Vol.2018-CVIM-212 No.17
2018/5/10

シーン全体のイベント理解のためのグループ行動認識

千藤 滉平[1,a)]　ムハンマド ハリス[1,b)]　浮田 宗伯[1,c)]

概要：
本研究では，「複数人の動作の認識結果」と「各動作に関係のある物体の検出結果」を統合的に参照することで，シーン全体で起きているイベントを理解するためのグループ行動認識を提案する．物体検出や人の動作認識は数多く研究され，それぞれの手法の精度は大きく向上してきているが，それらすべての関係性を考慮したグループ行動認識は，まだ検討が大きく進んでいない分野である．提案するグループ行動認識法では，まず人と物体の領域検出およびその領域のクラス識別を行う．識別クラスは，人領域においては各人の動作の種類であり，物体領域についてはその物体の種類である．提案手法では，これら各領域のクラス識別結果に加え，グループ行動認識のためには「各領域間の相対的な位置関係」も重要な情報である一方，その他背景領域は認識の汎化性向上のためには不要な情報であると仮定した．この仮定に基づき，検出領域のみから特徴量を抽出した特徴量マップを生成し，この特徴量マップとグループ行動ラベルとを対応付けて識別器を学習する．この学習には，特徴量マップを入力とする畳み込み深層学習を利用することで，行動や物体の特徴量と空間的な位置関係を頑健にモデル化する．評価実験のため，バスケットボールの動作である「ドリブル，パス，レイアップシュート，ジャンプシュート」の 4 種のグループ行動からなる動画データセットを用意した．実験では，個別の物体・行動検出の精度が平均 74.83 %である．この検出誤りを提案手法のグループ行動認識のフレームワークで考えることにより解決することができた．

キーワード：グループ行動認識　行動認識　物体検出　行動検出

1. はじめに

現在様々なチームスポーツの作戦において，選手のプレイなどの大量のデータを用いることによって戦術を練ることが多い．このように大量のデータを統計解析することによって新しい戦術を発見する手法を戦術解析と定義する．現状では，データ数が少なかったり，特定のスポーツ（バレーやバスケなど）に特化したものが多い．また様々なサービス [1],[2],[3] も始まったばかりのものも多く，より精度向上のため様々な研究が行われている．本研究の目的としては戦術解析にグループ行動認識の技術を用いることにより今まで以上のデータを扱うビックデータ解析を行うことで，今までとは価値観の違った戦術の発見をすることである．

この戦術解析をチームスポーツの試合で行うために，各選手の行動，配置状況，試合状況，選手の能力等の情報を得ることが必要である．これらの情報はすべて重要でデータ化しビッグデータ解析できるものばかりである．特にこの中で各選手の行動や配置状況はどこで誰が上手くいったのかなど条件の組み合わせが多く，そういった情報こそビッグデータ解析すべきであり，映像から自動的に大量にデータを集める価値のあるものである．本研究では以上の理由から，各選手の行動，配置状況の 2 つの情報を映像から自動的に抽出することに注力した．そして，この 2 つの情報を基にそれらすべての関係性を基にしたシーン全体の動作の情報を得るグループ行動認識を目標とする．

本研究では「複数人の動作の認識結果」と「各動作に関係のある物体の検出結果」を統合的に参照することで，シーン全体で起きているイベントを理解するためのグループ行動認識を提案する．本研究のグループ行動認識は以下の 3 つを新規性としてやっていく．

- 「複数人の動作の認識結果」と「各動作に関係のある物体の検出結果」を基にグループ行動認識を行う．
- 複数人の相対的な配置関係を陽に学習する．
- 上の 2 つ以外の情報であるその他背景領域は認識の汎化性向上のためには不要な情報であると仮定し排除する．

[1] 豊田工業大学院　先端工学専攻
Toyota technological Institute, Department of Advanced Science and Technology，Japan
[a)] sd18419@toyota-ti.ac.jp
[b)] muha.haris@gmail.com
[c)] ukita@toyota-ti.ac.jp

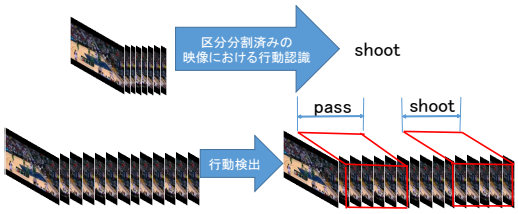

図 1 区分分割済みの映像における行動認識と時空間行動検出の違い．区分分割済みの映像における行動認識は 1 つのセグメント済みの動画に 1 つの行動が与えられ，それを候補ラベルのいずれかに分類する手法である．それに対して時空間行動検出は，どこでどういう動作がなされているのかという空間的な検出と，長い動画においてどの時間で行動が行われているのかという時間的な検出の両方を行っているものが時空間行動検出である．

2. 関連研究

本研究では時空間行動検出を行った結果を用いてグループ行動認識を行っている．この方法には区分分割済みの映像における行動認識や物体検出など様々な方法を組み合わせることによって成り立っている．本章では本研究の基礎技術である区分分割済みの映像における行動認識や物体検出から，それを拡張させた時空間行動検出やグループ行動認識について説明する．

2.1 区分分割済みの映像における行動認識

区分分割済みの映像における行動認識は図 1 の上のように 1 つのセグメント済みの動画に 1 つの行動が与えられ，それを候補ラベルのいずれかに分類する手法である．この手法は ILSVRC 2012 において CNN の手法が劇的な精度を出したため，様々な研究でディープニューラルネットが使われるようになった．区分分割済みの映像における行動認識の手法としては，文献 [4],[5],[6],[7], [8],[9] などがある．文献 [4] は CNN を使った行動認識を, 文献 [5] は RGB とオプティカルフローを CNN で処理した行動認識を, 文献 [6] は 3 次元の CNN を使った行動認識を, 文献 [7] は RNN を用いた行動認識を，文献 [8] は LSTM を使った行動認識，文献 [9] は LTC-CNN を用いた行動認識が行われている．最近の傾向として複数フレームを用いることによって認識精度を上げようとする傾向が多く見られる．

2.2 物体検出

物体検出の手法として, 文献 [10],[11],[12],[13],[14],[15] がある．物体検出を行う際に，文献 [10],[11] は選択的検索法 (selective search)[16] を, 文献 [12] は領域提案ネットワーク (RPN) を用いて検出領域の提案生成を行っている．また文献 [12] からは提案生成および物体検出のネットワークを 1 つにまとめたエンドツーエンドの学習で行われようになっている．これらに対して提案生成の過程をなくし，直接的に物体検出を行おうとしたものとして文献 [13],[14],[15] がある．文献 [13] は入力画像を領域分割し，その分割された領域ごとにバウンディングボックスを与えて検出している．この文献 [13] は 1 つの特徴量マップを使用しているが，文献 [14],[15] は異なる解像度の特徴量マップを使用することによって異なるアスペクト比，スケールの物体も検出できるようになった．また文献 [15] はバウンディングボックスを k-means 法を用いて最適なものを見つけるように改良した．このように物体検出の手法は年々，急速に改良されていることがわかる．

2.3 時空間行動検出

時空間行動検出は, 先ほど説明した区分分割済みの映像における行動認識（図 1 の上部）に比べて，どこでどういう動作がなされているのかという空間的な検出と，長い動画においてどの時間で行動が行われているのかという時間的な検出の両方を行っているもの（図 1 の下部）である．時空間行動検出の手法として，文献 [17],[18],[19],[20], [21],[22] がある．文献 [17] は選択的検索法 [16] を用いた行動検出を, 文献 [18] は R-CNN[10] を用いた行動検出を, 文献 [22] は Faster-R-CNN[12] を用いた行動検出を，文献 [19]，[20] は SSD[14] を用いた行動検出を行っている．また文献 [21] は検出結果をそのまま使うのではなく，フレームごとの検出を滑らかにつなげるためのアルゴリズムを使用したことによって検出精度が向上し，さらに文献 [20] は [21] のアルゴリズムを改良して，複数フレームごとの検出を滑らかにつなげるアルゴリズムに改良したことによってさらに検出精度を上げた．文献 [17] は，選択的検索法で見つけた候補ボックスを Dense Trajectories を使って複数フレームで追跡することによって，ロバストにマッチした候補ボックスを手に入れる．この手に入れた候補ボックスを教師なし学習の学習データとして用いることで教師なし学習を可能にしている．文献 [22] は弱教師学習のラベルとしてシーンのラベルのみ与えられ, 映像中で物体検出したものの中に特徴的な動きや類似する動きを見つけることによって，動画同士を比較し共通な動きを行動検出とすることによって弱教師学習を可能にしている．本研究ではグループ行動認識の精度を上げるためには高い検出精度を必要とため，文献 [20] を改良してグループ行動認識を行う．

2.4 グループ行動認識

グループ行動認識の手法として，文献 [23],[24],[25],[26],[27],[28] がある．文献 [23] は, 人々の相互作用と人々の特徴量レベルの相互作用を組み合わせることによってグループ行動認識を行っている．文献 [24] は, 個々の人を認識したうえで行動の役割ごとに関係性を導くことによってグループ行動認識を行っている．文献 [28] は非線形関数によるアンサンブル学習を行うこと

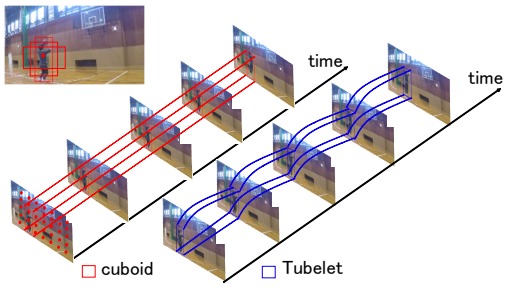

図 2 Anchor cuboid と Tubelet のイメージ図.Anchor cuboid は一定間隔に anchor cuboid の核がありその核ごとに異なるアスペクト比およびサイズの anchor cuboid が与えられる．また，Anchor cuboid は全てのフレームにおいてボックスのサイズ，アスペクト比，位置は不変であるのに対し，Tubelet はフレームごとにボックスのサイズ，アスペクト比，位置が変わる．

によって得る隠れ変数から，関係性を導くことでグループ行動認識を得る方法を行っている．文献 [26] は映像中からシーンや人々の特徴量を CNN で特徴量を抽出した後，RNN に与えることによってそれぞれの関係性を割り出すことによってグループ行動認識の精度を向上させている．文献 [25] はバウンディングボックスから抽出した人の特徴量とその特徴量を LSTM に与えたことにより出力される特徴量をプーリングし，さらに LSTM に与えることによって階層的なグループ行動認識を実現することによって精度を向上させている．文献 [27] は，グループ行動認識を行うためにその行動に最も関係している人たちのみ認識するという観点から，RNN を用いて特徴的な動きの特徴量を処理することによって弱教師学習によるグループ行動認識を行っている．

このように様々な観点でグループ行動認識が行われているが，本研究では個々の認識精度の向上からグループ行動認識の精度向上のため，[20] の手法を用いて，「複数人の動作」および「各動作に関係のある物体」をそれぞれ検出する．そしてそれらを用いて背景情報を削除することによって，認識に有益となる情報のみを用いたグループ行動認識を行う．

3. 個別検出の認識手法

本章では個別検出の認識手法である Action Tubelet Detector(ACT-detector)[20] を説明する．本章では ACT-detector の特徴，処理概要，学習について述べる．

3.1 ACT-detector の特徴

ACT-detector[20] とは動画に対する個別の行動検出の手法である．この手法の特徴として行動検出した結果を用いることによって動画における行動認識を行う点である．フレーム単位で行動認識を行うとそのフレームのどこに人がいるのかで見え方が変わるので，一般的に認識は難しくなる．しかし，ACT-detector による行動認識はフレームのどこに人がいるのかを位置を特定し，バウンディングボックスではみ出ることなく囲む．それによりフレーム全体で行う行動認識より人の見え方のバリエーションを減らすことができ，認識精度が向上する．またこの ACT-detector では行動検出のために図 2 の左図である Anchor cuboid を使って図 2 の右図である Tubelet を作成する．図 2 の左図である Anchor cuboid は全てのフレームにおいてボックスのサイズ，アスペクト比，位置は不変であるが，図 2 の右図である Tubelet はフレームごとにボックスのサイズ，アスペクト比，位置が変わるという特徴を持つ．

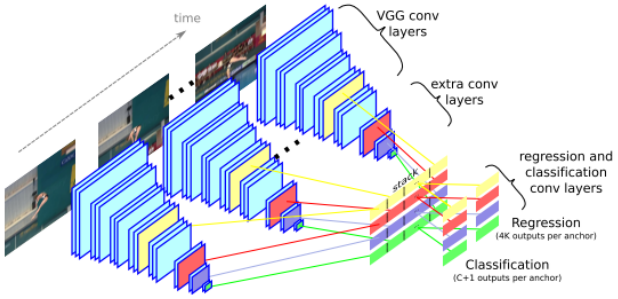

図 3 ACT-detector の構造. 動画全体から K のフレームにおいて特徴量を検出し，それを stack に貯めておく．それを基に回帰として $4 \times K$([バウンディングボックスの中心座標 xy およびその幅と高さ]×[Anchor cuboid に含まれるフレーム数]) の座標と分類として $C+1$(判別するクラス数+背景) クラスのスコアを出力する．そして次は 1 つずれた $K-1$ フレーム重なっている K フレームで同じ処理を繰り返す．この図は ACT-detector[20] の論文より引用．

3.2 ACT-detector の処理概要

ACT-detector は図 3 のように動画全体から K のフレームの画像として入力し，分類として「判別するクラス数 + 背景のスコア」を回帰として $4 \times K$「[バウンディングボックスの中心座標 xy およびその幅と高さ]×[Anchor cuboid に含まれるフレーム数]」の座標を Anchor cuboid ごとに出力する．この処理の概要は次の 3 つで行われる．

step1. 特徴量抽出 (3.2.1) 動画から一定間隔で K フレームを取り出し，SSD(Single Shot MultiBox Detector)[14] を用いて各フレームから異なる解像度の特徴量を抽出し stack に貯める．

step2. Tubelet の生成 (3.2.2) Anchor cuboid ごとに $4 \times K$ の座標と $C+1$ のスコアを出力し，K フレームによって作成された Tubelet ができる．

step3. Tubelet の連結 (3.2.3) K フレームごとに作成された Tubelet を滑らかにつなげることによって動画のフレーム数によって作成された Tubelet を作成する．

3.2.1 SSD による特徴量の抽出

図 3 を見ると ACT-detector は動画から一定間隔で K フ

レームを取り出し，その K フレームごとに SSD[14] を用いて特徴量を取り出す．SSD は VGG と extra の 2 段構成となっている．まず VGG で画像から特徴量を抽出し，その特徴量を extra で異なる解像度の特徴量マップに変化させてることで様々な物体検出を可能としている．それを K フレームすべて行い各フレームの異なる解像度における特徴量を図 3 の stack に貯める．この K フレームの出力が終わった後は 1 つずらした K フレームで同じ処理が行うことにより計算の効率性を向上させている．

3.2.2 Anchor cuboid ごとにクラス分類および座標の回帰し，K フレームの Tubelet を作成

まず図 2 の左上の画像のように異なる解像度において赤い点で示されている核を一定間隔でばらまく．この核は異なるアスペクト比，サイズのデフォルトボックスのセットの中心であり，これを K フレームに渡って拡張したものが Anchor cuboid となる．この Anchor cuboid の長さはどれだけ動く人を Anchor cuboid 内に収められるかで決まる．もし K が小さければ少ないフレーム数を基に検出位置を特定する必要があり，K が大きいと Anchor cuboid に動いている人がはみ出るので正確に位置特定が難しくなる．文献 [20] ではこの二つの最適値として $K = 6$ としている．この Anchor cuboid の周辺で位置を探すことにより人の動きを捉えた検出としてバウンディングボックスの中心座標およびその幅と高さを回帰し，Anchor cuboid ごとにクラスと背景のスコアを分類として出力する．これら

によって K フレームにおける Tubelet ができる．この K フレームの Tubelet は次の処理のために閾値 0.3 によって NMS を行い，スコアの高い順に 10 個の Tubelet を保持する．

3.2.3 K フレームの Tubelet を滑らかに結合

最後に先ほど保持した K フレームの Tubelet を滑らかにつなげることによって動画のフレーム数の Tubelet を作成する．この作成方法は文献 [21] のアルゴリズムを改良することによって実現している．この動画のフレーム数の Tubelet は以下の過程で作られる．

- 初期値として動画の最初の K フレームから作成された Tubelet の 10 個をリンクとして保持する．このリンクは 10 個できる．
- 以下の条件でリンクに K フレームの Tubelet を追加していく．(1) リンクにある Tubelet のスコアが高い順から次の K フレームの Tubelet 候補から選ばれていないもの．(2) 最もスコアの高いもの.(3) リンクにある最後に追加された Tubelet と Tubelet 候補の重なっている $K-1$ フレームのそれぞれの IoU の平均が 0.2 以上のもの．
- Tubelet 候補の長さが K フレームを満たさなくなったら終了する．
- あるリンク内において，フレームごとにみるとそれぞれの Tubelet には重なりがある．その重なりのあるフレームらの座標を平均することによってリンクごとに動画のフレーム数の Tubelet ができる．つまり 10 個の動画のフレーム数の Tubelet ができる．

3.3 ACT-detector の学習

学習データとして必要なものはフレーム分割された動画と各フレームの検出したいものに対する Tubelet のバウンディングボックスの座標，そのクラスである．まず，正例 P を Anchor cuboid のデフォルトボックスのセットの中で 1 つでも正解 Tubelet のバウンディングボックスとの IoU が 0.5 を超えているものと定義し，それ以外は負例 N とする．また $x_{\mathrm{ij}}^{y} \in$0,1 は Anchor cuboid a_i がラベル y における正解 Tubeletg_j との IoU が 0.5 を超えているものとする．そのとき損失関数は以下のように定義する．

$$L = \frac{1}{N}(L_{\mathrm{conf}} + L_{\mathrm{reg}}) \tag{1}$$

このとき $N = \sum_{\mathrm{i,j,y}} x_{\mathrm{ij}}^{y}$ とする．N は正例の数を示している．L_{conf} は分類誤差を示している．ここで，$\widehat{C}_i^y$ をラベル y の Anchor cuboida_i における分類スコアとすると分類誤差は次のようになる

$$L_{\mathrm{conf}} = -\sum_{\mathrm{i\in P}} x_{\mathrm{ij}}^{y} \log(\widehat{C}_i^y) - \sum_{\mathrm{i\in N}} \log(\widehat{C}_i^0) \tag{2}$$

L_{reg} は回帰誤差を示している．ここでの回帰は Tubelet の中心座標 (x, y) およびその幅と高さである．ここで $\widehat{r}_i^{x_k}$ をフレーム f_k 時における Anchor cuboida_i の x 座標の回帰座標と $\widehat{g}_j$ を正解の検出座標と定義すると回帰誤差は次のように定義する．

$$L_{\mathrm{reg}} = \frac{1}{K} \sum_{\mathrm{i\in P}} \sum_{\mathrm{c\in x,y,w,h}} x^y \sum_{\mathrm{K=1}}^{K} SmoothL1(\widehat{r}_i^{c_k} - \widehat{g}_{ij}^{c_k}) \tag{3}$$

$$g_{ij}^{x_k} = \frac{g_j^{x_k} - a_i^{x_k}}{a_i^{w_k}} \quad g_{ij}^{y_k} = \frac{g_j^{y_k} - a_i^{y_k}}{a_i^{h_k}} \tag{4}$$

$$g_{ij}^{w_k} = \log(\frac{g_j^{w_k}}{a_i^{w_k}}) \quad g_{ij}^{h_k} = \log(\frac{g_j^{h_k}}{a_i^{h_k}}) \tag{5}$$

4. 提案手法

先ほど説明した ACT-detector は個々の行動を認識する手法として使われている．本研究の最終目標としてチームスポーツの戦術解析を行いたいと考えている．よって本研究では，「複数人の動作の認識結果」と「各動作に関係のある物体の検出結果」のすべての関係性を基にしたシーン全体の動作の理解をするグループ行動認識へ拡張する．これから本研究の特徴および処理概要を説明する．

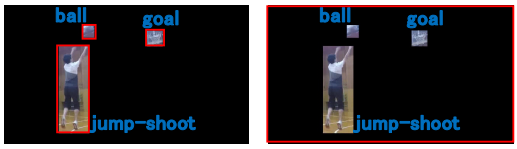

図 4 個々の行動を認識する最新の手法（ACT-detector）の行動認識結果からそのままグループ行動を選択する方法および提案手法によるグループ行動認識の違い．前者は左図のように検出結果を個別に認識の対象とするが，提案手法のグループ行動認識は右図のように個々の認識結果を踏まえてシーン全体の行動認識を行う．

4.1 本研究の特徴

4.1.1 複数人の動作と物体を基にしたグループ行動認識

現在行われているグループ行動認識のほとんどが個別の人の動作を認識し，それらの関係性を基にシーン全体の動作を認識する．本研究ではチームスポーツのグループ行動認識を取り扱う．チームスポーツでは個々の人の動作に加えてそれらに関連する物体も重要な情報となる．例えばバスケットボールではシュートする選手とゴールの位置関係によって得点が変わる．そのような理由より本研究では人に加えて物体を扱うことによってより高精度なグループ行動認識を目指す．

本研究では ACT-detector を用いて複数人の動作とそれに関連する物体を検出する．図 4 は個々の行動を認識する最新の手法（ACT-detector）の行動認識結果からそのままグループ行動を選択する方法と提案手法によるグループ行動認識の違いを示している．前者は図 4 の左のように個別の検出結果の中からシーン全体の認識に当てはまる動作をそのままグループ行動認識として利用する．図 4 の左の例では「ジャンプシュート」の検出結果のみをグループ行動認識としている．この場合もし個別の検出と正解検出例との IoU が低くなった場合，認識精度が低くなる．

それに対して提案手法のグループ行動認識は図 4 のように検出結果をすべて統合的に参照することでシーン全体のグループ行動認識を決めている．図 4 の右の例では個々の検出である「ジャンプシュート」，「ボール」，「ゴール」の 3 つの検出結果を統合的に参照した結果シーン全体では「ジャンプシュート」としている．この方法をとることによりたとえ「ジャンプシュート」の検出と正解検出例の IoU が低かったとしても他の検出である「ボール」，「ゴール」が正しく検出されていれば正しいグループ行動認識結果が期待できる．

4.1.2 複数人の相対的な配置関係を学習

絶対位置で考えた場合，カメラワーク等で移動したら人の見え方は変わってしまう．しかし，今回の自作データセットは図 5 のようにカメラを 3 視点から撮影した．また視点だけでなく，人の位置もシーケンスごとに異なるように撮影した．この視点と人の位置が異なるデータを実験に用いることで，提案する「複数の人の相対的な位置関係の特徴量化」が正しく機能するかどうかを確認できる．

図 5 3 視点による撮影．本研究の自作データセットは視点と人の位置が異なるデータを実験に用いることで，提案する「複数の人の相対的な位置関係の特徴量化」が正しく機能するかどうかを確認できる．

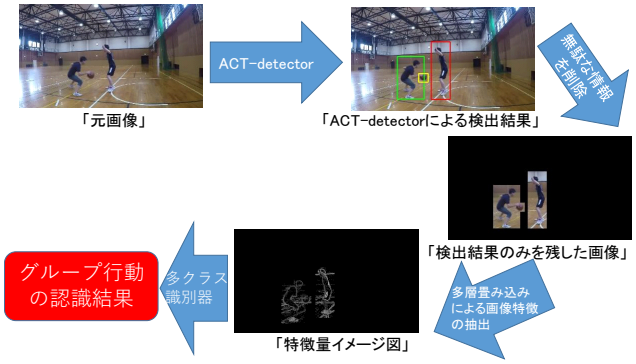

図 6 本研究の提案手法の説明図．まず ACT-detector によって検出し，検出結果のみ残しそれ以外の部分の情報をなくす．そして多層畳み込み層により検出結果の見え方の特徴量を抽出し，識別器にその特徴量を与えることによって分類する．

4.1.3 背景領域の排除による汎化性能の向上

本研究では「複数人の動作の認識結果」と「各動作に関係のある物体の検出結果」のその他背景情報は削除した図 6 の「検出結果のみを残した画像」を使用する．このようにした理由は，背景を含む画像を用いると畳み込み処理で得た特徴量マップには背景情報が含まれる．この情報はグループ行動認識には有益な情報とはならずむしろ過学習の要因になると考える．よって汎化性能向上のため，検出結果のみの残しそれ以外の画素値を 0 にした画像を作成することで背景情報を排除した．

4.2 本研究の処理概要

提案手法は図 6 の通りである．まず，図 6 のように，与えられた「元画像」から ACT-detector によって「複数人の動作」と「各動作に関係のある物体」の特徴量からそれぞれのバウンディングボックスを得る（「ACT-detector による検出結果」）．これによりの個別検出の認識結果が得られた．次に汎化性能向上のため，図 6 のように「検出結果のみを残した画像」を作成する．この作成方法は図 7 に示す．まず，与えられた動画から ACT-detector によって「複数人の動作」と「各動作に関係のある物体」のバウンディングボックスの座標と画素値を持たない元画像と同じサイズの画像を用意する．そして元画像を全探索し，探索座標がバウンディングボックス内なら元画像の画素値を画素値

```
検出領域(x1,y1),(x2,y2)(x1<x2,y1<y2)
画像のサイズ(x,y)

for  x_i in x:
    for  y_i in y:
        If x1≤ x_i ≤x2 & y1 ≤ y_i ≤y2:
            for  y_i in y:
                P2(x_i, y_i) = P1(x_i, y_i)
        Else:
            pass
```

図 7 検出領域のみを残すためのプログラムのアルゴリズム．まず，与えられた動画から ACT-detector によって「複数人の動作」と「各動作に関係のある物体」のバウンディングボックスの座標と画素値を持たない元画像と同じサイズの画像を用意する．そして元画像を全探索し，探索座標がバウンディングボックス内なら元画像の画素値を画素値なしの画像と同じ位置に代入する処理を行うことで，図 6 中央のような検出結果のみ残した画像を作成する．

なしの画像と同じ位置に代入する処理を行うことによって「検出結果のみの画像」ができる．そしてこの画像を用いて相対的な位置関係を抽出した特徴量（「特徴量イメージ図」）を抽出する．これは先ほど使用した ACT-detector の特徴抽出の部分を利用した．そのとき，異なるサイズの特徴量マップが得られるが，次に行う多クラス識別機に与えたところ特徴量マップが大きいもののほうが精度がよくなったため本研究では ACT-detector で得られる最も大きい特徴量マップ（$1024 \times 18 \times 18$）を使用した．図 4 のように，最後に先ほど得た特徴量マップを多クラス識別であるランダムフォレストに入力することでグループ行動認識を行った．また本研究でランダムフォレストを用いて理由として，決定木による複数の弱学習器を統合しているため汎化性能が高く，かつディープラーニングによる分類よりも処理時間が短く済むのでランダムフォレストを用いた．本研究ではこのようにグループ行動認識を行っている．

5. 実験結果および考察

この章では，本研究で使用した自作データセットの説明と，個々の行動を認識する最新の手法（ACT-detector）の行動認識結果からそのままグループ行動を選択する方法および文献 [25] に倣った max pooling によるグループ行動認識と提案手法のグループ行動認識の結果を比較する．

5.1 データセットの説明

本研究ではグループ行動認識を行うためにバスケットボールの主な動作であるドリブル，パス，ジャンプシュート，レイアップシュートの動画を撮影した．これら動画は，シーン全体としてのラベルとして「ドリブル」，「パス」，「ジャンプシュート」，「レイアップシュート」の 4 つを与え，さらに上の 4 つに加えて「ディフェンス」，「ゴール」，「ボール」，「レシーブ」の 8 つを与えた．ゴール方向の向きを正面と定義すると撮影方向は正面とそこから左右約 30

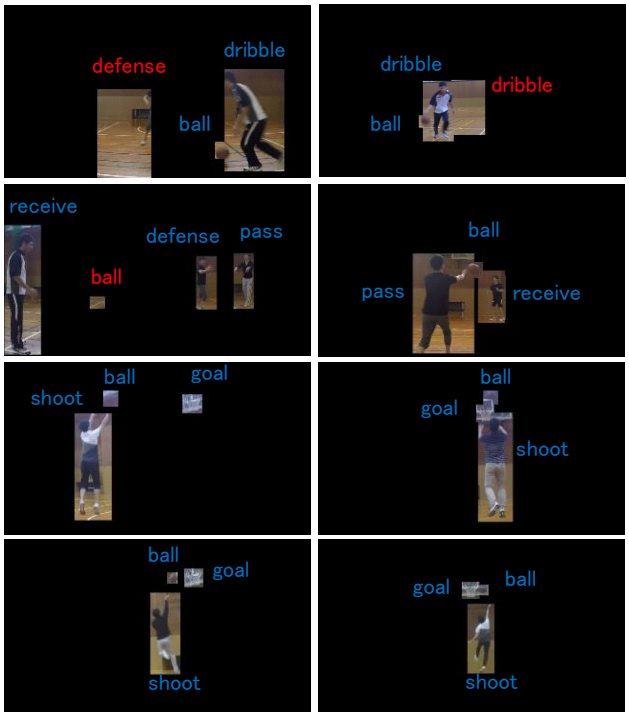

図 8 ACT-detector による検出結果の例．上からドリブル，パス，ジャンプシュート，レイアップシュートの画像の検出結果である．青字で示しているのが検出成功で，赤字で示しているのが検出失敗である．

度ずつずらした 3 視点で撮影した．また個々の行動を認識する最新の手法（ACT-detector）のトレーニングとして使用した画像は 19782 枚，テストは 8478 枚使用した．そして max pooling の手法と提案手法は個々の行動を認識する最新の手法（ACT-detector）のテスト結果をトレーニングデータとテストデータに分け，トレーニングデータとして画像 5934 枚，テストデータとして 2544 枚使用した．またバウンディングボックスのアノテーションは [29] のアノテーションツールを用いた．

5.2 個々の行動の認識によるグループ行動認識

まず，本研究では個々の行動認識の結果のすべてを一括に特徴量を取る手法（図 6 の多層畳み込みによる画像特徴の抽出）を用いることにより，そのグループ行動認識に影響を与えると考えられる動作の特徴量を一つにまとめた．これによってグループ行動認識を実現している．次に，個々の行動を認識する最新の手法（ACT-detector）の行動認識結果からそのままグループ行動を選択する方法の検出結果例（図 8）と認識結果（表 1）を示す．図 8 は上からシーン全体の正解ラベルが「ドリブル」，「パス」，「ジャンプシュート」，「レイアップシュート」の画像が順に並んでおり，青字が検出成功，赤字が検出失敗を示している．ACT-detector の現状としては図 8 のドリブルの左図やパスの左図のよう

表 1 個々の行動を認識する最新の手法（ACT-detector）の行動認識結果からそのままグループ行動を選択する方法の結果．「ドリブル」から「レシーブ」までの項が各動作の動作の認識率で 8 クラス平均がすべての検出結果の平均認識率であり，4 クラス平均が「ドリブル」,「パス」,「ジャンプシュート」,「レイアップシュート」の平均認識率．

ドリブル	パス	レイアップシュート
97.41	83.14	98.97
ジャンプシュート	ディフェンス	ゴール
74.84	86.83	72.83
ボール	レシーブ	
91.76	74.39	
8 class average	4 class average	
74.83	88.59	

表 2 max pooling を利用したグループ行動認識の結果を示す混合行列．行方向が学習器による予測結果を示しており，列方向が実際の動画の正解ラベルを示している．

	dribble	pass	lay-up-shoot	jump-shoot
dribble	1033	4	0	0
pass	26	255	1	1
lay-up-shoot	5	0	612	1
jump-shoot	4	0	6	596

に正しい検出領域を捉えれていなかったり，ドリブルの右図のように背景をドリブルと誤検出したりするのが現状である．表 1 は個々の行動を認識する ACT-detector の行動認識結果からそのままグループ行動を選択する方法の結果を示している．ドリブルからレシーブまでの項が各動作の動作の認識率で 8 クラス平均がすべての検出結果の平均認識率であり，4 クラス平均がシーン全体の認識を示す「ドリブル」,「パス」,「ジャンプシュート」,「レイアップシュート」の平均認識率を示している．表 1 を見るとボールなどの動きが速く小さいものなどは行動認識でも精度が悪くなっている．その理由として動きの速いものや小さい物体は画素が薄れてしまい背景と同化してしまうため検出に失敗してしまうと考えられる．「ドリブル」,「パス」,「ジャンプシュート」,「レイアップシュート」の 4 クラスの認識率の平均としては 88.59 %となっている．この値はシーン全体の認識を示す 4class であり，提案手法のグループ行動認識の結果と比較するために使用する．

5.3 max pooling を利用したグループ行動認識の結果

次に，文献 [25] に倣って max pooling を利用したグループ行動認識を行う．まず，図 6 のように元画像から ACT-detector による検出を行う．次にこの検出した結果ごとにクロップし，300 × 300 にリサイズした画像を作成する．そしてこのリサイズした画像を多層畳み込みによる画像特徴の抽出を行う．これで各フレームごとに検出した物体の数だけの特徴量ベクトルが得られる．そして文献 [25] の

表 3 本研究の提案手法であるグループ行動認識の結果を示す混合行列．行方向が学習器による予測結果を示しており，列方向が実際の動画の正解ラベルを示している．

	dribble	pass	lay-up-shoot	jump-shoot
dribble	1036	1	0	0
pass	0	283	0	0
lay-up-shoot	2	0	616	0
jump-shoot	0	0	0	606

図 9 提案手法の認識結果を可視化した例．左図が正解のグループ行動認識がパスであり，ACT-detector の検出が失敗しているが，提案手法の認識が成功している．右図が正解のグループ行動認識がレイアップシュートであるが，提案手法の認識がドリブルと判定し，失敗している例である．

手法に倣って，これら特徴量ベクトルを max pooling を行い，各画像に一つの特徴量ベクトルにする．この特徴量ベクトルをランダムフォレストに与えることによって分類を行う．そして，提案手法によるグループ行動認識の結果を示す混合行列（表 2）を示す．この混合行列は行方向が学習器による予測結果を示し，列方向は実際の動画の正解ラベルを示している．「ドリブル」,「パス」,「ジャンプシュート」,「レイアップシュート」の 4 クラスの認識率の平均としては 98.11 %となって，個々の個別認識の手法で失敗しているものも正しく認識できるようになっている．

5.4 提案手法によるグループ行動認識の結果

次に，提案手法によるグループ行動認識の結果を示す混合行列（表 3）を示す．「ドリブル」,「パス」,「ジャンプシュート」,「レイアップシュート」の 4 クラスの認識率の平均としては 99.88 %と max pooling の手法よりも 1.77 %改善することができた．この改善理由として，max pooling による手法は検出結果の特徴量ベクトルらを max pooling により一つの特徴量ベクトルに変換する必要があるため，すべての特徴量を残すことはできない．これに対して，提案手法は検出結果をすべて残した特徴量ベクトルを作成したことによってわずかだが，max pooling による手法よりも精度が向上したと考えられる．また本研究のデータセットは 3 視点で撮影を行い，位置もバリエーションが豊富ではあったが，動作に類似度が多く，大多数のパターンが学習データに含まれていたため認識精度が高かったと考えられる．本研究で使用したランダムフォレストの決定木の数は 10 で行った．

そして，図 9 に示しているのが提案手法の認識結果を可

視化した図である．図 9 の左図が ACT-detector の検出が失敗しているが，提案手法の認識が成功している例で図 9 の右図が提案手法の認識が失敗している例である．図 9 の左図のようにヒトの 1 部が上手く検出できていないような場合でもその他の上手くいった部分を基にしてグループ行動認識では上手くいっているケースなどがあり，このような部分から部分的に上手く検出できれば提案手法がうまく機能することがわかる．しかし，認識対象の動作が上手く検出できても，他のラベルと変わらないような図 9 の右図のケースではシーン全体でも認識を誤ることがある．こういったもののグループ行動認識の精度を上げるためにはフレーム単体ではなく動画において全体で考える必要があると考えられる．

6. むすび

本研究では，ACT-detector を改良してシーン全体で行動認識を行うグループ行動認識を行った．その結果としては提案手法のグループ行動認識が最も高い精度が出た．しかし，本研究で用いたデータセットはそれぞれの行動に曖昧性がないように作成し，またクラス数としても 4 分類なので比較的に分類が容易であると考えられる．今後の予定としては，NBA の試合動画や文献 [25] で使われているバレーボールのデータセットなどのより難易度の高いデータセットを使うことを検討している．そして敵チーム，味方チームの選手のグループ行動認識を行い，それらのグループ行動認識を統合することによってシーン全体ではどのようなプレイになるのかを階層ごとに認識することを考えている．それに加えて，カメラワークの動きがあっても行動認識ができるのか，戦術解析に有益な動作を行動認識できるのかなど様々な課題に取り組んでいく．

参考文献

[1] データスタジアム株式会社:, http://www.datastadium.co.jp/.
[2] 有限会社フィットネスアポロ社:, http://sportscode.jp/products/sportscode/.
[3] 株式会社富士通研究所:, http://www.fujitsu.com/jp/group/labs/resources/.
[4] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L.: Large-scale video classification with convolutional neural networks, *CVPR* (2014).
[5] Simonyan, K. and Zisserman, A.: Two-stream convolutional networks for action recognition in videos, *Advances in neural information processing systems* (2014).
[6] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M.: Learning spatiotemporal features with 3d convolutional networks, *ICCV, IEEE*, IEEE (2015).
[7] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z. and Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn), *arXiv preprint arXiv:1412.6632* (2014).
[8] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. and Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description, *CVPR* (2015).
[9] Varol, G., Laptev, I. and Schmid, C.: Long-term temporal convolutions for action recognition, *TPAMI* (2017).
[10] Girshick, R., Donahue, J., Darrell, T. and Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation, *CVPR* (2014).
[11] Girshick, R.: Fast r-cnn, *arXiv preprint arXiv:1504.08083* (2015).
[12] Ren, S., He, K., Girshick, R. and Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems* (2015).
[13] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A.: You only look once: Unified, real-time object detection, *CVPR* (2016).
[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C.: Ssd: Single shot multibox detector, *ECCV*, Springer (2016).
[15] Redmon, J. and Farhadi, A.: YOLO9000: better, faster, stronger, *arXiv preprint* (2017).
[16] Uijlings, J. R., Van De Sande, K. E., Gevers, T. and Smeulders, A. W.: Selective search for object recognition, *International journal of computer vision*, Vol. 104, No. 2, pp. 154–171 (2013).
[17] Marian Puscas, M., Sangineto, E., Culibrk, D. and Sebe, N.: Unsupervised tube extraction using transductive learning and dense trajectories, *ICCV* (2015).
[18] Peng, X. and Schmid, C.: Multi-region two-stream R-CNN for action detection, *ECCV*, Springer (2016).
[19] Saha, S., Singh, G., Sapienza, M., Torr, P. H. and Cuzzolin, F.: Deep learning for detecting multiple space-time action tubes in videos, *arXiv preprint arXiv:1608.01529* (2016).
[20] Kalogeiton, V., Weinzaepfel, P., Ferrari, V. and Schmid, C.: Action tubelet detector for spatio-temporal action localization, *ICCV* (2017).
[21] Singh, G., Saha, S., Sapienza, M., Torr, P. and Cuzzolin, F.: Online real-time multiple spatiotemporal action localisation and prediction, *CVPR* (2017).
[22] Chen, L., Zhai, M. and Mori, G.: Attending to Distinctive Moments: Weakly-supervised Attention Models for Action Localization in Video, *CVPR* (2017).
[23] Lan, T., Wang, Y., Yang, W., Robinovitch, S. N. and Mori, G.: Discriminative latent models for recognizing contextual group activities, *TPAMI*, Vol. 34, No. 8, pp. 1549–1562 (2012).
[24] Lan, T., Sigal, L. and Mori, G.: Social roles in hierarchical models for human activity recognition, *CVPR*, IEEE (2012).
[25] Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A. and Mori, G.: A hierarchical deep temporal model for group activity recognition, *CVPR*, IEEE (2016).
[26] Deng, Z., Vahdat, A., Hu, H. and Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition, *CVPR* (2016).
[27] Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K. and Fei-Fei, L.: Detecting events and key actors in multi-person videos, *CVPR* (2016).
[28] Hajimirsadeghi, H. and Mori, G.: Learning ensembles of potential functions for structured prediction with latent variables, *ICCV* (2015).
[29] vatic: , https://github.com/cvondrick/vatic.