポーツ映像では，ボールは小さく早く動くために検出が難しい．よって，一般的な物体検出によるボール検出の結果にはミスが多い．そこで，そのような検出ミスを減らすためにスポーツ映像中の人群の視線がボールに集中しやすいという事実を利用する．例えば，図 1 に示したバレーボールシーンのように，一般的なスポーツ映像では人群の視線がボールに集中している場合が多い．しかし，一般的なスポーツ映像では人群の視線を推定できるほど人物頭部が高解像度撮影されていない．一方，ボール検出と比べると人体の一部である頭部の検出や方位推定は容易である．その上，目線だけでボールを追うようなケースでなければ頭部方位によって視線を近似できる．そこで，本研究では「頭部集合の方位群によって近似した視線集中領域」を「一般的な物体検出によるボール検出の結果」と統合することで，ボール検出精度を向上する手法を提案する．

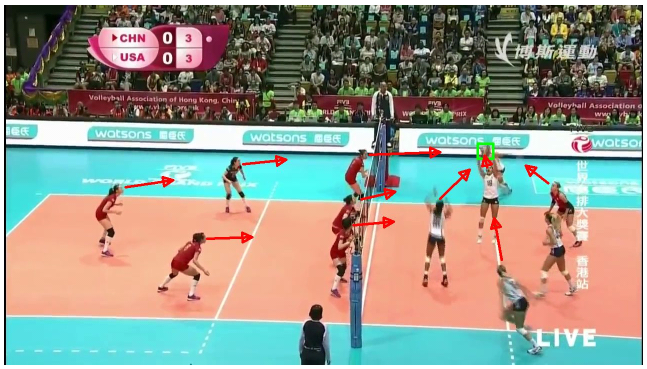

図 1 スポーツ映像におけるボールへの視線集中．バレーボールやバスケットボールなどのスポーツ映像では，プレーヤー群の視線がボールに集中しやすい．ボールの正解バウンディングボックスは緑色の四角形で示した．

本研究の貢献は，以下の 3 点である．

- 頭部集合の方位群により近似した視線集中領域によるボール検出手法を提案する．
- 頭部集合の方位群を表現する方法として，頭部方位分布の 2 次元マップ表現を提案する．全頭部方位を固定チャネル数のマップで表現することによって，各フレームに写っている対象人数が動的に変化するスポーツ映像に対しても適用可能な表現となっている．
- 「視線集中領域によるボール検出」と「一般的な物体

[1] 豊田工業大学
Toyota technological Institute, Nagoya, Aichi, 468-8511, Japan
[2] 兵庫県立大学
University of Hyogo, Kobe, Hyogo, 651-2197 , Japan
a) sd21431@toyota-ti.ac.jp
b) kawashima@sis.u-hyogo.ac.jp
c) ukita@toyota-ti.ac.jp

Vol.2021-CG-184 No.15
Vol.2021-DCC-29 No.15
Vol.2021-CVIM-227 No.15
2021/11/5

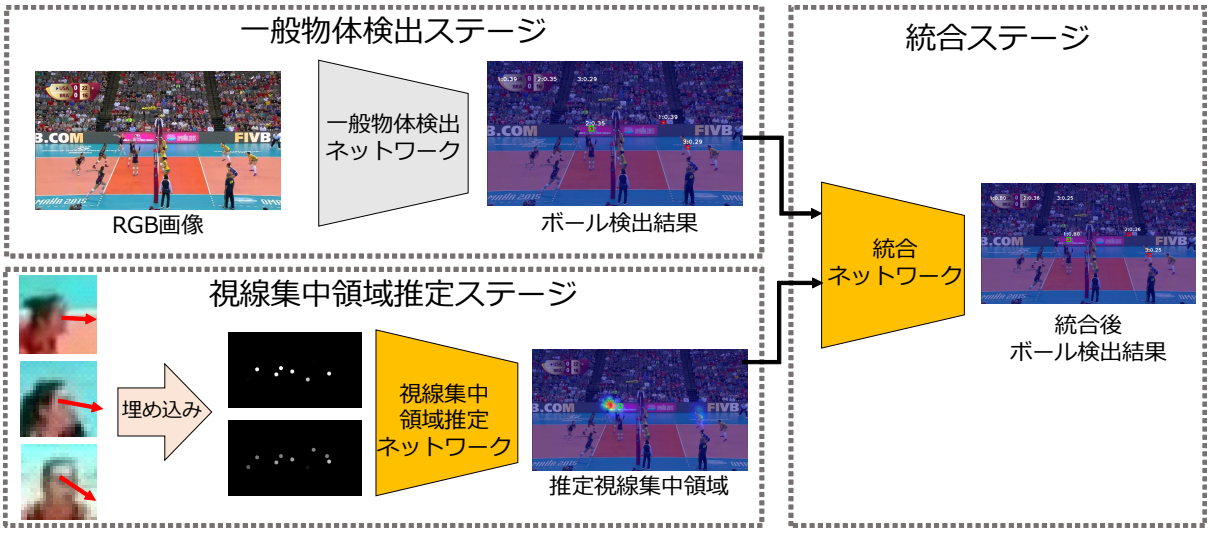

図 2 一般的な物体検出と視線集中領域を統合する提案ネットワーク．一般物体検出ステージで得られる一般物体検出によるボール検出の結果と，視線集中領域推定ステージ 2 で得られる視線集中領域を統合ステージで統合することでボール検出をする．

検出」をそれぞれを独立なサブネットワークで実行して，それらの結果を統合するフュージョンネットワークによって，ボール検出の精度を向上させる．

2. 関連研究

2.1 視線推定

視線推定とは，人の画像から視線方向を推定するタスクである．画像中の眼球姿勢から 2 次元視線方位を推定する手法 [12], [13] や，3 次元視線方位を推定する手法 [14], [15] がある．また，人の姿勢を用いて 2 次元視線方位を推定する手法 [16] や，画像中の人と物体位置の関係を考慮して 2 次元視線方位を推定する手法 [17] や，時系列情報を考慮した 3 次元視線方位を推定する手法 [18] などがある．これらの手法では人物頭部が高解像度撮影されている一方で，本研究で扱う一般的なスポーツ映像では視線を推定できるほど人物頭部が高解像度撮影されていないため，頭部方位を視線に近似することで視線推定をする．

2.2 頭部方位推定

頭部方位推定とは，人の画像から頭部方位を推定するタスクである．顔のランドマークから 3 次元頭部方位を推定する手法 [19], [20] や，顔のランドマークを用いずに顔画像から直接 3 次元頭部方位を推定する手法 [21] などがある．これらの手法では高解像度撮影された人物頭部画像を用いている一方，本研究で扱う一般的なスポーツ映像では人物頭部が高解像度撮影されていないため，低解像度頭部画像を扱った手法 [22] のように，2 次元頭部方位を推定する．

2.3 視線集中領域推定

視線集中領域推定に関する研究として，単一人物の視線方位から注視領域を推定する手法 [23] や，単一人物の視線方位と顕著性マップを統合することで視線推定をする手法 [17] がある．また，人群の視線と視線集中領域を同時最適化する手法 [24] もある．この手法は事前に得られた人群の視線と未知の視線集中領域を EM アルゴリズムにより同時最適化することで，人群の視線方向の精度を向上させている．この手法とは異なり，本研究では深層学習を用いることで学習ベースの視線集中領域推定をする．学習ベースのデータドリブンな手法を使うことで，スポーツ映像に特化した視線領域推定が可能である．

2.4 一般物体検出

一般物体検出とは画像中にある任意の物体を検出するタスクであり，盛んに研究されている．物体候補領域を検出した後に，そのクラスを推定する Faster-RCNN [25] や，物体候補領域の検出とそのクラス推定を同時にする YOLO [26]，物体のクラスと大きさをヒートマップで推定する CenterNet [27] などの手法がある．本研究では，視線集中領域をヒートマップとして推定することから，様々な一般物体検出手法と組み合わせることが容易である．提案手法では，CenterNet を一般的な物体検出手法の一例として使う．

2.5 マルチモーダルフュージョン

マルチモーダルフュージョンとは複数のモダリティを統合する手法である．RGB 画像を深度画像を統合する手法 [28] や，RGB 画像と人物姿勢を統合する手法 [29] など様々なモダリティの統合について研究されている．本研究では，共にヒートマップ形式であるボール検出結果と視線集中領域を統合する．ヒートマップは RGB 画像と比べて

Vol.2021-CG-184 No.15
Vol.2021-DCC-29 No.15
Vol.2021-CVIM-227 No.15
2021/11/5

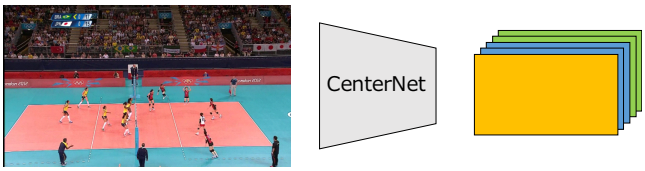

図 3 CenerNet [27] を用いたボール検出．CenterNet は RGB 画像を用いて，各ピクセルにおける物体の存在確率，サイズ，サイズのオフセットをヒートマップで予測する．物体の存在確率のヒートマップは 1 枚出力し，サイズとサイズのオフセットのヒートマップは画像座標上の x 軸と y 軸の両方についてを 2 枚ずつ出力する．

情報量が少ないため，様々なフュージョンネットワークの中でも比較的シンプルである Early fusion, Late fusion [30] を利用する．

3. 提案手法

我々が提案した 3 つのステージから構成されるネットワークを図 2 に示す．一般物体検出ステージでは，RGB 画像を入力にして一般的な物体検出によるボール検出をする．視線集中領域推定ステージでは，最初に頭部集合の方位群を推定し，その推定された頭部集合の方位群を 2 次元マップ形式に変換する．そのマップを入力にして，視線集中領域を推定する．統合ステージでは，一般物体検出ステージで得られたボール検出結果と，視線集中領域推定ステージで得られた視線集中領域を統合することで，最終的なボール検出結果を得る．

3.1 一般物体検出

一般物体検出によるボール検出には，CenterNet [27] を用いる．CenterNet は図 2 に示すように，RGB 画像を用いて，各ピクセルにおける物体の存在確率，サイズ，サイズのオフセットをヒートマップで予測する．物体の存在確率のヒートマップは物体のクラス数だけ出力する一方で，サイズとサイズのオフセットは全クラスで共通であり，画像座標上の x 軸と y 軸の両方についてを出力する．CenterNet は focal loss [31] を用いて学習される．

本研究において物体のクラス数はボールの 1 クラスのみであるため，図 3 に示すように物体の存在確率のヒートマップは 1 枚出力され，サイズとサイズのオフセットのヒートマップはそれぞれ 2 枚ずつ出力される．提案手法では，このようにして得られた 5 枚のヒートマップと，視線集中領域推定ステージで得られる視線集中領域のヒートマップ 1 枚を統合ステージで統合する．

3.2 頭部方位推定

図 2 に示した視線集中領域推定ステージの入力は頭部集合の方位群であり，図 4 のように 3 段階の処理で求められる．

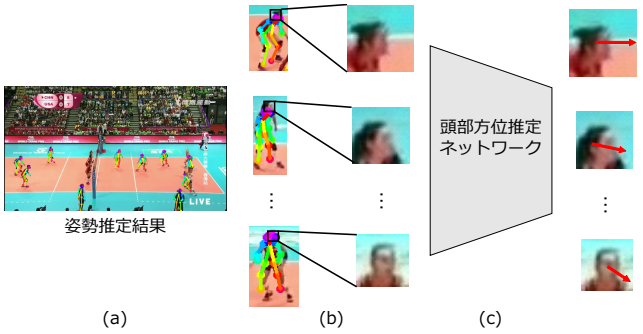

図 4 プレーヤーの頭部方位推定．(a) Openpose を用いて各プレーヤーの姿勢を推定する．(b) 推定された姿勢を基に，各プレーヤーの頭部領域をクロップする．(c) 頭部姿勢推定ネットワークを用いて，クロップされた頭部領域画像から頭部方位を推定する．

1 段階目 (図 4 (a)) では，OpenPose [32] を用いて RGB 画像上における人群の姿勢推定をする．2 段階目 (図 4 (b)) では，推定された人の姿勢を使って頭部領域をクロップする．使うキーポイントは鼻と首であり，それぞれの画像座標での位置を $K^{nose} = (K_x^{nose}, K_y^{nose})$，$K^{neck} = (K_x^{neck}, K_y^{neck})$ とすると，求めたい顔領域 $H = (H_{xmin}, H_{ymin}, H_{xmax}, H_{ymax})$ は鼻と首のキーポイントの距離 Dis を用いて以下のように計算される．

$$
\begin{aligned}
Dis &= ((K_x^{nose} - K_x^{neck})^2 + (K_y^{nose} - K_y^{neck})^2)^{0.5}, \\
H_{xmin} &= K_x^{nose} - Dis, H_{xmax} = K_y^{nose} + Dis, \\
H_{ymin} &= K_y^{nose} - Dis, H_{ymax} = K_y^{nose} + Dis,
\end{aligned} \tag{1}
$$

このようにしてクロップされた頭部領域画像を入力に，3 段階目 (図 4 (c)) では頭部方位推定ネットワークを用いて頭部方位を推定する．頭部方位推定ネットワークには，最初に畳み込み層 3 層による特徴量抽出を行い，Global Average Pooling (GAP) を適用し，最後に全結合層 1 層で頭部方位を推定するネットワーク 1 と，最初に ResNet による特徴量抽出を行い，その後 GAP を適用し，最後に全結合層 1 層で頭部方位を推定するネットワーク 2 を用いた．頭部方位は 2 つのスカラー値で推定され，それぞれ画像座標における頭部方位の x 軸方向成分と y 軸方向成分である．この頭部方位推定ネットワークは以下の Mean Squared Error loss により学習される．

$$
\mathcal{L}_h = \frac{1}{2}\left((G_x - E_x)^2 + (G_y - E_y)^2\right) \tag{2}
$$

G_x と G_y は，正解頭部方位ベクトルの画像座標における x 軸方向と y 軸方向の成分を示しており，E_x と E_y は，推定頭部方位ベクトルの画像座標における x 軸方向と y 軸方向の成分を示している．

Vol.2021-CG-184 No.15
Vol.2021-DCC-29 No.15
Vol.2021-CVIM-227 No.15
2021/11/5

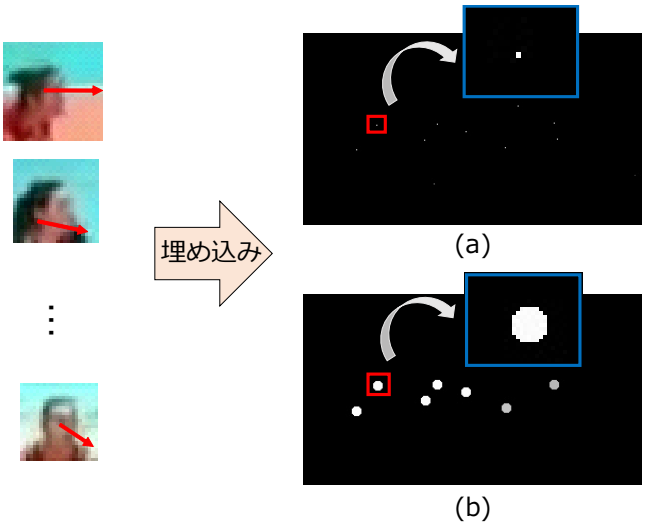

図 5 推定された頭部集合の方位群の画像への埋め込み方法．(a) 頭部方位情報を対応する頭部の中心座標にのみに埋め込む方法．(b) 頭部方法情報を対応する頭部の中心から一定範囲に値を埋め込む方法．

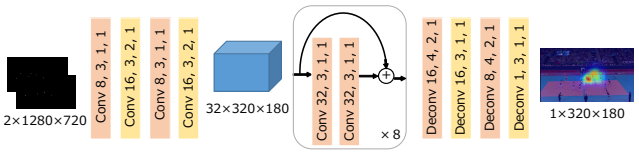

図 6 視線集中領域推定ネットワーク．頭部集合の方位群が埋め込まれた 2 枚のマップを用いて視線集中領域を推定する．Conv の後に続く 4 つの数字は，出力チャネル数，カーネルサイズ，dilation, padding の値を示す．

3.3 人群の頭部方位を用いた視線集中領域推定

このように 3.2 節で推定された頭部集合の方位群を用いて視線集中領域を推定する．頭部集合の方位群を用いた視線集中領域は，2 段階の処理により推定される．

1 段階目では， 3.2 節で得られた頭部集合の方位群を，2 枚のマップに埋め込む．1 枚目のマップの検出頭部座標には，推定頭部方位ベクトルの x 軸方向成分を，2 枚目のマップの検出頭部座標には，推定頭部方位ベクトルの y 軸方向成分を埋め込む（図 5（a））．ただし，この x，y 軸方向成分を検出頭部座標の 1 座標にだけ埋め込んでしまっては，頭部座標が少しずれただけでも完全に異なる表現になってしまう．そこで，頭部座標のずれを許容するためのデータ拡張として，検出頭部座標を中心とした一定範囲内に推定頭部方位ベクトルを埋め込む方法（図 5（b））を提案する．このように，頭部集合の方位群を 2 次元マップ上に埋め込むことで，認識対象人数が動的に変化するようなスポーツ映像にも適用可能な手法となる．

このようにして得られた 2 枚のマップを入力にし，2 段階目では図 6 に示す視線集中領域推定ネットワークを用いて視線集中領域をヒートマップで推定する．視線集中領域

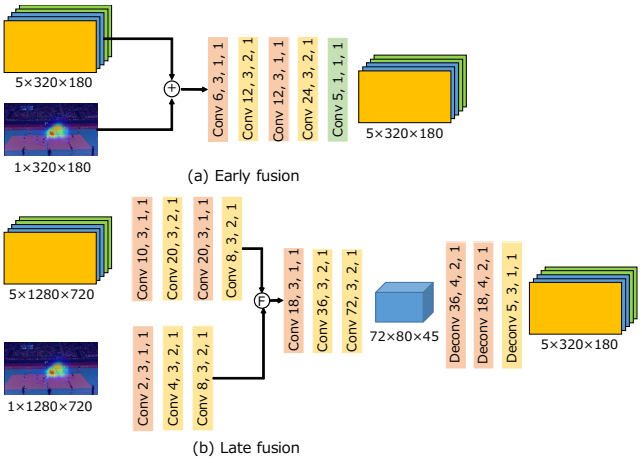

図 7 ボール検出と視線集中領域の統合ネットワーク．Conv の後に続く 4 つの数字は，出力チャネル数，カーネルサイズ，dilation, padding の値を示す．(a) ボール検出結果と視線集中領域を連結し，その後畳み込む Early fusion ネットワーク．(b) ボール検出結果と視線集中領域をそれぞれ畳み込んだ後，element-wise addition もしくは element-wise multiplication により統合する Late fusion ネットワーク．

推定ネットワークは以下の Binary Cross Entropy loss により学習される．

$$\mathcal{L}_b = \sum_i (-G_{x,y} log(s(E_{x,y})) \\ -(1-G_{x,y}) log(1-s(E_{x,y}))) \quad (3)$$

$E_{x,y}$ と $G_{x,y}$ は，推定された視線集中領域のヒートマップと正解ヒートマップの画像座標 (x, y) でのピクセル値を示している．$s(\cdot)$ はシグモイド関数である．視線集中領域の正解ヒートマップは，ボールの正解バウンディングボックスを基に生成する．ボールの正解バウンディングボックス座標を $P = (p_{xmin}, p_{ymin}, p_{xmax}, p_{ymax})$ とし，その中心座標を $(p_{xmid}, p_{ymid}) = (\frac{p_{xmin}+p_{xmax}}{2}, \frac{p_{ymin}+p_{ymax}}{2})$ と表すと，視線集中領域の正解ヒートマップは以下のガウス分布状の分布から生成する．

$$G_{x,y} = \exp\left(-\frac{(x-p_{xmid})^2 + (y-p_{ymid})^2}{2\sigma^2}\right), \quad (4)$$

σ は分布の分散を決めるパラメタであり，本研究では実験結果から 0.1 と定めた．

3.4 ボール検出と視線集中領域の統合

図 2 に示した統合ステージでは， 3.1 節で得られた一般的な物体検出によるボール検出の結果と， 3.3 節で得られた頭部集合の方位群によって近似した視線集中領域を統合し，最終的なボール検出結果を得る．統合のためには，図 7 に示した 2 種類のネットワークを提案する．図 7（a）に示した Early fusion ネットワークは， 3.1 節で得られた一般的な物体検出によるボール検出の結果と 3.3 節で推定され

Vol.2021-CG-184 No.15
Vol.2021-DCC-29 No.15
Vol.2021-CVIM-227 No.15
2021/11/5

表 1 頭部方位推定の角度誤差（deg）の比較．Resnet を用いたネットワーク 2 で，鼻，左肩，右肩のキーポイントを入力に用いた場合に最も角度誤差が小さかった．

ネットワーク	キーポイント	角度誤差（deg）
ネットワーク 1	なし	69.2
ネットワーク 1	あり	62.0
ネットワーク 2	なし	61.7
ネットワーク 2	あり	**58.5**

表 2 2 つの頭部集合の方位群の画像への埋め込み方法を用いた場合の視線集中領域推定結果の比較．頭部集合の方位群を頭部の中心から一定範囲に埋め込む場合の結果が 3 つの項目全てで誤差が小さかった．

埋め込み範囲	x 軸方向誤差	y 軸方向誤差	距離誤差
中心座標	44	15	50
一定範囲	**35**	**14**	**41**

た視線集中領域を連結し，その後畳み込むことで最終的なボール検出をする．一方で，図 7（b）に示した Late fusion ネットワークは，3.1 節で得られた一般的な物体検出によるボール検出結果と，3.3 節で推定された視線集中領域をそれぞれ畳み込み，その後 element-wise addition もしくは element-wise multiplication によって統合してから更に畳み込むことで，最終的なボール検出結果を得る．この統合のためのネットワークも，3.1 節で紹介した CenterNet と同様に focal loss [31] を用いて学習する．

4. 実験結果

4.1 データセットと性能評価方法

提案手法の評価のため，バレーボールデータセット [33] を使用した実験結果を示す．バレーボールデータセットは，55 のバレーボール動画から抽出された 4830 のシーケンスで構成される．1 シーケンスは 41 フレームで構成され，中心フレームには各プレーヤーの個人行動ラベルとバウンディングボックスがアノテーションされている．個人行動ラベルは 9 種類（waiting，setting，digging，falling，spiking，blocking，jumping，moving，standing）である．

本研究では，上述したデータセット [33] 標準のアノテーションに加えて，以下のアノテーションを追加で与えた．まず，全フレームに対してボールのバウンディングボックスを vatic [34] というツールを用いてアノテーションした．このアノテーションデータを式 (4) で利用し，視線集中領域の正解ヒートマップを生成する．頭部方位については，頭部領域からボールの中心座標への方向を正解頭部方位として自動で粗いアノテーションをした．ただし，スパイクやジャンプなどをしているプレーヤーはボールに注目していないことが多いため，ボールに注目していることが多い個人行動ラベルが “standing” と “waiting” であるプレーヤーのみにアノテーションをした．このようにアノテーションをすることで，誤差が小さいアノテーションデータのみで学習が可能である．このアノテーションデータを式 (2) で利用し，頭部方位推定ネットワークを学習する．しかし，テストには誤差ができるだけ小さいアノテーションデータを使うことが適切である．よって，自動アノテーション結果を人が目視で確認し，自動アノテーション結果と本来の頭部方位との誤差が明らかに大きいデータはテストに使用しない．また，テストデータの頭部方位に偏りがあるため，頭部方位が均等になるようにサンプリングすることで，様々な頭部方位に対する精度を評価する．

頭部方位推定の評価には [16], [17] のように頭部方位の角度誤差を用いた．視線集中領域はヒートマップのピークがボールの中心座標に近いかで評価する．最初に，ヒートマップを 0 から 1 に正規化した後に，閾値 0.9 で二値化する．その後，値が 1 になったピクセル群とボールの中心座標との距離をピクセルごとに計算する．最後に，そのピクセルごとに計算されたボールの中心座標との距離の平均を求める．このようにして求めた平均距離で視線集中領域推定を評価する．ボール検出の評価には，一般物体検出の評価に使われる Average Precision（AP）を用いる．また，本研究のボール検出では [9] と同様に画像上にボールが一つしかないという制約がある．よって，AP に加えてボール検出結果の中で最も信頼度の高い検出結果のみを考慮する Accuracy も評価に使う．

4.2 頭部方位推定

頭部方位推定の結果を表 1 に示す．頭部方位推定の精度向上のために頭部領域画像に加えて鼻，左肩，右肩のキーポイントを入力に用いる場合でも実験した．これらのキーポイントを用いることで，顔がカメラに正対しているか後頭部がカメラに正対しているか判別しにくい場合のヒントになるため，頭部方位推定精度の向上が可能である．ネットワーク 1 でキーポイントを入力に用いた場合（表 1 の 1 行目）と用いなかった場合（表 1 の 2 行目）を比較すると，キーポイントを用いた場合の方が角度誤差が小さかった．同様に，ネットワーク 2 でキーポイントを入力に用いた場合（表 1 の 3 行目）と用いなかった場合（表 1 の 4 行目）を比較すると，キーポイントを用いた場合の方が角度誤差が小さかった．よって，鼻，左肩，右肩のキーポイントを用いることで，顔がカメラに正対しているような頭部方位推定が難しい場合のヒントになり精度が向上したと考えられる．

4.3 人群の頭部方位を用いた視線集中領域推定

人群の頭部方位のマップへの埋め込み方法として，対応する頭部の中心座標にのみ値を埋め込む方法と（図 5 (a)），対応する頭部の中心から一定範囲に値を埋め込む方

Vol.2021-CG-184 No.15
Vol.2021-DCC-29 No.15
Vol.2021-CVIM-227 No.15
2021/11/5

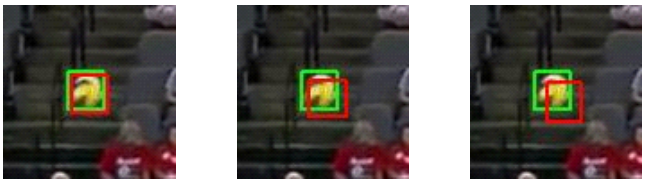

図 8 IoU が 0.75，0.50，0.25 での正解バウンディングボックスと検出バウンディングボックスとの重なり．正解バウンディングボックスは緑枠，検出バウンディングボックスは赤枠で囲った．

表 3 個人行動ラベルを入力に用いた視線集中領域推定の結果．頭部方位の埋め込み方法として頭部の中心から一定範囲に埋め込む方法を用いている．個人行動ラベルを用いた場合の方が 3 つの項目全てで誤差が小さかった．

個人行動ラベル	x 軸方向誤差	y 軸方向誤差	距離誤差
なし	35	14	41
あり	**28**	**13**	**34**

法（図 5 (b)）を用いた場合の視線集中領域推定結果を表 2 に示す．対応する頭部の中心から一定範囲に値を埋め込む場合が 3 つの項目全てで誤差が小さかった．この理由としては，頭部の中心から一定の範囲に値を埋め込むことで頭部位置のずれに対してネットワークが頑健になったことが考えられる．

また，4 節で述べたようにプレーヤーそれぞれには個人行動ラベルがアノテーションされている．その個人行動ラベルを視線集中領域推定ネットワークの入力に用いた場合と用いない場合の結果を表 3 に示す．個人行動ラベルを用いる場合が個人行動ラベルを用いない場合に比べて，3 つの項目全てで誤差が小さかった．特に，x 軸方向誤差，距離誤差は 41 から 34 へと改善した．これは，個人行動ラベルを用いることでスパイクやブロックなどで頭部の動きが安定しないプレーヤーを特定し，その頭部方位の重要度を低くするような学習がされるからだからと考えられる．

4.4 ボール検出

「推定された視線集中領域」と「一般物体検出によるボール検出結果」の統合ネットワークを変化させた際のボール検出結果を，表 4 に示す．IoU の閾値が 0.25 の場合（表 4 の 2，5 列目）は Early fusion での結果が最もよい．一方で，IoU の閾値が 0.50, 0.75 の場合（表 4 の 3，4，6，7 列目）は Late fusion（element-wise addition）での結果が最もよい．本研究ではボールという他の物体に比べて小さな物体を検出していることから，図 8 に示すように，IoU の閾値が 0.25 であっても十分にボールは検出されているといえる．よって，統合ステージにおける推定された視線集中領域と一般的な物体検出によるボール検出結果の統合には Late fusion（element-wise addition）を用いた場合の精度が最も良いといえる．

次に，一般物体検出によるボール検出結果と提案手法でのボール検出結果の比較を 5 に示す．提案手法の統合ステージでの統合ネットワークには表 4 で最も精度が良かった Late fusion（element-wise addition）を用いた．提案手法によるボール検出の結果は，一般物体検出によるボール検出の結果に比べて，大幅に精度が向上している．このように精度が向上した理由としては，一般的なボール検出手法によるボール検出の結果の信頼度が低い場合に，推定された視線集中領域が大きなヒントになるからだと考える．実際に，図 9 に示すように，一般物体検出によるボール検出結果（図 9（a））の信頼度が低い場合でも，視線集中領域推定結果（図 9（c））がおおよそボールの領域を推定していれば，提案手法によるボール検出（図 9（d））は成功している．

5. まとめ

スポーツ映像中の人群の視線がボールに集中しやすいという事実を利用し，一般的な物体検出によるボール検出の結果と，推定視線集中領域を統合することでボール検出の精度向上を目指した．実験結果から，視線集中領域を用いることでボール検出の AP，Accuracy が約 10%向上することを確認した．また，一般的な物体検出によるボール検出で正解ボールバウンディングボックスが 2 番目，3 番目に検出されているような場合に，視線集中領域を統合することで正しいボール検出が可能になることを定性的に示した．

今後の課題として，視線集中領域推定の更なる精度向上が挙げられる．特に，self-attention により頭部方位群の頭部方位それぞれに重み付けすることで，視線集中領域推定に重要な頭部方位のみを考慮することが考えられる．

参考文献

[1] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.
[2] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.
[3] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J. Black. On the integration of optical flow and action recognition. In *GCPR*, 2018.
[4] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.
[5] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019.
[6] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *CVPR*, 2019.

Vol.2021-CG-184 No.15
Vol.2021-DCC-29 No.15
Vol.2021-CVIM-227 No.15
2021/11/5

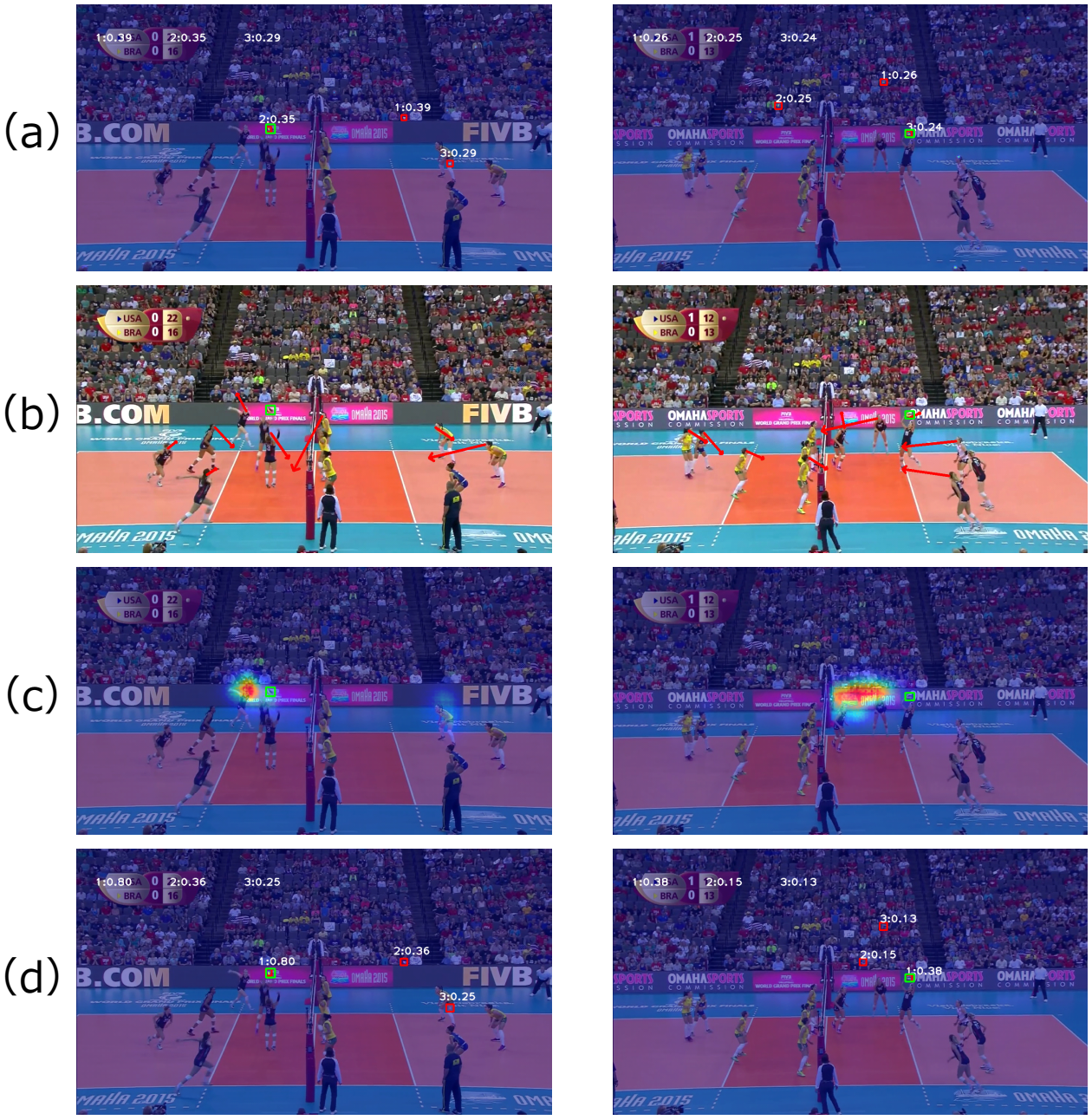

図 9 一般的な物体検出によるボール検出の結果と提案手法の比較．ボールの正解バウンディングボックスは緑色の四角形で示した．また，(a) と (d) では検出されたボールの中で信頼度が大きい順に 3 つを赤枠の四角形で示した．推定された視線集中領域を使うことで，ボール検出が正しく検出されるようになることがわかる．(a) CenterNet によるボール検出の結果．(b) 頭部集合の方位群の推定結果．(c) 視線集中領域の推定結果．(d) 提案手法によるボール検出の結果．

[7] Chihiro Nakatani, Kohei Sendo, and Norimichi Ukita. Group activity recognition using joint learning of individual action recognition and people grouping. In *MVA*, 2021.

[8] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *CVPR*, 2020.

[9] Jacek Komorowski, Grzegorz Kurzejamski, and Grzegorz Sarwas. Deepball: Deep neural-network ball detector. In *VISAPP*, 2019.

[10] Jacek Komorowski, Grzegorz Kurzejamski, and Grzegorz Sarwas. Footandball: Integrated player and ball detector. In *VISAPP*, 2020.

[11] Yu-Chuan Huang, I-No Liao, Ching-Hsuan Chen, Tsì-Uí Ik, and Wen-Chih Peng. Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. In *AVSS*, 2019.

[12] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Mach. Vis. Appl.*, 28(5-6):445–461, 2017.

[13] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini

表 4 様々な統合ネットワークを用いた場合のボール検出結果．IoU の閾値が 0.25 の場合は Early fusion の結果が最も良く，IoU の閾値が 0.50，0.75 の場合は Late fusion（element-wise addition）の結果が最も良い．

統合手法	AP_{25}	AP_{50}	AP_{75}	$Accuracy_{25}$	$Accuracy_{50}$	$Accuracy_{75}$
Early fusion	0.55	**0.50**	**0.15**	0.564	**0.503**	**0.139**
Late fusion (concat)	0.54	0.41	0.08	0.563	0.426	0.090
Late fusion (element-wise multiplication)	0.56	0.44	0.10	0.600	0.449	0.092
Late fusion (element-wise addition)	**0.57**	0.47	0.12	**0.606**	0.481	0.116

表 5 一般的な物体検出によるボール検出の結果と提案手法でのボール検出結果の比較．提案手法の統合ステージにおける統合ネットワークには Late fusion（element-wise addition）を用いている．全ての IoU の閾値において提案手法が CenterNet の結果を上回っている．

手法	AP_{25}	AP_{50}	AP_{75}	$Accuracy_{25}$	$Accuracy_{50}$	$Accuracy_{75}$
CenterNet	0.49	0.20	0.01	0.505	0.192	0.017
提案手法	**0.61**	**0.49**	**0.12**	**0.606**	**0.481**	**0.116**

Kannan, Suchendra M. Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2176–2184. IEEE Computer Society, 2016.

[14] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, 2018.

[15] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *ICCV*, 2019.

[16] Philipe A. Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. Gaze estimation for assisted living environments. In *WACV*, 2020.

[17] Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *NeurIPS*, 2015.

[18] Petr Kellnhofer, Adrià Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019.

[19] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[20] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1087–1096. Computer Vision Foundation / IEEE, 2019.

[21] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *CVPR*, 2018.

[22] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Fast and accurate head pose estimation via random projection forests. In *ICCV*, 2015.

[23] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, Lecture Notes in Computer Science, 2018.

[24] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *ICCV*, 2013.

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[26] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[27] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019.

[28] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *CVPR*, 2018.

[29] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[30] Seungyong Lee, Seong-Jin Park, and Ki-Sang Hong. Rdfnet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *ICCV*, 2017.

[31] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[32] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[33] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016.

[34] Carl Vondrick, Donald J. Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation - A set of best practices for high quality, economical video labeling. *Int. J. Comput. Vis.*, 101(1):184–204, 2013.